#7420

Optimizing proteomic data access and analysis in the cloud: Leveraging FireCloud's integration with the Proteomic Data Commons

Emily LaPlante, PhD[1], Alexander Baumann, PhD[1*],

Ratna R. Thangudu, PhD[2*], D. R. Mani, PhD[1*], Bing-Xing Huo, PhD[1]

[1]Broad Institute of MIT and Harvard,[2]ICF International
*These authors contributed equally to this work.

HIGHLIGHTS

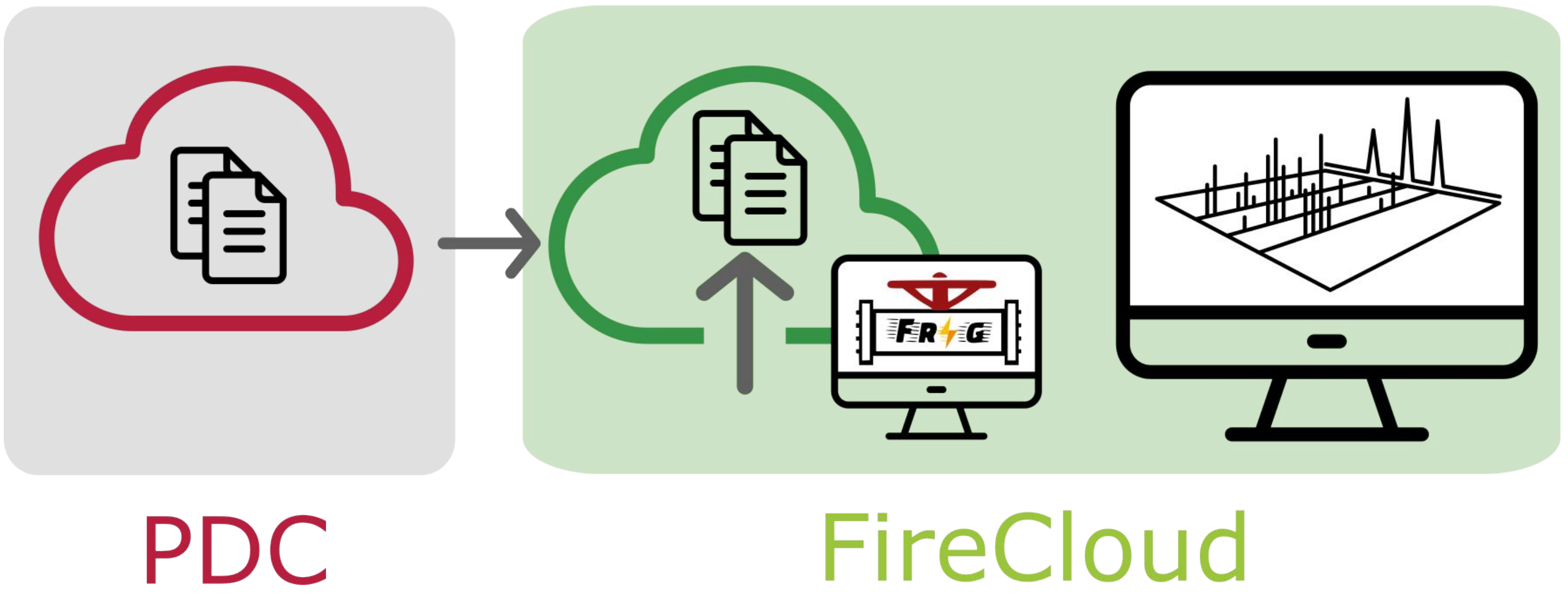
- Data from the NCI Proteomic Data Commons (PDC) can now be imported to FireCloud, the cloud workbench, for **secure sharing and scalable analysis**.
- Data and metadata are **automatically organized for FragPipe pipeline** to be run on exported data.
- **Typical workflows** for isobaric Tandem Mass Tag (TMT) or relative and absolute quantitation (iTRAQ) are available upon import. Workflows can be modified to accept any raw mass spectrometry data type.
- FireCloud hosts TARGET, TCGA, CPTAC and GDC data allowing **co-analysis of proteomics data** on the cloud.

SUMMARY

As a part of the Cancer Research Data Commons (CRDC), the cloud workbench, FireCloud, has expanded its capabilities to include accessing and analyzing the PDC data. FireCloud, powered by Terra, is a secure, scalable cloud-native platform which provides batch workflow execution, interactive analysis including data visualization, and more than 2,900 publicly available tools.

The integration of PDC and FireCloud enables researchers to leverage the data navigation and file-level search capabilities on the PDC web browser and export data to a cloud workspace. Data can then be analyzed on the cloud in pre-configured Fragpipe workflows. Workspaces can be made public to comply with the NIH Data Management and Sharing policy requirements.

GRAPHICAL ABSTRACT

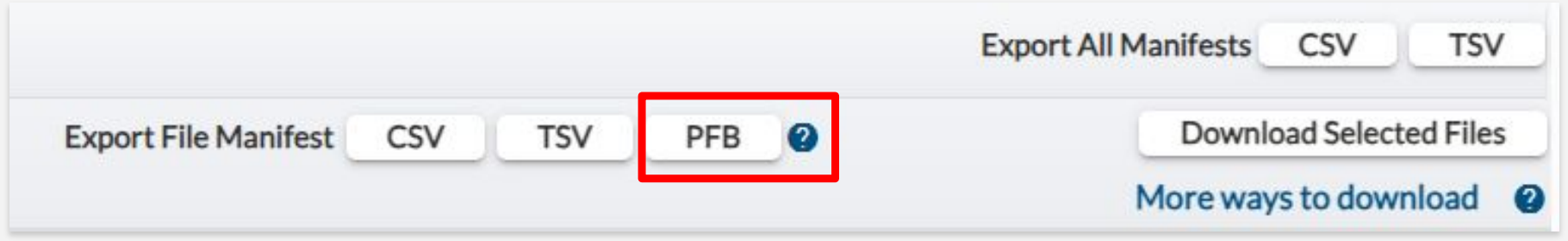


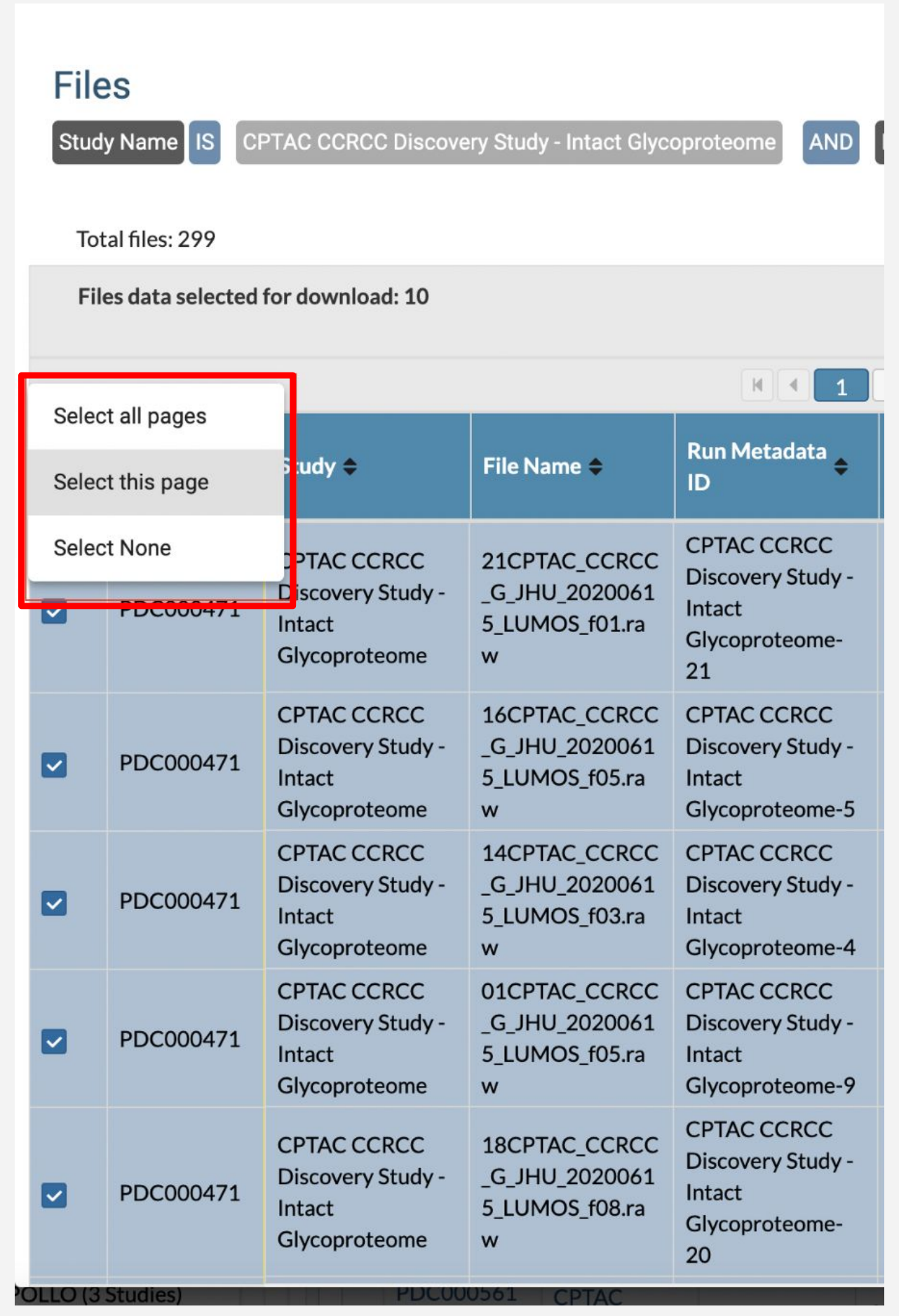
PDC primarily hosts mass spectrometry-based proteomic data from large consortia such as Clinical Proteomic Tumor Analysis Consortium (CPTAC), International Cancer Proteogenomics Consortium (ICPC), and Applied Proteogenomics Organizational Learning and Outcomes (APOLLO). Studies include proteome, phosphoproteome, glycoproteome, acetylome, and ubiquitylome data obtained using data dependent acquisition (DDA) or data independent acquisition (DIA) mass spectrometry-based approaches either by label-free or isobaric-labeling workflows using iTRAQ or TMT reagents.

Cancer proteomic data from PDC can be exported to the FireCloud analysis environment. In FireCloud, a workspace is configured to organize PDC metadata for Fragpipe and example workflows have been developed to analyze isobarically labeled MS data from Tandem Mass Tag (TMT) or Isobaric tags for relative and absolute quantitation (iTRAQ data).

PDC SUPPORTS EXPORT BY PORTABLE FORMAT FOR BIOINFORMATICS (PFB) TO FIRECLOUD

Data can be exported from a PDC study summary page or from the file tab of the data explorer. Select the PFB export button after selecting samples.



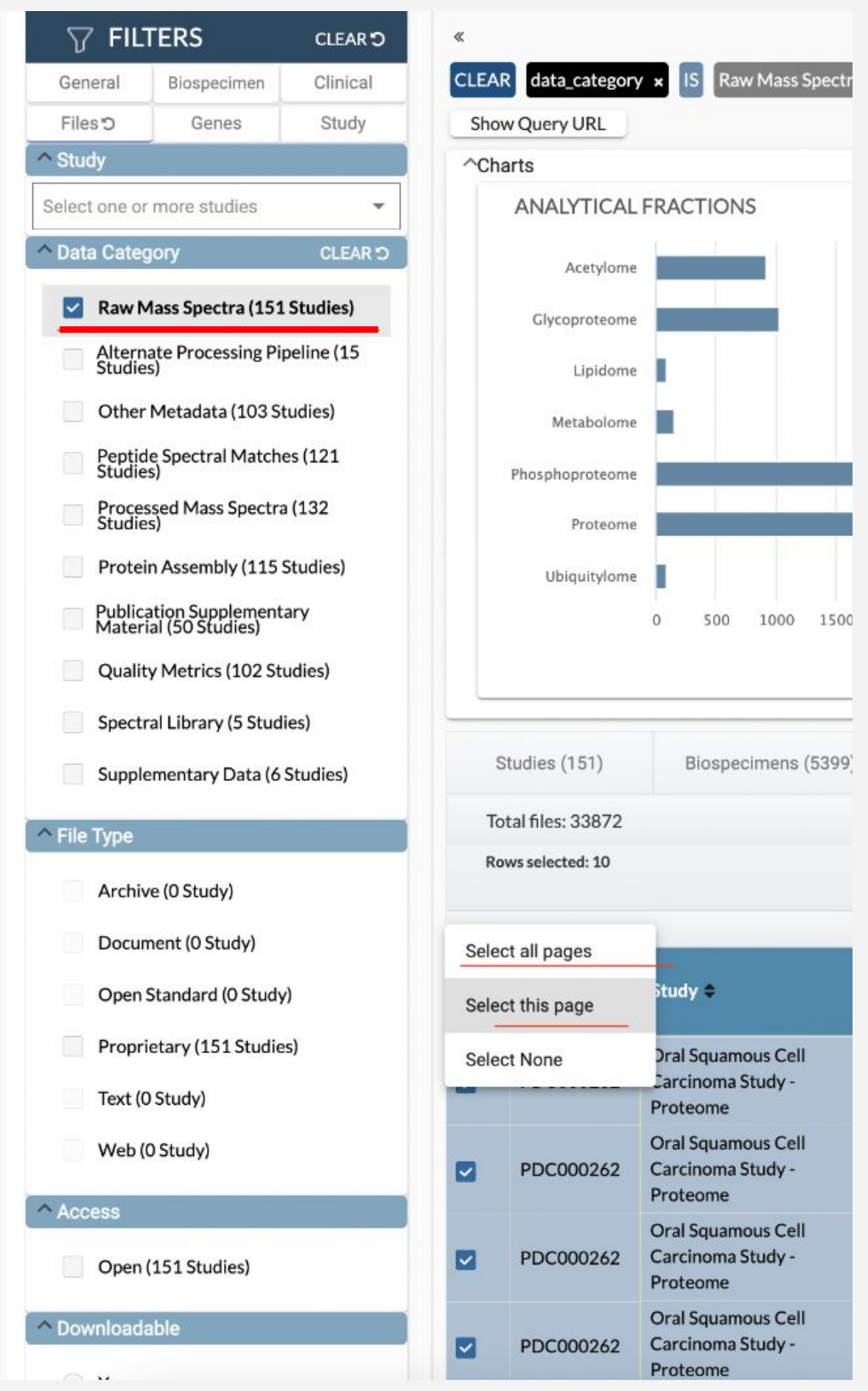

Files

Study Name IS CPTAC CCRCC Discovery Study - Intact Glycoproteome AND

Total files: 299

Files data selected for download: 10

		Study	File Name	Run Metadata ID
☑	PDC000471	CPTAC CCRCC Discovery Study - Intact Glycoproteome	21CPTAC_CCRCC_G_JHU_20200615_LUMOS_f01.raw	CPTAC CCRCC Discovery Study - Intact Glycoproteome-21
☑	PDC000471	CPTAC CCRCC Discovery Study - Intact Glycoproteome	16CPTAC_CCRCC_G_JHU_20200615_LUMOS_f05.raw	CPTAC CCRCC Discovery Study - Intact Glycoproteome-5
☑	PDC000471	CPTAC CCRCC Discovery Study - Intact Glycoproteome	14CPTAC_CCRCC_G_JHU_20200615_LUMOS_f03.raw	CPTAC CCRCC Discovery Study - Intact Glycoproteome-4
☑	PDC000471	CPTAC CCRCC Discovery Study - Intact Glycoproteome	01CPTAC_CCRCC_G_JHU_20200615_LUMOS_f05.raw	CPTAC CCRCC Discovery Study - Intact Glycoproteome-9
☑	PDC000471	CPTAC CCRCC Discovery Study - Intact Glycoproteome	18CPTAC_CCRCC_G_JHU_20200615_LUMOS_f08.raw	CPTAC CCRCC Discovery Study - Intact Glycoproteome-20



☑ Raw Mass Spectra (151 Studies)
□ Alternate Processing Pipeline (15 Studies)
□ Other Metadata (103 Studies)
□ Peptide Spectral Matches (121 Studies)
□ Processed Mass Spectra (132 Studies)
□ Protein Assembly (115 Studies)
□ Publication Supplementary Material (50 Studies)
□ Quality Metrics (102 Studies)
□ Spectral Library (5 Studies)
□ Supplementary Data (6 Studies)

□ Archive (0 Study)
□ Document (0 Study)
□ Open Standard (0 Study)
□ Proprietary (151 Studies)
□ Text (0 Study)
□ Web (0 Study)

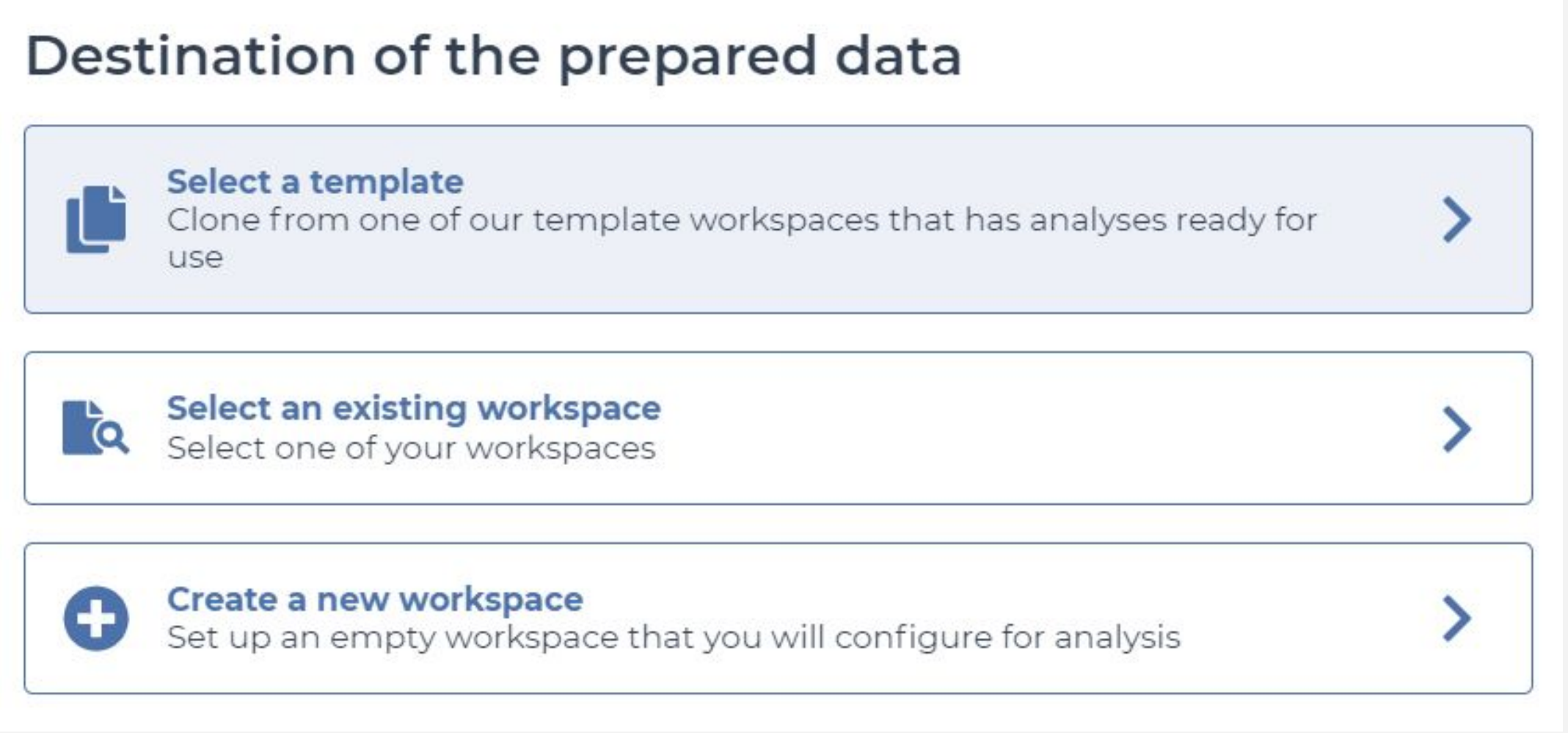
□ Open (151 Studies)

Total files: 33872

Rows selected: 10

		Study
☑	PDC000262	Oral Squamous Cell Carcinoma Study - Proteome
☑	PDC000262	Oral Squamous Cell Carcinoma Study - Proteome
☑	PDC000262	Oral Squamous Cell Carcinoma Study - Proteome

(1) Log into FireCloud (Terra) and import data into a template workspace, an existing workspace, or a new workspace.

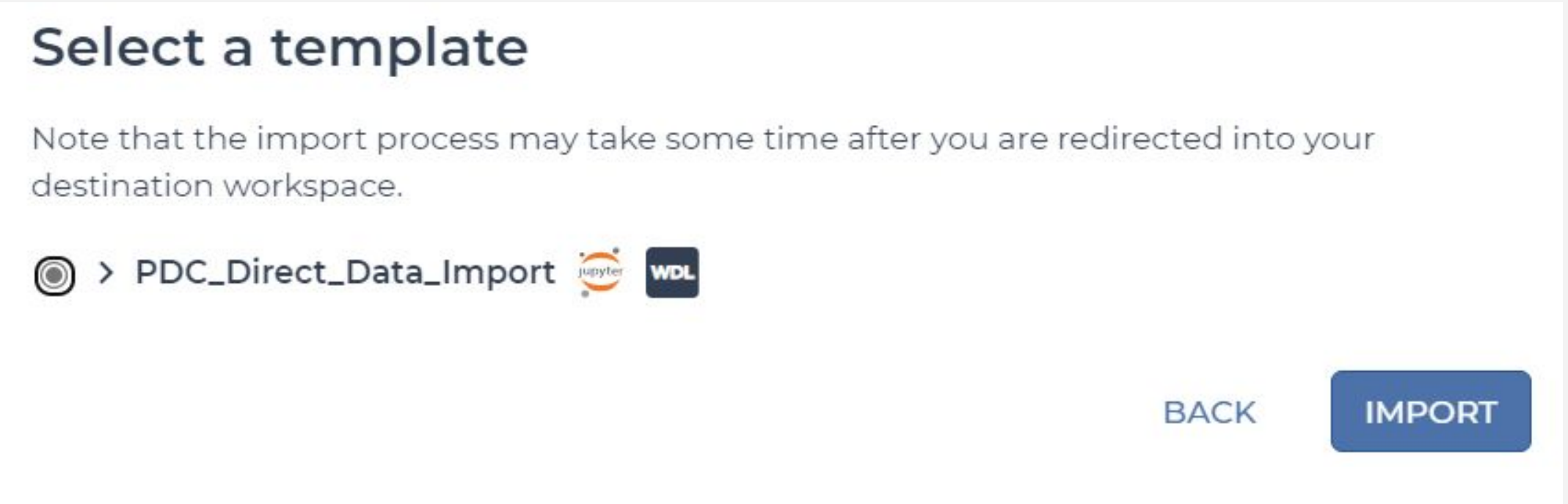

Destination of the prepared data

Select a template – Clone from one of our template workspaces that has analyses ready for use

Select an existing workspace – Select one of your workspaces

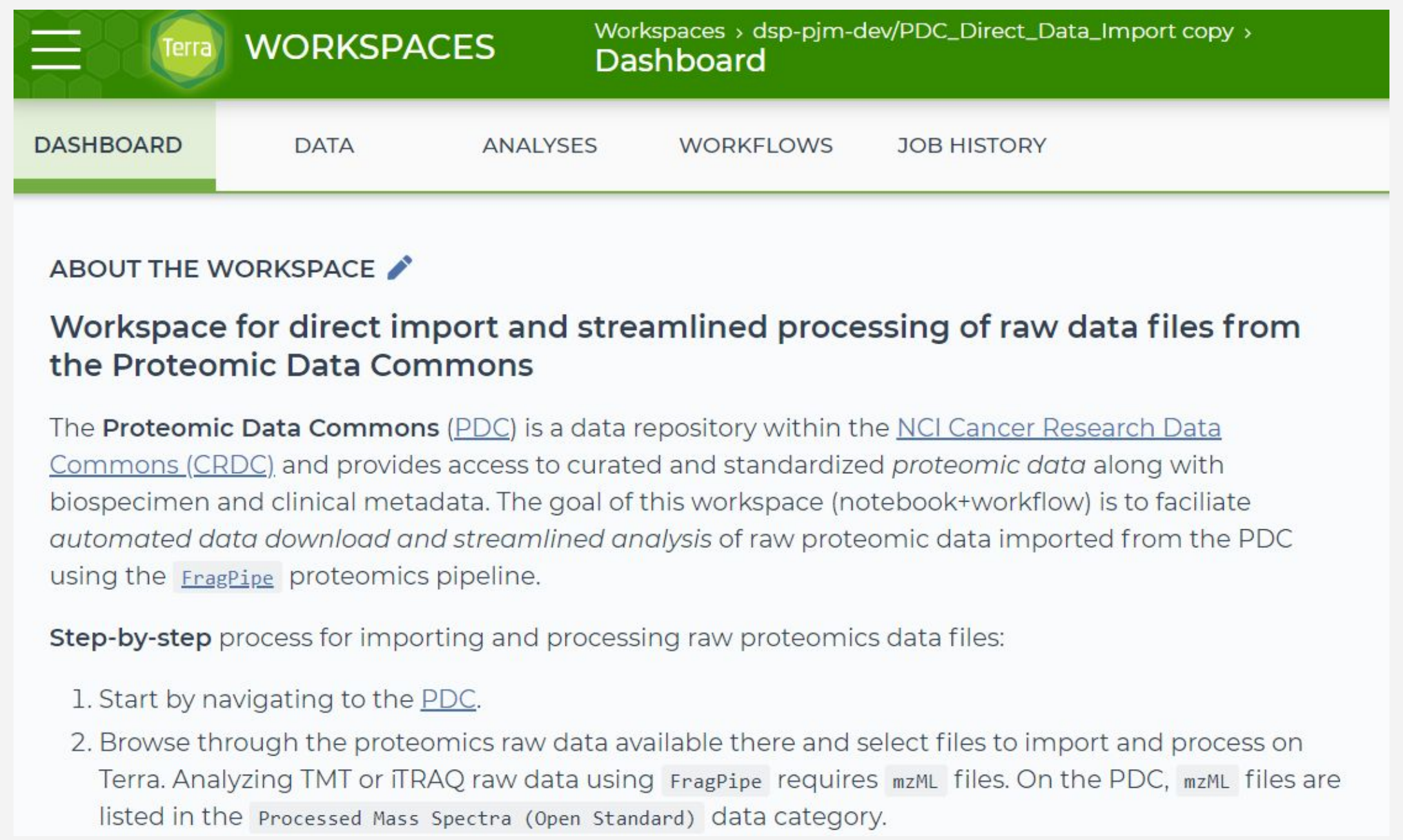
Create a new workspace – Set up an empty workspace that you will configure for analysis

A ***PDC_Direct_Data_Import* workspace template** has pre-configured proteomic analysis workflows for easy start.


Select a template

Note that the import process may take some time after you are redirected into your destination workspace.

(2) Simply follow the step by step instructions in the public ***PDC_Direct_Data_Import*** workspace Dashboard. This is available even if data is not imported from PDC.



ABOUT THE WORKSPACE

Workspace for direct import and streamlined processing of raw data files from the Proteomic Data Commons

The **Proteomic Data Commons** (PDC) is a data repository within the NCI Cancer Research Data Commons (CRDC) and provides access to curated and standardized *proteomic data* along with biospecimen and clinical metadata. The goal of this workspace (notebook+workflow) is to faciliate *automated data download and streamlined analysis* of raw proteomic data imported from the PDC using the `FragPipe` proteomics pipeline.

Step-by-step process for importing and processing raw proteomics data files:

1. Start by navigating to the PDC.
2. Browse through the proteomics raw data available there and select files to import and process on Terra. Analyzing TMT or iTRAQ raw data using `FragPipe` requires `mzML` files. On the PDC, `mzML` files are listed in the `Processed Mass Spectra (Open Standard)` data category.

(3) Using the Jupyter notebooks in the analysis tab and the WDL workflows in the Workflows tab, the instructions will guide you through running the FragPipe pipeline for standard data analysis.

Workspaces can be shared with collaborators and editors, as well as made publicly available.

CONCLUSIONS

As a part of CRDC, FireCloud, powered by Terra, provides a cloud environment for processing PDC data using any open-access or custom-made tools. Researchers can choose to import PDC data to a workspace that is already set up for processing, making preparation for downstream analysis simple. Proteomic data can be integrated with data from sources including direct upload, NCI data hosted on FireCloud (TARGET, TCGA, CPTAC), and other NCI data repositories accessible via Terra. On FireCloud, researchers have access to thousands of analysis methods in the form of workflows and tools from Dockstore, the Broad Methods Repository and featured workspaces in environments including R, Jupyter, and Galaxy. After analysis, results can be shared in a reproducible manner to meet the NIH Data Management and Sharing policy requirements.

ACKNOWLEDGEMENTS

This project is supported by National Cancer Institute contract 17X149F9 to Broad Institute Data Sciences Platform. We thank Drs. Jill S Barnholtz-Sloan, Erin Beck, Ina Felau, Shirish (Sam) Pathak for their insightful suggestions and critiques for this work.

KEY RESOURCES

PDC browser	https://pdc.cancer.gov
FireCloud	https://FireCloud.terra.bio/
PDC documentation	https://pdc.cancer.gov/pdc/cloud-data-analysis
FireCloud/Terra documentation	https://support.terra.bio/
Fragpipe documentation	https://fragpipe.nesvilab.org/