Step 1 - Downloading your sequencing data

Prepared by: Enrique Doster

Last modified: 2024-02-16

This brief guide is intended to introduce a VERO team member to how they will likely access sequencing data and download it for analysis.

If instead you need to access data coming from your sequencing core, you can review this tutorial for more information on transferring data from basespace, to the TAMU HPRC, and finally to the group's sharepoint for final storage.

Table of Contents

Preparing the data you're getting

- Some questions to ask yourself
    - What kind of samples are you analyzing?
        - Do you need to run qiime2 or AMR++?
        - What needs to be installed in the server?
    - How much storage space do I need to fit the raw samples?
    - How much storage space will I need to complete the analyses?
        - E.g. if running AMR++, you could try running the entire pipeline at once which will require 2-3X the storage space compared to the raw data, or you can run each piece of the pipeline individually to better manage your space.
    - Are my samples sequenced across multiple lanes

Downloading samples from sharepoint (onedrive) using Globus

The TAMU HPRC has more information about ways to transfer files here, but generally we recommend using Globus to transfer files between onedrive and TAMU's servers.

You'll need to log into globus and can follow these instructions for more information.

In short, you'll go to the File Manager and select the "TAMU AIP Azure OneDrive Globus Storage Collection" as one of your endpoints. From the starting point, you'll have to click the "up one folder" button to then see the "Shared" folder where you'll find the folders that were shared with you.

Set up your second endpoint as either "TAMU Grace - dtn" or "TAMU Terra - fnt", navigate to the right folder, and click "Start" on the side with the OneDrive folder to begin transfer toward the TAMU server. On the server, you'll typically use directory in your scratch space, like this "/scratch/user/enriquedoster/".

Special consideration for sequencing across multiple lanes

For shotgun metagenomic samples, we often split a sequencing pool into multiple lanes to improve sequencing depth or avoid lane effects. If that's the case, you'll need to combine ("concatenate"), these files systematically. For each sample, you'll need to concatenate the forward reads in the same order across the sequencing lanes and repeat the process for the reverse reads. Here's an example of a script you can use to concatenate sample reads found in different folders (remember to add the appropriate sbatch commands if you want to submit it as a job for bigger datasets).

For most cases, concatenation is fast and doesn't require additional HPRC resources. Therefore, you can run the script as a typically "bash" script and wait for it to finish before logging off that terminal.

To do this:

1. On the terminal, navigate to a good location with storage space
2. Put the example script above into your terminal (e.g. using nano, or uploading a text file)
3. This script uses positional arguments for each of the directories with your fastq samples (see below).
    - Change "/path/to/dir1" to the absolute path to your reads
    - do the same for the next lane and replace "/path/to/dir2", etc
    - only do this for your input directories
4. The script automatically outputs the concatenated samples into a new directory, "cat_reads" in your working directory.

For example:

```
bash concat_reads_multiple_directories.sh /path/to/dir1 /path/to/dir2 /path/to/dir3
```

Next steps

Depending on whether you're analyzing 16S reads or shotgun metagenomic reads (including target enriched reads), you'll be following similar steps but using different software.

Generally, the bioinformatics steps you are performing are:

1. QC evaluation
2. QC trimming
3. Contaminant removal (e.g. host DNA, mitochondrial DNA, etc)
4. Read classification (e.g. taxonomic, antimicrobial resistance, functional genes, etc)

Follow the relevant tutorials to use qiime2 for 16S analysis and AMR++ for the resistome.