BTEP course

Center for Cancer Research

BTEP

Bioinformatics Training & Education Program

Alexandra L Emmons Ph.D. & Joe Wu Ph.D.
BTEP/GAU/CCR/NCI/NIH - email ncibtep@mail.nih.gov
Bioinformatics Training and Education Program

Table of Contents

Course overview

Partek Flow at NIH

Course Overview

Partek Flow is a start-to-finish solution for analyzing high dimensional multi-omics sequencing data. It is a point-and-click software and is suitable for those who wish to avoid the steep learning curve associated with analyzing sequencing data through command line and/or code. At NIH, Partek Flow is hosted on the Biowulf *(https://partekflow.cit.nih.gov)* high performance computing cluster (HPC). Researchers interact with the software through a web browser using a URL supplied by Biowulf once a Biowulf and Partek Flow account has been set up. This enables investigators to take advantage of the compute power offered by HPC while using a graphical user interface to construct a sequencing data analysis workflow. Partek Flow enables the creation of publication quality visualizations.

This in-person and hands-on training will introduce participants to single cell RNA sequencing (move this to beginning) analysis on Partek Flow using a single human PBMC sample count matrix. Participants will also learn how to access and methods for transferring data to the Partek Flow server at NIH. Skills acquired from this class are applicable to analyzing other types of sequencing data in Partek Flow.

Date: March 5, 2024 Time: 2 - 4 PM Location: Building 35A Room 620/630

Learning Objectives

After this class, participants will

- Become familiar analyzing single cell RNA sequencing data using Partek Flow including
    - Importing of data into a Partek Flow project
    - QA/QC, filtering, and normalizing of single cell RNA data
    - Performing cell type classification based on gene expression
    - Performing differential expression and pathway analysis
    - Producing visualization (PCA, UMAP, tSNE, dotplot, volcano plot, hierarchical clustering etc.)
- Know how to access Partek Flow
- Know how to sign onto the NIH Partek Flow server
- Be able to transfer data from NCI CCR Sequencing Facility Data Management Environment to their Biowulf Partek Flow folder

Link to documents from Partek

Class documents provided by Partek *(https://bioinformatics.ccr.cancer.gov/btep/wp-content/uploads/sites/2/PartekFlowSCTrainingHandout_MAR_2024.pdf)*

Accessing Partek Flow at NIH and tips for data transfer

Learning objectives

After consulting this guide, participants will

- Know how to access Partek Flow at NIH.
- Be able to transfer data from NCI CCR Sequencing Facility Data Management Environment to their Biowulf Partek Flow folder.

Instructions for accessing Partek Flow

NCI researchers can find instructions for accessing Partek Flow at https://bioinformatics.ccr.cancer.gov/btep/partek-flow-bulk-and-single-cell-rna-seq-data-analysis/ *(https://bioinformatics.ccr.cancer.gov/btep/partek-flow-bulk-and-single-cell-rna-seq-data-analysis/)*. But the things needed are

- A Biowulf (The High Performance Computing cluster) account — see here for information about how to obtain a HPC account *(https://hpc.nih.gov/docs/accounts.html)*.
- A /data directory on Biowulf with enough disk space to hold their Partek Flow files — please fill out this online form *(https://hpc.nih.gov/dashboard/storage_request.php)* if you do not already have a /data directory or if you require more disk space.
- A Partek Flow account created for them — please contact staff@hpc.nih.gov.

Once these steps have been accomplished, Partek Flow is available at https://partekflow.cit.nih.gov/flow *(https://partekflow.cit.nih.gov/flow)*.

The Partek Flow folder on Biowulf

HPC staff will create a folder called "PartekFlow" in the user's Biowulf data directory. This folder will hold all Partek Flow projects.

Transferring data from NCI CCR Sequencing Facility to Partek Flow on Biowulf

Those researchers who used the NCI CCR Sequencing Facility *(https://bioinformatics.ccr.cancer.gov/docs/resources-for-bioinformatics/raw_data_from_cores/)* to get sequencing done will receive a link to their data. This data can be transferred to the "PartekFlow" folder on Biowulf using Globus. The steps for setting up a Globus endpoint for the Biowulf "PartekFlow" folder can be found at https://partekflow.cit.nih.gov/#upload_globus *(https://partekflow.cit.nih.gov/#upload_globus)*. The embedded PDF shows how to connect the sequencing facility's data management environment to a Globus endpoint.

For those who have not setup a Globus account, refer to https://hpc.nih.gov/docs/globus/setup.php *(https://hpc.nih.gov/docs/globus/setup.php)* for instructions.

Tip

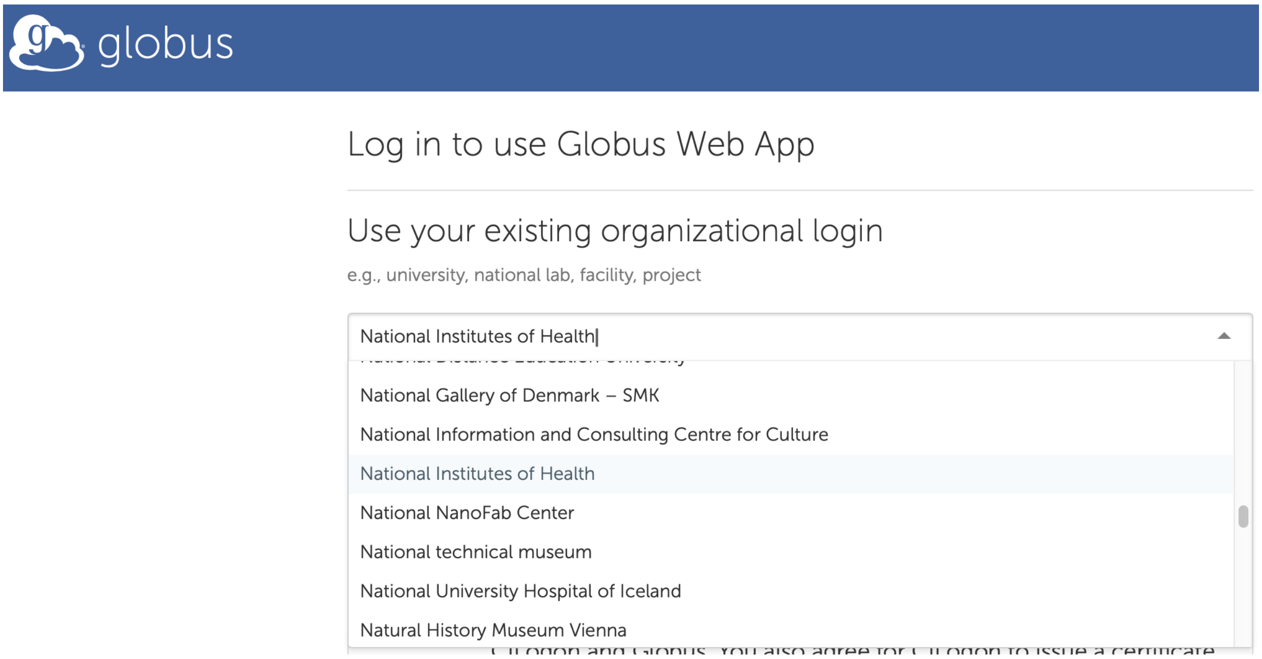
If following the Biowulf instructions for creating a Globus endpoint for the "PartekFlow" folder, it will be a good idea to use subdirectories for data generated for different experiments. This exercise will use a subdirectory called fnl_example_single_cell_fastq.

Sign onto Globus

Information regarding Globus and how to obtain it can be found at https://hpc.nih.gov/docs/globus/setup.php *(https://hpc.nih.gov/docs/globus/setup.php)*.

For those Globus already setup, goto https://www.globus.org *(https://www.globus.org)* to log in by clicking on the "LOG IN" icon at the top right of the pages.

After clicking on the log in button, select organziational affiliation, which is National Institutes of Health in this example.



globus

Log in to use Globus Web App

Use your existing organizational login

e.g., university, national lab, facility, project

National Institutes of Health|

National Gallery of Denmark – SMK

National Information and Consulting Centre for Culture

National Institutes of Health

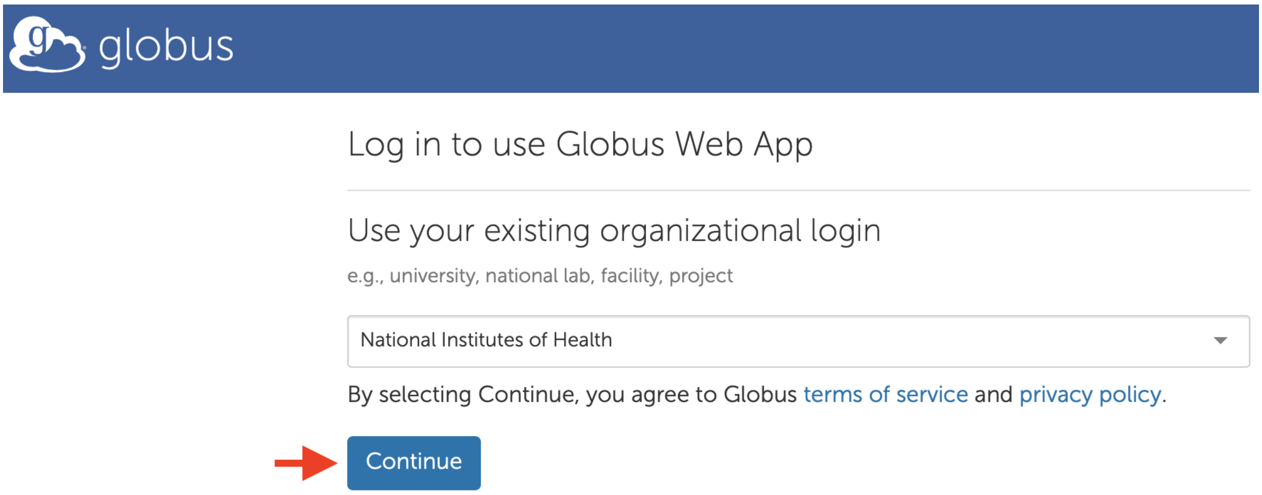
National NanoFab Center

National technical museum

National University Hospital of Iceland

Natural History Museum Vienna

Click on "Continue" when the organizational affiliation has been selected.



globus

Log in to use Globus Web App

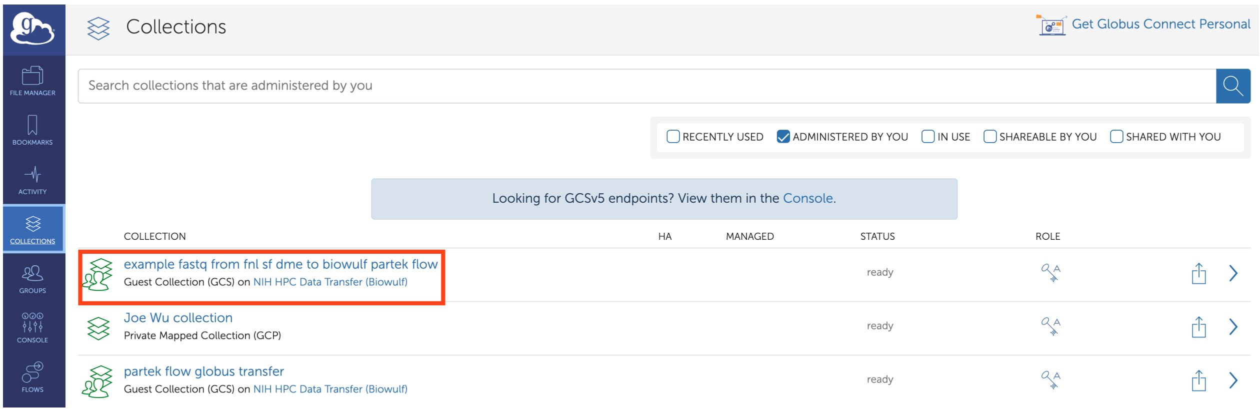
Use your existing organizational login

e.g., university, national lab, facility, project

National Institutes of Health

By selecting Continue, you agree to Globus terms of service and privacy policy.

After clicking "Continue", users will be brought the to Globus interface where file transfers are managed. Clicking on "COLLECTIONS" and then check "ADMINISTER BY YOU" will reveal several endpoints including on that points to that for the instructor's Biowulf "PartekFlow" folder labeled "example fastq from fnl sf dme to biowulf partek flow" to see the overview. Note the UUID.



Collections

Search collections that are administered by you

☐ RECENTLY USED ☑ ADMINISTERED BY YOU ☐ IN USE ☐ SHAREABLE BY YOU ☐ SHARED WITH YOU

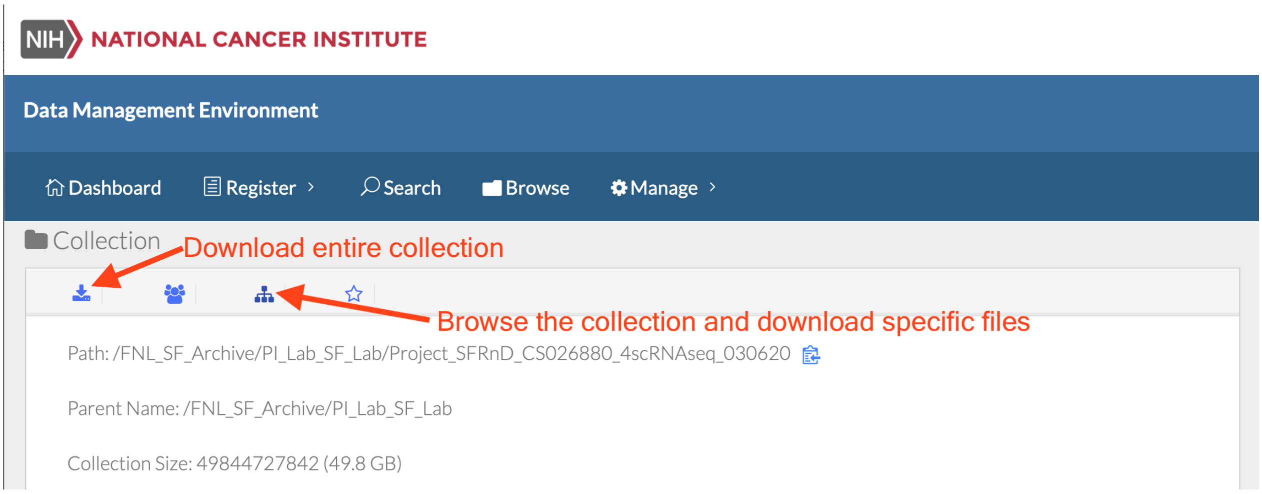
Looking for GCSv5 endpoints? View them in the Console.

COLLECTION	HA	MANAGED	STATUS	ROLE
example fastq from fnl sf dme to biowulf partek flow Guest Collection (GCS) on NIH HPC Data Transfer (Biowulf)			ready	
Joe Wu collection Private Mapped Collection (GCP)			ready	
partek flow globus transfer Guest Collection (GCS) on NIH HPC Data Transfer (Biowulf)			ready	

NCI CCR Sequencing Facility Data Management Environment

The NCI CCR Sequencing Facility will send researchers a link to the data, which is stored in their Data Management Environment (DME). Again, instructions for connecting DME to a Globus endpoint on Biowulf are in the embedded PDF.

Users are able to download an entire collection of data or browse the collection and download a subset.



Download entire collection

Browse the collection and download specific files

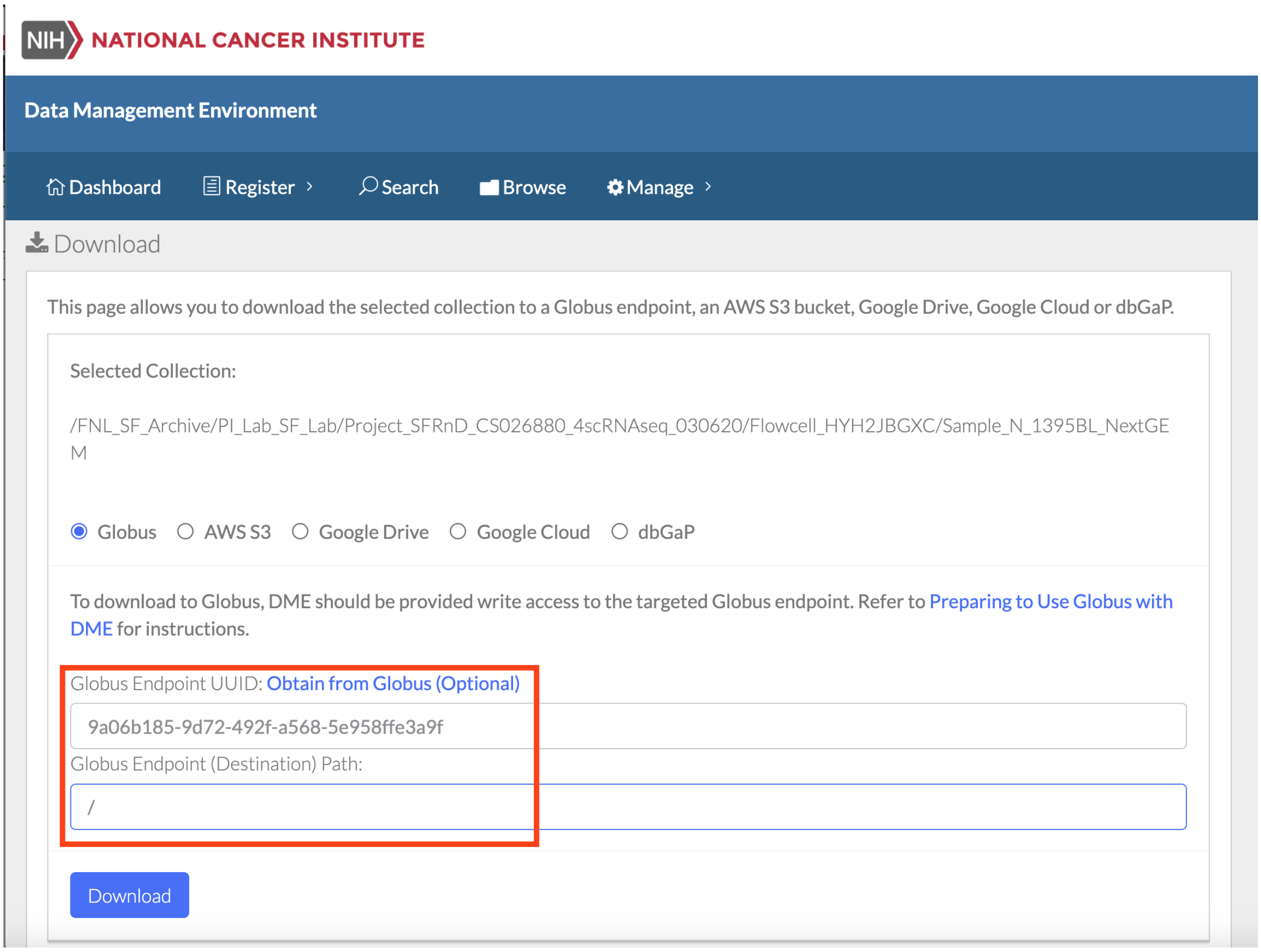
Path: /FNL_SF_Archive/PI_Lab_SF_Lab/Project_SFRnD_CS026880_4scRNAseq_030620

Parent Name: /FNL_SF_Archive/PI_Lab_SF_Lab

Collection Size: 49844727842 (49.8 GB)

This example will browse the collection and download the N_1395BL_NextGEM_count.tar.

At the subsequent page, enter the UUID for the Biowulf Partek Flow Globus endpoint and "/" for the path. Then click "Download".



This page allows you to download the selected collection to a Globus endpoint, an AWS S3 bucket, Google Drive, Google Cloud or dbGaP.

Selected Collection:

/FNL_SF_Archive/PI_Lab_SF_Lab/Project_SFRnD_CS026880_4scRNAseq_030620/Flowcell_HYH2JBGXC/Sample_N_1395BL_NextGEM

● Globus ○ AWS S3 ○ Google Drive ○ Google Cloud ○ dbGaP

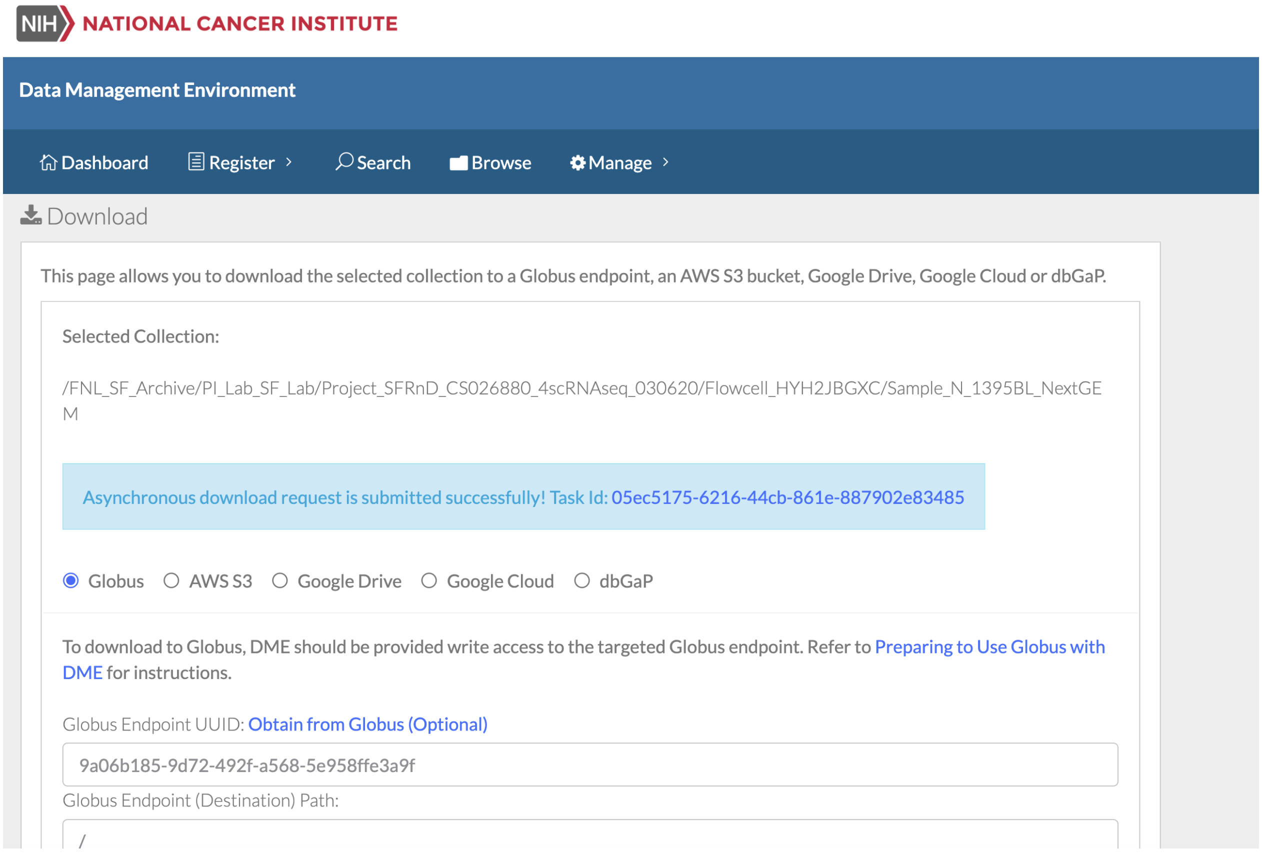
To download to Globus, DME should be provided write access to the targeted Globus endpoint. Refer to Preparing to Use Globus with DME for instructions.

Globus Endpoint UUID: Obtain from Globus (Optional)

9a06b185-9d72-492f-a568-5e958ffe3a9f

Globus Endpoint (Destination) Path:

/



Download

This page allows you to download the selected collection to a Globus endpoint, an AWS S3 bucket, Google Drive, Google Cloud or dbGaP.

Selected Collection:

/FNL_SF_Archive/PI_Lab_SF_Lab/Project_SFRnD_CS026880_4scRNAseq_030620/Flowcell_HYH2JBGXC/Sample_N_1395BL_NextGEM

Asynchronous download request is submitted successfully! Task Id: 05ec5175-6216-44cb-861e-887902e83485

● Globus ○ AWS S3 ○ Google Drive ○ Google Cloud ○ dbGaP

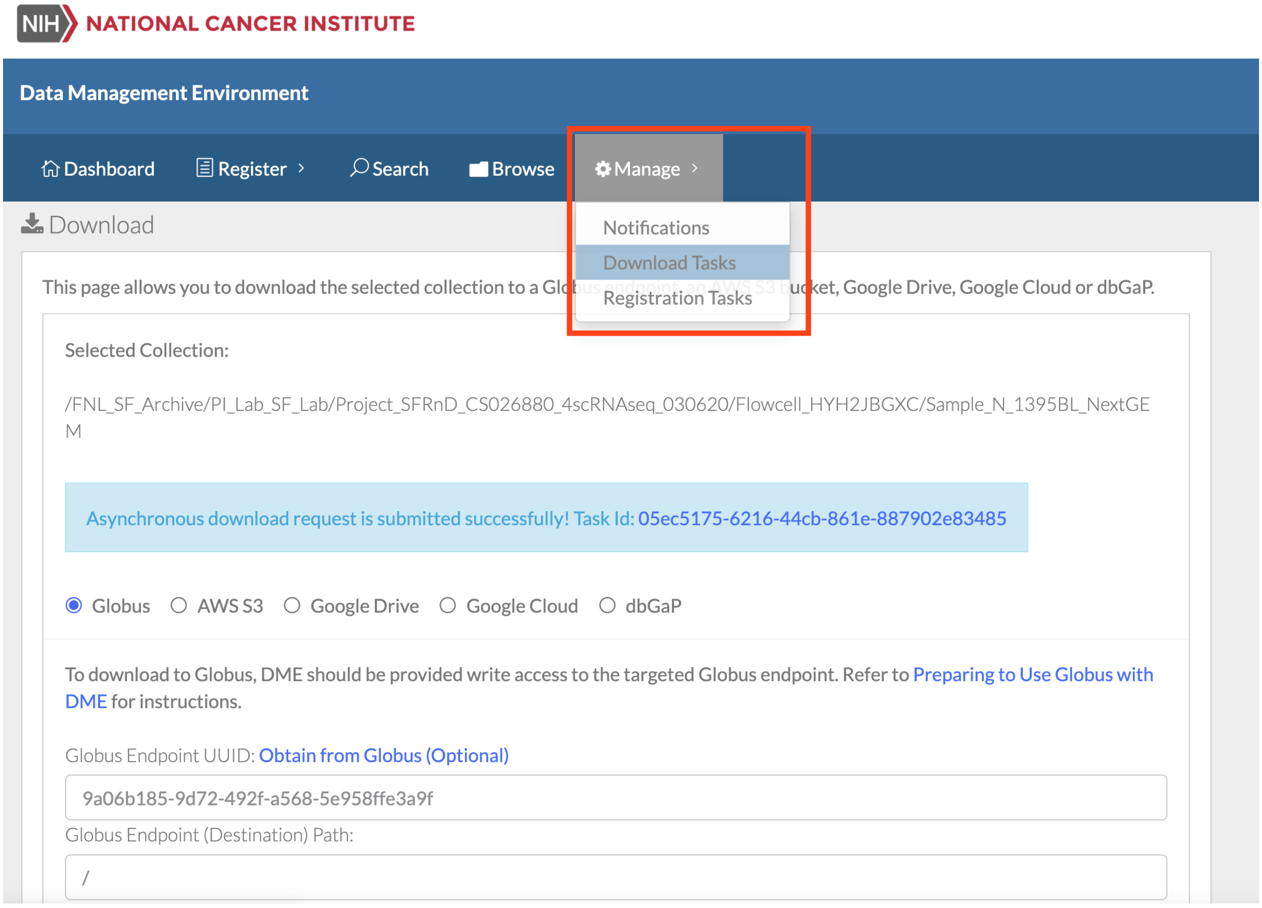
To download to Globus, DME should be provided write access to the targeted Globus endpoint. Refer to Preparing to Use Globus with DME for instructions.

Globus Endpoint UUID: Obtain from Globus (Optional)

9a06b185-9d72-492f-a568-5e958ffe3a9f

Globus Endpoint (Destination) Path:

/



Download

This page allows you to download the selected collection to a Globus endpoint, an AWS S3 bucket, Google Drive, Google Cloud or dbGaP.

Selected Collection:

/FNL_SF_Archive/PI_Lab_SF_Lab/Project_SFRnD_CS026880_4scRNAseq_030620/Flowcell_HYH2JBGXC/Sample_N_1395BL_NextGEM

Asynchronous download request is submitted successfully! Task Id: 05ec5175-6216-44cb-861e-887902e83485

● Globus ○ AWS S3 ○ Google Drive ○ Google Cloud ○ dbGaP

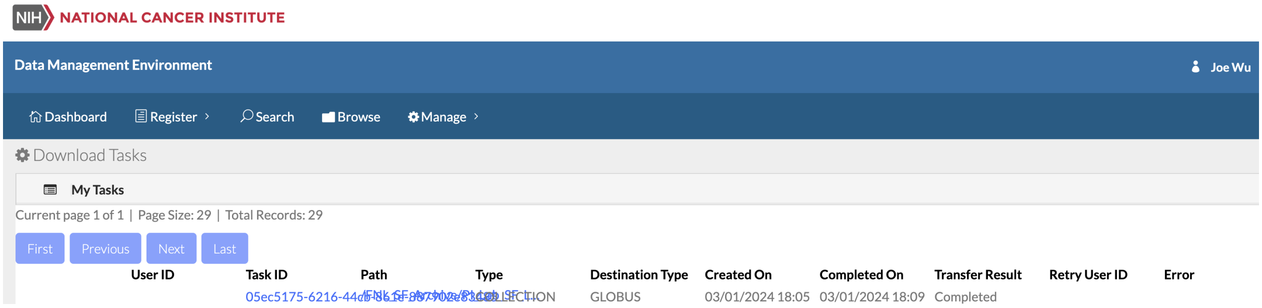
To download to Globus, DME should be provided write access to the targeted Globus endpoint. Refer to Preparing to Use Globus with DME for instructions.

Globus Endpoint UUID: Obtain from Globus (Optional)

9a06b185-9d72-492f-a568-5e958ffe3a9f

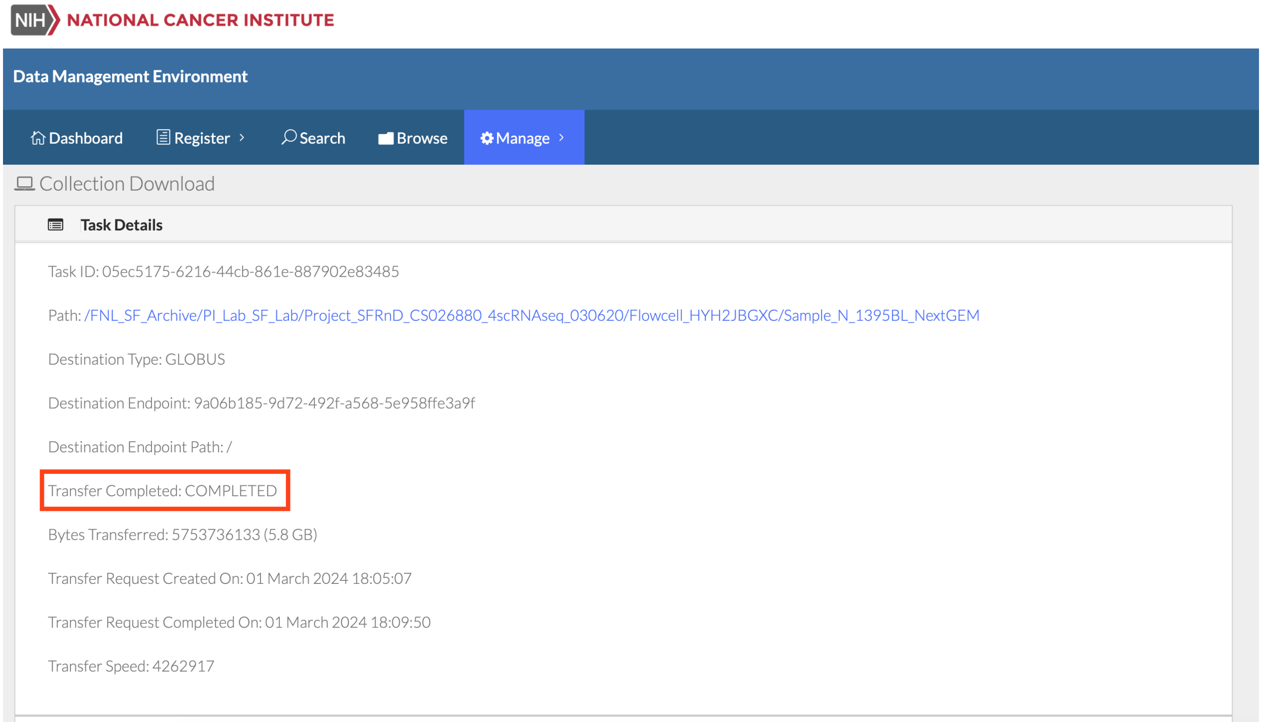
Globus Endpoint (Destination) Path:

/



Current page 1 of 1 | Page Size: 29 | Total Records: 29

User ID	Task ID	Path	Type	Destination Type	Created On	Completed On	Transfer Result	Retry User ID	Error
	05ec5175-6216-44cb-861e-887902e83485	/FNL_SF_Archive/PI_Lab_SF_L...	COLLECTION	GLOBUS	03/01/2024 18:05	03/01/2024 18:09	Completed		



Task ID: 05ec5175-6216-44cb-861e-887902e83485

Path: /FNL_SF_Archive/PI_Lab_SF_Lab/Project_SFRnD_CS026880_4scRNAseq_030620/Flowcell_HYH2JBGXC/Sample_N_1395BL_NextGEM

Destination Type: GLOBUS

Destination Endpoint: 9a06b185-9d72-492f-a568-5e958ffe3a9f

Destination Endpoint Path: /

Transfer Completed: COMPLETED

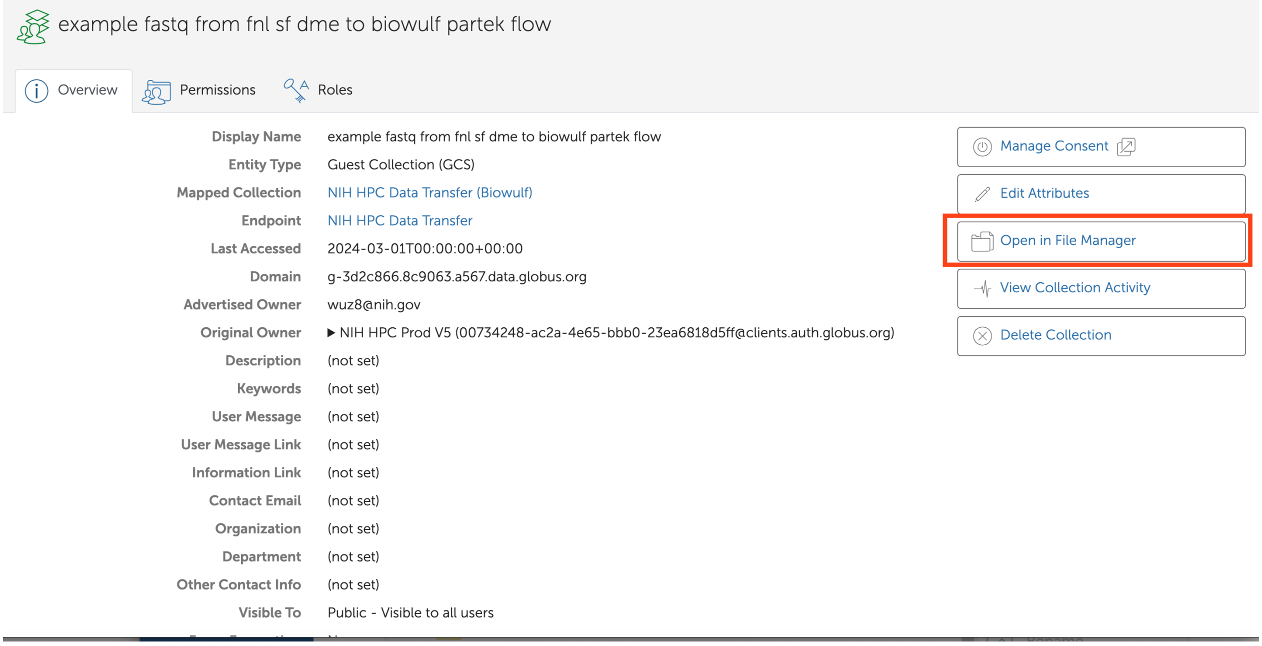
Bytes Transferred: 5753736133 (5.8 GB)

Transfer Request Created On: 01 March 2024 18:05:07

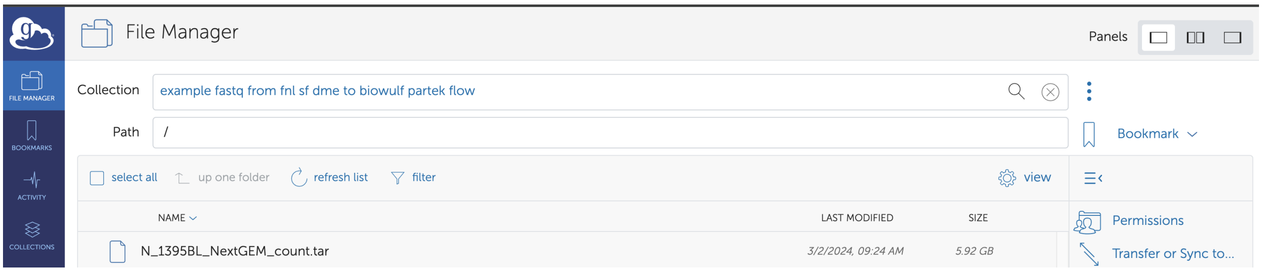
Transfer Request Completed On: 01 March 2024 18:09:50

Transfer Speed: 4262917

Globus will also send an email to the user's NIH email account after transfer has been completed.



Display Name	example fastq from fnl sf dme to biowulf partek flow
Entity Type	Guest Collection (GCS)
Mapped Collection	NIH HPC Data Transfer (Biowulf)
Endpoint	NIH HPC Data Transfer
Last Accessed	2024-03-01T00:00:00+00:00
Domain	g-3d2c866.8c9063.a567.data.globus.org
Advertised Owner	wuz8@nih.gov
Original Owner	▶ NIH HPC Prod V5 (00734248-ac2a-4e65-bbb0-23ea6818d5ff@clients.auth.globus.org)
Description	(not set)
Keywords	(not set)
User Message	(not set)
User Message Link	(not set)
Information Link	(not set)
Contact Email	(not set)
Organization	(not set)
Department	(not set)
Other Contact Info	(not set)
Visible To	Public - Visible to all users



Collection: example fastq from fnl sf dme to biowulf partek flow

Path: /

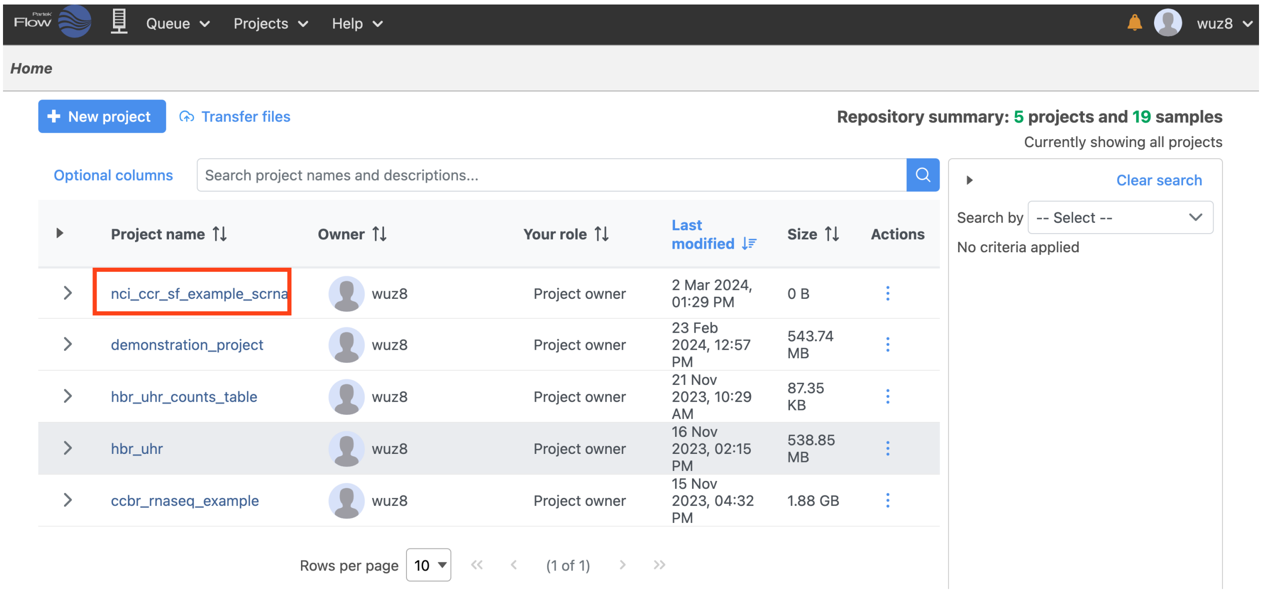
NAME	LAST MODIFIED	SIZE
N_1395BL_NextGEM_count.tar	3/2/2024, 09:24 AM	5.92 GB

These files will show up on Biowulf as well. These files will need to be unpacked using `tar -xvf`.

```
[wuz8@biowulf fnl_example_single_cell_fastq]$ ls -1 *.tar
N_1395BL_NextGEM_count.tar
```

Importing data to Partek Flow project

Log into Partek Flow at https://partekflow.cit.nih.gov/flow *(https://partekflow.cit.nih.gov/flow)*. This example will use the nci_ccr_sf_example_scrna project, so click on it.

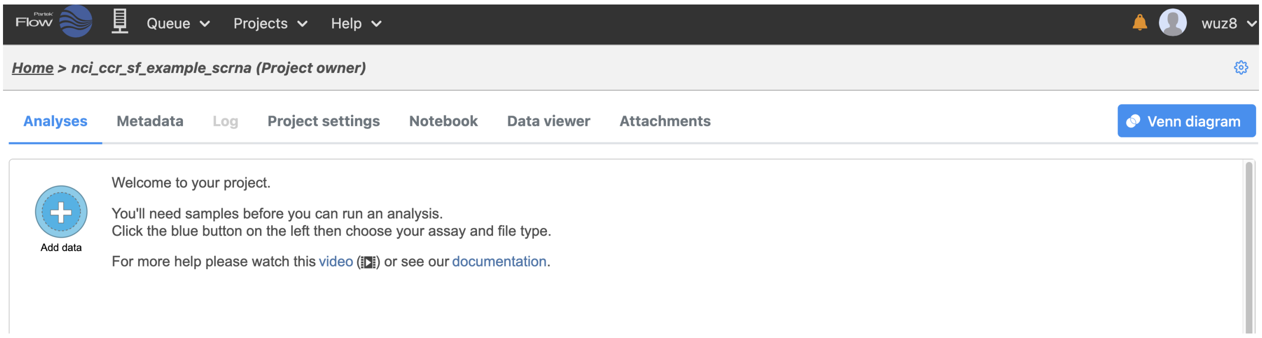


Repository summary: 5 projects and 19 samples

Currently showing all projects

Project name	Owner	Your role	Last modified	Size	Actions
nci_ccr_sf_example_scrna	wuz8	Project owner	2 Mar 2024, 01:29 PM	0 B	⋮
demonstration_project	wuz8	Project owner	23 Feb 2024, 12:57 PM	543.74 MB	⋮
hbr_uhr_counts_table	wuz8	Project owner	21 Nov 2023, 10:29 AM	87.35 KB	⋮
hbr_uhr	wuz8	Project owner	16 Nov 2023, 02:15 PM	538.85 MB	⋮
ccbr_rnaseq_example	wuz8	Project owner	15 Nov 2023, 04:32 PM	1.88 GB	⋮

Click on "Add data".

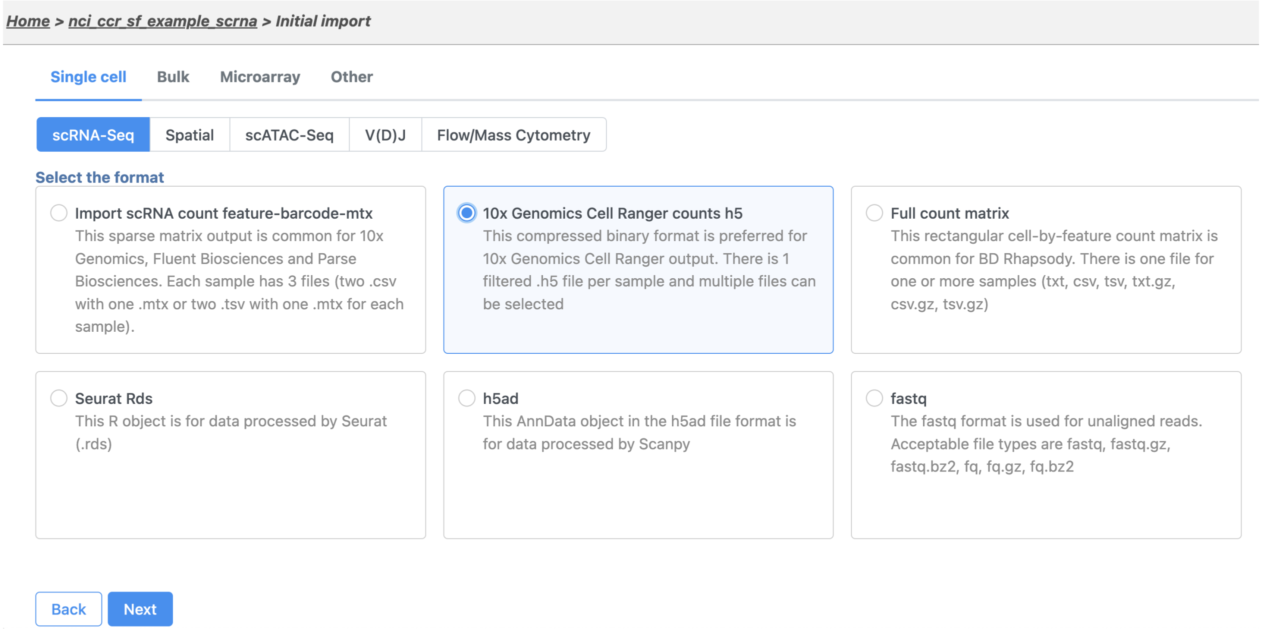


Welcome to your project.

You'll need samples before you can run an analysis.
Click the blue button on the left then choose your assay and file type.

For more help please watch this video or see our documentation.

Select Single cell, scRNA-Seq, and check 10x Genomics Cell Ranger counts h5. Then click Next.



Home > *nci_ccr_sf_example_scrna* > *Initial import*

Single cell | Bulk | Microarray | Other

scRNA-Seq | Spatial | scATAC-Seq | V(D)J | Flow/Mass Cytometry

Select the format

○ Import scRNA count feature-barcode-mtx
This sparse matrix output is common for 10x Genomics, Fluent Biosciences and Parse Biosciences. Each sample has 3 files (two .csv with one .mtx or two .tsv with one .mtx for each sample).

● 10x Genomics Cell Ranger counts h5
This compressed binary format is preferred for 10x Genomics Cell Ranger output. There is 1 filtered .h5 file per sample and multiple files can be selected

○ Full count matrix
This rectangular cell-by-feature count matrix is common for BD Rhapsody. There is one file for one or more samples (txt, csv, tsv, txt.gz, csv.gz, tsv.gz)

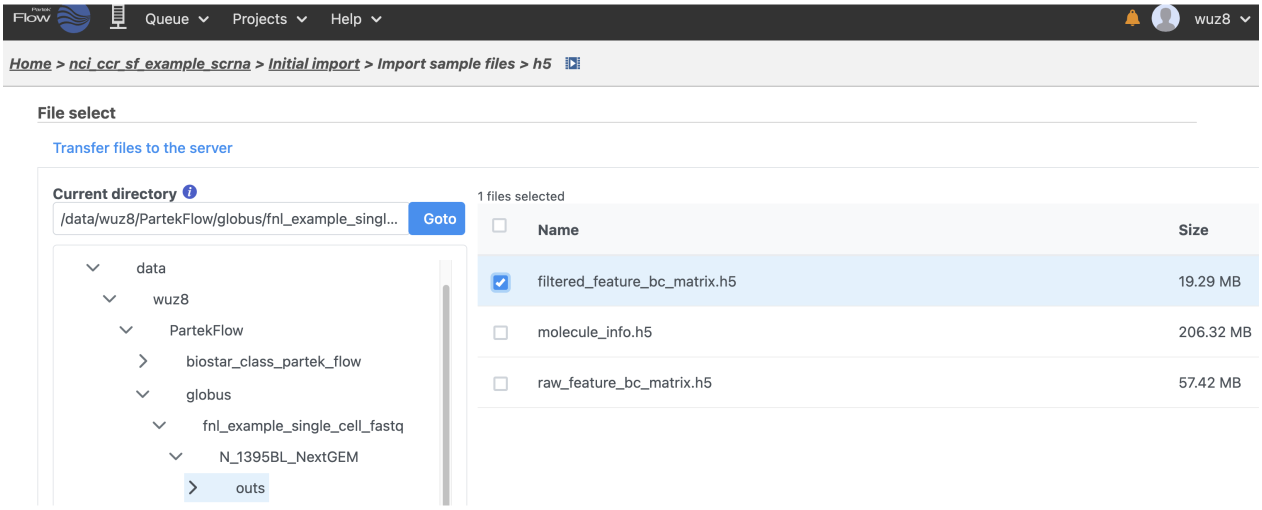
○ Seurat Rds
This R object is for data processed by Seurat (.rds)

○ h5ad
This AnnData object in the h5ad file format is for data processed by Scanpy

○ fastq
The fastq format is used for unaligned reads. Acceptable file types are fastq, fastq.gz, fastq.bz2, fq, fq.gz, fq.bz2

Back | Next

Navigate PartekFlow, globus, fnl_example_single_cell_fastq, N_1395BL_NextGEM, outs and select the filtered_feature_bc_matrix.h5 file and then click Next at the bottom of the screen.



Home > *nci_ccr_sf_example_scrna* > *Initial import* > *Import sample files* > *h5*

File select

Transfer files to the server

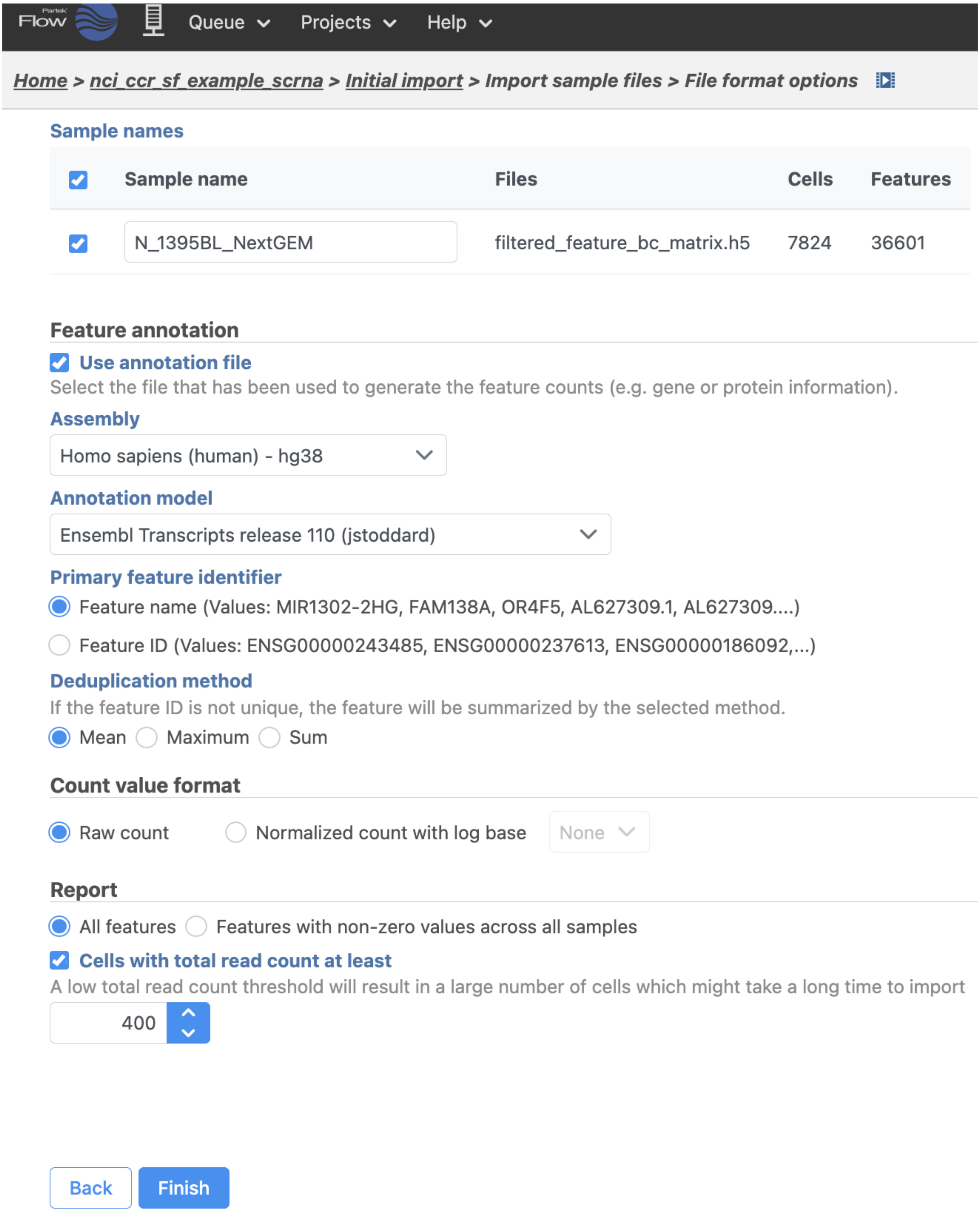
Current directory
/data/wuz8/PartekFlow/globus/fnl_example_singl... Goto

data
wuz8
PartekFlow
biostar_class_partek_flow
globus
fnl_example_single_cell_fastq
N_1395BL_NextGEM
outs

1 files selected

□	Name	Size
☑	filtered_feature_bc_matrix.h5	19.29 MB
□	molecule_info.h5	206.32 MB
□	raw_feature_bc_matrix.h5	57.42 MB

In the subsequent page, provide an informative sample name and select the appropriate assembly. Then click Finish.



Home > nci_ccr_sf_example_scrna > Initial import > Import sample files > File format options

Sample names

☑	Sample name	Files	Cells	Features
☑	N_1395BL_NextGEM	filtered_feature_bc_matrix.h5	7824	36601

Feature annotation

☑ Use annotation file

Select the file that has been used to generate the feature counts (e.g. gene or protein information).

Assembly: Homo sapiens (human) - hg38

Annotation model: Ensembl Transcripts release 110 (jstoddard)

Primary feature identifier

● Feature name (Values: MIR1302-2HG, FAM138A, OR4F5, AL627309.1, AL627309....)

○ Feature ID (Values: ENSG00000243485, ENSG00000237613, ENSG00000186092,...)

Deduplication method

If the feature ID is not unique, the feature will be summarized by the selected method.

● Mean ○ Maximum ○ Sum

Count value format

● Raw count ○ Normalized count with log base None

Report

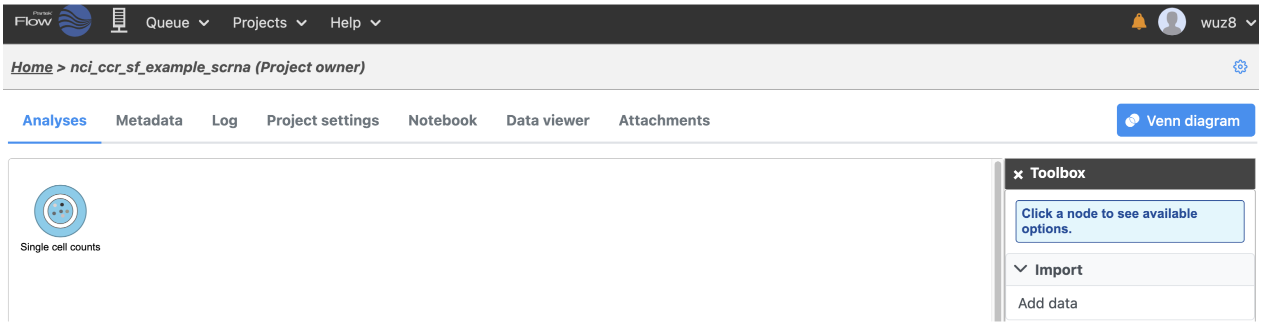
● All features ○ Features with non-zero values across all samples

☑ Cells with total read count at least

A low total read count threshold will result in a large number of cells which might take a long time to import

400

When import is done there will be a "Single cell counts" data node in the Analyses window.


Flow
Queue
Projects
Help
wuz8
Home > nci_ccr_sf_example_scrna (Project owner)
Analyses
Metadata
Log
Project settings
Notebook
Data viewer
Attachments
Venn diagram
Single cell counts
Toolbox
Click a node to see available options.
Import
Add data