

# Running ImmunoVerse on CGC

Guangyuan(Frank) Li, PhD  
NYU Grossman School of Medicine  
Last update: 2025.05.08

# 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.sbggenomics.com/home>
2. eRA common should work just fine, otherwise, please follow the links to create an account using your email



CANCER GENOMICS CLOUD  
SEVEN BRIDGES

**Log in**

 Log in with eRA Commons

[Log in with username and password](#)

New to the CGC? [Create an account](#)

# Step 2: Create a project

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

The screenshot shows the 'Create a project' dialog box overlaid on the 'Projects' page. The 'Create a project' dialog has fields for 'Name' (with a red arrow pointing to it labeled 'Name it'), 'Billing group' (set to 'Pilot Funds (li2g2uc)' with a red arrow pointing to it labeled 'Billing group'), and 'Execution settings' (checkboxes for 'Spot instances' and 'Reuse'). The 'Spot instances' checkbox is checked and has a red arrow pointing to it with the text 'You need a billing account either ideally from your lab, or you can apply \$300 pilot fund (https://www.cancergenomicscloud.org/cgc-apply-for-collaborative-funds)'. The 'Reuse' checkbox is unchecked. The main 'Projects' page lists several existing projects: 'deploy\_pipelines', 'CONTROLLED NeoVerse\_pancancer\_2', 'NeoVerse\_development\_project', 'CONTROLLED NeoVerse\_pancancer', and 'Demonstration: Building an App'. A green button '+ Create Project' is visible on the left of the main page. Red arrows on the main page point to the 'Create Project' button and the 'Spot instances' checkbox in the dialog, both labeled 'uncheck'.

Projects

deploy\_pipelines  
Created by:li2g2uc · Jan 13, 2025, 10:37

**CONTROLLED** NeoVerse\_pancancer\_2  
Created by:li2g2uc · Sep 28, 2024, 23:01

NeoVerse\_development\_project  
Created by:li2g2uc · May 15, 2024, 12:15

**CONTROLLED** NeoVerse\_pancancer  
NeoVerse

Created by:li2g2uc · Mar 15, 2024, 16:33

Demonstration: Building an App  
Created by:rowan\_beck\_era · Jul 18, 2023, 13:06

+ Create Project View all projects

Public Data and Apps

Analyze

331061

publicly available files

Create a project

Name

https://cgc.sbggenomics.com/u/li2g2uc/

**CONTROLLED** This project will contain controlled data.

General information Advanced settings

Billing group

Pilot Funds (li2g2uc)

You need a billing account either ideally from your lab, or you can apply \$300 pilot fund (<https://www.cancergenomicscloud.org/cgc-apply-for-collaborative-funds>)

Location

AWS (us-east-1)

Execution settings

Spot instances  
Spot instances can significantly reduce the cost of your task execution if results are not needed urgently. [Learn more](#)

Reuse  
Automatic reuse of precomputed results can significantly reduce the time and cost of your task execution. [Learn more](#)

Cancel Create

# Step 2: Create a project (cont'd)

deploy\_pipelines  
Created by:li2g2uc · Jan 13, 2025, 10:37

**CONTROLLED** NeoVerse\_pancancer\_2  
Created by:li2g2uc · Sep 28, 2024, 23:01

NeoVerse\_development\_project  
Created by:li2g2uc · May 15, 2024, 12:15

**CONTROLLED** NeoVerse\_pancancer  
NeoVerse  
Created by:li2g2uc · Mar 15, 2024, 16:33

Demonstration: Building an App  
Created by:rowan\_beck\_era · Jul 18, 2023, 13:06

[+ Create Project](#) [View all projects](#)

Public Data and Apps

Analyze

**331061**

### Create a project

Name

**CONTROLLED** This project will contain controlled data.  ⓘ

[General information](#) [Advanced settings](#)

**Network Access settings ⓘ**

Block network access  
Execution within the project won't have network access.

Allow network access Allow network Allow network  
Execution will have unrestricted network access.

**Download Restriction settings ⓘ**

⚠ Download Restriction settings cannot be modified after the project has been created.

Download restriction will be applied to all the files that are imported to the project.

No

[Cancel](#) [Create](#)

# Step 3: Navigate your project

The screenshot shows the ImmunoVerse project interface. On the left, the 'Overview' tab is selected, displaying a summary of the project's status. It includes sections for 'Description', 'All the files', 'ImmunoVerse Overview', and 'Tools available'. The 'Tools available' section lists 14 different pipelines. On the right, the 'Analysis' tab is selected, showing a list of pipeline runs. The first run is 'COMPLETED' (rescore\_pipeline run - 02-06-25 15:37:21), while the subsequent five are 'ABORTED' or 'FAILED'.

When running each app for specific input, you create an instance or task

Description

All the files

ImmunoVerse Overview

Apps built from docker containers

Users upload the RNA-Seq raw data, and immunopeptidome data if available, and use the following pre-built pipelines to (1) Profile all genetic aberrations, (2) Generate sample-specific search space, (3) Run immunopeptidome search. You don't need to write any code to finish all the analysis (amazing isn't it). We believe separating each module can maximize the flexibility in real-world scenarios, we are also working on connecting all dots into a single workflow in the future.

Tools available

1. Gene Expression Pipeline (all samples keeping pair order, 10GB RAM each done in 30min)
2. Alignment Pipeline (all samples keeping pair order, 30GB RAM each done in 5h)
3. Splicing Intron Pipeline (all samples, 2GB RAM each done in 30min)
4. Transposable Element Pipeline (all samples, 10GB RAM each in 5h)
5. Pathogen Pipeline (all samples, 100GB RAM each done in 10min)
6. Variant Pipeline
  - 6.1 RNA Variant Pipeline (all samples, 2GB RAM each done in 90min)
  - 6.2 VEP pipeline (batch by sample, 5GB RAM each done in 30min)
7. Gene Fusion Pipeline (batch by both sample and pair-end)
8. HLA type pipeline
  - 8.1 decompress pipeline (batch by file, 2GB RAM each done in 5min)
  - 8.2 optype pipeline (batch by sample and pair-end, 20GB RAM each done in 20min)
9. Summarization Pipeline
10. (optional) bam\_to\_fastq pipeline
11. (optional) circular RNA pipeline
12. Immunopeptidome Pipeline
  - 12.1 MaxQuant Pipeline
  - 12.2 msconvert pipeline
  - 12.3 Rescore pipeline
  - 12.4 HLA binding pipeline

Members

sharma28

Manage members

Analysis

Tasks Data Studio

Completed rescore\_pipeline run - 02-06-25 15:37:21

Submitted by: li2g2uc - Feb 6, 2025 10:37

Aborted rescore\_pipeline run - 02-06-25 15:09:28

Submitted by: li2g2uc - Feb 6, 2025 10:09

Aborted rescore\_pipeline run - 02-06-25 14:53:12

Submitted by: li2g2uc - Feb 6, 2025 9:53

Aborted rescore\_pipeline run - 02-06-25 14:31:20

Submitted by: li2g2uc - Feb 6, 2025 9:31

Failed rescore\_pipeline run - 02-06-25 14:12:29

Submitted by: li2g2uc - Feb 6, 2025 9:12

# Upload and Download

# Upload files

Dashboard **Files PREMIUM** Apps Tasks Data Studio Interactive Apps

deploy\_pipelines ⓘ

Interactive Browsers Settings Notes

Root

Search Extension Tags Paired-end + Clear filters

<input type="checkbox"/>	Name	Size	Extension	Task ID
<input type="checkbox"/>	ensembl_protein.fasta <small>reference</small>	11.36 MiB	FASTA	-
<input type="checkbox"/>	gencode.v36.annotation.gtf <small>reference</small>	1.29 GiB	GTF	-
<input type="checkbox"/>	gene_model.txt <small>reference</small>	52.57 MiB	TXT	-
<input type="checkbox"/>	GRCh38.d1.vd1.fa <small>reference</small>	2.94 GiB	FA	-
<input type="checkbox"/>	GRCh38.d1.vd1.gencode.v36.annotation.star-fusion-1.12.0-CTAT-index-archive.tar <small>reference</small>	44.87 GiB	TAR	12f76e64-e489-... Feb. 02, 2025 15:...
<input type="checkbox"/>	hg19_maxquant_combined_txt	-	-	6ea7ebaa-448b-... Feb. 05, 2025 03:...

Easiest way, just drag the file from your computer

New Folder + Add files ...

Public Files

Projects

Your Computer

FTP / HTTP

GA4GH Data Repository Service (DRS)

Data Tools

Volumes

Import from a manifest file

Jan. 25, 2025 13:... -

Feb. 02, 2025 15:... -

Feb. 05, 2025 03:... -

# Upload files (cont'd)

Add files to "deploy\_pipelines"

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



Import from an FTP or HTTP(S) server: ?

Paste the link of the file(s) you want to import

`https://genome.med.nyu.edu/public/yarmarkovichlab/ImmunoVerse/normal/  
normal_intron.txt`

or  on your computer containing the links

Add tags

Resolve naming conflicts:

Skip ▾

Import

# Download

Dashboard   Files **PREMIUM**   Apps   Tasks   Data Studio   Interactive Apps   **deploy\_pipelines** ⓘ   Interactive Browser

Root  

1 item selected  

**download**

Name	Size	Extension	Task ID	Created on
<input checked="" type="checkbox"/> ensembl_protein.fasta <small>reference</small>	11.36 MiB	FASTA	-	Jan. 13, 2025 10:...
<input type="checkbox"/> gencode.v36.annotation.gtf <small>reference</small>	1.29 GiB	GTF	-	Jan. 13, 2025 10:...
<input type="checkbox"/> gene_model.txt <small>reference</small>	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 workflows
- Two files are present twice, please remove the prefix as final clean
  - \_1\_hg38.fa
  - \_1\_Eensemle\_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

The screenshot shows the Immuniverse platform interface. At the top, there is a navigation bar with links for Home, Projects, Data, Public Apps, Public Projects, and Developer. Below the navigation bar is a secondary menu with Dashboard, Files PREMIUM, Apps, Tasks, Data Studio, and Interactive Apps. The main title "test\_immunoverse" is displayed above the file list. On the left, there is a "Root" folder icon. On the right, there are buttons for "New Folder" and "+ Add files". A dropdown menu is open, listing options: "Public Files" (highlighted with a yellow border and a red arrow pointing to it), "Projects", "Your Computer", "FTP / HTTP", "GA4GH Data Repository Service (DRS)", "Data Tools", "Volumes", and "Import from a manifest file". A banner at the bottom states "Files are the basis of every analysis." and provides a link to learn more about different ways to add files.

Home Projects ▾ Data ▾ Public Apps ▾ Public Projects Developer ▾

Dashboard **Files PREMIUM** Apps Tasks Data Studio Interactive Apps

test\_immunoverse ⓘ

New Folder + Add files ⋮

Public Files

Projects

Your Computer

FTP / HTTP

GA4GH Data Repository Service (DRS)

Data Tools

Volumes

Import from a manifest file

Files are the basis of every analysis.

New folder + Add files

learn more about different ways to add files.

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

Add files to "test\_immunoverse"

Please contact me if you don't have access to this project

Search for project: deploy\_pipelines

Search, Extension, Tags, Paired-end, Clear filters

Copy to Project

Name ▲

Name	Size	Ext
ensembl_protein.fasta	11.36 MiB	FASTA
gencode.v36.annotation.gtf	1.29 GiB	GTF
gene_model.txt	52.57 MiB	TXT
GRCh38.d1.vd1.fa	2.94 GiB	FA
GRCh38.d1.vd1.gencode.v36.annotation.s	44.87 GiB	TAR
hg19_maxquant_combined_txt	-	-
hg38.fa	3.05 GiB	FA
hg38.knownGene.gtf	564.57 MiB	GTF
hg38_rmsk_TE.gtf.loclnd	883.82 MiB	LOCIN

Tags filter: reference

Apply button

Red arrows highlight the 'deploy\_pipelines' project, the 'reference' tag in the dropdown, and the 'Apply' button.

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

Add files to "test\_immuneverse"

Search for project: deploy\_pipelines

Root Exclude subfolders

Search: Extension: Tags: reference Paired-end + Clear filters

Copy to Project

63 items

Name	Path	Size
annot.renamed.txt.gz	Files / ImmunoV...	10.08
reference ImmunoVerse_data		
annot.renamed.txt.gz.tbi	Files / ImmunoV...	2.74 M
reference ImmunoVerse_data		
blacklist_splicing.txt	Files / ImmunoV...	25.59
reference ImmunoVerse_data		
canonical.txt	Files / ImmunoV...	2.50 M
reference ImmunoVerse_data		
chrLength.txt	Files / star_hg3...	13.83
reference star_hg38_index		
chrName.txt	Files / star_hg3...	65.86
reference star_hg38_index		
chrNameLength.txt	Files / star_hg3...	79.69
reference star_hg38_index		
chrStart.txt	Files / star_hg3...	29.85
reference star_hg38_index		
contigs.txt	Files / ImmunoV...	136.68
reference ImmunoVerse_data		
cosmic_prelift.bed.gz	Files / ImmunoV...	1.63 M
reference ImmunoVerse_data		
cosmic_prelift.bed.gz.tbi	Files / ImmunoV...	43.22
reference ImmunoVerse_data		
ensembl_protein.fasta	Files	11.36 M
reference		

Demonstration: Building an App

NeoVerse\_pancancer\_2

NeoVerse\_development\_proj...

NeoVerse\_pancancer

ImmunoVerse

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

The screenshot shows the QIIME 2 interface with the following details:

- Header:** Home, Projects, Data, Public Apps, Public Projects, Developer.
- Sub-Header:** Dashboard, Files (PREMIUM), Apps, Tasks, Data Studio, Interactive Apps, test\_immunoverse, Interactive Browsers, Settings, Notes.
- Toolbar:** Root, Search, Extension, Tags, Add filters.
- File List:** A list of files and folders on the left, many of which are tagged with "reference" and "ImmunoVerse\_data".
- Create New Folder Dialog:** A modal window titled "Create New Folder" is open.
  - A note says: "Folders can't be subsequently renamed."
  - The "Name" field contains "ImmunoVerse\_data" (highlighted with a red box).
  - The "Path" field contains "Files /".
  - Buttons at the bottom are "Cancel" and "Create" (highlighted with a red box).
- Table:** A table on the right lists files and their details:

Extension	Task ID	Create
FASTA	-	Feb. 06
FASTA	-	Feb. 06
FA	-	Feb. 06
K2D	-	Feb. 06
TXT	-	Feb. 06
TXT	-	Feb. 06
-	-	Feb. 06
-	-	Feb. 06
-	-	Feb. 06
156 bytes	TXT	Feb. 06
323.92 MiB	BED	Feb. 06
11.36 MiB	FASTA	Feb. 06

# Organize a bit (create a folder named ImmunoVerse\_data and move all files tagged as ImmunoVerse\_data to this folder)

The screenshot shows a cloud storage interface with a search and filter overlay. The main table lists various files and their details. A modal dialog is open, showing a dropdown menu for filtering by tag. The tag 'ImmunoVerse\_data' is selected, highlighted with a blue border. A red arrow points from the text 'ImmunoVerse\_data' in the dropdown to the tag itself. Another red arrow points to the 'Apply' button at the bottom right of the modal.

Name	Size	Extension	Task ID	Create
ImmunoVerse_data	-	-	-	Feb. 06
_2_ensembl_protein.fasta	11.36 MiB	FASTA	-	Feb. 06
reference ImmunoVerse_data	11.36 MiB	FASTA	-	Feb. 06
_1_ensembl_protein.fasta	3.05 GiB	FA	-	Feb. 06
reference	64 bytes	K2D	-	Feb. 06
_1_hg38.fa	1.37 GiB	TXT	-	Feb. 06
reference ImmunoVerse_data	96.94 MiB	TXT	-	Feb. 06
opts.k2d	23.20 GiB	-	-	Feb. 06
reference kraken2_db	1.46 GiB	-	-	Feb. 06
normal_erv.txt	156 bytes	TXT	-	Feb. 06
SA	323.92 MiB	BED	-	Feb. 06
reference star_hg38_index	64 items			
SAindex				
reference star_hg38_index				
splice_erv_db.txt				
reference ImmunoVerse_data				
tcga_tmp_prelift.bed				
reference ImmunoVerse_data				

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

The screenshot shows a file management interface with a 'Move' dialog box open. The dialog box is titled 'Move' and asks 'Move 31 selected items?'. It contains a tree view under 'Files' with a blue selection box around a folder named 'ImmunoVerse\_data'. A red arrow points to this folder. Below the tree view are 'Tags' settings, including a checked checkbox for 'Keep preexisting tags' and a text input field for adding new tags. At the bottom of the dialog are 'Cancel' and 'Move' buttons, with the 'Move' button highlighted by a red arrow. The background shows a list of 31 selected items, each with a 'reference' tag and a 'ImmunoVerse\_data' tag. The items include various file types like FASTA, TXT, BED, and TBI.

Move

Move 31 selected items?

Files

ImmunoVerse\_data

Tags

Keep preexisting tags

Add new tags by separating them with enter key

New folder

Cancel Move

Files

Size	Extension	Task II
11.36 MiB	FASTA	-
1.37 GiB	TXT	-
96.94 MiB	TXT	-
156 bytes	TXT	-
323.92 MiB	BED	-
16.84 MiB	BED	-
73 bytes	TXT	-
43.22 KiB	TBI	-
75 bytes	TXT	-
10.08 GiB	TXT.GZ	-
136.68 KiB	TXT	-

31 items

Organize a bit (create another folder named star\_hg\_index and move all files tagged as star\_hg38\_index to this folder)

The screenshot shows a cloud storage interface with a "Move" dialog box open over a list of files.

**Move Dialog:**

- Header:** Move
- Message:** Move 16 selected items?
- File Selection:** A tree view shows a folder named "star\_hg38\_index" under "ImmunoVerse\_data".
- Buttons:** New folder, Cancel, Move

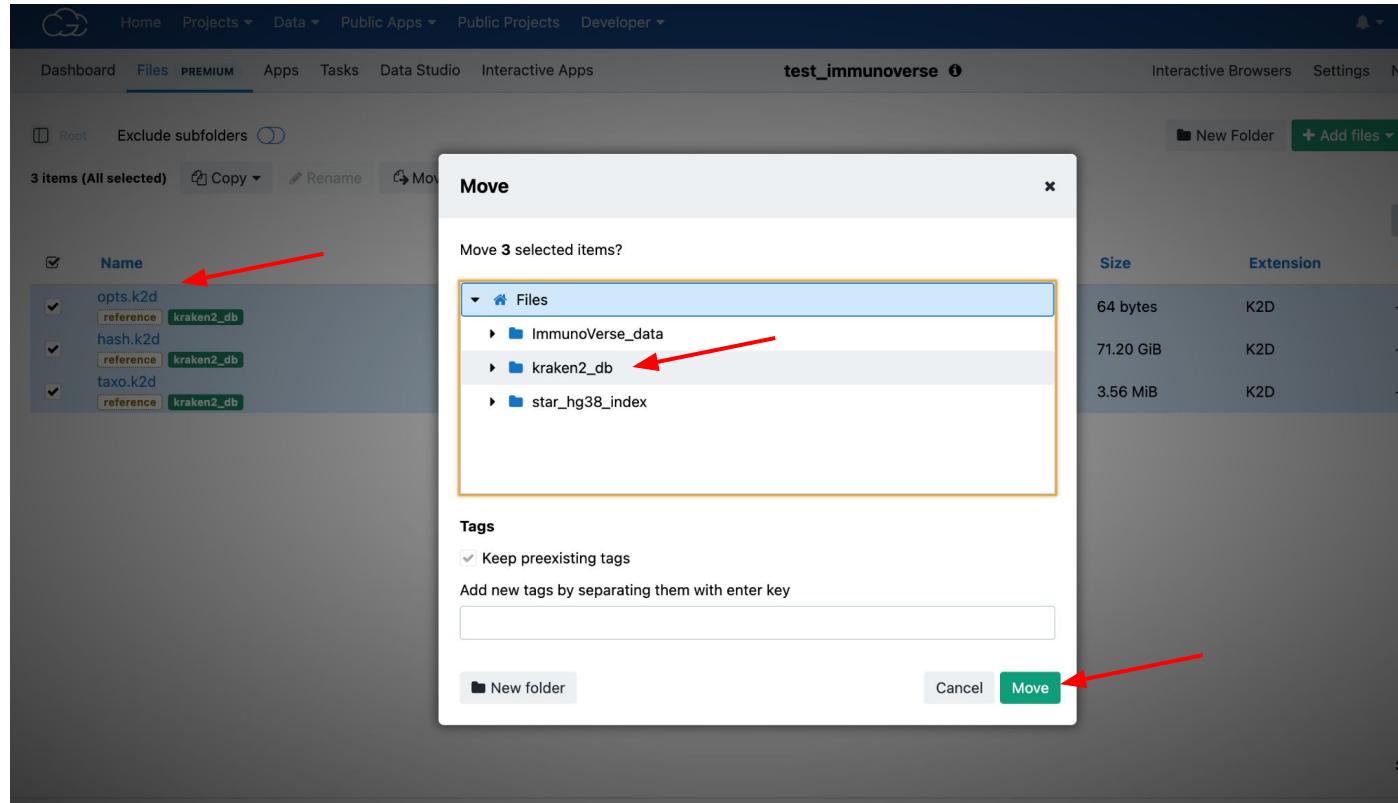
**List of Files:**

Size	Extension	Task II
23.20 GiB	-	-
1.46 GiB	-	-
29.85 KiB	TXT	-
14.88 MiB	TAB	-
10.83 MiB	TXT	-
79.69 KiB	TXT	-
972 bytes	TXT	-
3.63 GiB	-	-
467.08 KiB	OUT	-
65.86 KiB	TXT	-
48.57 MiB	TAB	-

**UI Elements:**

- A red arrow points to the "Name" checkbox in the list of selected items.
- A red arrow points to the "star\_hg38\_index" folder in the file selection tree.
- A red arrow points to the "Move" button in the dialog.

Organize a bit (create a folder named kraken2\_db, and move all files tagged as kraken2\_db to this folder)



# Final clean

Root Exclude subfolders

Search bar: **ensembl**  Extension  Tags   Clear filters

Name
<a href="#">_1_ensembl_protein.fasta</a> <small>reference Immunoverse_data</small>

A red arrow points from the text "Click the file and rename" to the file name `_1_ensembl_protein.fasta`.

Root Exclude subfolders

Search bar: **hg38**  Extension  Tags   Clear filters

Name
<a href="#">star_hg38_index</a>
<a href="#">_1_hg38.fa</a> <small>reference</small>

A red arrow points from the text "Click the file and rename" to the file name `_1_hg38.fa`.

Click the  
file and  
rename

Dashboard Files Apps Tasks Data Studio Interactive Apps

Files

**\_1\_hg38.fa**

**REFERENCE**

3.0 GiB (3,273,481,150 bytes) · Created on May 8, 2025 15:58 (Eastern Day

Metadata Raw View

# Copy my app/apps

The screenshot shows the QGIS application interface. The top navigation bar includes links for Home, Projects, Data, Public Apps, Public Projects, and Developer. A red arrow points to the 'Apps' tab, which is currently selected. The main content area displays a search bar with placeholder text 'Search names and description', and several filter dropdowns: Category: All, Toolkit: All, Language and version: All, Status: Available, and Cost Estimator: All. To the right of these filters is a 'Create app' button and a 'Add apps' button with a plus sign. A red arrow points to the 'Add apps' button. A context menu is open from the 'Add apps' button, listing three options: 'Public apps' (which is highlighted with an orange border), 'Projects', and 'My created apps'. The central part of the screen shows a large 'No apps' message with a sub-message stating 'No apps with the given search term can be found.' Below this is a 'Clear filters' button.

# Copy my app/apps

Add apps to test\_immunoverse

Search Search names and description Category: All Toolkit: All Language and version: All

Name	Type	Modified by	Modified on	
deploy_pipelines	Tool	li2g2uc	Feb 06, 2025 13:31	
NeoVerse_pancancer_2	Tool	li2g2uc	Feb 06, 2025 10:37	
NeoVerse_development_project	Tool	li2g2uc	Feb 05, 2025 18:28	
NeoVerse_pancancer	Tool	li2g2uc	Feb 04, 2025 18:22	
Demonstration: Building an App	Tool	li2g2uc	Feb 04, 2025 12:03	
STAR-Fusion Build Fusio	Tool	li2g2uc	Jan 31, 2025 11:33	
SBG Decompressor CWL	Tool	li2g2uc	Jan 27, 2025 16:28	
variant_pipeline	Tool	li2g2uc	Jan 26, 2025 11:14	
Variant Effect Predictor	Tool	li2g2uc	Jan 24, 2025 16:17	
OptiType	Tool	li2g2uc	Jan 24, 2025 15:47	
STAR-Fusion	Tool	li2g2uc	Jan 24, 2025 15:23	
alignment_pipeline	Tool	li2g2uc	Jan 24, 2025 14:52	
gene_pipeline	Tool	li2g2uc	Jan 14, 2025 14:02	
splicing_intron_pipeline	Tool	li2g2uc	Jan 13, 2025 12:12	
telocal_pipeline	Tool	li2g2uc	Jan 13, 2025 12:12	

# Finished setting things up!!!

Dashboard Files PREMIUM Apps Tasks Data Studio Interactive Apps **test\_immunoverse** ⓘ Interactive Browsers Settings Notes

Root New Folder + Add files ...

15 items (All selected) Copy Rename Move Metadata Edit tags Download ...

<input checked="" type="checkbox"/>	Name	Size	Extension	Task ID	Create
<input checked="" type="checkbox"/>	kraken2_db	-	-	-	Feb. 06
<input checked="" type="checkbox"/>	star_hg38_index	-	-	-	Feb. 06
<input checked="" type="checkbox"/>	ImmunoVerse_data	-	-	-	Feb. 06
<input checked="" type="checkbox"/>	hg38.fa	3.05 GiB	FA	-	Feb. 06
<input checked="" type="checkbox"/>	reference				
<input checked="" type="checkbox"/>	ensembl_protein.fasta	11.36 MiB	FASTA	-	Feb. 06
<input checked="" type="checkbox"/>	reference				
<input checked="" type="checkbox"/>	gencode.v36.annotation.gtf	1.29 GiB	GTF	-	Feb. 06
<input checked="" type="checkbox"/>	reference				
<input checked="" type="checkbox"/>	GRCh38.d1.vd1.fa	2.94 GiB	FA	-	Feb. 06
<input checked="" type="checkbox"/>	reference				
<input checked="" type="checkbox"/>	GRCh38.d1.vd1.gencode.v36.annotation.star-fusion-1.12.0-CTAT-index-archive.tar	44.87 GiB	TAR	12f76e64-e489...	Feb. 06
<input checked="" type="checkbox"/>	reference				
<input checked="" type="checkbox"/>	uniprot_reviewed_curated_addition.fasta	14.20 MiB	FASTA	-	Feb. 06
<input checked="" type="checkbox"/>	reference				
<input checked="" type="checkbox"/>	homo_sapiens_vep_112_GRCh38.tar.gz	25.46 GiB	TAR.GZ	-	Feb. 06
<input checked="" type="checkbox"/>	reference				
<input checked="" type="checkbox"/>	hg38_rmsk_TE.gtf.locind	883.82 MiB	LOCIND	-	Feb. 06
<input checked="" type="checkbox"/>	reference				

15 items

# 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

DRAFT gene\_pipeline run - 02-06-25 18:49:10 ⚡

Last update by li2g2uc on Feb. 6, 2025 13:49

App: gene\_pipeline - Revision: 6

Task Inputs Execution Settings

Inputs

- Batching Off
- ensembl\_protein Change selection
  - ensembl\_protein.fasta
- fastq\_gz\_files Change selection
  - 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
- kallisto\_index Change selection
  - kallisto\_index
- nuorf Change selection
  - nuorf.fasta
- uniprot\_isoform Change selection
  - uniprot\_reviewed\_curated\_addition.fasta

App Settings

Output Settings

Parameter	Value	Notes
cores	3	Keep it
outdir	.	
strand	no	Whether your library is stranded or not, no is always safe if you don't know
gene_fasta	No value	
isoform_fasta	No value	
nuorf.fasta	No value	
tpm_result	No value	

Consistent with number of samples

Remember, select file in the order, sample1\_R1, sample1\_R2, sample2\_R1, sample2\_R2...

# Gene expression

Task Inputs    Execution Settings 

**Spot Instances** Off 

Spot instances can significantly reduce the cost of your task execution if results are not needed urgently.

[Learn more](#)

---

**Memoization (WorkReuse)** Off 

Automatic reuse of precomputed results can significantly reduce the time and cost of your task execution.

[Learn more](#)

---

**Elastic Disk BETA** Off 

Automatic extension of attached disk space will allow task execution to continue if the original disk capacity becomes insufficient.

[Learn more](#)

---

**Instance type**

**App default** Instance not defined, will be automatically selected 

**Custom** Select an instance from the list 

This setting overrides the instance set by the app developer and the instance selection from any previous run of this task. [Learn more](#).

Instance: r5.4xlarge (16vCPUs, 128GB RAM)   4096 

Attached storage (GB) 

Price: \$1.008 per hour

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

# Gene expression

Dashboard Files Apps Tasks Data Studio Interactive Apps **deploy\_pipelines** ⓘ Interactive Browsers Settings Notes

DRAFT gene\_pipeline run - 02-06-25 18:49:10 ⚙️ Last update by li2g2uc on Feb. 6, 2025 13:49 App: gene\_pipeline - Revision: 6

Task Inputs Execution Settings

Inputs

- ensembl\_protein
  - ensembl\_protein.fasta
- fastq\_gz\_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
  - HN21-10181\_R2.fastq.gz
  - ...and 1 more item
- kallisto\_index
  - kallisto\_index
- nuorf
  - nuorf.fasta
- uniprot\_isoform
  - uniprot\_reviewed\_curated\_addition.fasta

App Settings

- cores
- outdir
- strand

Show non-default ⓘ Output Settings

- 3
- gene\_fasta
  - HN19-9674\_gene.fasta
  - HN20-9844\_gene.fasta
  - HN21-10181\_gene.fasta
- isoform\_fasta
  - HN19-9674\_isoform.fasta
  - HN20-9844\_isoform.fasta
  - HN21-10181\_isoform.fasta
- nuorf\_fasta
  - HN19-9674\_nuorf.fasta
  - HN20-9844\_nuorf.fasta
  - HN21-10181\_nuorf.fasta
- tpm\_result
  - HN19-9674\_gene\_tpm.txt
  - HN20-9844\_gene\_tpm.txt
  - HN21-10181\_gene\_tpm.txt

Once finishing setup, click run

Once done, you will have fasta (canonical protein sequences with expressed gene), isoform fasta (isoform protein that are expressed), nuorf fasta (cryptic orfs), tpm (gene to tpm in each sample)

# Alignment

Copy this pipeline

The screenshot shows a pipeline run interface with the following details:

- Completed Pipeline Run:** alignment\_pipeline run - 01-25-25 18:19:27
- Inputs:**
  - 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
  - sequence:
    - GRCh38.d1.vd1.fa
  - star\_index:
    - star\_hg38\_index- App Settings:** cores, outdir
- Output Settings:** 3 samples, each producing a .bam and .bai file. The samples are labeled HN19-9674, HN20-9844, and HN21-10181.

Annotations in red text and arrows:

- An arrow points to the completed pipeline run title with the text: "30GB RAM for each sample, so you need an instance with more than 3 cores and 90GB RAM, will done in 4 hours".
- An arrow points to the "fastq\_files" input section with the text: "Remember, select file in the order, sample1\_R1, sample1\_R2, sample2\_R1, sample2\_R2...".
- An arrow points to the "Output Settings" section with the text: "Consistent with number of samples".
- An arrow points to the "bam" output section with the text: "Keep it".

# Splicing and intron retention

COMPLETED splicing\_intron\_pipeline run - 01-25-25 22:32:13

Get support

View stats & logs

Edit and rerun

Executed on Jan. 25, 2025 17:35 by li2g2uc

Spot Instances: Off

Memoization (WorkReuse): Off

Price: \$1.97

Duration: 41 minutes

App: splicing\_intron\_pipeline - Revision: 0

Each bam take 2GB RAM, you don't need 40 cores,  
in this case, three files I will select 3 cpus, almost  
any instance should work, done in 30 min

## Inputs

### bam\_files

HN19-9674\_secondAligned.sortedByCoord.out.bam  
HN20-9844\_secondAligned.sortedByCoord.out.bam  
HN21-10181\_secondAligned.sortedByCoord.out.bam

### gene\_model

gene\_model.txt

### reference

gencode.v36.annotation.gtf

### sequence

hg38.fa

## App Settings

Show non-default

cores

40

outdir

.

strand

no

## Output Settings

### intron\_peptide

HN19-9674\_secondAligned.sortedByCoord.out\_intron\_peptid...  
HN20-9844\_secondAligned.sortedByCoord.out\_intron\_peptid...  
HN21-10181\_secondAligned.sortedByCoord.out\_intron\_pepti...

### intron\_result

HN19-9674\_secondAligned.sortedByCoord.out\_intron.txt  
HN20-9844\_secondAligned.sortedByCoord.out\_intron.txt  
HN21-10181\_secondAligned.sortedByCoord.out\_intron.txt

### splicing\_result

HN19-9674\_secondAligned.sortedByCoord.out\_splicing.txt  
HN20-9844\_secondAligned.sortedByCoord.out\_splicing.txt  
HN21-10181\_secondAligned.sortedByCoord.out\_splicing.txt



# 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 | Price: \$1.13 | Duration: 40 minutes

▼ 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

COMPLETED telocal\_pipeline run - 01-25-25 22:35:57

Executed on Jan. 25, 2025 17:37 by li2g2uc

Spot Instances: Off | Memoization (WorkReuse): Off | Price: \$10.79 | Duration: 4 hours, 12 minutes

App: telocal\_pipeline - Revision: 0

Each bam take 10GB RAM, if you set cores=3, then you need 30GB RAM, each finishes in 5h

Inputs	App Settings	Output Settings
bam_files	cores: 20	output: HN19-9674_secondAligned.sortedByCoord.out_TElocal_out.c...
	outdir: .	HN20-9844_secondAligned.sortedByCoord.out_TElocal_out.c...
	strand: no	HN21-10181_secondAligned.sortedByCoord.out_TElocal_out....
te_local_gene		
hg38.knownGene.gtf		
te_local_te		
hg38_rmsk_TE.gtf.locInd		

# Gene fusion

BATCH 3 STAR-Fusion run - 02-06-25 19:11:07 Last update by li2g2uc on Feb. 6, 2025 14:11 App: STAR-Fusion - Revision: 0

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

Task Inputs Execution Settings

Inputs

Batching  On

Input files \*

Batch by: File metadata

This task will be batched by file metadata (Sample ID) and this will create 3 groups.

Hn19 (2 items)   
Hn20 (2 items)   
Hn21 (2 items)

CTAT genome lib archive

Batch by: None

GRCh38.d1.vd1.gencode.v36.annotation.star-fusion-1.1...

Execution Settings

App Settings

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

Output Settings

Fusion predictions

Fusion predictions abridged

FusionInspector HTML fusions summary

FusionInspector fusion predictions

STAR-Fusion output archive

Chimeric read filtering parameters: Min non-specific multimapping read percentage

Downstream analysis of fusion candidates: Denovo reconstruct

Downstream analysis of fusion candidates: Examine coding effect

Downstream analysis of fusion candidates: Extract

Get support Discard Run

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

Root Exclude subfolders

2 items selected Copy Rename Move Metadata Edit tags Download ...

New Folder Add files ...

Name	Path	Size	Extension	Task II
<input checked="" type="checkbox"/> HN19-9674_R1.fastq.gz Input	Files	1.92 GiB	FASTQ.GZ	-
<input checked="" type="checkbox"/> HN19-9674_R2.fastq.gz Input	Files	1.97 GiB	FASTQ.GZ	-
<input type="checkbox"/> HN20-9844_R1.fastq.gz Input	Files	1.60 GiB	FASTQ.GZ	-
<input type="checkbox"/> HN20-9844_R2.fastq.gz Input	Files	1.64 GiB	FASTQ.GZ	-
<input type="checkbox"/> HN21-10181_R1.fastq.gz Input	Files	1.79 GiB	FASTQ.GZ	-
<input type="checkbox"/> HN21-10181_R2.fastq.gz Input	Files	1.83 GiB	FASTQ.GZ	-

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

Dashboard   Files   PREMIUM   Apps   Tasks   Data Studio

Root   Exclude subfolders

2 items selected   Copy   Rename   Move

Name

- HN19-9674\_R1.fastq.gz   input
- HN19-9674\_R2.fastq.gz   input
- HN20-9844\_R1.fastq.gz   input
- HN20-9844\_R2.fastq.gz   input
- HN21-10181\_R1.fastq.gz   input
- HN21-10181\_R2.fastq.gz   input

**Update metadata values**

Luminary  Enter value

Primary site  Enter value

Disease type  Enter value

Age at diagnosis  Enter value

Vital status  Enter value

Days to death  Enter value

Sample ID  HN19

Sample UUID  Enter value

Sample type  Enter value

Aliquot ID  Enter value

Aliquot UUID  Enter value

**Custom metadata**

sbg\_public\_files\_category  Enter value

species  Enter value

HN19 HN19

Cancel   Save

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

The screenshot shows a file manager interface with a dark theme. On the left, a list of files is displayed under the 'Root' directory. Several files are selected, indicated by a checked checkbox next to each file name. Red arrows point from the text 'By doing that, when you supply all fastq.gz files, the program will figure out how to pair, and how to parallelize' to the selected files. In the center, a modal window titled 'Update metadata values' is open. The window contains a warning message: '⚠ You can edit metadata only on files. Read more'. Below this is a 'Metadata schema' section with various fields and dropdown menus. One dropdown menu for 'Paired-end' has the value '1' selected, which is also highlighted with a red arrow. At the bottom right of the modal are 'Cancel' and 'Save' buttons.

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)

COMPLETED **variant\_pipeline run - 01-26-25 16:14:21** ↗

Executed on Jan. 26, 2025 11:15 by li2g2uc

Each file only takes 2GB ram, so based on this to select the instance, will finish in 2h

Spot Instances: Off ⓘ | Memoization (WorkReuse): Off ⓘ | Price: \$2.60 ⓘ | Duration: 1 hour, 40 minutes ⓘ

App: variant\_pipeline - Revision: 2

**Inputs** ↗

- bam\_files ↗
  - HN19-9674\_secondAligned.sortedByCoord.out.bam
  - HN20-9844\_secondAligned.sortedByCoord.out.bam
  - HN21-10181\_secondAligned.sortedByCoord.out.bam
- sequence ↗
  - GRCh38.d1.vd1.fa

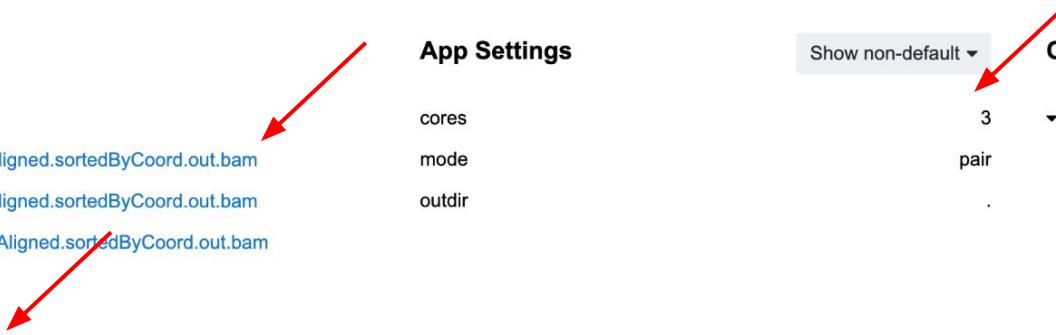
**App Settings**

Show non-default ▾

cores	3
mode	pair
outdir	.

**Output Settings** ↗

- vcf ↗
  - HN19-9674\_variants.vcf
  - HN20-9844\_variants.vcf
  - HN21-10181\_variants.vcf



# Variants (step 2 is to run variant effect predictor)

COMPLETED Variant Effect Predictor run - 01-26-25 20:40:30

Executed on Jan. 26, 2025 15:40 by ll2g2uc

Spot Instances: Off | Memoization (WorkReuse): Off | Price: \$0.43 | Duration: 27 minutes

App: Variant Effect Predictor - Revision: 0

Inputs	App Settings	Output Settings
Chromosome synonyms	Add 1000 genomes phase 3 global allele frequency	False Compressed (bgzip/gzip) output
No files selected	Add APPRIS identifiers	False Optional file with VEP warnings and errors
Custom annotation - BigWig sources only	Add CCDS transcript identifiers	False HN21-10181_variants.vep.vcf_warnings.txt
No files selected	Add Ensembl protein identifiers	False Output summary stats file
Custom annotation sources	Add GA4GH Variation Representation Specification	False HN21-10181_variants.vep.vcf_summary.html
No files selected	Add HGVS identifiers	False VEP output file
Fasta file(s) to use to look up reference sequence	Add MANE Select identifiers	False HN21-10181_variants.vep.vcf
No files selected	Add MANE Select or MANE Plus Clinical identifiers	False
GFF annotation file	Add UniProt-associated database identifiers	False
No files selected	Add a flag indicating if the transcript is canonical	False
GTF annotation file	Add allele frequency from continental 1000 genomes populations	False
No files selected	Add biotype of transcript or regulatory feature	False
Input VCF	Add cDNA, CDS and protein positions (position/length)	False
HN21-10181_variants.vcf	Add gene symbols where available	False
NCBI BAM file for correcting transcript models	Add genomic HGVS identifiers	False
No files selected	Add gnomAD allele frequencies	False
Optional config file	Add gnomAD allele frequencies from genome populations	False
No files selected	Add miRNA report	False
Species cache file	Add reference allele in the output	False
homo_sapiens_vep_112_GRCh38.tar.gz	Add transcript support level	False
dbNSFP database file	Add transcript version	False

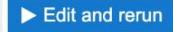
You can either just run one vcf at a time or batch by file

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

Red arrows point to the "Input VCF" and "Species cache file" sections.

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

t..

**COMPLETED** **SBG Decompressor CWL1.0 run - 01-27-25 21:29:01: file: HN21-10181\_R1.f...** 

Executed on Jan. 27, 2025 16:32 by [li2g2uc](#)

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

App: SBG Decompressor CWL1.0 - Revision: 0

Inputs	App Settings	Show non-default	Output Settings
 <b>Input Archive File</b>   HN21-10181_R1.fastq.gz	 Flatten Outputs 	False	 <b>Output Files</b>   HN21-10181_R1.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)

..

**COMPLETED OptiType run - 01-27-25 22:09:29: sample\_id: HN19**

Executed on Jan. 27, 2025 17:10 by li2g2uc

Spot Instances: Off | Memoization (WorkReuse): Off | Price: \$0.12 | Duration: 13 minutes

App: OptiType - Revision: 0

Inputs

Input file(s) Type of data

HN19-9674\_R2.fastq  
HN19-9674\_R1.fastq

Either c4.x2large or c5.4xlarge or r5.16xlarge, you can specify the pair and sample and batch by file metadata

Will be done in 20mins

Show non-default

Output Settings

rna

Config output  
HN19\_config.ini

Coverage plot  
HN19.coverage\_plot.pdf

HLA 4-digits results  
HN19.result.tsv

HLA 8-digits results  
HN19.result\_type.tsv

HLA Types

Get support | View stats & logs | Edit and rerun



Will be done in 20mins

HN19-9674\_R2.fastq

HN19-9674\_R1.fastq

HN19\_config.ini

HN19.coverage\_plot.pdf

HN19.result.tsv

HN19.result\_type.tsv

HN19.Types

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

The screenshot shows a file management interface with the following details:

- Header:** Dashboard, Files PREMIUM, Apps, Tasks, Data Studio, Interactive Apps, deploy\_pipelines (with an info icon), Interactive Browsers, Settings, Notes.
- Breadcrumbs:** Root / result (highlighted with a red arrow).
- File List:** 75 items (All selected). The table includes columns: Name, Size, Extension, Task ID, and Created.
- Actions:** Copy, Rename, Move, Metadata, Edit tags, Download, and a three-dot menu.
- Buttons:** New Folder, Add files, and a three-dot menu.

<input checked="" type="checkbox"/>	Name	Size	Extension	Task ID	Created
<input checked="" type="checkbox"/>	HN19-9674_gene.fasta	10.75 MiB	FASTA	add7309e-434...	Jan. 14, 2024
<input checked="" type="checkbox"/>	HN19-9674_gene_tpm.txt	1.38 MiB	TXT	add7309e-434...	Jan. 14, 2024
<input checked="" type="checkbox"/>	HN19-9674_isoform.fasta	5.69 MiB	FASTA	add7309e-434...	Jan. 14, 2024
<input checked="" type="checkbox"/>	HN19-9674_nuorf.fasta	16.18 MiB	FASTA	add7309e-434...	Jan. 14, 2024
<input checked="" type="checkbox"/>	HN19-9674_R1.fastq	9.39 GiB	FASTQ	88fddae5-1d1e...	Jan. 27, 2024
<input checked="" type="checkbox"/>	HN19-9674_R2.fastq	9.39 GiB	FASTQ	56da2aa1-53fc...	Jan. 27, 2024
<input checked="" type="checkbox"/>	HN19-9674_secondAligned.sortedByCoord.out.bam	3.94 GiB	BAM	a198756c-09ce...	Jan. 25, 2024
<input checked="" type="checkbox"/>	HN19-9674_secondAligned.sortedByCoord.out.bam.bai	3.96 MiB	BAI	a198756c-09ce...	Jan. 25, 2024
<input checked="" type="checkbox"/>	HN19-9674_secondAligned.sortedByCoord.out_intron.txt	295.28 KiB	TXT	ba97093b-cc2...	Jan. 25, 2024
<input checked="" type="checkbox"/>	HN19-9674_secondAligned.sortedByCoord.out_intron_peptide.txt	989.38 KiB	TXT	ba97093b-cc2...	Jan. 25, 2024
<input checked="" type="checkbox"/>	HN19-9674_secondAligned.sortedByCoord.out_splicing.txt	7.65 MiB	TXT	ba97093b-cc2...	Jan. 25, 2024

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

COMPLETED **summarization\_pipeline run - 02-04-25 19:22:35** ↗

Executed on Feb. 4, 2025 14:22 by li2g2uc

Spot Instances: Off | Memoization (WorkReuse): Off | Price: \$0.03 | Duration: 48 minutes

App: summarization\_pipeline - Revision: 6

**Inputs** ↗ Take 1h to finish

- db ↗
- ImmunoVerse\_data ↗
- intdir ↗
- result ↗

**App Settings**

Show non-default ↗

outdir

**Output Settings** ↗ Keep it

fastas ↗

- HN19-9674\_Abelson\_murine\_leukemia\_virus\_UP000147198.fasta
- HN19-9674\_Kirsten\_murine\_sarcoma\_virus\_UP000242176.fasta
- HN19-9674\_Mus\_musculus\_mobilized\_endogenous\_polytropic...
- HN19-9674\_TE\_self\_translate.fasta
- HN19-9674\_intron.fasta

...and 20 more items

1 core, 100GB should be safe

Till now, you had sample-specific aberrations and search space (fastas)

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

COMPLETED **maxquant\_pipeline run - 02-04-25 23:22:42**

Executed on Feb. 4, 2025 18:23 by li2g24 | Get support | View stats & logs | Edit and rerun

Always use instance with more than 20 cores and 100GB RAM

Spot Instances: Off | Memoization (Off) | Duration: 8 hours, 57 minutes

App: maxquant\_pipeline - Revision: 4

Take about 5-24 hours

**Inputs**

- fasta\_dir
  - test.fasta
- immuno\_dir
  - test\_immuno

**App Settings**

- hla\_class
- outdir
- peptide\_fdr
- sample\_run\_name
- technology

Show non-default **Output Settings**

- 1 output\_same\_immuno\_dir
  - hg19\_maxquant\_combined.txt
- 1 hg19
- 1 orbitrap

A folder with all maxquant tabular results

You can use specific fdr like 0.01 or 0.05, or if you want to do rescore later, it requires whole PSM lists (so fdr=1)

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

COMPLETED **msconvert\_pipeline run - 02-05-25 14:51:27** 

Executed on Feb. 5, 2025 09:51 by li2g2uc

Spot Instances: Off  | Memoization (WorkReuse): Off  | Price: \$0.24  | Duration: 8 minutes 

App: msconvert\_pipeline - Revision: 2

C4.2xlarge should be fine, 10mins

**Inputs**   
raw\_file   
20240110\_E\_OdinLC\_IC\_PDX\_HD\_19.raw 

**App Settings**  
outdir 

Show non-default 

**Output Settings**   
mzml   
20240110\_E\_OdinLC\_IC\_PDX\_HD\_19.mzML

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

COMPLETED rescore\_pipeline run - 06-15-25 19:39:29 ↗ Find an instance with more than 64GB RAM should be safe

Executed on June 15, 2025 15:40 by li2g2uc

Spot Instances: Off ⓘ | Memoization (WorkReuse): Off ⓘ | Price: \$0.06 ⓘ | Duration: 3 minutes ⓘ

App: rescore\_pipeline - Revision: 20

Inputs	App Settings	Output Settings
input1 ⓘ	mode ⓘ	rescore
hg19_maxquant_combined_txt	run_rescore ⓘ	output_mirror_plot
input2 ⓘ	technology ⓘ	False
test_mzml	orbitrap	output_ms2pip_prediction
		▼ output_rescore ⓘ
		msmsScans_new.txt

Inputs:

- input1: hg19\_maxquant\_combined\_txt
- input2: test\_mzml

Output Settings:

- rescore
- output\_mirror\_plot
- False
- output\_ms2pip\_prediction
- ▼ output\_rescore: msmsScans\_new.txt

Red arrows point from the "input1" and "input2" sections in the "Inputs" list to the corresponding "hg19\_maxquant\_combined\_txt" and "test\_mzml" files respectively.

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

# Immunopeptidome analysis (Step 4 is the HLA binding prediction)

COMPLETED **hla\_binding\_pipeline run - 02-06-25 19:08:42**

Executed on Feb. 6, 2025 14:09 by li2g2uc

Spot Instances: Off | Memoization (WorkReuse): Off | Price: \$0.08 | Duration: 2 minutes

App: hla\_binding\_pipeline - Revision: 3

**Inputs**   
hla\_type   
test\_hla\_type.txt  
rescored.txt   
test\_msmsScans\_new.txt

**App Settings**

Setting	Value
hla_class	1
outdir	.
sample_run_name	test

**Output Settings** Unknown file name

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 [🔗](#)

Executed on June 15, 2025 16:02 by li2g2uc

Spot Instances: Off [?](#) | Memoization (WorkReuse): Off [?](#) | Price: \$0.05 [?](#) | Duration: 5 minutes [?](#)

App: rescore\_ms2pip - Revision: 2

Any instance could work, lightweight

**Inputs** [📄](#)

- input1 [?](#) [📄](#)
  - peptides.csv
- input2 [?](#)
  - No files selected

**App Settings**

mode [?](#)

ms2pip\_intensity

**Show non-default ▾**

**Output Settings** [📄](#)

output\_mirror

**output\_ms2pip\_prediction** [📄](#)

- ms2pip\_prediction\_NEVTTEIRF\_2.png
- ms2pip\_prediction\_NEVTTEIRF\_2.tsv
- ms2pip\_prediction\_VTYNYPVHY\_2.png
- ms2pip\_prediction\_VTYNYPVHY\_2.tsv

**Inputs** [📄](#)

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

**csv file**

**App Settings**

mode [?](#)

ms2pip\_intensity

**Show non-default ▾**

**Output Settings** [📄](#)

output\_mirror

**output\_ms2pip\_prediction** [📄](#)

- ms2pip\_prediction\_NEVTTEIRF\_2.png
- ms2pip\_prediction\_NEVTTEIRF\_2.tsv
- ms2pip\_prediction\_VTYNYPVHY\_2.png
- ms2pip\_prediction\_VTYNYPVHY\_2.tsv

**Files**

ms2pip\_prediction\_NEVTTEIRF\_2.png

25.2 kB (25,774 bytes) - Produced on June 15, 2025 16:07 (Eastern Daylight Time), by rescore\_ms2pip run.

Metadata Raw View Preview

**Figure**

MS<sup>2</sup>PIP prediction for NEVTTEIRF/2

Intensity

m/z

# 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

mass\_analyzer

mode

scan

Show non-default

Output Settings

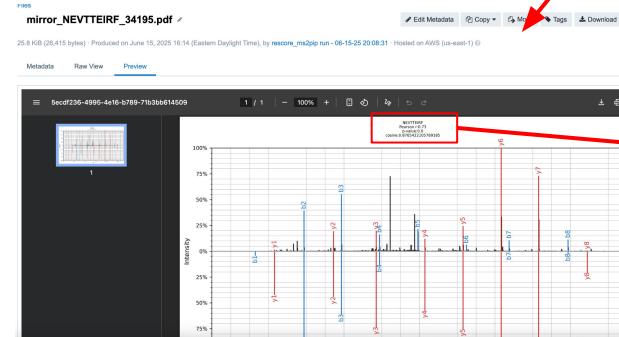
output\_mirror

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