Getting Started with Partek Flow

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

Transfer data to Partek Flow server using command line

Transfer data to Partek Flow server using Flow web GUI

Import data to Partek Flow projects

Partek Flow Bioinformatics Biowulf Globus Data transfer from NCI CCR Sequencing Facility

As far as the documentation goes, it should say that the licenses are NCI only. People can send a message if they want to ask beyond that.

Getting Started with Partek Flow at NIH

Learning Objectives

After this class, participants will

- Know how to acquire access to Partek Flow
- Become familiar with approaches for transferring data to the NIH Partek Flow server such as Globus
- Be able to import data into a Partek Flow project

What is Partek Flow?

Partek Flow 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. It enables the analysis of high dimensional multi-omics sequencing data including DNA, RNA, single cell RNA, ATAC/ChIP, and spatial transcriptomics. At NIH, Partek Flow is hosted on the Biowulf High Performance Computing (HPC) cluster. Researchers interact with Partek Flow through a web browser using a URL supplied by Biowulf after a Biowulf and Partek Flow account has been established. 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 also enables the creation of publication quality visualizations.

> **Note**
>
> NCI holds a license for Partek Flow.

Instructions for Obtaining Access to 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/)*. Here is what is needed:

- 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)*.

- `/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 require more disk space.
- Partek Flow account — please contact staff@hpc.nih.gov.

Once these steps have been accomplished, use https://partekflow.cit.nih.gov/flow *(https://partekflow.cit.nih.gov/flow)* to sign onto the NIH Partek Flow server. NIH single sign-in enables users to authenticate and connect to the Partek Flow server. Alternatively, users can log on by supplying username and password. However, the password may different from the one in which the user signs connects to Biowulf with as HPC staff allow for the researcher to set up something different upon Partek Flow account creation. The username though, is the user's NIH username.
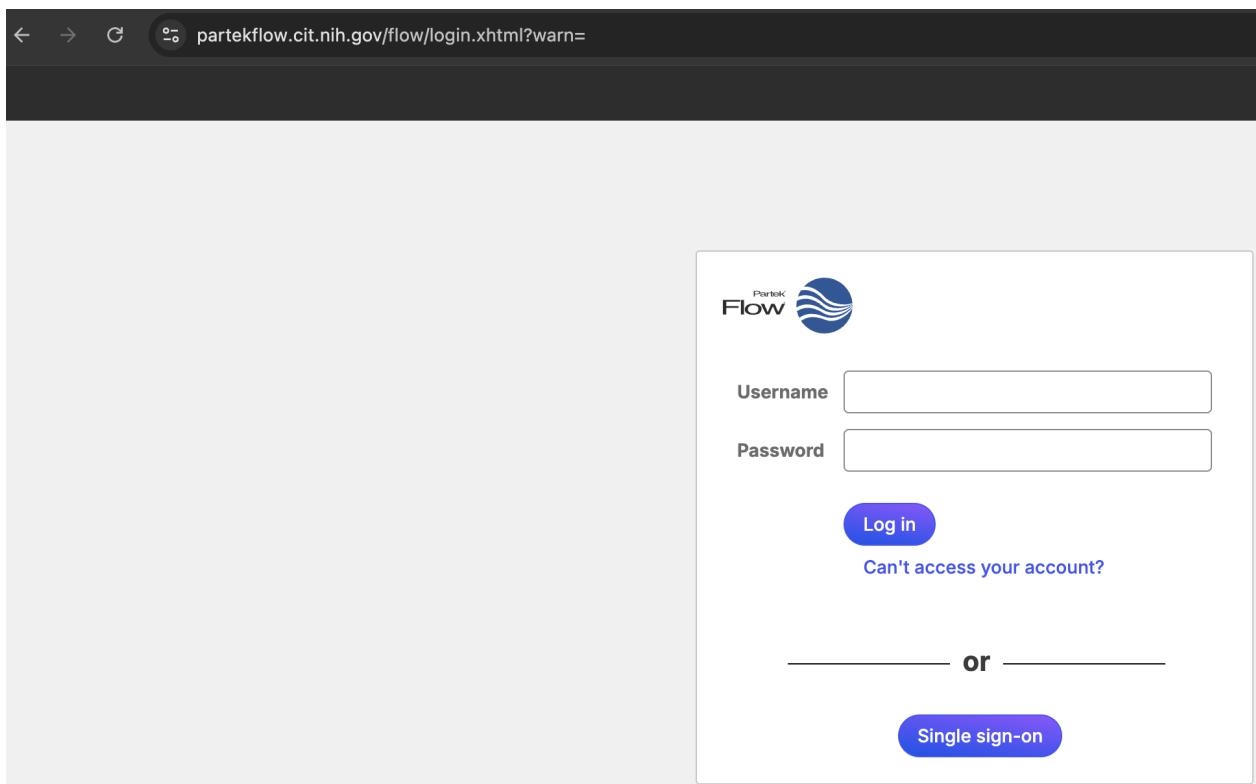

Overview of the Partek Flow User Interface

Upon logging in, users will be taken to Partek Flow's main page. In this page users can access

- Help
- View a list of existing projects
- Create new project
- Transfer files (this does not use Globus)
- Access user profile

Users can access help documentation, how-to videos and information for contacting support under the "Help" menu. The email for Partek support is support@partek.com.


Where applicable, users can find links to how-to video and documentation for specific tasks, such as importing data.

The settings section is divided into five tabs, which are expandable to reveal more options.

Click on the user profile tab at the top right to reveal the option for changing settings.
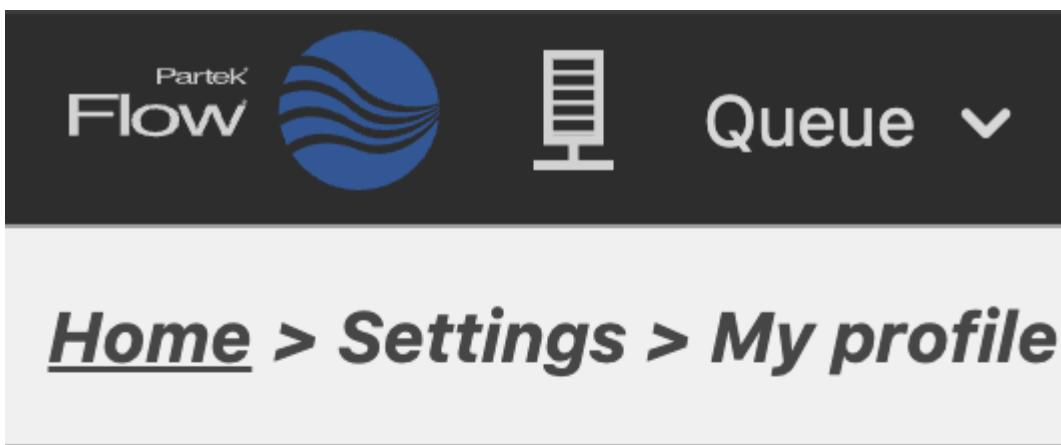
User profiles can be viewed and edited under "My Profile" in the "Personal" settings section. Note the information regarding the user's Biowulf `/data` folder storage usage and quota.

Users can access an example bulk and single cell RNA sequencing dataset under the "System" setting tab.

An important feature in the settings is that users can manage genomic references and annotations by clicking on "Library files" under "Components".

Users can view "Systems resources" under the "Usage" settings. Here, users are informed of how many tasks are running and waiting in the queue. A visualization of Biowulf Partek Flow worker nodes is also provided.

partekflow.cit.nih.gov/flow/settings/resource-management.xhtml

Flow | Queue ⌄ | Projects ⌄ | Help ⌄

Home > Settings > System resources

> Personal
> System
> Components
> Access
∨ Usage
  System queue
  System resources

⎙ View queued tasks

Queue status

Running tasks 1
Waiting tasks 0

Time estimates are being continuously updated and will become more accurate.

Licensing

Available worker licenses 48 out of 50

Active workers

Name	CPU usage	Memory usage	Machine capacity	Uptime	Type
hpcpartek.cit.nih.gov			12 cores | 63 GB RAM		Server
cn2287	load w: 0.0 | s: 0.0	w: 0.0 | s: 7.3 GB	56 cores | 252 GB RAM	1 day	Worker
cn3168	load w: 1.0 | s: 0.0	w: 5.6 | s: 8.0 GB	56 cores | 252 GB RAM	07:55:54	Worker

▪ Worker usage   ▫ System usage   �X Worker limit

Class overview slides

I

Tips on Biowulf for Partek Flow Users

Even though Partek Flow runs on Biowulf, the NIH Unix-based high performance computing cluster, users do not need to be experts in command line to use this software. However, there are somethings that are useful to know because Partek Flow projects, input, and output are all stored on Biowulf.

Signing onto Biowulf

To sign onto Biowulf, open a Terminal if using MacOS or Command Prompt if using Windows 10 or above. Once opened, at the prompt, do the following to sign on. The breakdown of the command is as follows.

- `ssh`: Connect to remote computer (ie. Biowulf)
- `user`: User's NIH username
- `@biowulf.nih.gov`: The remote computer to connect to (ie. NIH Biowulf)

```
ssh user@biowulf.nih.gov
```

Users will then be asked to enter a password for signing onto Biowulf. This is the password that the user utilizes to sign onto NIH applications. The password will not appear as the user is typing but keep typing and hit enter when done.

```
user@biowulf.nih.gov's password:
```

Changing into the User's Data Folder

Upon signing onto Biowulf, users will land in the `/home/user` directory (again replace user with user's NIH username). This directory only has 16 GB of storage space and cannot be increased, thus users will need to perform analyses in the `/data/user` folder. Further, a folder called `/PartekFlow` in the `/data/user` directory is created when the Partek Flow account is activated for the user.

To check which folder on Biowulf a user is currently in, use the `pwd` command.

Upon signing onto Biowulf:

```
pwd
```

```
/home/user
```

To change into the user's `data` folder do the following.

```
cd /data/user
```

Biowulf staff will create a folder called `PartekFlow` in the user's `/data/user` directory upon activating a Partek Flow account. This folder contains all Partek Flow projects, inputs, outputs, and data Uploads.

List the contents of the `/data/user` folder.

```
ls /data/user
```

```
PartekFlow
```

Change into the `PartekFlow` folder.

```
cd /data/user/PartekFlow
```

The `PartekFlow` folder houses project folders (only if users have projects), which contain input and output as well a folder called `uploads` that stores data that has been loaded to the NIH Partek Flow server.

```
Project_RNAsequencing_1234
uploads
```

Note

Most likely, Partek Flow users will not have to use the command line to do anything Partek Flow related as tasks such as creating projects, deleting projects, uploading files, etc. can be done in the software or using Globus.

Transferring Data to the NIH Partek Flow Server Using Globus

Globus

NCI CCR researchers will likely use the NCI CCR Sequencing Facility (or SF) *(https:// bioinformatics.ccr.cancer.gov/docs/resources-for-bioinformatics/raw_data_from_cores/)* for sequencing projcts. Data generated from the NCI CCR SF will be stored within its Data Manangement Environment (DME). Researchers can use Globus to transfer data from NCI CCR SF DME to the NIH Partek Flow server and the steps for accomplishing this are described below.

> **Note**
>
> The staff at Biowulf has created detailed documents for Globus, which can be found at https://hpc.nih.gov/docs/globus/setup.php *(https://hpc.nih.gov/docs/globus/setup.php).*

Step 1: Logging into Globus

Goto https://www.globus.org *(https://www.globus.org)* to log in by clicking on the "LOG IN" icon at the top right corner of the page.
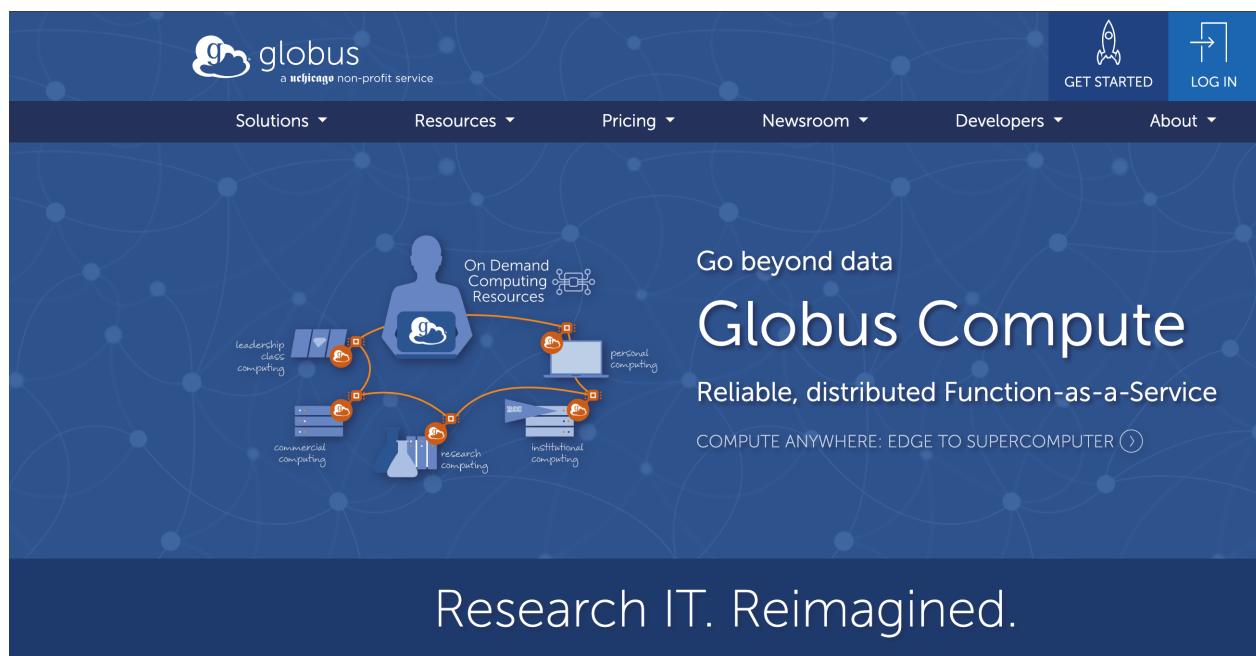
In the next page, select organziational affiliation from the drop down menu (in this example it is National Institutes of Health).

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

Subsequently, users will be taken to the NIH authentication page. Click on "Sign in" to authenticate using PIV card.
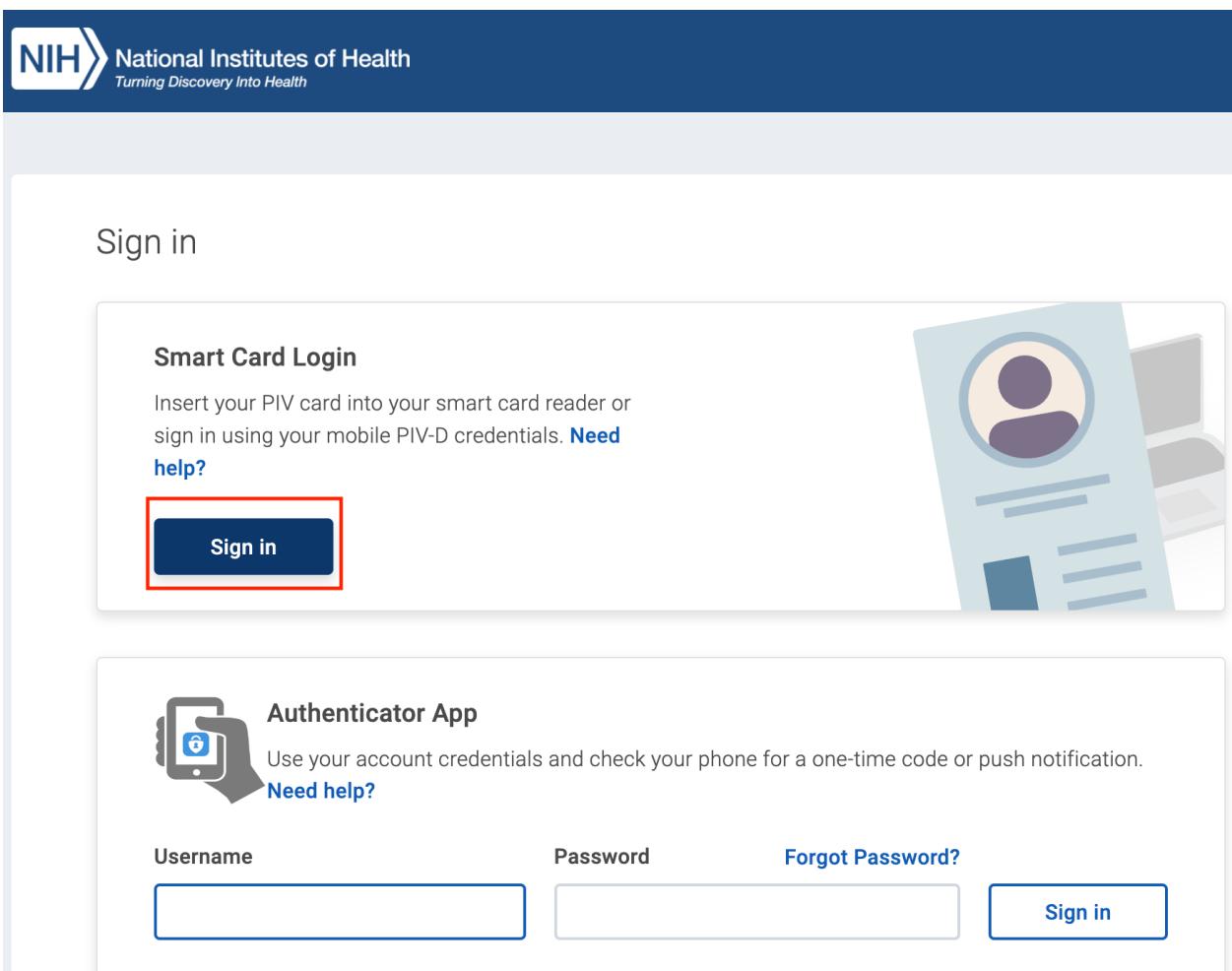
Next, select the appropriate PIV card certificate (usually the one with the user's name followed by "- A (Affiliate)) and a pop-up will appear to take the user's NIH pin.
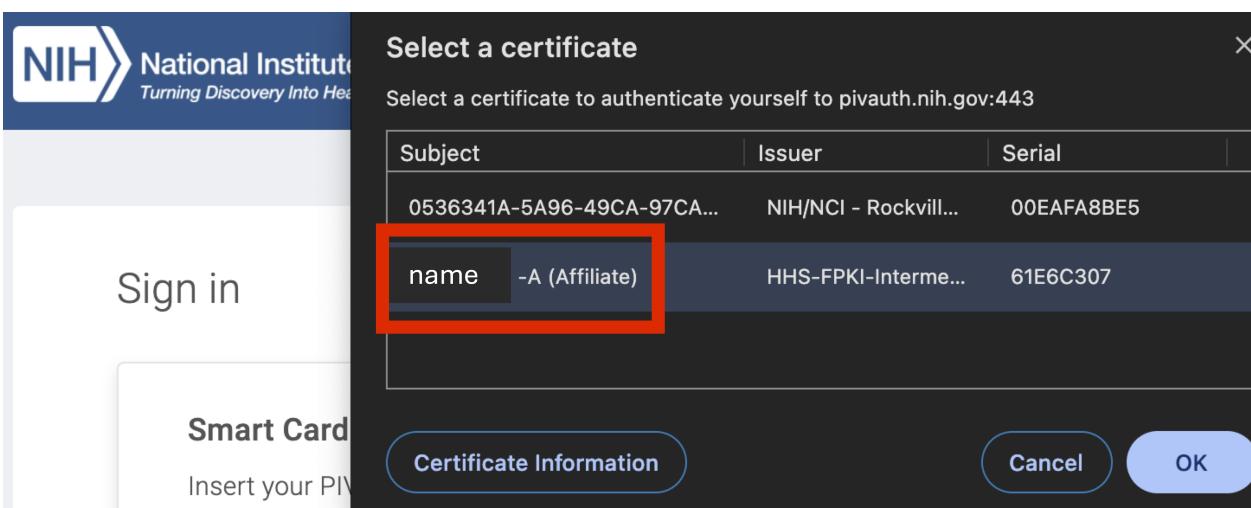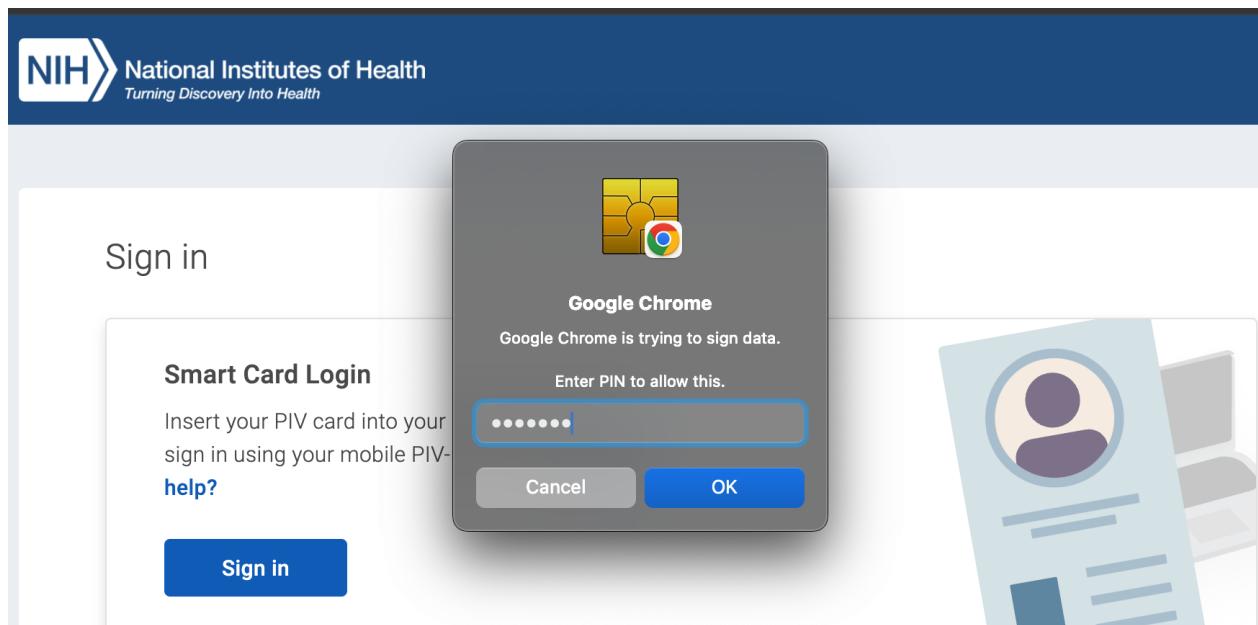
In the next page, scroll to the bottom and click "I Agree".

Here's the information to be released

First name: First
Last name: Last
E-mail address: First.Last@nih.gov
NetID: userid@nih.gov

The privacy policy of the service you're connecting to details things like why the service requires this information, how long the information will be retained, who the information will be shared with, etc. In general, the information is needed to facilitate your access, it will not be shared, and it will be retained for a limited time. You can review the service's privacy policy here:
https://www.cilogon.org/privacy

Do you agree to release the information listed above to this service?

Please select your sharing preference from the options below and click on the I Agree button.

🔘 **Ask me again at next login**

   ○ I agree to send my information this time, but I want to be prompted again the next time I access a service that requests this information.

○ **Ask me again if information to be provided to this service changes**

   ○ I agree to send the information listed above to this service now and in the future but I wanted to prompted if the information the service requires changes.

○ **Send now and in the future - Do not ask me again**

   ○ I agree to release the information listed above to this service and to any service that asks for the same information. I want to be prompted if a service asks for different information.

[I Do Not Agree]   [I Agree] ⬅

Users will then be brought to the Globus interface where file transfers are managed. Click on "COLLECTIONS" to see the recently used Globus data transfer endpoints here. The "NIH HPC Data Transfer (Biowulf)" *(https://hpc.nih.gov/docs/globus/setup.php#endpoints)* endpoint points to the content available on the cluster.
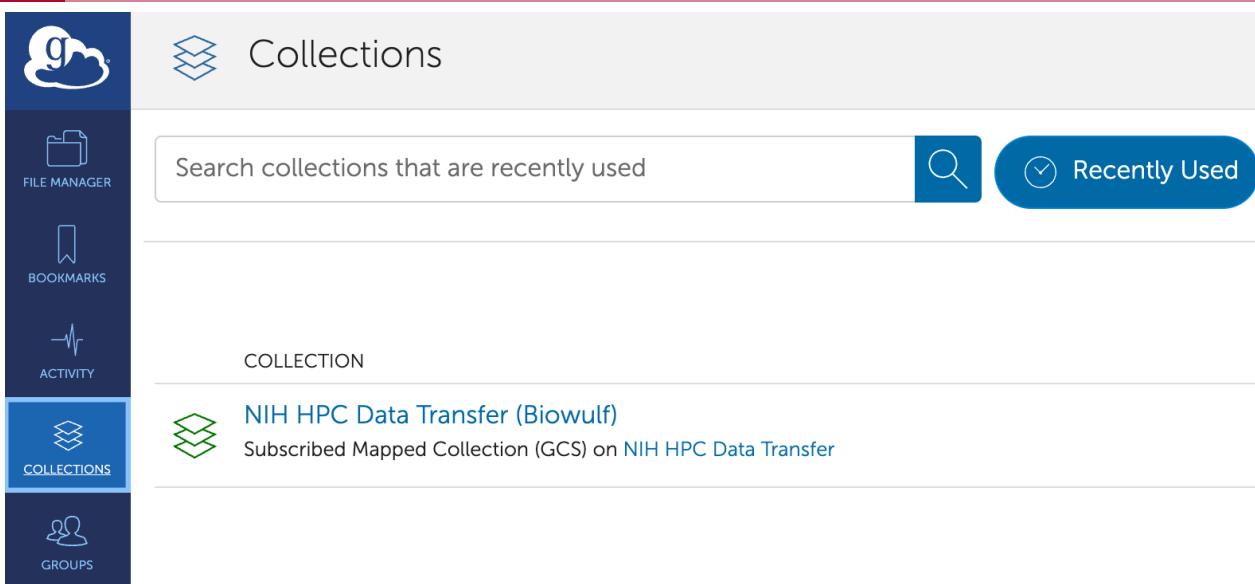
Step 2: Setting up a Globus Endpoint to the Partek Flow Server

Click on the "NIH HPC Data Transfer (Biowulf)" endpoint and then "Open in File Manager". This will take users to their Biowulf `/home` directory. Username is the Biowulf user name for the specific user.
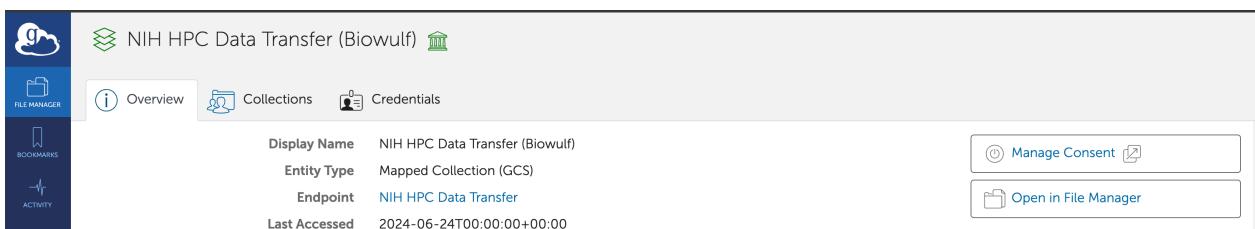
Recall

The user's Biowulf `/home` directory is not suitable for analyzing data. To conduct analysis, use the `/data` directory.

To goto the user's `/data/directory`, replace `/home/username` in the box labeld "Path" with `/data/username`.

After switching to the user specific `/data` directory, find and click into the `PartekFlow` folder. Recall that the `PartekFlow` folder will exist only if the user has contacted Biowulf staff about activating a Partek Flow account.
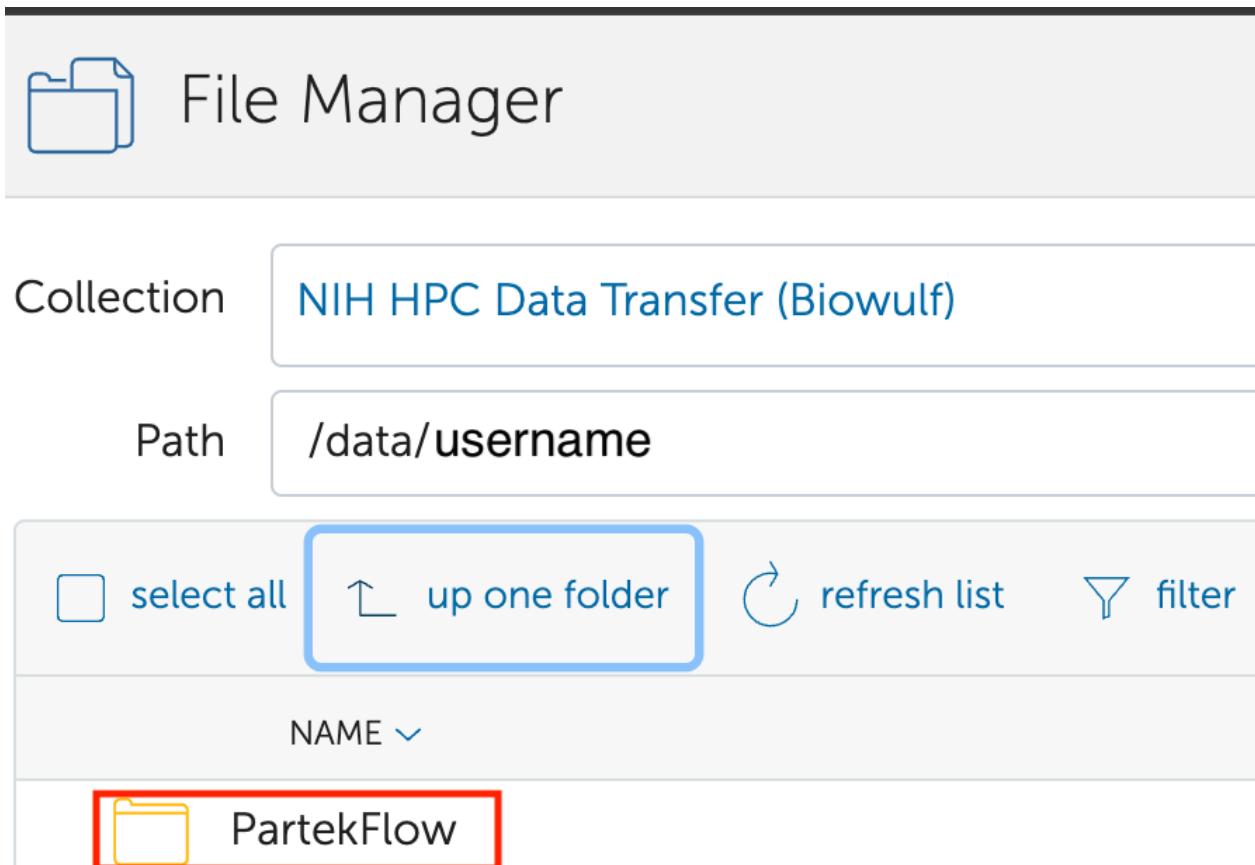
Next, click on "New Folder" to make a folder called "globus" to store data uploaded to the Partek Flow server via Globus.

Then, click into the `globus` folder and create one named `example_data_transfer`.

Go back to the "NIH HPC Data Transfer (Biowulf)" endpoint and click on the "COLLECTIONS" tab after the `example_data_transfer` folder has been created. From there, click on "Add Guest Collection".

In the "Add Guest Collection" menu, click "Browse" to select the folder in which the endpoint will reference (ie. `/data/username/PartekFlow/globus/example_data_transfer`). Enter a display name and description for the endpoint and then "Create Collection" when ready.

The user will then be taken to a page for setting up sharing between the endpoint and the location where the data is (ie. NCI CCR Sequencing Facility DME).

Leave the entry in the box labeled "Path" as "/". Make sure to mark the "Write" permission box because the NCI CCR Sequencing Facility DME has to write the data into this Globus endpoint. Then click "Select a Group" to choose the group in which to share `/data/username/PartekFlow/globus/example_data_transfer` with.

EXAMPLE DATA TRANSFER
Add Permissions – Share With

Path   /   ✓   Browse

Share With   ○ user – share with specific individuals
              ● group – make data accessible to members of a group
              ○ all users – make data accessible to all logged in users of Globus
              ○ public (anonymous) – make data accessible to everyone

Group   Select a Group

Permissions   ☑ read
              ☑ write

Add Permission   Cancel

In the drop down menu, select "HPCDME-PROD-App-Accts-Pool-FNLCR" to return to the "Add Permissions" page. Hit "Add Permission" when ready.

EXAMPLE DATA TRANSFER
Select a Group

Search by group name or select one below

HPCDME   ⊗

HPCDME-PROD-App-Accts-Pool-CCBR

HPCDME-PROD-App-Accts-Pool-HITIF
Pool of Globus app accounts for the NIH NCI DOC known as HiTIF

HPCDME-DEV-App-Accts-Pool-FNLCR
HPC DME Globus Group for FNLCR

HPCDME-PROD-App-Accts-Pool-FNLCR
Pool of Globus app accounts for the NIH NCI DOC of the Sequencing Facility at FNLCR

Click "Done" to finish the adding permission process.

The user will then be returned to the "example data transfer" endpoint and see that this endpoint has been shared with "HPCDME-PROD-App-Accts-Pool-FNLCR".

Click on the "Overview" tab and scroll to the bottom of the page. Take note of the UUID, which tells Sequencing Facility DME where to send data. The UUID for each Globus endpoint will be different.

Step 3: Downloading Data from the NCI CCR SF DME

Copy the link to the data provided by the NCI CCR SF and sign in with user specific NIH credentials and the page below will be shown. Users can download all data or browse through their data.

This example will click on the tab for browsing data and download FASTQ files in the folder labeled "Sample_N_1395BL_NextGEM". To download, just click on the "down arrow" corresponding to this folder in the column labeled "Download".

After clicking on the download arrow, users will be taken to the dialogue page shown below. Be sure to select Globus for the "Transfer Type", go back to Globus and copy then paste the endpoint UUID (this will ensure that the data gets transferred to the right place), and finally, leave the path as "/" as it was set when creating the "example data transfer" Globus Endpoint. When ready, hit "Download".

If all goes well, users will see a message indicating that the data transfer request has been submitted successfully.

Click on "Manage" and then "Download Tasks" to check download progress. Each download is assigned a task ID.

Go back to the "example data transfer" endpoint on Globus and click on "Open in File Manager".

The data will populate in the `data/username/PartekFlow/globus/example_data_transfer` folder (ie. the folder pointed to by the "example data transfer" endpoint) as the download proceeds.
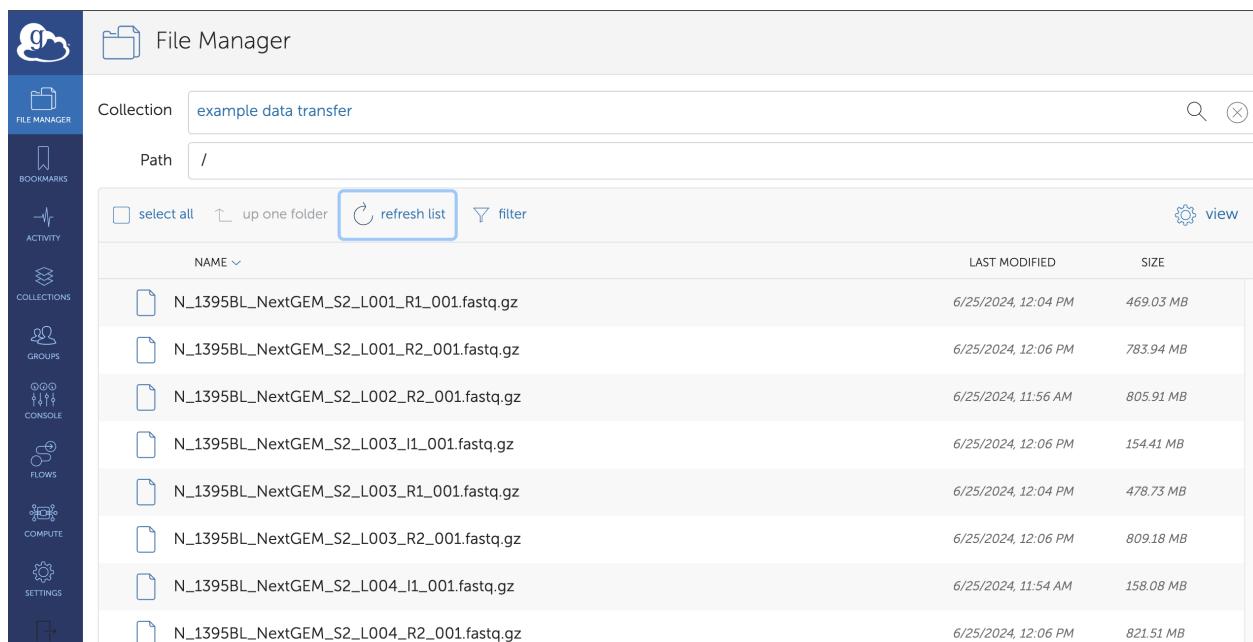
These changes are also reflected on Biowulf. Again, replace username with the user's Biowulf user name.

```
ls /data/username/PartekFlow/globus/example_data_transfer
```

```
N_1395BL_NextGEM_S2_L001_I1_001.fastq.gz
N_1395BL_NextGEM_S2_L001_R1_001.fastq.gz
N_1395BL_NextGEM_S2_L001_R2_001.fastq.gz
N_1395BL_NextGEM_S2_L002_I1_001.fastq.gz
N_1395BL_NextGEM_S2_L002_R1_001.fastq.gz
N_1395BL_NextGEM_S2_L002_R2_001.fastq.gz
N_1395BL_NextGEM_S2_L003_I1_001.fastq.gz
N_1395BL_NextGEM_S2_L003_R1_001.fastq.gz
N_1395BL_NextGEM_S2_L003_R2_001.fastq.gz
N_1395BL_NextGEM_S2_L004_I1_001.fastq.gz
N_1395BL_NextGEM_S2_L004_R1_001.fastq.gz
N_1395BL_NextGEM_S2_L004_R2_001.fastq.gz
```

Transferring Data to the NIH Partek Flow Server Using Command Line

> **Note**
>
> Command line uploads can be used when the files are already present on Biowulf. -- Biowulf *(https:// partekflow.cit.nih.gov/#upload_commandline)*

Copy from User's Biowulf `data` Folder to Partek Flow `uploads` Folder

Copying from user's Biowulf `/data` directory to the `uploads` subfolder in `PartekFlow` is allowed. Be sure to be in an interactive session for this by using the `sinteractive` command. For instance:

To transfer the folder `hbr_uhr_fastq_download` in `/data/username` to the `uploads` subfolder in `PartekFlow`, do the following. Replace username with the user's Biowulf user name.

```
cp -r /data/username/hbr_uhr_fastq_download /data/username/PartekFlo
```

Note that the Biowulf user owns the `hbr_uhr_fastq_download PartekFlow`. This enables users to delete the folder. The `ls` command lists directory content and the `-l` option list the content in long or detailed format.

```
[wuz8@cn4304 username]$ ls -l /data/username/PartekFlow/uploads/
drwxr-s---+ 2 username     partekfl 4096 May 28 11:55 hbr_uhr_fastq_c
```

It is possible to change into the `upload` subfolder in `/data/username/PartekFlow` and make a directory using the `mkdir` command.

```
cd /data/username/PartekFlow/uploads
```

```
[username@cn4304 uploads]$ mkdir hbr_uhr_fastq
```

Again the user owns the directory, and thus can delete it. Creating folders is a good way to keep data organized for analyses.

```
[wuz8@cn4304 uploads]$ ls -l
drwxrws---+ 2 username     partekfl 4096 May 28 13:12 hbr_uhr_fastq
```

Tip

File transfer using command line:

- Copy from data folder to Partek Flow uploads folder is **OK**
- Copy from data folder to Partek Flow directory is **OK**
- Copy from data folder to Partek Flow project folder is **not OK**
- Copy from Partek Flow uploads folder to project folder is **not OK**

Caution

While the above command line methods for data transfer are valid, users should remember that: "Virtually all file transfer activities should be run from within the web interface, rather than from the command line. We have implemented a permissions policy on users' PartekFlow directories to prevent inadvertent file removal mistakes that breaks the old way of moving files around." -- Biowulf staff

`scp` from Personal Computer to the Partek Flow `uploads` Folder.

The `scp` commands enables users to securely transfer data to and from personal computer to a remote computer such as a high performance computing system. To transfer data from personal computer to the Partek Flow `uploads` folder do the following.

Note

Helix is the node on the Biowulf system that is intended for data transfer.

```
scp local_file username@helix.nih.gov:/data/username/PartekFlow/uploa
```

Enter user's Biowulf password.

For example, the following command construct will transfer the folder `hbr_uhr_fastq` on local computer that contains FASTQ files to the Partek Flow `uploads` folder on Biowulf. The `-r` option in `scp` will copy a folder and its content.

```
scp -r hbr_uhr_fastq_files username@helix.nih.gov/data/username/Parte
```

```
ls -l /data/username/PartekFlow/uploads
```

Again, the Biowulf account user owns the directory and can modify and delete it.

```
drwxr-s---+ 2 username  partekfl 4096 Jun 25 21:13 hbr_uhr_fastq_file
```

Transferring Data to the NIH Partek Flow Server Using the Web Tool

Using the Partek Flow web tool to transfer data

If users have data stored on a personal computer, the Partek Flow web tool can be used to transfer data onto the server. Upon signing in to Partek Flow, a button labeled "Transfer files" is available. Click on this.
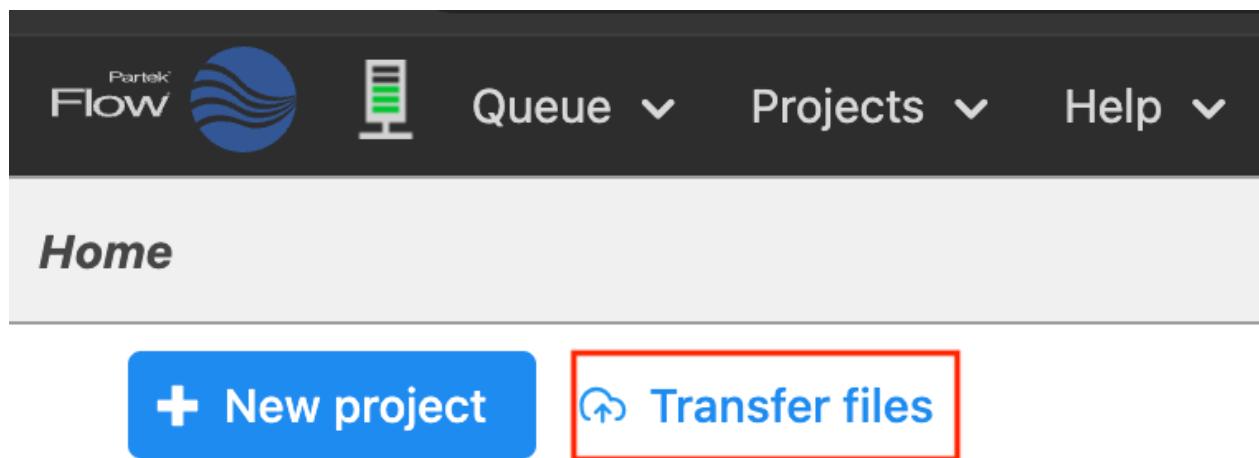
This takes the users to the "Transfer files" section of the settings page. Click on the blue "Transfer files" tab at the top.
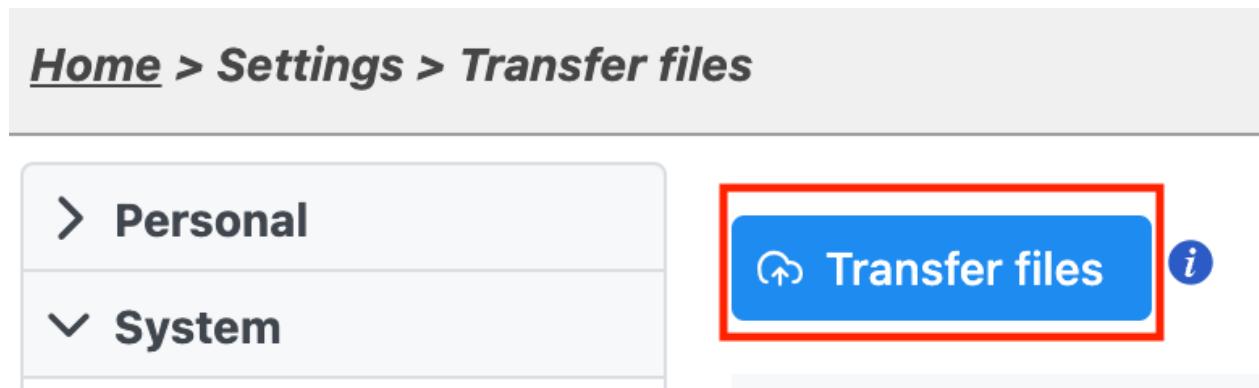
A dialogue box that assists the user with data transfer appears.

> **Note**
>
> Users can only transfer to the Partek Flow `uploads` folder or subfolders within `uploads`.

The current "Upload directory" as indicated at the bottom of the file transfer dialogue box is `/data/username/PartekFlow/uploads`. Click on the magnifying glass next to it to create new folders within Partek Flow `uploads` to store data uploaded for specific projects. It is a good idea to create a folder to store data for each project.
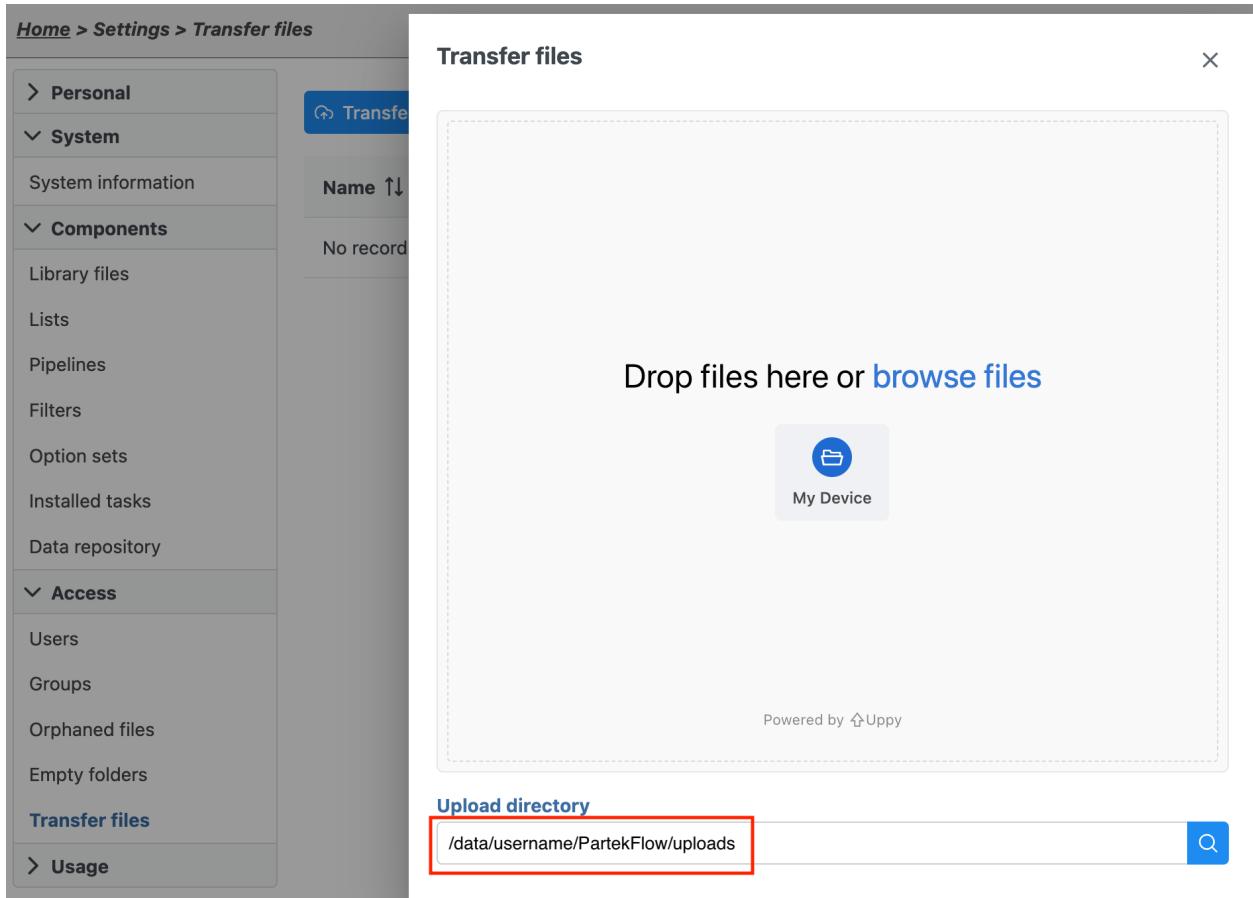
To create a new folder, click on the blue "Create new folder" tab.
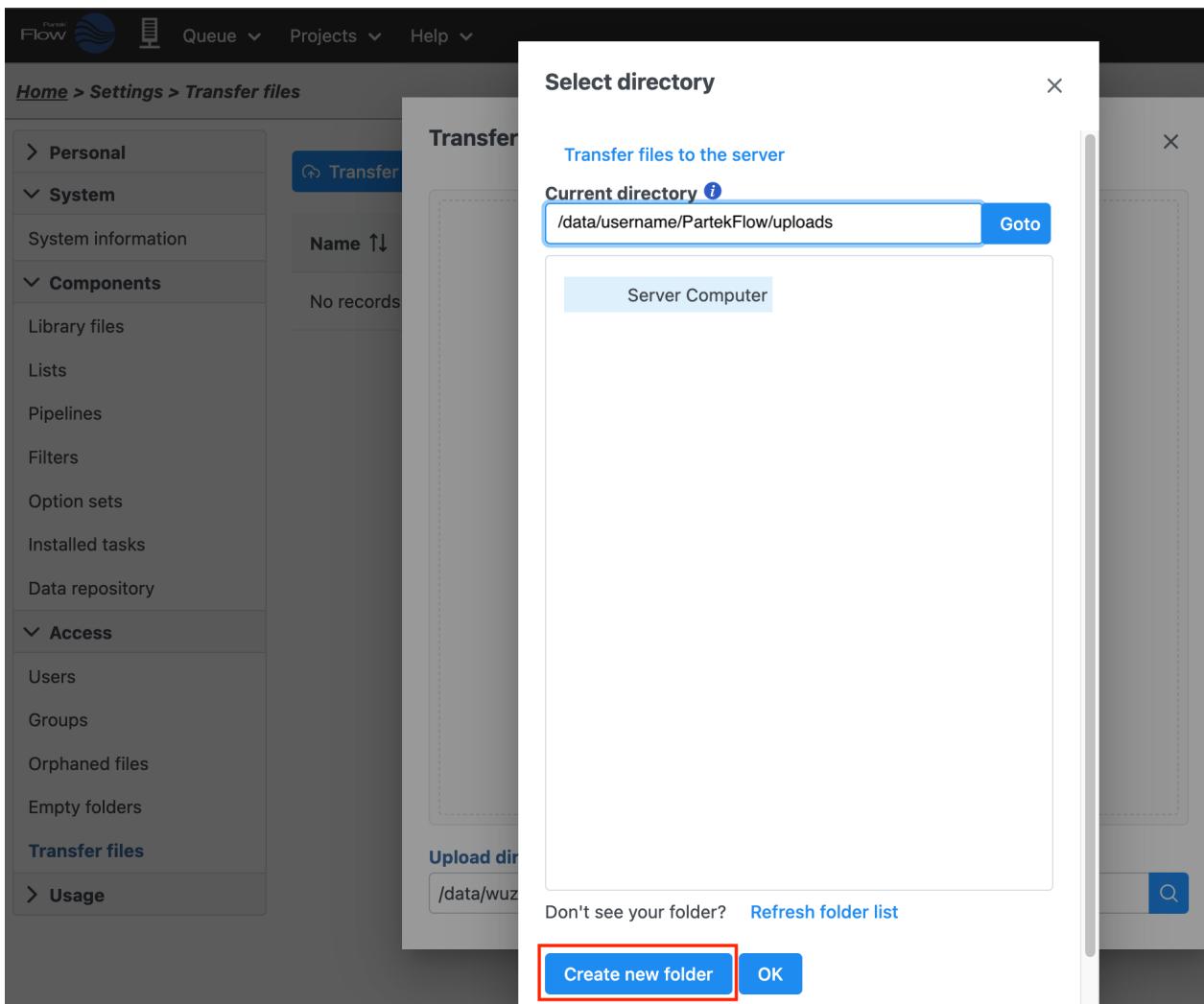
Enter the name of the folder (ie. hbr_uhr_fastq) and click the blue "Create" tab when ready.
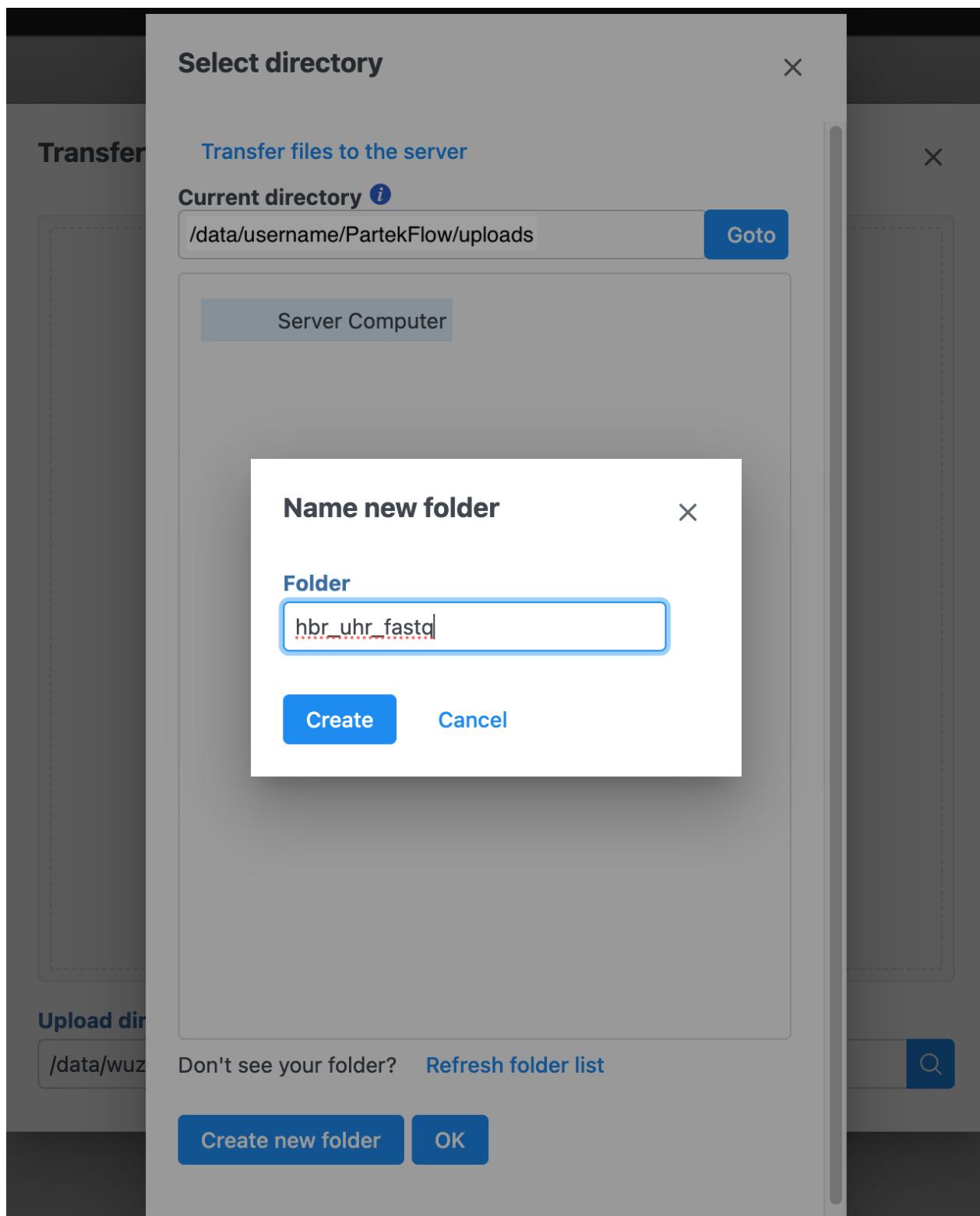
The "Current directory" is updated to reflect that the user is now in `/data/username/PartekFlow/uploads/hbr_uhr_fastq`.
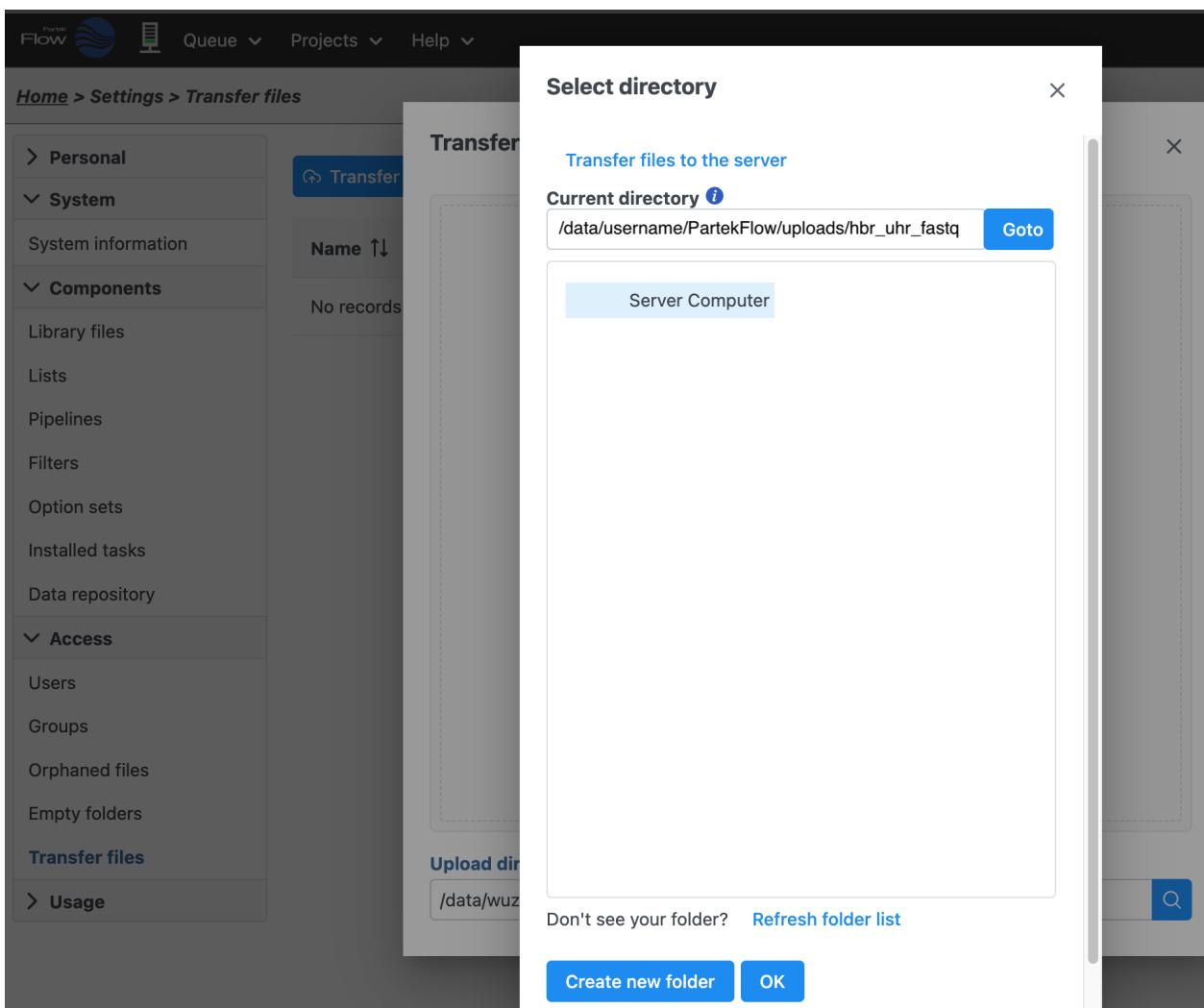
Click "Ok" to return to the file transfer dialogue box. Note that the "Upload directory" has been updated to `/data/username/PartekFlow/uploads/hbr_uhr_fastq` as well.

Twelve FASTQ files will be transferred to the folder `hbr_uhr_fastq` in `uploads`. Users can either drag and drop or browse to select files for upload. Click the green "Upload" button when ready (the number shown inside this button depends on the number of items to be transferred).

Transfer files   ✕

Cancel	12 files selected	+ Add more

UHR_1_R1.fq
53 MB

HBR_1_R2.fq
28 MB

HBR_3_R2.fq
30 MB

Upload 12 files
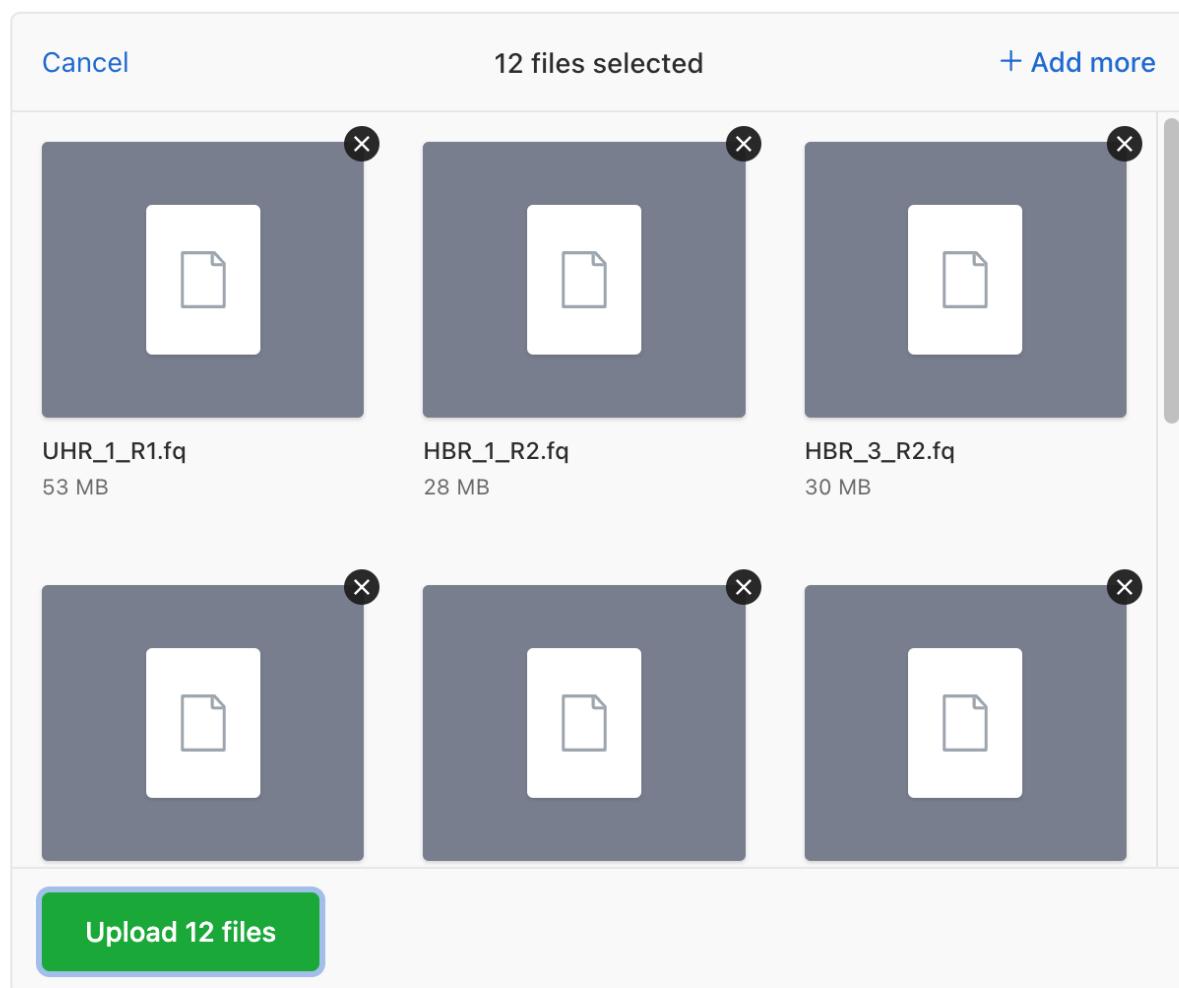
Upload directory

/data/username/PartekFlow/uploads/hbr_uhr_fastq   🔍

The warning message below will appear once the data transfer has commenced. Click "I understand".

A bar indicating the transfer status will also appear.

After the transfer is complete, users can click on the arrow next to the `/data/username/PartekFlow/upload` folder to view it subfolders and contents of subfolders.

> **Note**
>
> The user will not own contents transferred via the web tool. Thus, the only way to delete is using the "trash can" under the "Action" column. If user attempts to delete via command line, the message "Permission denied" will appear.

Listing the contents of the `/data/username/PartekFlow/uploads` folder on Biowulf will reveal the `hbr_uhr_fastq` subdirectory.

```
[wuz8@cn4304 username]$ ls -l /data/username/PartekFlow/uploads/
drwxrws---+ 2 partekfl partekfl 4096 May 28 14:26 hbr_uhr_fastq
```

An attempt to remove the `hbr_uhr_folder` will result in a message saying "Permission denied". In the `rm` command below, `-r` tells `rm` to remove a folder and `f` forces `rm` to delete without asking for confirmation.

```
[wuz8@cn4304 username]$ rm -rf /data/username/PartekFlow/uploads/hbr_
rm: cannot remove '/data/username/PartekFlow/uploads/hbr_uhr_fastq/hb
```

Importing Data to the Partek Flow Projects

Data import into Partek Flow projects

After data has been transferred to the NIH Partek Flow server, the next step is to import them into a project. Sign onto Partek Flow and hit the "New project" button click to create a new project called project1. The project will then appear in the project table. Click on project1 to enter the analysis interface.
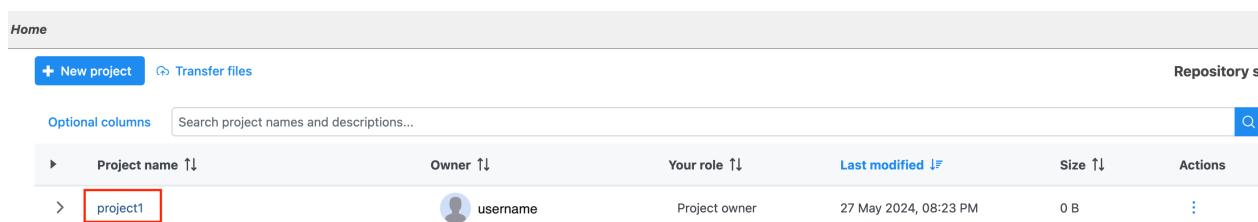
Click on the round button labeled "Add data" to begin adding data to the project.
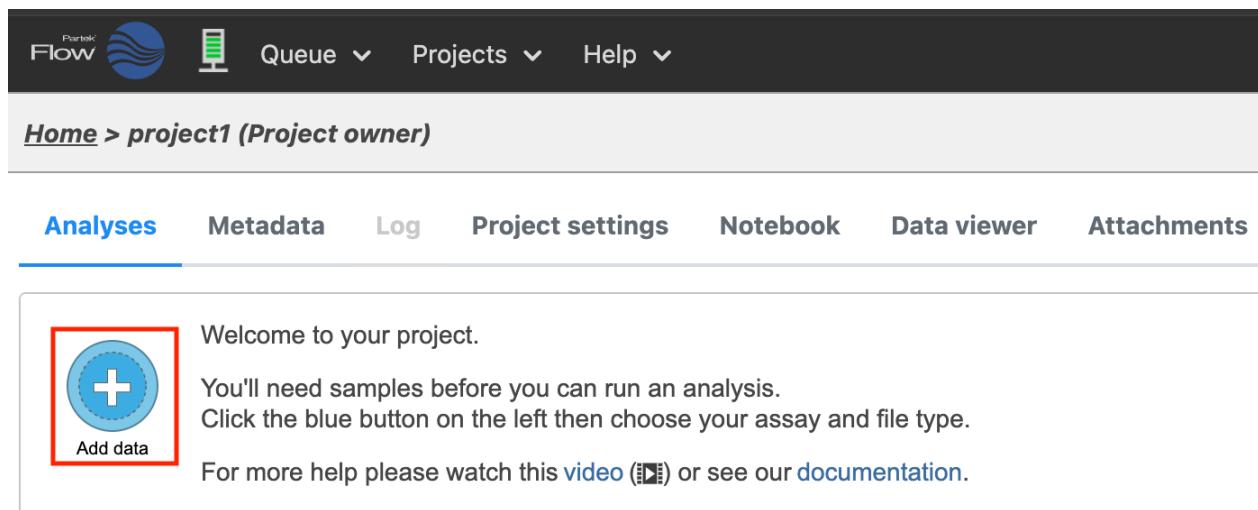
Users will be greeted with a range of options for data import. In this example, click "Bulk" and then select "RNA-Seq". FASTQ files will be imported. Click "Next" when ready.

In the subsequent page, the current directory is displayed in the file menu on the left.

> **Note**
>
> Users will only be able to access content in the `/data/username/PartekFlow` folder once signed on. So make sure to put the data there using one the methods described in this tutorial.

This example imports the FASTQ file in `uploads/hbr_uhr_fastq`. Be sure to check the box (labeled "Name") to select all files.

Scroll to the bottom of the data import page and select the data type (ie. mRNA) and click "Finish".

A rectangle labeled "Importing" will appear as data is being imported into a project.

Tip

The time it takes to import or to complete a task in Partek Flow depends on the queue. If the queue is long, then it's time for a coffee break.
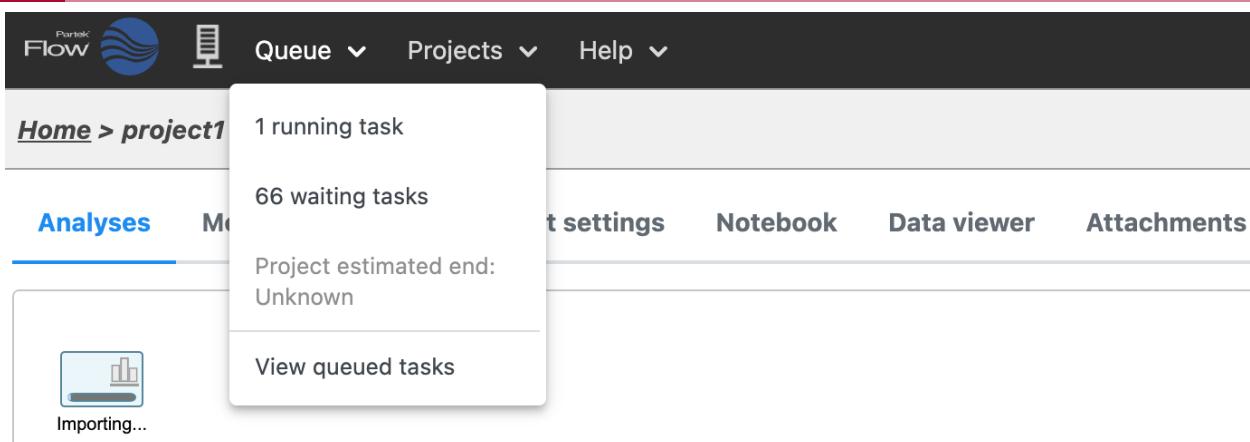
Upon successfully importing data, there will be a circular data node in the project.
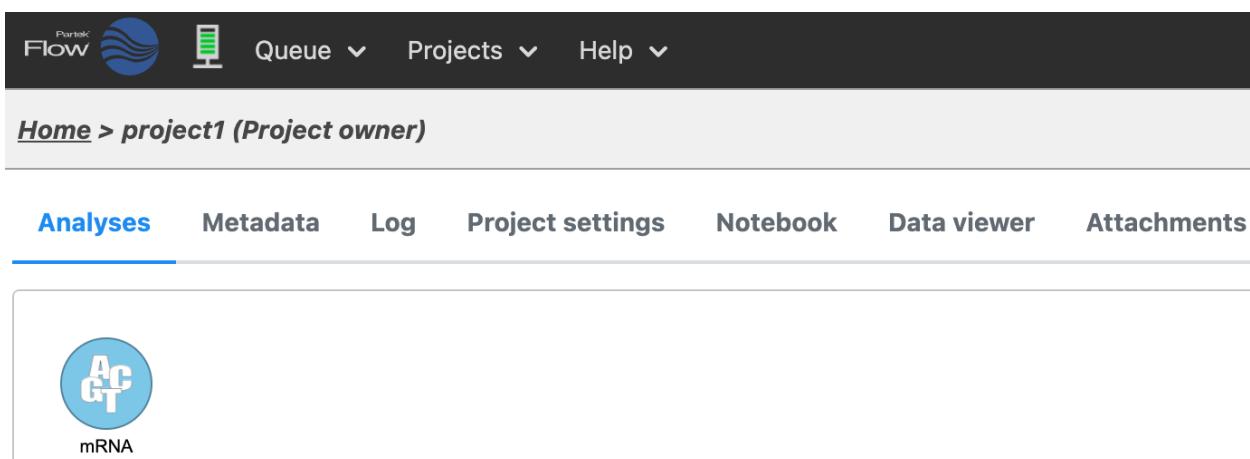
Right click on the data node to either delete this node, files, or change data type.
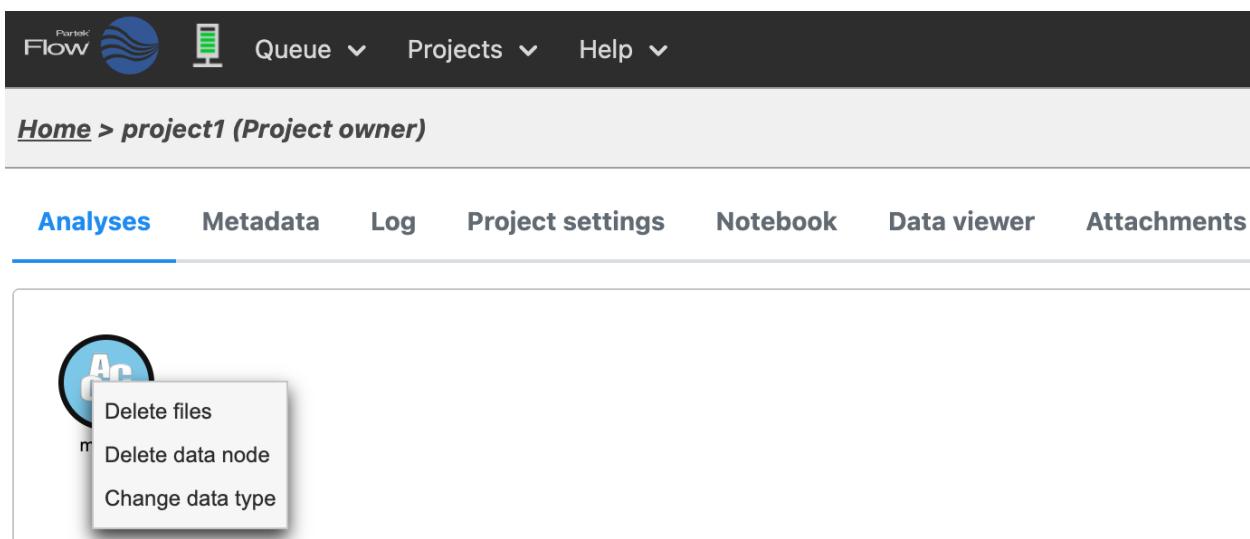
Next, click on the "Metadata" tab to add metadata for the samples. Under the "Sample attributes" sections, users can manually add metadata or assign it from a file. This example will demonstrate manually adding metadata by clicking on "Manage".

In the next page, select "Add new attribute".

In the dialogue box that appears, name this attribute (think of an attribute as a variable) "Treatment" so that users can distinguish the treatment group in which each sample belong. Be sure that "Categorical" is selected as this attribute type (ie. treatment group is categorical). Click "Add" when ready.
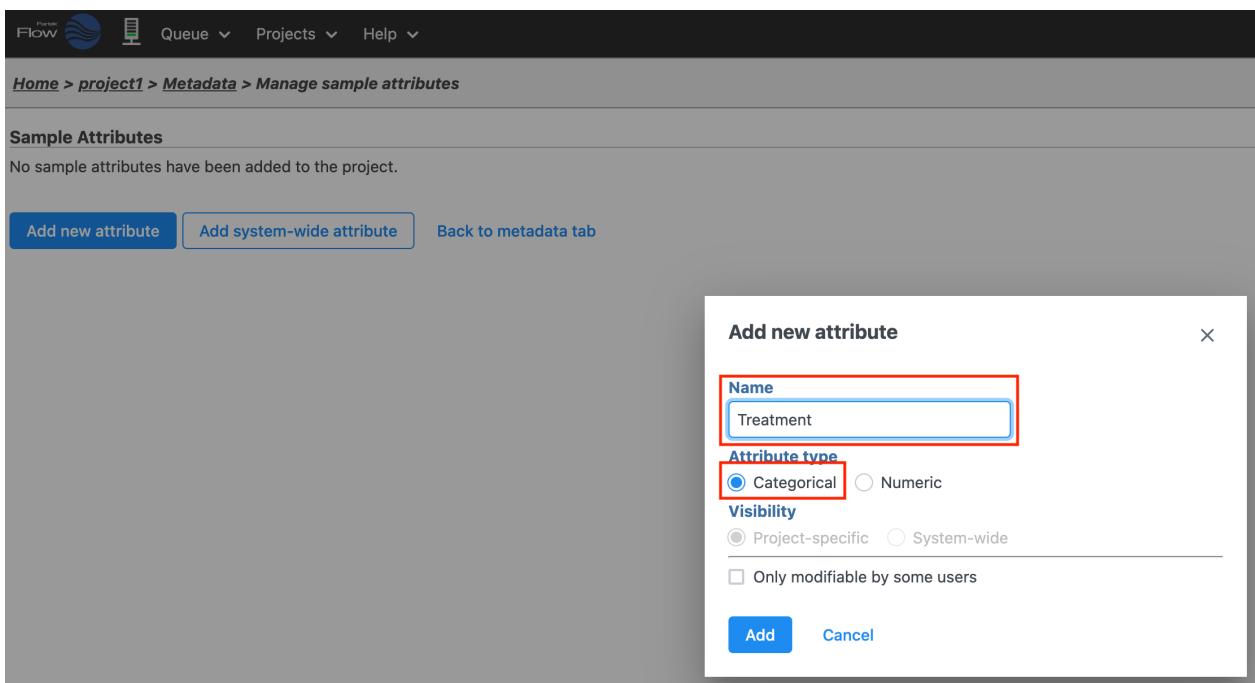
Then, under the "New category" box, enter the first treatment group (ie. HBR). Be sure to click the green "+" to add it.
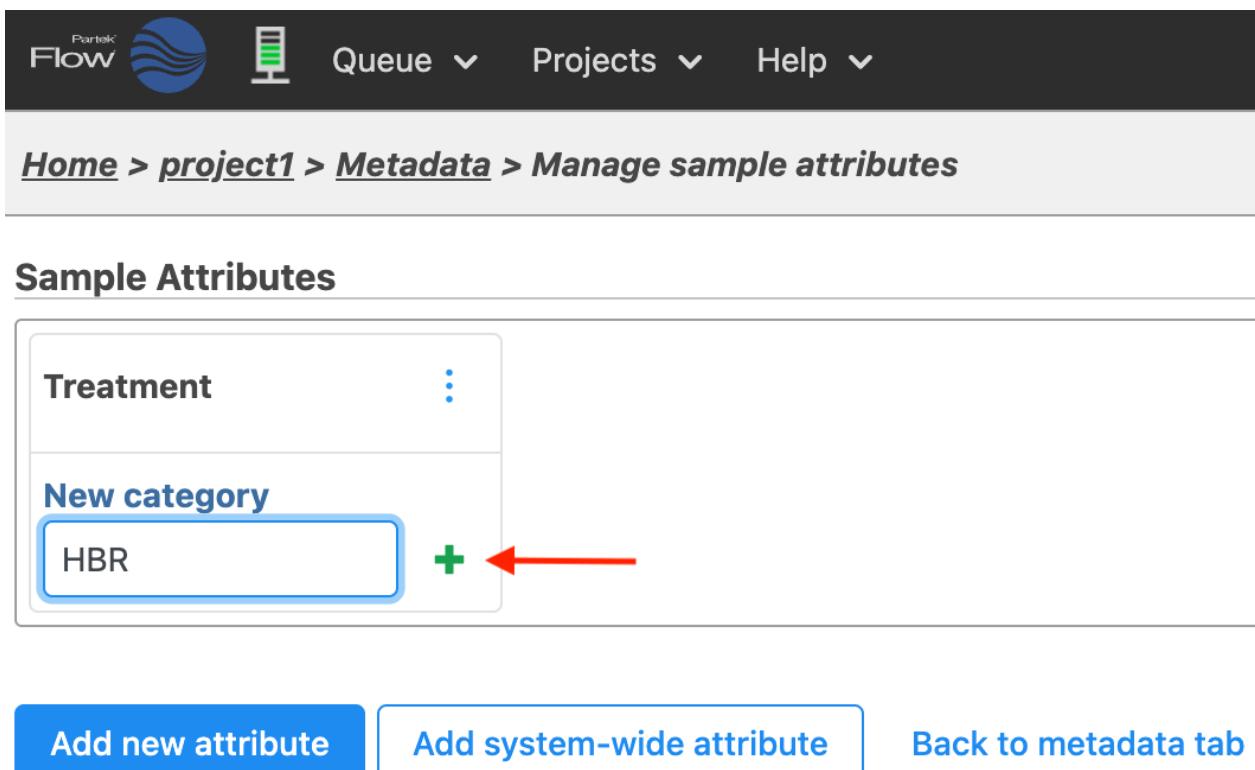
After that, add the second treatment group (ie. UHR) and click on the green "+" to add. Click the "Back to metadata tab" when finishing adding treatment groups.
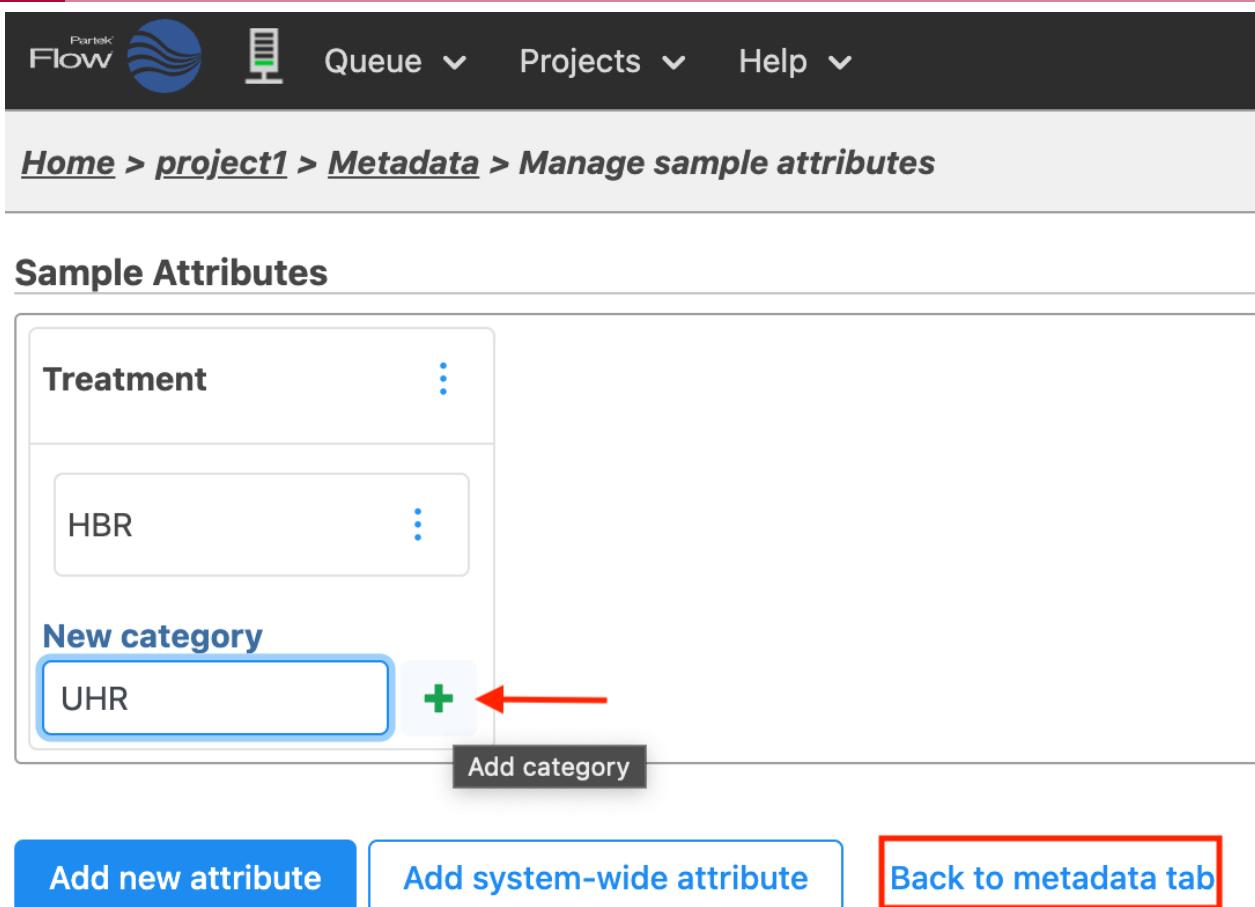
Hit the "Assign values" tab.

Then, select the appropriate treatment for each sample using the drop down box. Click on "Apply changes" when finished and then the "Analyses" tab to go back to the analysis.
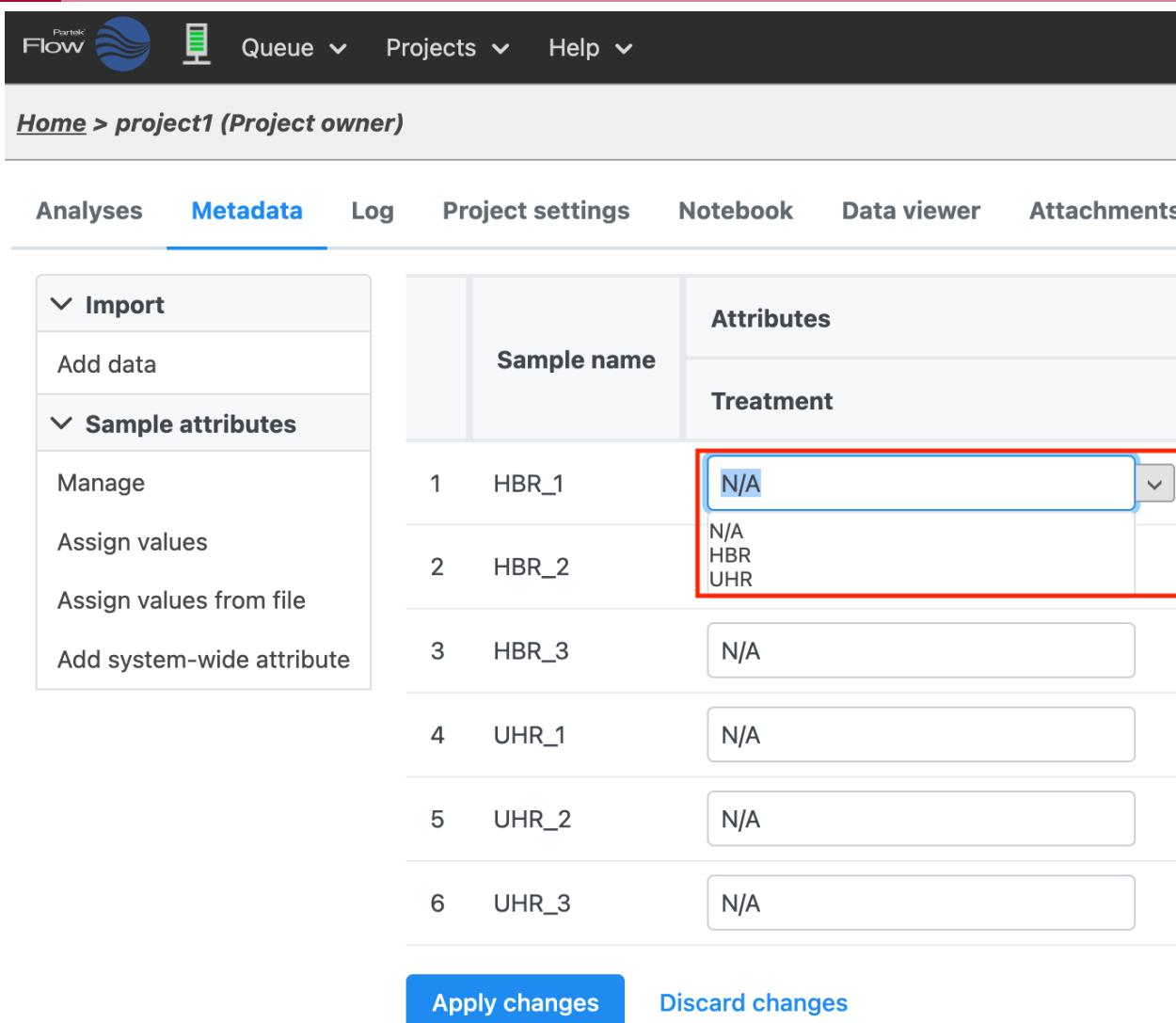
Click on a data node to see a menu of tasks that can be performed. For instance, the first step in analyzing high throughput sequencing data from FASTQ files is to perform pre-alignment QC.
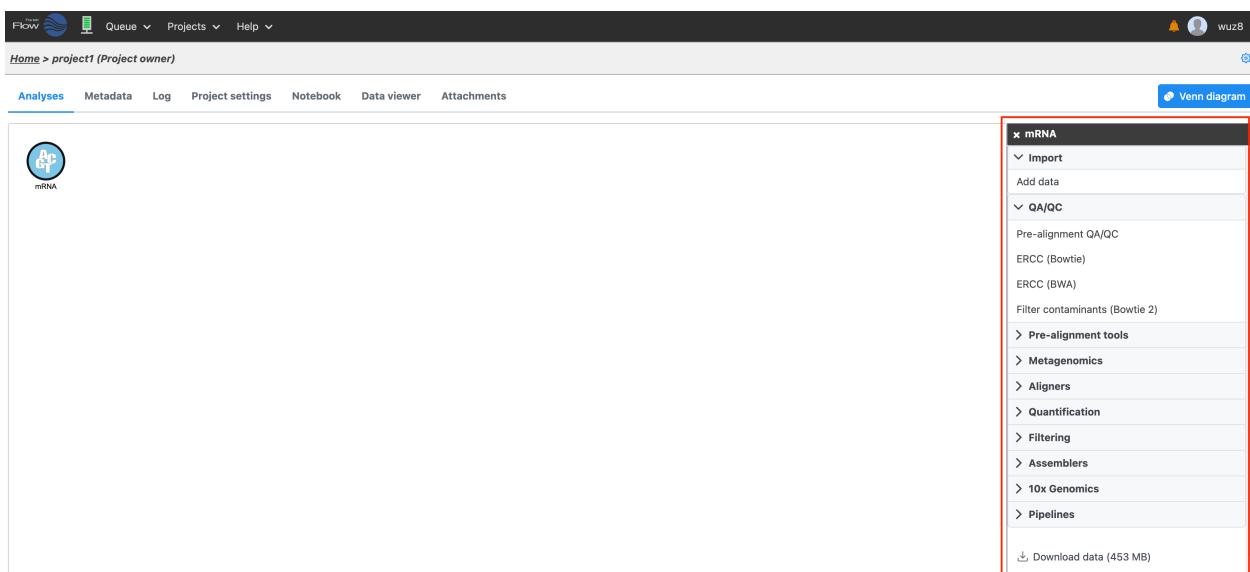
Note

In Partek Flow, data nodes are round and task nodes are rectangular. Running tasks are light blue rectangles with a status bar inside. Once the task is complete, the rectangle turns into a darker shade of blue.