pyRAD on HPC

What is pyRAD?

pyRAD is a software package for population genetic data analysis. It is used for filtering, clustering, and analyzing high-throughput restriction site-associated DNA (RAD-seq) data. pyRAD includes tools for handling large data sets, detecting SNPs, and performing downstream population genetic analysis. Some of the use cases are,

1. Demultiplexing: separating sequences from different individuals based on barcode information.
2. Quality filtering: removing low-quality sequences and reducing errors in the data.
3. Aligning sequences: aligning sequences to a reference genome or building a consensus sequence from RAD loci.
4. Detection of genetic variation: identifying single nucleotide polymorphisms (SNPs) and other types of genetic variation in the data.
5. Phylogenetic analysis: inferring relationships between individuals and populations based on genetic data.
6. Population genetics analysis: calculating measures of diversity, structure, and migration between populations.

Links:

Official Paper

GitHub

Versions Available:

The following versions are available on the cluster:

- pyRAD 3.0.4

How to load pyRAD?

To load pyRAD, use the following commands:

```
#Load the pyRAD module
module load bio/pyrad/3.0.4
```

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 pyRAD as loaded module. Users must load a conda environment along with this software with NumPy and SciPy.

```
#Load the python2 module
module load python/python2/2.7
```

How to use pyRAD?

To use pyRAD, users will need to have your RAD-seq data in a compatible format, such as fastq or fasta. You may need to convert or reformat the data if it is not in the correct format.

```
#create a parameter file
pyRAD -n
```

This will create a parameter file. Edit the parameter file for desired result.

Here is a slurm script to submit job to the scheduler,

```
#!/bin/bash

#SBATCH --job-name=pyRAD_run
#SBATCH --output=pyRAD_run.out
#SBATCH --error=pyRAD_run.err
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=4
#SBATCH --mem-per-cpu=8GB
#SBATCH --time=2:00:00
#SBATCH --p main
#SBATHC --qos main

# load necessary modules
module load python/python2/2.7
 module load bio/pyrad/3.0.4
# run the command
pyRAD -p params_file.txt
```

params_file.txt is a file that specifies the parameters for running the **pyRAD** software. It contains information about the input data, analysis settings, and output options for a given **pyRAD** analysis.

For example, the **params_file.txt** might contain the following information:

```
## input data
# raw read files in fastq format
input_data: /path/to/read_files/*.fastq

## analysis settings
# minimum depth of coverage required for a locus to be included in the
final data set
min_cov: 5
# maximum number of missing data allowed for a locus
max_missing: 0.3
# minimum number of individuals required for a locus to be included in
the final data set
min_ind: 3

## output options
# location of the final output files
outdir: /path/to/output/
# prefix for output files
outname: my_output
```

This **params_file.txt** specifies the input data as all **.fastq** files in the directory **/path/to/read_files/**, sets the minimum depth of coverage to 5, the maximum missing data to 0.3, and the minimum number of individuals to 3, and specifies the output directory as **/path/to/output/** with the output files prefix named "my_output".

The specific parameters and their acceptable values for a given **pyRAD** analysis can be found in the **pyRAD** manual or documentation.

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