Molly Mitchell, Ph.D
Lakshmi Thsaliki, MS
bphl-sebioinformatics@flhealth.gov

HiPerGator Analysis Reference Guide

This document provides a basic overview of the steps involved in setting up your compute environment on HiPerGator (HPG) to performing next-generation sequencing analysis. The following steps are meant to be followed with assistance from the Southeast Region Bioinformatics Regional Resource Lead (BRR). Information in this document may change as software or pipelines are updated. It is strongly recommended for any new data analyst to complete new HPG user training with the BRR and/or review the slides and recordings for each training at https://github.com/StaPH-B/southeast-region. New HPG user training includes Introduction to High Performance Computing (HPC)/HiperGator (HPG), Introduction to Linux – Part 1, and Introduction to the BaseSpace Command Line Interface. Additional walk-throughs with the BRR include compute environment set up and tool installation (as outlined in this document) and analysis pipeline trainings. Please contact the BRR at bphl-sebioinformatics@flhealth.gov for further assistance and one-on-one tutorials, as needed.

Account setup and HiPerGator access

Each person performing data analysis needs to have their own HPG account. Please see the document, “How to Request and Access Your HiPerGator Account” for detailed instructions for account setup and access (https://github.com/StaPH-B/southeast-region/blob/master/hipergator/20220216_SoutheastRegion_HiPerGatorAccountAccess.pdf).

Software installation

In order to access HPG, you need two software programs, Putty and WinSCP, installed by your IT department on your Windows computer. Putty allows you to login to HPG and work in the command-line via a terminal. WinSCP allows you to transfer files from your local Windows computer to HPG. Note – There are other programs besides Putty and WinSCP that do the exact same thing. Discuss these options with your IT department. Putty and WinSCP are free software that are usually approved for use by most IT departments.

The first time a user logs into HPG to perform any analysis, they need to setup their computing environment.

Set up your python computing environment

1. Downloaded and install Anaconda

```
$ wget https://repo.anaconda.com/archive/Anaconda3-2022.05-Linux-x86_64.sh
$ ./ Anaconda3-2022.05-Linux-x86_64.sh
```

Molly Mitchell, Ph.D
Lakshmi Thsaliki, MS
bphl-sebioinformatics@flhealth.gov

Install Illumina's BaseSpace Command Line Interface

See the document "Instructions to Install and Configure BaseSpace CLI" for detailed instructions on how to install and use the BaseSpace CLI.

Set up your directory for storing and using Singularity images on HiPerGator

```
$ mkdir /blue/bphl-<state>/<user>/singularity/
$ ln -s /blue/bphl-<state>/<user>/singularity/ ~/.singularity
```

By default, Singularity will try to store your images in /home/<user>/.singularity/. However, these files can be quite large and may exceed your home directory storage quota. The above commands will make a symbolic link to this directory so the files can actually be stored on the /blue drive and use your storage quota. Period is given as it is a hidden directory.

Set up your directory for using NCBI on HiPerGator

```
$ mkdir /blue/bphl-<state>/<user>/ncbi/
$ ln -s /blue/bphl-<state>/<user>/ncbi/ ~/ncbi
```

By default, NCBI will try to store your images in /home/<user>/ncbi/. However, these files can be quite large and may exceed your home directory storage quota. The above commands will make a symbolic link to this directory so the files can actually be stored on the /blue drive and use your storage quota. Period is not required as this is a regular directory.

To install BioPython in base conda environment

The command is "**pip install biopython"**

To install Pandas in base conda environment

The command is **"pip install pandas"**

Data/file transfers

Sequencing data (e.g., fastq files) can be transferred to HiPerGator, at least, one of two ways.

Transfer fastqs directly from BaseSpace to HiPerGator

Molly Mitchell, Ph.D
Lakshmi Thsaliki, MS
bphl-sebioinformatics@flhealth.gov

To transfer fastq files from BaseSpace to HiPerGator, use the following command (installation of the BaseSpace CLI is required):

```
$ bs download project -n <projectname> --extension=fastq.gz -o <output dir>
```

Transfer files from local computer to HiPerGator

To transfer fastqs (or other files) from your local Windows computer to HiPerGator, use WinSCP. The hostname for HiPerGator is “hpg.rc.ufl.edu”. Enter your username and password, accordingly. To transfer files, use the drop-down menus to navigate to your directory with your files on your local computer, and navigate to the directory on HiPerGator where you want to transfer your files to. Drag and drop your files.

Storage

To check the available space in your orange or blue storage, you can navigate to the command line and enter:

blue_quota - to check storage in blue
orange_quota – to check storage in orange

Folder structure and best practices

/home/<user>/

When you log into HiPerGator, you will be located in your home directory (/home/<user>/). This space has limited storage and is not high-performance. Do NOT run jobs or store data in this location. Your home directory is used to keep very minor files, install software, etc.

/blue/bphl-<state>/

This is your state’s group directory. You will find individual user directories and your state’s share and public-share directories here.

/blue/bphl-<state>/share/

This is a shared space for all bphl-<state> users. All files in the ‘share’ directory can be viewed by the group’s users but are private to all other HiPerGator users.

Molly Mitchell, Ph.D
Lakshmi Thsaliki, MS
bphl-sebioinformatics@flhealth.gov

We recommend that you transfer your sequencing data (and any other shared files) to /blue/bphl-<state>/share/ and make a “data” folder to keep a master copy of your data that you have transferred to HPG. This will eliminate redundant copies in individual users’ folders.

/blue/bphl-<state>/public-share/

This directory has the same permissions as /blue/bphl-<state>/share/ with the exception that bphl-florida users have access to this folder. This space is used to transfer data and other files to the BRR in bphl-florida. Additionally, this is where the BRR will share files, scripts, pipelines, databases, etc. to bphl-<state> users.

/blue/bphl-<state>/<user>/

This is your user directory within your state group. This is where you will you perform your analyses. We recommend the following best practices to keep your files organized.

1) Make a directory specifically for the project or analysis you are performing.
2) Within your project folder, make a new “fastqs” folder. Copy the fastq files located in /blue/bphl-<state>/share/ to this new folder.
3) All scripts/pipelines will be located in /blue/bphl-<state>/public-share/scripts/. Copy the pipeline you wish to use to your project folder.
    a. Note – This is not necessary for the pipeline to run, but it provides quick documentation for yourself to remember what you ran if you need to go back to troubleshoot (especially if updates were made to a prior version of a pipeline).

srun4 for interactive session

- For quick and small interactive jobs you can use the command "srun4" which will start an interactive session using 1 cpu 4gb RAM for 8 hours.
- For testing out tools in the command line that require more resources like space and time, you can use the below command

```
srun --qos=bphl-umbrella --account=bphl-umbrella --cpus-per-task=<number of cpus> --mem=<memAmount>gb --time=<time limit> --pty bash -i
```

- Time limit should be specified in the above command, by default it is ten minutes and at the end of the time it will kill your job and take you back to a login node.

Analysis scripts/pipelines

Molly Mitchell, Ph.D
Lakshmi Thsaliki, MS
bphl-sebioinformatics@flhealth.gov

Below is a list of pipelines and scripts that are currently available for use on HPG. Prior to your first use, please speak with the BRR to learn more about how each pipeline/script works.

Pipelines

- FLAQ (FLorida Assembly Quality) – generates annotated, de novo assemblies and quality metrics for WGS of bacterial species

- FLAQ-AMR (FLAQ-Antimicrobial Resistance) – generates annotated, de novo assemblies and quality metrics for WGS of bacterial species; determines species identification; determines ST using MLST schemes from PubMLST; identifies AMR genes, virulence genes, and plasmids; performs serotyping of *E. coli*, if applicable (more species to come).

- FLAQ-SC2 (FLAQ-SARS-CoV-2) – generates SARS-CoV-2 consensus assemblies from ARTIC V1, V2, or V3 targeted amplicon sequencing using Illumina (e.g., Nextera XT or Flex) and non-Illumina (e.g., PrimalSeq or MN Tailed) library prep. Outputs variant file and final report with quality metrics (including a PASS/FAIL quality flag based on public repository submission criteria). Automatically generates flags if indels or internal stop codons are present, indicating the need for manual review. Individual scripts also are available to prepare and format assemblies for batch submissions to GISAID and NCBI's Genbank, and to remove indels/SNPs that are likely PCR or sequencing artifacts/errors prior to submission).

- Targeted Amplicon Variant Calling and Consensus Sequence Generation – identify variants and generate a consensus sequence for HIV genotyping (will work for any target/amplicon – just need to provide a reference sequence).

- FL-cgSNP (Core-Genome SNP analysis) – reference-free method for pan-genome analysis to identify core genes shared by all isolates and generates a multiple sequencing alignment, pairwise SNP matrix, and maximum-likelihood phylogenetic tree.

- hqSNP (High-Quality SNP analysis) – reference-based method using CDC's Lyveset pipeline to identify hqSNPs among isolates; generates a pairwise SNP matrix, and maximum-likelihood phylogenetic tree

In Development:

- FLAQ-Lp (FLAQ-*Legionella pneumophila*) – includes CDC's Lp Species ID Tool and CDC's Lp Serotyping Tool (individual scripts using these tools are already available on HiPerGator).

- FLAQ-USC (FLAQ-Unknown Species Classification) – Generates a consensus species identification (or nearest neighbor classification) for difficult to identify isolates based on

Molly Mitchell, Ph.D
Lakshmi Thsaliki, MS
bphl-sebioinformatics@flhealth.gov

culture alone (individual scripts using the pipeline tools are already available on HiPerGator).

Individual Tool and Scripts

- Quick Species ID (screen against RefSeq database)
- Species ID, contamination check, and metagenomic classification
- Run CDC's Lp Species ID Tool and Lp Serotyping Tool locally
- Pull out gene sequence of interest from assembly (e.g., pull out AR genes of interest)
- Merge fastqs from multiple lanes on a NextSeq into one R1 and one R2 file (compatible with BioNumerics)
- Download fastqs from NCBI's Sequence Read Archive
- Scripts to run any tool individually (e.g., SeqSero2, mlst, abricate, etc.)
- Run any tool at
- Batch runs (i.e., run >2 analysis scripts at one time)
- **Custom scripts and pipelines as requested or needed**