0014\_MiSeq\_Analysis

Walker Symonds-Orr

2023-08-30

First, access the cluster. Make sure your PIV card is inserted.

At the terminal, type:

ssh orrwe@ai-submit1.niaid.nih.gov

Before using the cluster, you need to create your own node, so you are not using common resources on the cluster. You will get automated (and even) emails from IT if you fail to do this!

qrsh

qrsh has other arguments as well, which can be used to control the size of the node you get, for example.

Before going forward: remember that many of the directories on the cluster are shared, and that it’s very important to treat it with respect and not do anything of which consequence you are uncertain.

## Demultiplexing

The first step in sequence analysis is de-multiplexing, which means taking bcl binary files from the sequencer and converting them to fastq files while at the same time identifying which adapter index belongs to which sample.

Identify the directory that contains your sequencing data. First, you will want to connect to LOCUS where our data is stored. In finder, press ⌘K (Go > Connect to Server). Then type

smb://locusdata.niaid.nih.gov/

and select LVD\_QVE.

Within this, select the folder for Sequencing Data and navigate to your sequencing run.

create a new directory to house your demultiplexed sequences called “demux” or something to that effect.

blc2fastq is a “module”, which is like an application you can use on the cluster. Before using it, you must load it. To see available modules, type

module avail #followed by a string representing a search of the modules

If you don’t supply an argument afterwards, this command will produce a long list of modules. Type bcl and you will get a list of bcl2fastq modules. Choose the most current iteration by typing:

module load bcl2fastq2/2.20

Remember that you can use tab-complete!

bcl2fastq module is now loaded on your node. You can type

bcl2fastq help

to get a sense of what arguments this function takes.

When you are ready to demultiplex the dna, you will type

bcl2fastq -i "input\_directory" -R "run\_directory" -o "output\_directory"

The arguments in quotes will be replaced with the specific directories at hand.

The input directory [-i] will be inside the directory with the sequencing run data created by the MiSeq. This will be under ./Data/Intensities/BaseCalls. You can copy the folder as pathname or use tab-complete. If you copy as pathname make sure to change the beginning of the address from /Volumes/LVD\_QVE to /hpcdata/lvd\_qve.

An example from a recent sequencing run I did is:

/hpcdata/lvd\_qve/Sequencing\_Data/QVEU\_Seq\_0087\_Miseq\_Capsid\_Deletion\_Library\_EV71\_WalkerOrr/QVEU\_Seq0087\_WO\_WB\_04\_Capsiddel\_Replicationdel\_DIMPLE/230820\_M02211\_0114\_000000000-KD2BY/Data/Intensities/BaseCalls

The Run directory [-R] (note the capital R) will be the root directory from the sequencing run.

For the output directory [-o], use the demux folder you just created.