BUSCO

from QC to gene prediction and phylogenomics

BUSCO : Assessing genomic data quality

Stéphanie ROBIN, BIPAA platform, INRAE /
GenOuest platform, Rennes

Context

Genomics resources : genome assemblies & annotations
=> Metrics and tools are necessary to perform quality assessments

Examples :
- Identify redondancies in a draft genome assembly due to a technical issues
- Quality required before performing comparative analyses

Existing tool to evaluate quality :
- contigs/scaffold counts and contig/scaffold N50 values

⇒ genome assembly contiguity

- BUSCO => completeness and redundancy in terms of expected gene content :
  - Assembled genomes / transcriptomes / annotated protein-coding gene sets,
  - Prokaryotic and eukaryotic data

BUSCO : Benchmarking Universal Single-Copy Orthologs

http://busco.ezlab.org

Bioinformatics, 31(19), 2015, 3210–3212
doi: 10.1093/bioinformatics/btv351
Advance Access Publication Date: 9 June 2015
Applications Note

OXFORD

Genome analysis

BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs

Felipe A. Simão†, Robert M. Waterhouse†, Panagiotis Ioannidis, Evgenia V. Kriventseva and Evgeny M. Zdobnov*

BUSCO : open-source software, with sets of Benchmarking Universal Single- Copy Orthologs

⇒ Quantitative assessment of genome assembly , annotation completeness, assembled transcriptomes based on evolutionarily informed expectations of gene content

BUSCO : Benchmarking **U**niversal **S**ingle-**C**opy **O**rthologs

- Completeness in terms of gene content / assembly and genome annotation
- Genes to be found only in single- copy in a genome

Metrics to describe genome, gene set or transcriptome completeness in BUSCO :

- **C : complete** : lengths are within two standard deviations of the BUSCO group
- **[D : duplicated]** : complet genes found with more than one copy
- **F : fragmented** : Genes only partially recovered
- **M : missing genes** : not recovered
- **n : number of genes used**

BUSCO : Benchmarking Universal Single-Copy Orthologs

Table 1. Assessment of fruitfly (*D. mela,*), nematode worm (*C. eleg,*), human (*H. sapi,*), owl limpet (*L. giga,*), and fungus (*A. nidu,*) genome assemblies (upper row) and gene sets (lower row) in BUSCO notation (C:complete [D:duplicated], F:fragmented, M:missing, n: gene number)

Species	Size	BUSCO notation assessment results
D. mela	139 Mbp	C:98% [D:6.4%], F:0.6%, M:0.3%, n:2 675
	13 918 genes	C:99% [D:3.7%], F:0.2%, M:0.0%, n:2 675
C. eleg	100 Mbp	C:85% [D:6.9%], F:2.8%, M:11%, n:843
	20 447 genes	C:90% [D:11%], F:1.7%, M:7.5%, n:843
H. sapi	3 381 Mbp	C:89% [D:1.5%], F:6.0%, M:4.5%, n:3 023
	20 364 genes	C:99% [D:1.7%], F:0.0%, M:0.0%, n:3 023
L. giga	359 Mbp	C:89% [D:2.3%], F:4.3%, M:5.8%, n:843
	23 349 genes	C:90% [D:13%], F:7.8%, M:2.1%, n:843
A. nidu	30 Mbp	C:98% [D:1.8%], F:0.9%, M:0.2%, n:1 438
	10 534 genes	C:95% [D:7.3%], F:3.8%, M:0.9%, n:1 438

Genome assembly less complete than genome annotation (H. sapiens)
⇒ limitations of the BUSCO gene prediction step

Genome annotation less complete than genome assembly
⇒ the annotated gene set may be missing some BUSCO gene matches that are in fact present in the genome (A. nidulans)

More 'missing' BUSCOs may also be reported for species that are highly derived with respect to the assessment clade—even with high-quality genomes (C. elegans)

BUSCO : Benchmarking Universal Single-Copy Orthologs, v4/v5

OXFORD JOURNALS Molecular Biology and Evolution

Mol Biol Evol. 2021 Oct; 38(10): 4647–4654.
Published online 2021 Jul 28. doi: 10.1093/molbev/msab199

PMCID: PMC8476166
PMID: 34320186

BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes

Mosè Manni,[1,2] Matthew R Berkeley,[1,2] Mathieu Seppey,[1,2] Felipe A Simão,[1,2] and Evgeny M Zdobnov[1,2]

The latest versions of the BUSCO datasets (*_odb10; Manni et al., 2021) include 67 eukaryotic, 83 bacterial, 16 archaeal, and 27 viral datasets

Usages :

- Single input file (either a genome assembly, annotated gene set, or transcriptome assembly) with a known taxonomic origin
- Input sequence without specifying a dataset for the assessment, which enables the evaluation of sequences with unknown taxonomic origin.
- Multiple inputs, metagenomic bins or MAGs from both prokaryotic and eukaryotic species

Use in Galaxy : inputs

Busco assess genome assembly and annotation completeness (Galaxy Version 5.3.2+galaxy0)

Sequences to analyse

15: Funannotate predict annotation on data 4, data 11, and data 8: protein sequences

Can be an assembled genome or transcriptome (DNA), or protein sequences from an annotated gene set.

Mode

annotated gene sets (protein)

(--mode)

Auto-detect or select lineage?

Select lineage

Let BUSCO decide the best lineage automatically, or select from known lineage

Lineage

Mucorales

(--lineage_dataset)

Which outputs should be generated

Select/Unselect all

× short summary text × list with missing IDs × summary image

Use in Galaxy : outputs

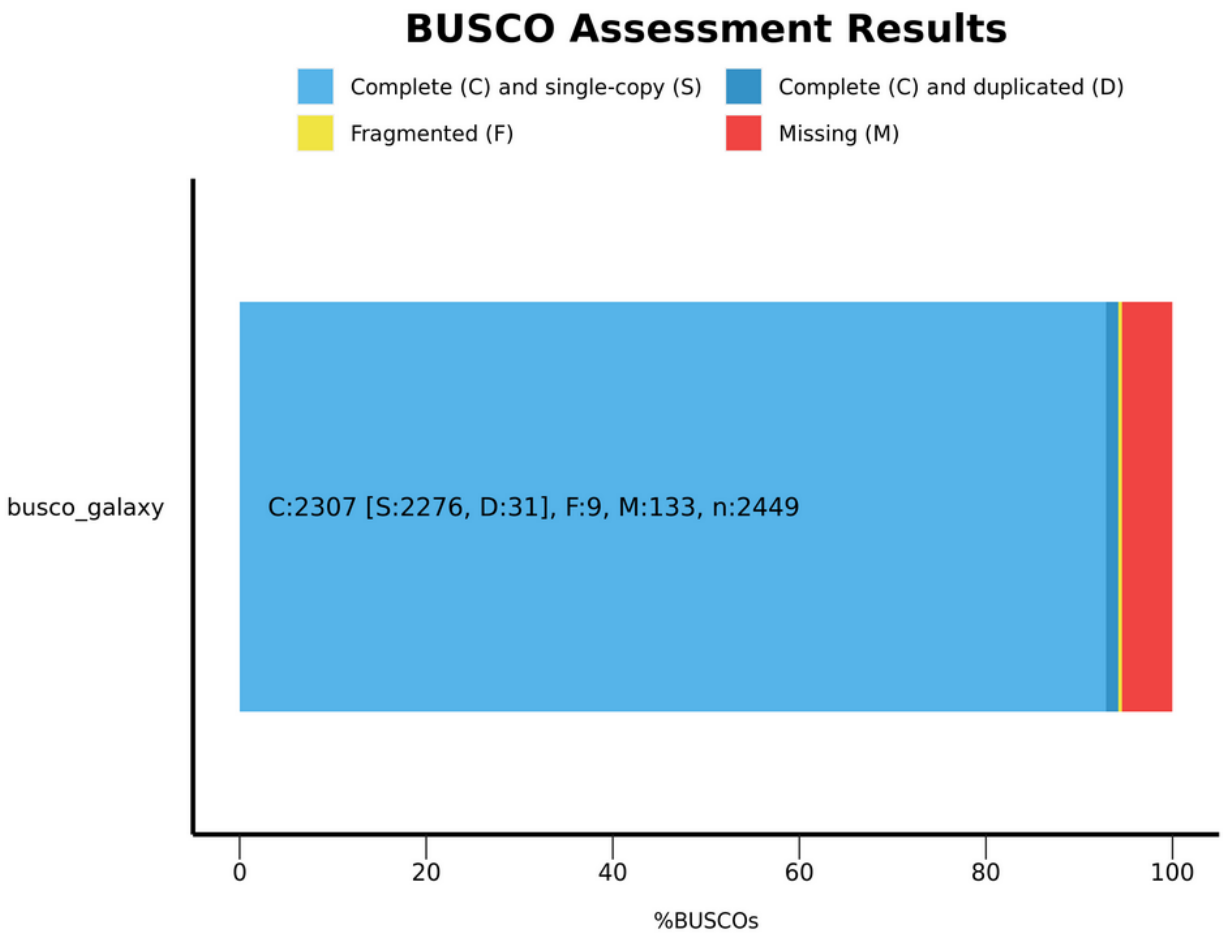
1st and 4th outputs : Short summary & Summary image

```
# BUSCO version is: 5.3.2
# The lineage dataset is: mucorales_odb10 (Creation date: 2020-08-05, number of genomes: 15, number of BUSCOs: 2449)
# Summarized benchmarking in BUSCO notation for file /shared/ifbstor1/galaxy/datasets/002/231/dataset_2231094.dat
# BUSCO was run in mode: proteins

        ***** Results: *****

        C:94.2%[S:92.9%,D:1.3%],F:0.4%,M:5.4%,n:2449
        2307    Complete BUSCOs (C)
        2276    Complete and single-copy BUSCOs (S)
        31      Complete and duplicated BUSCOs (D)
        9       Fragmented BUSCOs (F)
        133     Missing BUSCOs (M)
        2449    Total BUSCO groups searched

Dependencies and versions:
        hmmsearch: 3.1
```



Use in Galaxy : outputs

2nd and 3rd outputs : Full table & missing BUSCOs

BUSCO version is: 5.3.2

The lineage dataset is: mucorales_odb10 (Creation date: 2020-08-05, number of genomes: 15, number of BUSCOs: 2449)

# Busco id	Status	Sequence	Score	Length	OrthoDB url	Description
1at4827	Complete	FUN_011506-T1	8536.1	4025	https://www.orthodb.org/v10?query=1at4827	dynein heavy chain
2at4827	Complete	FUN_008069-T1	7507.0	4513	https://www.orthodb.org/v10?query=2at4827	Midasin
10at4827	Complete	FUN_003811-T1	3234.9	3035	https://www.orthodb.org/v10?query=10at4827	Phosphatidylinositol 3-/4-kinase, catalytic domain
26at4827	Complete	FUN_006426-T1	4983.1	2167	https://www.orthodb.org/v10?query=26at4827	Pre-mRNA-processing-splicing factor 8
27at4827	Complete	FUN_004720-T1	4380.4	2584	https://www.orthodb.org/v10?query=27at4827	Vacuolar protein sorting-associated protein 13
28at4827	Complete	FUN_005679-T1	4275.7	2083	https://www.orthodb.org/v10?query=28at4827	FKBP12-rapamycin binding domain
38at4827	Complete	FUN_011596-T1	3288.2	1818	https://www.orthodb.org/v10?query=38at4827	DNA polymerase epsilon catalytic subunit
49at4827	Complete	FUN_007107-T1	3511.5	2062	https://www.orthodb.org/v10?query=49at4827	armadillo-type protein
53at4827	Complete	FUN_000501-T1	3938.3	1944	https://www.orthodb.org/v10?query=53at4827	pre-mRNA splicing factor

BUSCO version is: 5.3.2

The lineage dataset is: mucorales_odb10 (Creation date: 2020-08-05, number of genomes: 15, number of BUSCOs: 2449)

Busco id
10014at4827
10103at4827
10252at4827
10312at4827
10446at4827
10527at4827
10555at4827
10618at4827
10668at4827
10974at4827
10977at4827