Argonne_Seed_seq_processing_June2023

Abby Sulesky-Grieb

2023-06-20

Code used on the MSU HPCC for QIIME2 workflow

Copy demultiplexed sequences into working space from raw_sequence directory

```
#!/bin/bash -login

#SBATCH --time=03:59:00
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --mem=30G
#SBATCH --job-name copy_sequences
#SBATCH --mail-user=suleskya@msu.edu
#SBATCH --mail-type=BEGIN,END

######## Job code

cp *.fastq /mnt/research/ShadeLab/WorkingSpace/Sulesky/New_Argonne_Seed_Sequencing/input_data

echo -e "\n `sacct -u suleskya -j $SLURM_JOB_ID --format=JobID,JobName,Start,End,Elapsed,NCPUS,ReqMem`
scontrol show job $SLURM_JOB_ID
```

submit job: sbatch copy_seqs.sb

Rename files for Figaro to run correctly

rename using the util-linux command “rename”

use module load to load util-linux

```
module spider util-linux

# choose most recent version

module spider util-linux/2.37

# follow instructions to load

module load GCCcore/11.2.0
module load util-linux/2.37

module list #check which packages are loaded
```

Copy sequencing files into figaro_input directory: /mnt/research/ShadeLab/WorkingSpace/Sulesky/New_Argonne_Seed_Sequ

Use ‘rename’ to make files match illumina naming convention: SampleName_S1_L001_R1_001.fastq

Example: > rename G0_ G0 *.fastq

this code takes all files that end in *fastq, replaces “G0_” with “G0”

had to run various lines to remove extra underscores in sample IDs and add L001 to all rename goes really quickly, ~1 second per command

Figaro for trim parameters

make figaro output folder in workingspace directory /New_Argonne_Seed_Sequencing/figaro

go to home directory where figaro is installed to run figaro: cd /mnt/home/suleskya/figaro-master/figaro

Run figaro as a job:

nano figaro_argonne_june2023.sb

```
#!/bin/bash -login
########## SBATCH Lines for Resource Request ##########

#SBATCH --time=3:59:00
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH --mem=64G
#SBATCH --job-name figaro
#SBATCH -A shade-cole-bonito
#SBATCH --mail-user=suleskya@msu.edu
#SBATCH --mail-type=BEGIN,END

########## Command Lines for Job Running ##########

conda activate figaro

python figaro.py -i /mnt/research/ShadeLab/WorkingSpace/Sulesky/New_Argonne_Seed_Sequencing/figaro/figa

conda deactivate

echo -e "\n `sacct -u suleskya -j $SLURM_JOB_ID --format=JobID,JobName,Start,End,Elapsed,NCPUS,ReqMem` \
scontrol show job $SLURM_JOB_I
```

sbatch figaro_argonne_june2023.sb

check job progress with “sq” command

Once job is finished, check the figaro output:

cd

> less trimParameters.json { “trimPosition”: [191, 84], “maxExpectedError”: [2, 2], “readRetentionPercent”: 94.64, “score”: 92.63569499836179 },
>
> control Z to exit “less”

we will truncate the sequences at forward 191 and reverse 84, which will merge 94.64 percent of the reads

Import data into Qiime2 format

Have to rename files in input-data for Qiime2 input with the same method as above (base) -bash-4.2$ rename G0__ G0. *.fastq (base) -bash-4.2$ rename G1__ G1.* .fastq (base) -bash-4.2$ rename G2__ G2. *.fastq (base) -bash-4.2$ rename Shade__ Shade* .fastq (base) -bash-4.2$ rename _Fina Fina *.fastq (base) -bash-4.2$ rename _R _L001_R *.fastq

have to zip the fastqs with gzip to .gz (did this in a job, it took 3 hours)

```
#!/bin/bash -login
########## SBATCH Lines for Resource Request ##########

#SBATCH --time=3:59:00
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH --mem=64G
#SBATCH --job-name import_data
#SBATCH -A shade-cole-bonito
#SBATCH --mail-user=suleskya@msu.edu
#SBATCH --mail-type=BEGIN,END

########## Command Lines for Job Running ##########

conda activate qiime2-2022.8

qiime tools import \
  --type 'SampleData[PairedEndSequencesWithQuality]' \
  --input-path input_data \
  --input-format CasavaOneEightSingleLanePerSampleDirFmt \
  --output-path demux-paired-end.qza

conda deactivate

echo -e "\n `sacct -u suleskya -j $SLURM_JOB_ID --format=JobID,JobName,Start,End,Elapsed,NCPUS,ReqMem`
scontrol show job $SLURM_JOB_I
```

Successful, saved data as demux-paired-end.qza, can use this file in dada2

Denoise and merge

```
#!/bin/bash -login
########## SBATCH Lines for Resource Request ##########

#SBATCH --time=24:00:00
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=32
#SBATCH --mem=100G
#SBATCH --job-name dada2
#SBATCH -A shade-cole-bonito
#SBATCH --mail-user=suleskya@msu.edu
#SBATCH --mail-type=BEGIN,END

########## Command Lines for Job Running ##########

conda activate qiime2-2022.8
```

```
qiime dada2 denoise-paired \
        --i-demultiplexed-seqs demux-paired-end.qza \
        --p-trunc-len-f 191 \
        --p-trunc-len-r 84 \
        --o-table table.qza \
        --o-representative-sequences rep-seqs.qza \
        --o-denoising-stats denoising-stats.qza

qiime metadata tabulate \
  --m-input-file denoising-stats.qza \
  --o-visualization denoising-stats.qzv

qiime feature-table summarize \
  --i-table table.qza \
  --o-visualization table.qzv \
  --m-sample-metadata-file 230126_Shade_Seeds_16s_DG.txt

qiime feature-table tabulate-seqs \
  --i-data rep-seqs.qza \
  --o-visualization rep-seqs.qzv

conda deactivate

echo -e "\n `sacct -u suleskya -j $SLURM_JOB_ID --format=JobID,JobName,Start,End,Elapsed,NCPUS,ReqMem`
scontrol show job $SLURM_JOB_I
```

job successful! took 12 hr 24 minutes

Taxonomy assignment with silva

Assign taxonomy using Silva

Download reference seqs from qiime2.org: wget https://data.qiime2.org/2022.8/common/silva-138-99-515-806-nb-classifier.qza

job: nano classify-silva-taxonomy.sb

```
#!/bin/bash -login
########## SBATCH Lines for Resource Request ##########

#SBATCH --time=08:00:00
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=32
#SBATCH --mem=64G
#SBATCH --job-name taxonomy
#SBATCH -A shade-cole-bonito
#SBATCH --mail-user=suleskya@msu.edu
#SBATCH --mail-type=BEGIN,END

########## Command Lines for Job Running ##########

conda activate qiime2-2022.8
```

```
qiime feature-classifier classify-sklearn \
  --i-classifier silva-138-99-515-806-nb-classifier.qza \
  --i-reads rep-seqs.qza \
  --o-classification taxonomy.qza

qiime metadata tabulate \
  --m-input-file taxonomy.qza \
  --o-visualization taxonomy.qzv

qiime tools export \
  --input-path taxonomy.qza \
  --output-path phyloseq

qiime tools export \
  --input-path table.qza \
  --output-path phyloseq

biom convert \
  -i phyloseq/feature-table.biom \
  -o phyloseq/otu_table.txt \
  --to-tsv

conda deactivate

echo -e "\n `sacct -u suleskya -j $SLURM_JOB_ID --format=JobID,JobName,Start,End,Elapsed,NCPUS,ReqMem`
scontrol show job $SLURM_JOB_I
```

Export OTU table to new directory phyloseq qiime tools export
–input-path table.qza
–output-path phyloseq

OTU tables exports as feature-table.biom so convert to .tsv biom convert
-i phyloseq/feature-table.biom
-o phyloseq/otu_table.txt
–to-tsv

download otu table, manually change “#OTU ID” column header to “OTUID”

download taxonomy file and manually change Feature ID to OTUID in taxonomy.tsv. change taxonomy and OTU tables to csv format.

these files are now ready to export to R and run using phyloseq

Make phylogenetic tree with de novo alignment

align reads de novo to make multiple sequence alignment (MSA) mask alignment to reduce ambiguity make tree with fasttree root tree to midpoint pipeline below will do all of these steps with default settings and save everything to output directory export makes newick format file for the trees, or .qza can be used in iTOL

```
#!/bin/bash -login
########## SBATCH Lines for Resource Request ##########

#SBATCH --time=08:00:00
#SBATCH --nodes=1
#SBATCH --ntasks=1
```

```
#SBATCH --cpus-per-task=32
#SBATCH --mem=64G
#SBATCH --job-name fasttree
#SBATCH -A shade-cole-bonito
#SBATCH --mail-user=suleskya@msu.edu
#SBATCH --mail-type=BEGIN,END

########## Command Lines for Job Running ##########

conda activate qiime2-2022.8

qiime phylogeny align-to-tree-mafft-fasttree \
  --i-sequences rep-seqs.qza \
  --output-dir mafft-fasttree-output

qiime tools export \
  --input-path mafft-fasttree-output/tree.qza \
  --output-path mafft-fasttree-output

qiime tools export \
  --input-path mafft-fasttree-output/rooted_tree.qza \
  --output-path mafft-fasttree-output/exported-rooted-tree/

conda deactivate

echo -e "\n `sacct -u suleskya -j $SLURM_JOB_ID --format=JobID,JobName,Start,End,Elapsed,NCPUS,ReqMem`
scontrol show job $SLURM_JOB_I
```