Bucknell High Performance Research Computing Network
BUSCO Documentation

Ava Warfel

Program Location:
BisonNet

Version:
4.1.3

BisonNet Module:
assembly_assesment

Program Description:
BUSCO is a program that can be used to assess the quality of a genomic assembly by looking at the completeness of the assembly.

Authors/Background:
BUSCO stands for Benchmarking Universal Single-Copy Orthologs and was developed by the EM Zdobnov Group from the University of Geneva Medical School and the Swiss Institute of Bioinformatics

Input Document Types:
-FASTA file

Standard Output Information:
BUSCO outputs contain 4 standard categories: Complete and single-copy, Complete and duplicated, Fragmented, or Missing.

-The complete and single-copy output number (C) describes complete matches in orthologs that occur once.
-The complete and duplicated output number (D) describes complete matches in orthologs that occur twice.
-The fragmented output number (F) describes partially matched orthologs.
-The missing output number (M) describes orthologs without a match.

Troubleshooting Common Errors:
If you receive an error message when running BUSCO, start with the following:

-Check for errors in script
-Check the path to files
-Check the location of the config file
-Make sure you have specified the correct input mode, genome, proteins or transcriptome

If these do not resolve the issue, additional help can be found at the BUSCO issues board:

https://gitlab.com/ezlab/busco/-/issues

Running BUSCO on BisonNet:

Mandatory Arguments:

```
busco -i [SEQUENCE_FILE] -l [LINEAGE] -o [OUTPUT_NAME] -m [MODE] [OTHER OPTIONS]
```

Optional Arguments (Obtained directly from author's documentation):

```
-i FASTA FILE, --in FASTA FILE
                        Input sequence file in FASTA format. Can be an assembled genome or transcriptome (DNA), or
protein sequences from an annotated gene set.
  -o OUTPUT, --out OUTPUT
                        Give your analysis run a recognisable short name. Output folders and files will be labelled with
this name. WARNING: do not provide a path
  -m MODE, --mode MODE  Specify which BUSCO analysis mode to run.
                        There are three valid modes:
                        - geno or genome, for genome assemblies (DNA)
                        - tran or transcriptome, for transcriptome assemblies (DNA)
                        - prot or proteins, for annotated gene sets (protein)
  -l LINEAGE, --lineage_dataset LINEAGE
                        Specify the name of the BUSCO lineage to be used.
  --auto-lineage        Run auto-lineage to find optimum lineage path
  --auto-lineage-prok   Run auto-lineage just on non-eukaryote trees to find optimum lineage path
  --auto-lineage-euk    Run auto-placement just on eukaryote tree to find optimum lineage path
  -c N, --cpu N         Specify the number (N=integer) of threads/cores to use.
  -f, --force           Force rewriting of existing files. Must be used when output files with the provided name already
exist.
  -r, --restart         Continue a run that had already partially completed.
  -q, --quiet           Disable the info logs, displays only errors
  --out_path OUTPUT_PATH
                        Optional location for results folder, excluding results folder name. Default is current working
directory.
  --download_path DOWNLOAD_PATH
                        Specify local filepath for storing BUSCO dataset downloads
  --datasets_version DATASETS_VERSION
                        Specify the version of BUSCO datasets, e.g. odb10
  --download_base_url DOWNLOAD_BASE_URL
                        Set the url to the remote BUSCO dataset location
  --update-data         Download and replace with last versions all lineages datasets and files necessary to their
automated selection
  --offline             To indicate that BUSCO cannot attempt to download files
  --metaeuk_parameters METAEUK_PARAMETERS
                        Pass additional arguments to Metaeuk for the first run. All arguments should be contained within
a single pair of quotation marks, separated by commas. E.g. "--param1=1,--param2=2"
  --metaeuk_rerun_parameters METAEUK_RERUN_PARAMETERS
                        Pass additional arguments to Metaeuk for the second run. All arguments should be contained within
a single pair of quotation marks, separated by commas. E.g. "--param1=1,--param2=2"
  -e N, --evalue N      E-value cutoff for BLAST searches. Allowed formats, 0.001 or 1e-03 (Default: 1e-03)
  --limit REGION_LIMIT  How many candidate regions (contig or transcript) to consider per BUSCO (default: 3)
  --augustus            Use augustus gene predictor for eukaryote runs
  --augustus_parameters AUGUSTUS_PARAMETERS
                        Pass additional arguments to Augustus. All arguments should be contained within a single pair of
quotation marks, separated by commas. E.g. "--param1=1,--param2=2"
  --augustus_species AUGUSTUS_SPECIES
                        Specify a species for Augustus training.
  --long                Optimization Augustus self-training mode (Default: Off); adds considerably to the run time, but
can improve results for some non-model organisms
  --config CONFIG_FILE  Provide a config file
  -v, --version         Show this version and exit
  -h, --help            Show this help message and exit
  --list-datasets       Print the list of available BUSCO datasets
```

On BisonNet:
(For best results, it is recommended to run the script in an interactive session on the queue)

1. Enter an interactive session
2. Create a script:
    Script Components:
        1. Hash-bang (Tells the computer you are writing in bash)
        2. Enter the required arguments to use the queue
        3. Load the BUSCO module
        4. Copy the config file for augustus
        5. Command to Run BUSCO
3. Run the script that was created using the command “sh {insert_script_name_here}”

Example Script:

```
#!/bin/bash

#SBATCH -p short # partition (queue)
#SBATCH -N 1 # (leave at 1 unless using multi-node specific code)
#SBATCH -n 1 # number of cores
#SBATCH --mem-per-cpu=8192 # memory per core
#SBATCH --job-name="myjob" # job name
#SBATCH -o slurm.%N.%j.stdout.txt # STDOUT
#SBATCH -e slurm.%N.%j.stderr.txt # STDERR
#SBATCH --mail-user=username@bucknell.edu # address to email
#SBATCH --mail-type=ALL # mail events (NONE, BEGIN, END, FAIL, ALL)

#Description: This script can be used to run BUSCO

#Usage: runbusco.sh input_assembly database type

#load the module on BisonNet
module load assembly_assesment

#access the config file by copying it
cp -r /software/apps/augustus/current/config .
export AUGUSTUS_CONFIG_PATH=”./config/”

#run BUSCO
busco -i ~/Projects/genome -l {Insert your lineage here} -o genome_BUSCO_output -m genome
```

Sources:

EZ Lab. (2020). User guide BUSCO v5.beta.1. Retrieved November 10, 2020, from https://busco.ezlab.org/busco_userguide.html

Bucknell University. (2020). BisonNet. Retrieved November 10, 2020, from http://bisonnet.bucknell.edu/