SPAdes on HPC

What is SPAdes?

SPAdes (St. Petersburg genome assembler) is a popular de novo genome assembly algorithm that is designed to handle various types of data including Illumina and PacBio long-read sequencing technologies. It was developed at the Saint Petersburg State University in Russia and is available as an open-source software.

SPAdes uses a hybrid approach to genome assembly that combines both de Bruijn graph-based and overlap-based techniques. It starts with the construction of a de Bruijn graph from the short-read data and then extends the contigs using long-read data or paired-end reads. It also includes error correction and scaffolding steps to improve the accuracy and contiguity of the resulting assembly.

SPAdes can be used for the assembly of bacterial, viral, and eukaryotic genomes and has shown to produce high-quality assemblies with low error rates. It is a popular choice for researchers in the field of genomics due to its accuracy, speed, and ability to handle complex datasets.

Links:

GitHub

User Guide

Versions Available:

The following versions are available on the cluster:

- SPAdes genome assembler v3.14.0

How to load SPAdes?

To load SPAdes, use the following commands:

```
#Load the SPAdes module
module load bio/spades/3.14.0
```

To verify if the module is loaded correctly, use the following command,

```
# List all the module loaded in the environment
module list
```

In a fresh environment, this should show only spades module loaded.

How to use SPAdes?

To use spades, see the following command line manual liked here. All the commands are accessible from spades.py script. All the commands and options are described in the above manual.

Bring up the help section,

```
# Help section to see commands and options
spades.py --help
```

Here, a run on a sample test file is demonstrated,

```
#Copy the test file to home folder
cp -r /share/apps/spades/SPAdes-3.14.0-Linux/share/spades/test_dataset
~
```

Now use the following slurm script to run spades on the test files,

```
#!/bin/bash
#SBATCH --job-name=spades
#SBATCH --output=spades.log
#SBATCH -p main
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=16
#SBATCH --mem=24G
 #SBATCH --qos main
#Load the module

module load bio/spades/3.14.0
#Run spades
spades.py --pe1-1 ~/test_dataset/ecoli_1K_1.fq.gz --pe1-2
~/test_dataset/ecoli_1K_2.fq.gz -o spades_out -t $SLURM_CPUS_PER_TASK
```

In this example, the script requests one node with 16 CPUs and 64GB of memory from the **main** partition. The **-t** option for SPAdes specifies the number of threads to use, which is set to the number of CPUs requested from Slurm (**$SLURM_CPUS_PER_TASK**).

Where to find help?

If you are confused or need help at any point, please contact OIT at the following address.

https://ua-app01.ua.edu/researchComputingPortal/public/oitHelp