GenomeQC

Please send questions (after reading this user guide) to: john.portwood@ars.usda.gov

Introduction:

GenomeQC generates descriptive summaries with intuitive graphics for genome assembly and structural annotations. It also benchmarks user supplied assemblies and annotations against the publicly available reference genomes of their choice. It is optimized for small and medium sized genomes (<2.5 Gb) and has pre-computed results for several maize genomes.

Software set up:

1) The GenomeQC web application is implemented in R shiny (version 1.5.9) and Python (version 3.6) and is freely available at https://genomeqc.maizegdb.org for non-profit academic use under the GPL license.

2) All scripts used in the web-application are available from the public GitHub project https://github.com/HuffordLab/GenomeQC

3) **Additional Programs:** In addition to the R shiny package and Python, the application requires:
    a) BUSCO v3.0.2 (https://gitlab.com/ezlab/busco) software and its dependencies
        i) NCBI BLAST+ v2.28.0 (https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+),
        ii) HMMER v3.1b2 (http://hmmer.org/)
        iii) Augustus v3.2.1 (http://bioinf.uni-greifswald.de/augustus/)

    b) gffread program version: 0.9.12 (http://ccb.jhu.edu/software/stringtie/gff.shtml#gffread_dl)

    c) The following R and Python packages:

R packages			
tools	R.utils	shinyWidgets	DT
seqinr	tidyverse	shinyBS	promises
Biostrings	gridExtra	reshape	future
stringr	grid	cowplot	

Python packages			
sys	traceback	Bio.Blast.Applications	email.mime.text
os	subprocess	iglob	email.mime.application
Bio	Statistics	pandas	email.mime.multipart
re	Numpy	plotly.offline	smtplib
argparse	collections	plotly.graph_objs	

d) For contamination analysis: NCBI UniVec database (ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec/) and a modified version of the taxify script (Note: No need to install blobtools. The modified version of the taxify script can be found with other scripts of the GenomeQC scripts folder on github.) from blobtools (https://zenodo.org/record/845347#.XDOf9s9KjVo).

Web Server Interface and Organization: The Interface is organized into three sections:

1. Compare reference genomes:
   This section outputs various pre-computed assembly and annotation metrics from a user-selected list of reference genomes.
   Input: The side panel in this section takes the following input parameters:

   Mandatory parameters:

   Reference genomes
   In this input field the user can select among the list of reference genomes for which they would like to compare assembly and annotation statistics. These metrics are pre-computed based on the same set of parameters and software implemented in GenomeQC.

   Email address
   GenomeQC generates and emails BUSCO plots for user-selected reference genomes and annotations. In this input field, the user must provide an email address to which they would like to receive these plots.

Output: The main panel is organized into four output tabs:

Assembly Metrics table
This tab displays multiple measures of genome assembly quality including basic statistics such as N50/NG50 and L50/LG50, which provide a standard measure of assembly contiguity.

Brief definitions of contiguity metrics:

- L50: A scaffold/contig L50 is calculated as the number of sequences whose sum of lengths make up 50% or more of the total assembly length.
- N50: A scaffold/contig N50 value is the length of the shortest scaffold/contig in the list of L50 sequences.
- LG50: A scaffold/contig LG50 is calculated as the number of sequences whose sum of lengths makes up 50% or more of the estimated genome size.
- NG50: The scaffold/contig NG50 is calculated in the same manner as N50 but is based on estimated genome size rather than total assembly length and is therefore a more comparable metric across assemblies.

The contiguity of a genome assembly improves as the N50/NG50 value increases and the L50/LG50 value decreases.

Assembly NG(X) plot
The most commonly reported contiguity values are at the 50% threshold, but NG(X) plots across thresholds (1- 100%) provide a more complete picture (Bradnam KR, Fass JN, Alexandrov A, et al., 2013). An NG(X) graph is a very useful visualization of the full distribution of contig/scaffold lengths across different genome assemblies.This tab displays the NG(X) plot for the pre-computed reference genomes. The x-axis of the plot spans the various thresholds (1-100%) and the y-axis spans the corresponding log-scaled scaffold/contig lengths (NG values) at these thresholds. Genome assemblies with larger scaffold/contig lengths across NG(X) thresholds are more contiguous.

Annotation Metrics Table
This tab outputs various annotation statistics (for the selected reference genomes) used to characterize gene structural annotations (Yandell, Mark et al., 2012). The table below provides a full list of these statistics.

Genome annotation statistics	
Number of gene models	Average exon length

Maximum gene length	Average number of exons per gene model
Minimum gene length	Number of transcripts
Average gene length	Average number of transcripts per gene model
Number of exons	Number of gene models less than 200 bp length

Assembly and Annotation busco plots
This tab generates and emails precomputed BUSCO scores for a user-selected set of reference genomes. The set of universally distributed single-copy orthologs (BUSCO) is quite useful for assessing the completeness of assemblies and annotations in terms of evolutionary informed expected gene content. The BUSCO pipeline characterizes completeness of the genic fraction of an assembly by detailing the proportion of genes that are complete and single copy, duplicated, missing, and fragmented. The BUSCO query gene set includes a set of orthologs that are single copy in more than 90% of the species in a given lineage of the tree of life (Kriventseva et al. 2008). Complete genome assemblies and annotations are expected to contain a high proportion of complete BUSCO genes. More information on BUSCO assessment can be found here: https://busco.ezlab.org/

2. Analyze your genome assembly:
   This section provides the user the option to perform analysis on their genome assembly as well as benchmark their analysis with the pre-computed reference genomes.
   Input: The side panel in this section takes the following input parameters:

   Mandatory parameters:

   Email address
   In this input field the user must provide the email address to which they would like to receive the BUSCO and contamination plots.

   Name of your genome assembly
   Here the user needs to provide a name for their analysis job. The users are required to input just one word (e.g. maize). This user input will be used to label the output plots and tables.

   Estimated Genome Size
   In this input field, the user needs to provide an estimate of genome size in megabases (Mb) (e.g. estimated genome size for maize is 2200 Mb). Please note, this is the estimate of the actual **biological** genome and not the sequence assembly. The value in this input field is used to calculate NG values at different thresholds.

BUSCO datasets
This parameter is required to calculate the BUSCO scores for the user-uploaded genome assembly. Here the user needs to specify which BUSCO lineage data to be used for assembly completeness assessment (e.g. for plants dataset (Embryophyta odb9) should be selected for analysis of a plant genome assembly like maize. The BUSCO dataset is formed by selecting a set of orthologs from OrthoDB v9 database which are found as single copy in more than 90% of the species in the phylogenetic tree of life.
A high quality genome assembly is expected to contain a higher number of complete and single copy BUSCO genes (C&S) and a lower number of missing (M) or fragmented (F) BUSCO genes.
More information on BUSCO datasets can be found here: https://busco.ezlab.org/

AUGUSTUS species
Here the user needs to select the AUGUSTUS species parameters that should be used for gene prediction to identify and classify the BUSCO matches. BUSCO assessment of genome assembly involves constructing gene models from the candidate regions identified by tblastn searches against the BUSCO consensus sequences. These gene predictions are then used by HMMER (http://hmmer.org/) which classifies the matches of gene predictions with the BUSCO lineage profiles as complete and single copy (C&S), duplicated (D), fragmented (F) or missing (M). BUSCO pipeline uses AUGUSTUS *de novo* gene predictor (http://augustus.gobics.de/) to construct the gene models.
More information on augustus species selection can be found here: https://busco.ezlab.org/

Genome assembly file
User need to upload a genome assembly file in compressed (.gz) FASTA format. Maximum upload limit for the assembly file is 1Gb.

Optional parameters:

Reference genomes
In this input field the user can select among the list of reference genomes for which they would like to benchmark their assembly statistics against the reference genomes. These reference values are pre-computed using the same set of parameters and softwares which the tool use to compute for the user uploaded files.

Output: The main panel is organized into four output tabs:

Assembly Metrics table

This tab calculates and outputs various quantitative measures of genome assembly assessment for the uploaded assembly and benchmark the metrics against the selected reference genomes. For more information on these statistics refer to: https://en.wikipedia.org/wiki/N50,_L50,_and_related_statistics

Assembly NG(X) plot
This tab calculates and outputs an NG(X) plot for the uploaded genome assembly. X-axis of the plot represents the various thresholds (1-100%) and the y axis represents the corresponding log-scaled scaffold/contig lengths (NG values) at these thresholds. By analogy, higher the NG(X) curve of a genome assembly, better the quality of assembly in terms of contiguity. Refer this article - Bradnam KR, Fass JN, Alexandrov A, et al., 2013, for more information on NG(X) plots.

Assembly and Contamination busco plots
This tab generates and email the BUSCO and contamination plots for the uploaded genome assembly and compares it with the pre-computed values of the user-selected reference genomes. A high quality genome assembly and annotations is expected to contain a higher number of complete and single copy BUSCO genes (C&S) and a lower number of missing (M) or fragmented (F) BUSCO genes. More information on BUSCO assessment can be found here: https://busco.ezlab.org/. For contamination analysis, megablast is used to quickly identify segments of the assembled genome sequences which may of vector origin or contain adaptor, linkers and primer sequences used in cloning cDNA or genomic DNA (https://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/).

3. Analyze your genome annotation:
This section provides the user t to perform analysis on their genome annotations as well as benchmark their analysis with the pre-computed reference genomes.
Input: The side panel in this section takes the following input parameters:

Mandatory parameters:

Email address
In this input field the user must provide the email address to which they would like to receive the BUSCO plot.

Name of your genome annotation
Here the user needs to provide a name for their analysis job. The users are required to input just one word (e.g. maize). This user input will be used to label the output plots and tables.

BUSCO datasets
This parameter is required to calculate the BUSCO scores for the user-uploaded genome annotation. Here the user needs to specify which BUSCO lineage data to be used for transcriptome completeness assessment (e.g. for plants dataset (Embryophyta

odb9) should be selected for analysis of a plant genome annotations like maize). BUSCO dataset is formed by selecting a set of orthologs from OrthoDB v9 database which are found as single copy in more than 90% of the species in the phylogenetic tree of life. A high quality genome annotations is expected to contain a higher number of complete and single copy BUSCO genes (C&S) and a lower number of missing (M) or fragmented (F) BUSCO genes.
More information on BUSCO datasets can be found here: https://busco.ezlab.org/

Genome assembly file
User need to upload a genome assembly file in compressed (.gz) FASTA format. Maximum upload limit for the assembly file is 1Gb.

Genome annotation file
User need to upload a genome annotation file in compressed (.gz) generic feature format. More information on the generic feature format can be found here: https://useast.ensembl.org/info/website/upload/gff3.html

Note :BUSCO analysis of gene structural annotations requires that every contig or chromosome name found in the 1st column of the input GFF file must have a corresponding sequence entry in fasta file as shown here: https://www.ncbi.nlm.nih.gov/genome/167?genome_assembly_id=161521
As an alternative, you could upload corresponding transcript fasta file in addition to the GFF file.

Optional parameters

Reference genomes
In this input field the user can select among the list of reference genomes for which they would like to benchmark their annotation statistics against the reference genomes. These reference values are pre-computed using the same set of parameters and softwares which the tool use to compute for the user uploaded files.

Transcript file
The user needs to upload transcript set (DNA nucleotide sequences) file in compressed (.gz) FASTA format. This file is used calculate annotation BUSCO scores. Currently the tool is configured to use the information from transcripts fasta file if the user uploads it. if the user does not upload the transcripts file, the tool will check if the information in the first column of gff file corresponds to the headers in the fasta file as shown in the example. If there is discrepancy, it will print an error message, else the job will be submitted.

Output: The main panel is organized into four output tabs:

Annotation Metrics table

This tab calculates and outputs various quantitative measures of genome annotation assessment for the uploaded annotations and benchmark the metrics against the selected reference genomes.

Assembly busco plot

This tab generates and email the BUSCO plot for the uploaded genome annotations and compares it with the pre-computed values of the user-selected reference genomes. More information on BUSCO assessment can be found here: https://busco.ezlab.org/.

Reference Genomes information:

Please note: we didn't have exon information for Brassica_rapa_v3 annotations

BUSCO dataset used: embryophyta_odb9

MaizeB73_v4_con
Estimated genome size (Mb): 2200
BUSCO species used: Maize

1) Assembly File:
B73_v4_scaffolds.fasta.gz
Fasta files (they were combined into a single file) used for Assembly metrics table and BUSCO score calculation.
Placed scaffolds: This file was created by splitting the AGP file based on the WGS coordinate information. (Source of AGP files: ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/plant/Zea_mays/all_assembly_versions/GCA_000005005.6_B73_RefGen_v4/GCA_000005005.6_B73_RefGen_v4_assembly_structure/Primary_Assembly/assembled_chromosomes/AGP/).

Unplaced scaffolds (source: ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/plant/Zea_mays/all_assembly_versions/GCA_000005005.6_B73_RefGen_v4/GCA_000005005.6_B73_RefGen_v4_assembly_structure/Primary_Assembly/unplaced_scaffolds/FASTA/) were combined with the placed scaffolds extracted from the AGP file to generate the final scaffold file for B73_v4.

2**) Annotation file:**
Zea_mays.AGPv4.36.gff3.gz
GFF file used for Annotation metrics table calculation.
Source of this file: ftp://ftp.ensemblgenomes.org/pub/plants/release-36/gff3/zea_mays/

3) Zea_mays.AGPv4.dna.toplevel.fa.gz
This fasta file was used to extract transcript sequences using gffread.
Source of this file: ftp://ftp.ensemblgenomes.org/pub/plants/release-36/fasta/zea_mays/dna/

4) **Transcript file:**
B73_v4_transcripts.fasta.gz
This file is the output from the gffread program (used for extracting transcripts sequences from gff files).
./gffread -w B73_v4_transcripts.fasta -g Zea_mays.AGPv4.dna.toplevel.fa Zea_mays.AGPv4.36.gff3

MaizeW22_NRgene_con

Estimated genome size (Mb): 2200
BUSCO species used: maize

1) Assembly file:
Zm-W22-REFERENCE-NRGENE-2.0_scaffolds.fasta.gz

Fasta files (they were combined into a single file) used for Assembly metrics table and BUSCO score calculation.
Placed scaffolds: This file was created by splitting the AGP file based on the WGS coordinate information. (Source of AGP files: ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/plant/Zea_mays/all_assembly_versions/GCA_001644905.2_Zm-W22-REFERENCE-NRGENE-2.0/GCA_001644905.2_Zm-W22-REFERENCE-NRGENE-2.0_assembly_structure/Primary_Assembly/assembled_chromosomes/AGP/). Unplaced scaffolds (source: ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/plant/Zea_mays/all_assembly_versions/GCA_001644905.2_Zm-W22-REFERENCE-NRGENE-2.0/GCA_001644905.2_Zm-W22-REFERENCE-NRGENE-2.0_assembly_structure/Primary_Assembly/unplaced_scaffolds/FASTA/) were combined with the placed scaffolds extracted from the AGP file to generate the final scaffold file for W22_NRgene.

2) Zm-W22-REFERENCE-NRGENE-2.0.gff.gz
GFF file used for Annotation metrics table calculation.
Source of this file: https://ftp.maizegdb.org/MaizeGDB/FTP/W22_v2.0/

3) **Transcript file:**
Zm-W22-CDS-NRGENE-2.0.fasta.gz
This transcript file is used for BUSCO annotation score calculation.
Source of this file: https://ftp.maizegdb.org/MaizeGDB/FTP/W22_v2.0/

MaizeMo17_CAU_con

Estimated genome size (Mb): 2200
BUSCO species used: Maize

1) Assembly file:
Zm-Mo17-Scaffolds-CAU-1.0.fa.gz (scaffolds)
Fasta file used for Assembly metrics table and BUSCO score calculation.
Source of this file: provided by the Mo17 Sequencing Group (State Key Laboratory of Agrobiotechnology and National Maize Improvement Center, Department of Plant Genetics and Breeding, China Agricultural University, Beijing, China).

2) Annotation file:
Zm-Mo17-REFERENCE-CAU-1.0_mgdb.gff3.gz
GFF file used for Annotation metrics table calculation.
Source of this file: https://ftp.maizegdb.org/MaizeGDB/FTP/Mo17-CAU/

3) Zm-Mo17-REFERENCE-CAU-1.0.fsa.gz
This fasta file was used to extract transcript sequences using gffread.
Source of this file: https://ftp.maizegdb.org/MaizeGDB/FTP/Mo17-CAU/

4) **Transcript file:**
Zm-Mo17-REFERENCE-CAU-1.0_transcripts.fasta.gz
This transcript file is used for BUSCO annotation score calculation.
./gffread -w Zm-Mo17-REFERENCE-CAU-1.0_transcripts.fasta -g Zm-Mo17-REFERENCE-CAU-1.0.fsa Zm-Mo17-REFERENCE-CAU-1.0_mgdb.gff3

Sorghum_bicolor_NCBIv3_con

Estimated genome size (Mb): 730
BUSCO species used: Maize

1) Assembly file:
ABXC03.1.fsa_nt.gz and ABXC03.2.fsa_nt.gz (contig files)
Fasta files (they were combined into a single file) used for Assembly metrics table and BUSCO score calculation.
Source of this file: ftp://ftp.ncbi.nlm.nih.gov/sra/wgs_aux/AB/XC/ABXC03/

2) Annotation file:
GCF_000003195.3_Sorghum_bicolor_NCBIv3_genomic.gff.gz
GFF file used for Annotation metrics table calculation.
Source of this file: https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/003/195/GCF_000003195.3_Sorghum_bicolor_NCBIv3/

3) GCF_000003195.3_Sorghum_bicolor_NCBIv3_genomic.fna.gz

This file was used to extract transcript sequences using gffread.
Source of this file: https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/003/195/GCF_000003195.3_Sorghum_bicolor_NCBIv3/

4) **Transcript file:**
GCF_000003195.3_Sorghum_bicolor_NCBIv3_transcripts.fasta.gz
This file is the output from the gffread program (used for extracting transcripts sequences from gff files).
./gffread -w GCF_000003195.3_Sorghum_bicolor_NCBIv3_transcripts.fasta -g GCF_000003195.3_Sorghum_bicolor_NCBIv3_genomic.fna GCF_000003195.3_Sorghum_bicolor_NCBIv3_genomic.gff

Rice_japonica_Nipponbare4.0_chr

Estimated genome size (Mb): 430
BUSCO species used: rice

1) Assembly file:
GCF_000005425.2_Build_4.0_genomic.fna.gz
Fasta file used for Assembly metrics table and BUSCO assembly score calculation.
Source of this file: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/425/GCF_000005425.2_Build_4.0/
2) Annotation file:
GCF_000005425.2_Build_4.0_genomic.gff.gz
GFF file used for Annotation metrics table calculation.
Source of this file: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/425/GCF_000005425.2_Build_4.0/

3) **Transcript file:**
GCF_000005425.2_Build_4.0_transcripts.fasta.gz
This transcript file is used for BUSCO annotation score calculation.This file is the output from the gffread program (used for extracting transcripts sequences from gff files).
./gffread -w GCF_000005425.2_Build_4.0_transcripts.fasta -g GCF_000005425.2_Build_4.0_genomic.fna GCF_000005425.2_Build_4.0_genomic.gff

Arabidopsis_thaliana_TAIR10.1_chr

Estimated genome size (Mb): 135
BUSCO species used: arabidopsis

1) Assembly file:
GCF_000001735.4_TAIR10.1_genomic.fna.gz
Fasta file used for Assembly metrics table and BUSCO assembly score calculation.
Source of this file: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/735/GCF_000001735.4_TAIR10.1/

2) Annotation file:
GCF_000001735.4_TAIR10.1_genomic.gff.gz
GFF file used for Annotation metrics table calculation.
Source of this file: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/735/GCF_000001735.4_TAIR10.1/

3) **Transcript file:**
GCF_000001735.4_TAIR10.1_transcripts.fasta.gz
This transcript file is used for BUSCO annotation score calculation.This file is the output from the gffread program (used for extracting transcripts sequences from gff files).
./gffread -w GCF_000001735.4_TAIR10.1_transcripts.fasta -g GCF_000001735.4_TAIR10.1_genomic.fna GCF_000001735.4_TAIR10.1_genomic.gff

Brassica_rapa_v3_chr

Estimated genome size (Mb): 443
BUSCO species used: arabidopsis

1) Assembly file:
Brapa_sequence_v3.0.fasta.gz
Fasta file used for Assembly metrics table and BUSCO score calculation.
Source of this file: http://brassicadb.org/brad/datasets/pub/BrassicaceaeGenome/Brassica_rapa/V3.0/

2) Annotation file:
Brapa_genome_v3.0_genes.gff3.gz
GFF file used for Annotation metrics table calculation.
Source of this file: http://brassicadb.org/brad/datasets/pub/BrassicaceaeGenome/Brassica_rapa/V3.0/

3) **Transcript file:**
Brapa_genome_v3.0_cds.fasta.gz
This transcript file is used for BUSCO annotation score calculation.

Source of this file: http://brassicadb.org/brad/datasets/pub/BrassicaceaeGenome/Brassica_rapa/V3.0/

Brachypodium_distachyon_v3_con

Estimated genome size (Mb): 355
BUSCO species used: wheat

1) Assembly file:
ADDN03.1.fsa_nt.gz
Fasta file used for Assembly metrics table and BUSCO score calculation.
Source of this file: ftp://ftp.ncbi.nlm.nih.gov/sra/wgs_aux/AD/DN/ADDN03/

2) Annotation file:
GCF_000005505.3_Brachypodium_distachyon_v3.0_genomic.gff.gz
GFF file used for Annotation metrics table calculation.
Source of this file: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/505/GCF_000005505.3_Brachypodium_distachyon_v3.0/

3)GCF_000005505.3_Brachypodium_distachyon_v3.0_genomic.fna.gz
This file is used to extract transcript sequences using gffread.
Source of this file: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/505/GCF_000005505.3_Brachypodium_distachyon_v3.0/

4) **Transcript file:**
GCF_000005505.3_Brachypodium_distachyon_v3.0_transcripts.fasta.gz
This transcript file is used for BUSCO annotation score calculation.
This file is the output from the gffread program (used for extracting transcripts sequences from gff files).
./gffread -w GCF_000005505.3_Brachypodium_distachyon_v3.0_transcripts.fasta -g GCF_000005505.3_Brachypodium_distachyon_v3.0_genomic.fna GCF_000005505.3_Brachypodium_distachyon_v3.0_genomic.gff

Setaria_italica_v2_chr

Estimated genome size (Mb): 490
BUSCO species used: maize

1) Assembly file:
ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/plant/Setaria_italica/all_assembly_versions/GCF_000263155.2_Setaria_italica_v2.0/GCF_000263155.2_Setaria_italica_v2.0_genomic.fna.gz

2) Annotation file:
ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/plant/Setaria_italica/all_assembly_versions/GCF_000263155.2_Setaria_italica_v2.0/GCF_000263155.2_Setaria_italica_v2.0_genomic.gff.gz

3) Transcript file:
GCF_000263155.2_Setaria_italica_v2.0_transcripts.fasta.gz
This transcript file is used for BUSCO annotation score calculation.This file is the output from the gffread program (used for extracting transcripts sequences from gff files).
./gffread -w GCF_000263155.2_Setaria_italica_v2.0_transcripts.fasta -g GCF_000263155.2_Setaria_italica_v2.0_genomic.fna GCF_000263155.2_Setaria_italica_v2.0_genomic.gff

Glycine_max_v2.1_con

Estimated genome size (Mb): 1100
BUSCO species used: arabidopsis

1) Assembly file:
ACUP03.1.fsa_nt.gz and ACUP03.2.fsa_nt.gz (contig files)
Fasta files (they were combined into a single file) used for Assembly metrics table and BUSCO score calculation.
Source of this file: ftp://ftp.ncbi.nlm.nih.gov/sra/wgs_aux/AC/UP/ACUP03/

2) Annotation file:
GCF_000004515.5_Glycine_max_v2.1_genomic.gff.gz
GFF file used for Annotation metrics table calculation.
Source of this file: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/004/515/GCF_000004515.5_Glycine_max_v2.1/

3) GCF_000004515.5_Glycine_max_v2.1_genomic.fna.gz
This fasta file was used to extract transcript sequences using gffread.

Source of this file: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/004/515/GCF_000004515.5_Glycine_max_v2.1/

4) **Transcript file:**
GCF_000004515.5_Glycine_max_v2.1_transcripts.fasta.gz
This file is the output from the gffread program (used for extracting transcripts sequences from gff files).
./gffread -w GCF_000004515.5_Glycine_max_v2.1_transcripts.fasta -g GCF_000004515.5_Glycine_max_v2.1_genomic.fna GCF_000004515.5_Glycine_max_v2.1_genomic.gff

Expected runtime for BUSCO analysis:

BUSCO analysis of genome assemblies and annotations is a computationally intensive job and the expected run time depends on the size of assemblies and annotation set. The following lists the expected run time for different genomes:

Genomes up to 200 Mb: up to 2 hours
Genomes between 200Mb and 400 Mb: 3-4 hours
Genomes between 400Mb and 700 Mb: 4-8 hours
Genomes between 700Mb and 1.5 Gb: 8-24 hours
Genomes greater than 1.5 Gb: >1 day

Best practices:

1. Please upload compressed files only. Also, the current limit of file upload is 1Gb (compressed file size). Please don't upload any files that exceed this limit.
2. For computing the BUSCO scores for gff files, please make sure the headers in the fasta file matches with the sequence names in the first column of the gff file. As an example, please look at these example files GCF_000005845.2_ASM584v2_genomic.fna.gz and GCF_000005845.2_ASM584v2_genomic.gff.gz here: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/845/GCF_000005845.2_ASM584v2/
3. Please submit one job at a time, especially when you are uploading large files.
4. Please don't forget to provide your email address if you want BUSCO and contamination results.
5. Please click the blue icon on right side of each input field to get more information on selecting BUSCO datasets and AUGUSTUS species for computing the BUSCO scores.
6. Please click on each tab from left to right one at a time, explore the tab completely, download its results and then move on to next tab. Click the BUSCO and contamination tab at the very end only.
7. Please provide just one word in the input fields for the name of assembly and annotation.