# README

Jose Angel Sanchez Gomez

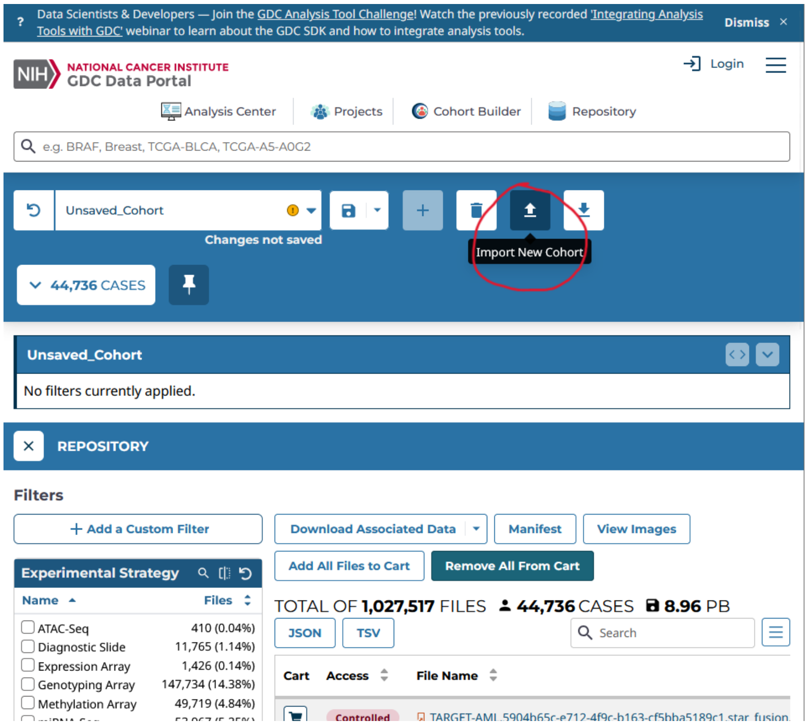
2024-10-01

INSTRUCTIONS FOR RECOVERING CLINICAL DATA:

0) Download the github repository “HubEstimationCodeSubmission” into your local environment.

1) Access the TCGA online data repository: https://portal.gdc.cancer.gov/analysis_page?app=Downloads

2) Select the option “Import New Cohort.”



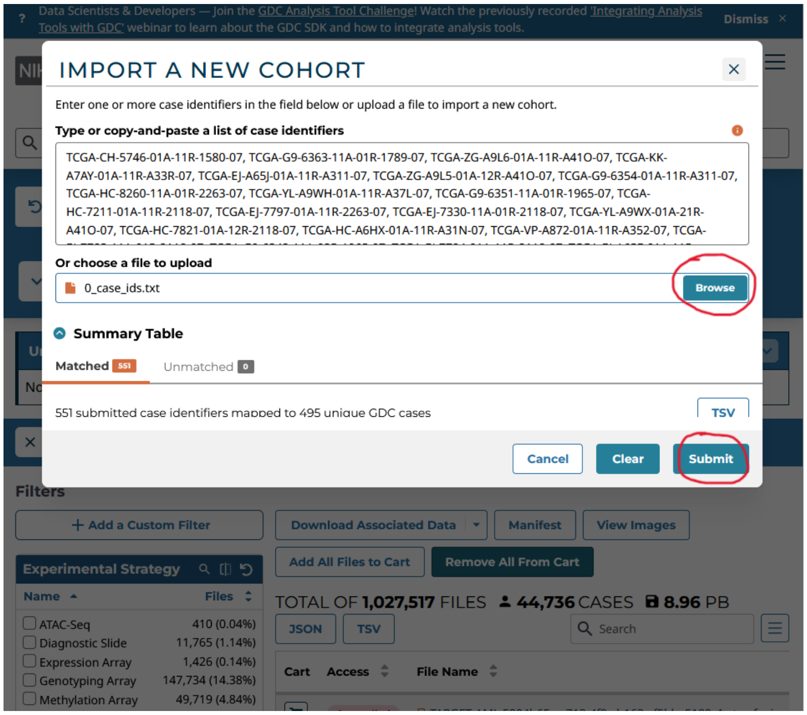
Data Scientists & Developers — Join the GDC Analysis Tool Challenge! Watch the previously recorded 'Integrating Analysis Tools with GDC' webinar to learn about the GDC SDK and how to integrate analysis tools.

No filters currently applied.

TOTAL OF 1,027,517 FILES 44,736 CASES 8.96 PB

| Name | Files |
|---|---|
| □ ATAC-Seq | 410 (0.04%) |
| □ Diagnostic Slide | 11,765 (1.14%) |
| □ Expression Array | 1,426 (0.14%) |
| □ Genotyping Array | 147,734 (14.38%) |
| □ Methylation Array | 49,719 (4.84%) |

3) Select the option “Choose a file to upload” and browse through your directories to upload the file “./realdata_4/200_ProcessingClinicalData/200_case_ids.txt”



Data Scientists & Developers — Join the GDC Analysis Tool Challenge! Watch the previously recorded 'Integrating Analysis Tools with GDC' webinar to learn about the GDC SDK and how to integrate analysis tools.

Enter one or more case identifiers in the field below or upload a file to import a new cohort.

Type or copy-and-paste a list of case identifiers

TCGA-CH-5746-01A-11R-1580-07, TCGA-G9-6363-11A-01R-1789-07, TCGA-ZG-A9L6-01A-11R-A41O-07, TCGA-KK-A7AY-01A-11R-A33R-07, TCGA-EJ-A65J-01A-11R-A311-07, TCGA-ZG-A9L5-01A-12R-A41O-07, TCGA-G9-6354-01A-11R-A311-07, TCGA-HC-8260-11A-01R-2263-07, TCGA-YL-A9WH-01A-11R-A37L-07, TCGA-G9-6351-11A-01R-1965-07, TCGA-HC-7211-01A-11R-2118-07, TCGA-EJ-7797-01A-11R-2263-07, TCGA-EJ-7330-11A-01R-2118-07, TCGA-YL-A9WX-01A-21R-A41O-07, TCGA-HC-7821-01A-12R-2118-07, TCGA-HC-A6HX-01A-11R-A31N-07, TCGA-VP-A872-01A-11R-A352-07, TCGA-

Or choose a file to upload

0_case_ids.txt

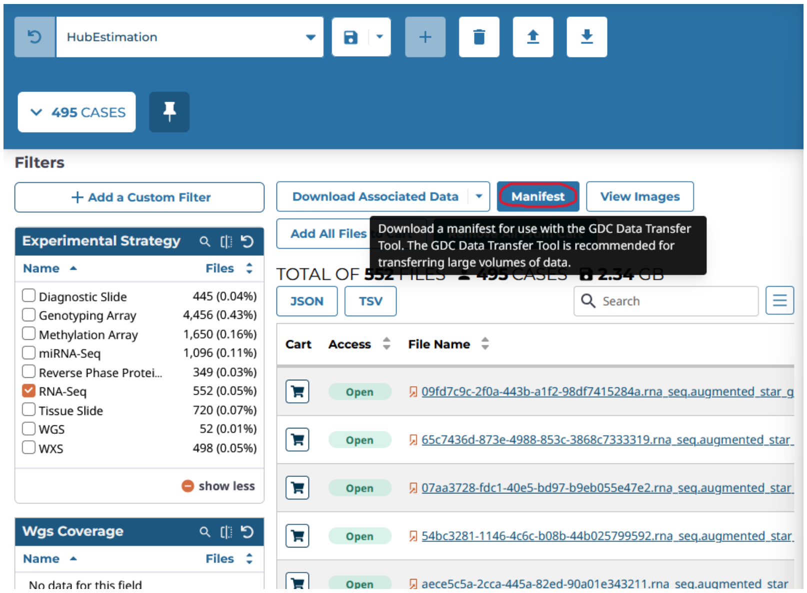
551 submitted case identifiers mapped to 495 unique GDC cases

TOTAL OF 1,027,517 FILES 44,736 CASES 8.96 PB

Experimental Strategy

| Name | Files |
|---|---|
| □ ATAC-Seq | 410 (0.04%) |
| □ Diagnostic Slide | 11,765 (1.14%) |
| □ Expression Array | 1,426 (0.14%) |
| □ Genotyping Array | 147,734 (14.38%) |
| □ Methylation Array | 49,719 (4.84%) |

4) Click on “cases” which provides “additional cases details and features.” Here, you will be provided the option to download clinical data files in “.tsv” format.



Download a manifest for use with the GDC Data Transfer Tool. The GDC Data Transfer Tool is recommended for transferring large volumes of data.

TOTAL OF 552 FILES 495 CASES 2.34 GB

Experimental Strategy

| Name | Files |
|---|---|
| □ Diagnostic Slide | 445 (0.04%) |
| □ Genotyping Array | 4,456 (0.43%) |
| □ Methylation Array | 1,650 (0.16%) |
| □ miRNA-Seq | 1,096 (0.11%) |
| □ Reverse Phase Protei... | 349 (0.03%) |
| ☑ RNA-Seq | 552 (0.05%) |
| □ Tissue Slide | 720 (0.07%) |
| □ WGS | 52 (0.01%) |
| □ WXS | 498 (0.05%) |

Wgs Coverage

| Name | Files |
|---|---|
| No data for this field | |

| Cart | Access | File Name |
|---|---|---|
| 🛒 | Open | 09fd7c9c-2f0a-443b-a1f2-98df7415284a.rna_seq.augmented_star_g |
| 🛒 | Open | 65c7436d-873e-4988-853c-3868c7333319.rna_seq.augmented_star_ |
| 🛒 | Open | 07aa3728-fdc1-40e5-bd97-b9eb055e47e2.rna_seq.augmented_star_ |
| 🛒 | Open | 54bc3281-1146-4c6c-b08b-44b025799592.rna_seq.augmented_star_ |
| 🛒 | Open | aece5c5a-2cca-445a-82ed-90a01e343211.rna_seq.augmented_star |

5) After downloading, you will have a zip file containing “clinical.tsv”, “exposure.tsv”, “family_history.tsv”, “follow_up.tsv”, and “pathology_detail.tsv”. Extract them so these files are located in the directory “./realdata_4/200_ProcessingClinicalData/”.

6) In order to use the data in the file “clinical.tsv” we first require some processing. To do this, load the file “clinical.tsv” into R using the RMarkdown “./realdata_4/200_ProcessingClinicalData/211_CleaningClinicalData.Rmd”. This file provides explanations on how the data is being cleaned and pre-processed for further analysis.

7) The resulting processed clinical data obtained through the file “211_ImportingClinicalData.Rmd” is saved in the file “./realdata_4/200_ProcessingClinicalData/212_clinical_reduced.csv”.