**Sequence Data Transfer from Basespace to TAMU Sharepoint/OneDrive**

Prepared by: Lee Pinnell

February 15, 2023

This document outlines the steps required to install Basespace CLI, then use it along with Globus Connect to move the files from the Basespace server to TAMU’s personal/lab Sharepoint/OneDrive space via either TAMU’s TERRA and/or GRACE clusters. Steps 1 and 2 you only need to do once to set-up Basespace CLI. Step 3 onwards are required each time you want to download data.

**STEP 1: INSTALL BASESPACE CLI**

The command-line version of Basespace allows you to use a bunch of different functions via terminal. The installation is very simple, and usage is straight-forward as well! I recommend installing the binary in either your personal bin/ or your groups bin/. Follow the instructions on the BS CLI [webpage](https://developer.basespace.illumina.com/docs/content/documentation/cli/cli-overview). Briefly, use wget to download the Linux binary where you want to store it. In the example below I store it in my personal bin on Terra/Grace:

# move to my personal bin

cd /scratch/user/ljpinnell/bin/

# create directory to store binary and navigate to it

mkdir basespace

cd basespace

# download the binary

wget https://launch.basespace.illumina.com/CLI/latest/amd64-linux/bs

# modify permissions on the binary

chmod u+x bs

After successful installation you should be able to run the the bs command. I always add new programs to my PATH instead of having to feed my commands the absolute PATH to executables, so I do that next.

cd /home/ljpinnell

nano .bash\_profile

Then within nano I add a comment line and the actual path:

## adding basespace to my path

PATH=$PATH:/scratch/user/ljpinnell/bin/basespace/

Exit and save the bash profile, then source it so the new path will be recognized:

source .bash\_profile

You should now be able to run bs from anywhere without giving the whole path.

**STEP 2: AUTHENTICATE YOUR BASESPACE CLI ACCOUNT**

This links your CLI version to your BS account, and is super simple:

bs auth

It will provide you with a link to copy into a browser. From there, login with your info and it should connect your account. If successful it will give you a nice greeting in terminal.

*Note: You can only link to one account at a time. But, if you need to switch accounts use the*

*‘--force’ flag to authenticate to a different account. You can switch back and forth by using the ‘--force’ flag as needed.*

**STEP 3: IDENTIFY AND DOWNLOAD THE PROJECT FASTQ FILES**

We want to list the projects associated with our account:

bs list projects

This will output a table that has the ‘Name’, ‘Id’, and ‘TotalSize’ as columns. You will use the ‘Id’ number as input into the actual download command in the next step. Identify the ‘Id’ (Project ID) that you want to download, highlight it, and copy it for pasting into the next command. In this case, we are interested in the actual FASTQ read files so we give it the extension flag for FASTQ. I output my reads to a directory called reads/ inside my project directory. In the example below my project is called ‘20230215\_Bovine\_LA\_16S’, and the Project ID is ‘358155798’. You’ll replace the number with the one you copied above and the project directory with a meaningful title. The command will create a directory called reads for you, so don’t create it beforehand. These commands are happening in my /scratch/user/ljpinnell/ directory:

# make project directory and move into it

mkdir 20230215\_Bovine\_LA\_16S

cd 20230215\_Bovine\_LA\_16S

# download the FASTQs from this project on basespace

bs download project -i 358155798 --extension=fastq.gz -o reads

This will take a couple of mins to start download, but when it does you will see it working in your terminal. Once it finishes download, I clean up the output. It will put the R1 and R2 files for each sample in the project into a separate sub-directory within reads/. I agglomerate all my reads into one directory (reads/), then remove all the empty subdirectories. This is happening in the project directory (20230215\_Bovine\_LA\_16S):

# move all my fastq.gz files into the reads directory, then remove the now empty sub-directories

mv reads/\*/\*.fastq.gz reads/

cd reads

rmdir \*

You’ll get a ton of warnings says all the fastq.gz files aren’t directories. Ignore it! Now you should have all your fastq.gz files and one .json file in your [PROJECTID]/reads/ directory.

This completes the Basespace part of the process.

**STEP 4: USE GLOBUS CONNECT TO MOVE FILES TO SHAREPOINT**

To quickly and easily move files from Terra or Grace to SharePoint site, Globus is the best way. Before transferring via Globus it’s easiest to have your destination on OneDrive set-up in advance.

1. As a manager of the SharePoint site, create a new library for you to store your sequence data in (ours is called ‘Sequence\_Data’).
2. Within that library create a directory for the sequencing project. Following the previous example, I’d create one called ‘20230215\_Bovine\_LA’ with a 16S/ subdirectory (this is in case we have other types of sequencing data that I would store in a different sub-directory, i.e., AMR-TE/).
3. After creating the project directory, share the overall directory (i.e., 20230215\_Bovine\_LA) with yourself by selecting share from the three dots beside the directory. Find yourself by your last name in the fill-in menu option and press ‘Send’
4. As far as I can tell you need to share it with your personal account in order to find the directory in Globus

Go to app.globus.org and login with your TAMU ID. Once logged in, you need to find the endpoints for OneDrive (SharePoint) and the cluster your data is on. At some point in this process you will need to login again using your NetID and Duo.

1. Go to the ‘Collections’ tab and search for ‘TAMU AIP Azure OneDrive Globus Storage Collection’. That is where your SharePoint site is accessible.
2. Click on the name of the collection (in this case TAMU AIP Azure OneDrive Globus Storage Collection), which will bring up a new page.
3. Select the ‘Open in File Manager’ link which will bring up the File Manager tab and have the TAMU AIP Azure OneDrive Globus Storage Collection open on the left side of the page.
4. From the blue bar, click on ‘up one folder’, which will change the directories listed below
5. Move into the directory ‘Shared’ by clicking on the directory name
6. This should now show you the directory you shared with yourself in C) above. Click on that to enter into it.

You are all set on the SharePoint end to receive the files. We will now bring the cluster (in this case Terra) up on the right side of the page:

1. On the right side, click on the ‘Search’ bar. Type in ‘TAMU terra-ftm’. That is the name of the collection for Terra. *TAMU grace-dtn* will work for accessing Grace.
2. Click on TAMU terra-ftn. Login using your username and password. You will receive a Duo prompt. Now you should have your /home/ directory listed below on the right side of the File Manager page.
3. Put the path to the directory containing the reads into the ‘Path’ bar near the top of the right side and under ‘TAMU terra-ftn’. In the example above it would be: /scratch/user/ljpinnell/20230215\_Bovine\_LA/. Hit Enter and you should see all your sequences below.

Now everything is set to move files from the terra (right side) to SharePoint (left side).

1. Make sure you have the proper directory listed for your destination (SharePoint). It can be a pain to move things once inside SharePoint.
2. From the middle of the page, drop down the ‘Transfer & Timer Options’ box. Label the transfer something meaningful so it’s easy to figure out which transfer is which.
3. Use ‘select al’l on the right side to select all of your sequence reads
4. Hit the ‘Start’ button on the right side. After a few seconds a pop-up box should say your transfer has started successfully.

To check on the status of transfers select the ‘Activity’ tab the vertical bar on the left side of the page. Here you can see how your transfers are going.

That’s it, your files should now be in Sharepoint/OneDrive!

**Notes:**

* **MiSeq generated data (i.e., 16S) will finish within a couple of minutes.**
* **HiSeq/NovaSeq runs (i.e., metagenomics, AMR-TE) runs will take hours.**
* **OneDrive will almost certainly throttle the speed at some point and may also timeout. The timeount may be the result of a few different errors related to OneDrive throttling speeds and your progress bar will turn yellow. That is okay and in my experience happens nearly every time the sum of the data is 1TB or bigger ! Just let it keep going. Eventually it will finish.**