Pronghorn Homework Module

All personal directories are located: /data/gpfs/assoc/biol\_bids-2/

Setting up

Submission script is located:

/data/gpfs/assoc/biol\_bids-2/atram.slurm.sh

Copy this file to your personal directory.

In your personal directory you will need to create two directories:

mkdir temp

mkdir atram\_out

Below is an example of the submission script and the bold sections within square brackets should be changed accordingly.

*atram.slurm.sh*

#!/usr/bin/bash -l

#SBATCH --job-name=**[give your job a descriptive name]**

#SBATCH --account=cpu-s5-biol\_bids-2

#SBATCH --partition=cpu-core-0

#SBATCH --nodes=1

#SBATCH --ntasks=1

#SBATCH --cpus-per-task=1

#SBATCH --time=14-00:00

#SBATCH --mem-per-cpu=3500M

#SBATCH --output=hostname\_%j.out

#SBATCH --error=hostname\_%j.err

#SBATCH --mail-type=ALL

#SBATCH --mail-user=**[your email]**

singularity exec /apps/atram/aTRAM.sif python3 /aTRAM/atram.py -b /data/gpfs/assoc/biol\_bids-2/Course\_content/atram\_files/AlspHabad/AlspHabad -t /data/gpfs/assoc/biol\_bids-2/**[netID]**/temp -Q /data/gpfs/assoc/biol\_bids-2/Course\_content/atram\_files/query\_files/**[Number of genes]**\_proteins.fasta -a trinity -o /data/gpfs/assoc/biol\_bids-2/**[netID]**/atram\_out/

Target files

All target files are found here: /data/gpfs/assoc/biol\_bids-2/Course\_content/atram\_files/query\_files/

Each file begins with the number of genes it contains. In class you will assemble 2 genes. For your homework you will assemble 20 genes.

Processors

You will need to select the number of processors you want to use per job. You can change the number of tasks and CPUs per task in the SBATCH options in the submission script. Pronghorn has 32 CPUs per node. Pronghorn will allocate the resources you request to aTRAM as it runs. You can start with a single task and CPU per task, measure the gene assembly rate, and change the computational resources from there.

Submitting

Submit the job using:

sbatch atram.slurm.sh

Queue

To check on the status of your job, use squeue in the command line. This will pull up a list of all of the jobs currently running or waiting in the queue. You will see your job by the job name and your user name. If your job is waiting in the queue, the time will say 0:00. Once it starts running the timer will start. You can check on your jobs only with squeue -u <USERNAME>

**For your homework you will assemble 20 genes.**

**Answer the following questions:**

1. What is the purpose of High Performance Computing (HPC) and when might you use an HPC cluster?

High performance computing provides users with the memory resource of several CPUs to enable a faster and more efficient way of performing a task. HPCs are used when running jobs that involve a large dataset (such as transcriptomic data). In addition, HPC is ideal for running jobs that can be split into parallel processes.

2. What is a slurm script? Print the path of your slurm script here (including file name).

A slurm script is a “job control” file that contains all the details about the job to run such as job name, computational resources to be allocated, associations and other information. In addition, slurm script may (or may not) contain bash script to run the task.

/data/gpfs/assoc/biol\_bids-2/eadeyemi/atram.slurm.sh

3. What is the path to your gene assemblies on pronghorn (i.e., output files).

/data/gpfs/assoc/biol\_bids-2/eadeyemi/atram.slurm.sh

4. How many CPUs per task did you ask for?

Two

5. How many genes assembled?

19 genes

6. Did all genes assemble? Why or why not?

No. A reason could be that the program did not find a gene that matches with high confidence to one sequence of the 20 provided.

7. How long did it take for them to assemble? What processes might have slowed this down?

36:03. A considerable amount of computational resources is required to assemble 20 protein sequences, and this process might have been a little overwhelming to perform on a single CPU.