Annotation structurale: Validation des prédictions avec BUSCO

EBAII Assembly & Annotation - Roscoff Juin 2024

Romane LIBOUBAN
GenOuest platform, Rennes

Présentation BUSCO

- Évaluation de l'exhaustivité des génomes, des transcriptomes et des gènes codant pour les protéines
- Basée sur la recherche de gènes orthologues universels, présents en un seul exemplaire dans les organismes d'un même clade

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

1. **Base de données d'orthologues universels** préalablement construite
2. **Comparaison des séquences**
3. **Evaluation de la complétude :** nombre de gènes orthologues retrouvés
4. **Rapports détaillés**

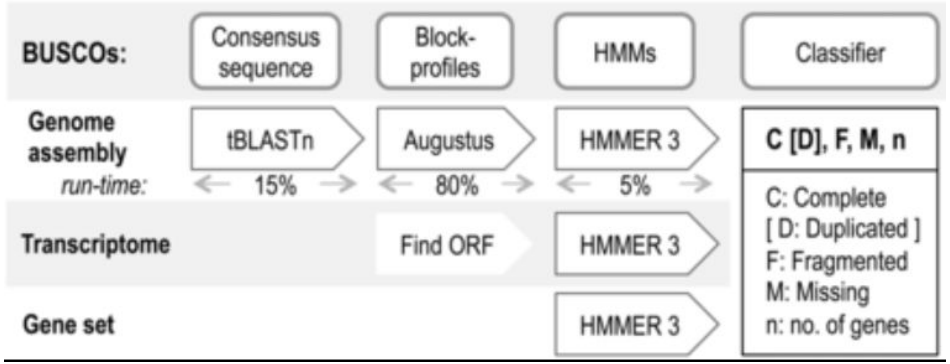
```
***** Results: *****

C:95.4%[S:95.3%,D:0.1%],F:0.0%,M:4.6%,n:8338
7953    Complete BUSCOs (C)
7946    Complete and single-copy BUSCOs (S)
7       Complete and duplicated BUSCOs (D)
0       Fragmented BUSCOs (F)
385     Missing BUSCOs (M)
8338    Total BUSCO groups searched
```

https://gitlab.com/ezlab/busco

Fonctionnement

- BUSCO utilise
  - **BLAST+** pour la recherche de séquences
  - **Augustus** pour la prédiction des gènes basés sur les profils de blocs
  - **HMMER** pour la recherche des profils issue des modèles de Markov cachés



BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs (https://doi.org/10.1093/bioinformatics/btv351)

BUSCO's limits

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

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)
https://academic.oup.com/bioinformatics/article/31/19/3210/211866

BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs (https://doi.org/10.1093/bioinformatics/btv351)

- Assemblage du génome moins complet que l'annotation (*H. sapiens*)
- Annotation du génome moins complète que l'assemblage (*A. nidulans*)
- Présence de missing peuvent aussi être signalés (*C. elegans*)

Used BUSCO

- Input avec une origine taxonomique connue
  - Assemblage du génome
  - Groupe de gènes annotés
  - Assemblage du transcriptome
- Outputs
  - Résumé court
  - Rapport complet: informations détaillées sur chaque gène orthologue analysé
  - Liste des gènes manquants
  - GFF file

BUSCO on Galaxy

- Input : génome assemblé, transcriptome ou séquences protéiques
- Sélectionner le mode en fonction de l'input
- Sélectionner la base de données de la lignée
- Plusieurs outputs

Output example

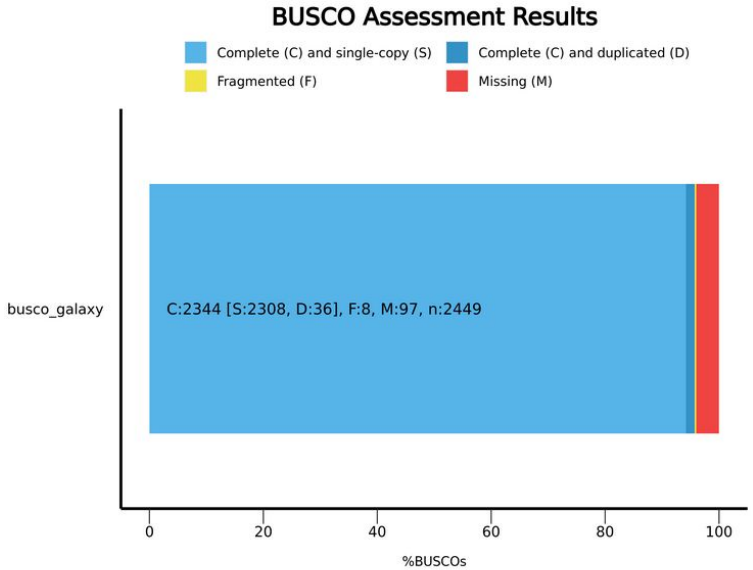
```
# BUSCO version is: 5.5.0
# The lineage dataset is: mucorales_odb10 (Creation date: 2020-08-05, number of genomes: 15,
number of BUSCOs: 2449)
# Summarized benchmarking in BUSCO notation for file /shared/ifbstor1/galaxy/datasets2/b/0/e
/dataset_b0e39757-06c9-4b80-8f94-51f94917c4d2.dat
# BUSCO was run in mode: euk_genome_met
# Gene predictor used: metaeuk

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

        C:95.7%[S:94.2%,D:1.5%],F:0.3%,M:4.0%,n:2449
        2344    Complete BUSCOs (C)
        2308    Complete and single-copy BUSCOs (S)
        36      Complete and duplicated BUSCOs (D)
        8       Fragmented BUSCOs (F)
        97      Missing BUSCOs (M)
        2449    Total BUSCO groups searched

Assembly Statistics:
        1425    Number of scaffolds
        1425    Number of contigs
        48545885        Total length
        0.000%  Percent gaps
        222 KB  Scaffold N50
        222 KB  Contigs N50


Dependencies and versions:
        hmmsearch: 3.1
        bbtools: 39.01
        metaeuk: 6.a5d39d9
        busco: 5.5.0
```



Output example

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10
# BUSCO version is: 5.5.0									
# The lineage dataset is: mucorales_odb10 (Creation date: 2020-08-05, number of genomes: 15, number of BUSCOs: 2449)									
# Busco id	Status	Sequence	Gene Start	Gene End	Strand	Score	Length	OrthoDB url	Description
1at4827	Complete	scaffold_162:43606-57994	43606	57994	+	8533.0	4025	https://www.orthodb.org/v10?query=1at4827	dynein heavy chain
2at4827	Complete	scaffold_67:146757-163028	146757	163028	+	7507.1	4522	https://www.orthodb.org/v10?query=2at4827	Midasin
10at4827	Complete	scaffold_21:270496-257489	270496	257489	-	2586.2	2260	https://www.orthodb.org/v10?query=10at4827	Phosphatidylinositol 3-/4-kinase, catalytic domain
26at4827	Complete	scaffold_45:149426-156552	149426	156552	+	4983.1	2167	https://www.orthodb.org/v10?query=26at4827	Pre-mRNA-processing-splicing factor 8
27at4827	Complete	scaffold_28:280672-270243	280672	270243	-	4343.7	2591	https://www.orthodb.org/v10?query=27at4827	Vacuolar protein sorting-associated protein 13

Column 1
BUSCO version is: 5.5.0
The lineage dataset is: mucorales_odb10 (Creation date: 2020-08-05, number of genomes: 15, number of BUSCOs: 2449)
Busco id
10103at4827
10252at4827
10312at4827
10527at4827
10555at4827
10618at4827
10668at4827
10977at4827
11139at4827

Compleasm

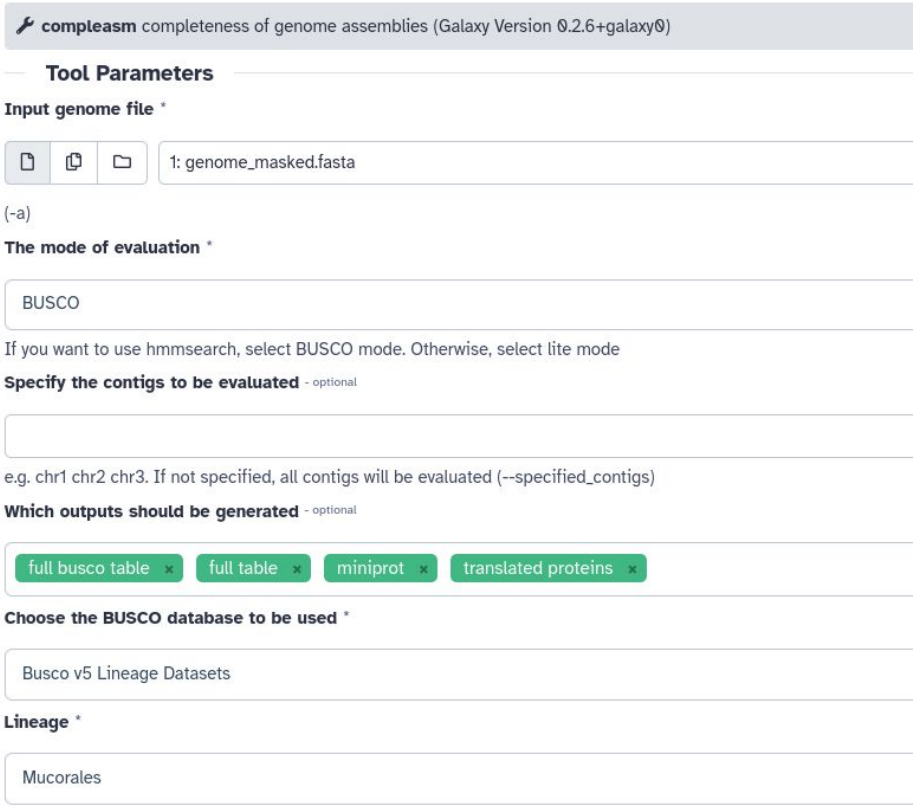
Compleasm

- Décrit comme plus rapide et plus précis que BUSCO
- On s'attend à ce que les résultats soient similaires à ceux de BUSCO

- https://github.com/huangnengCSU/compleasm

Compleasm on Galaxy

- Input : génome assemblé, transcriptome ou séquences protéiques
- Sélectionner la base de données de la lignée
- Plusieurs outputs



Output

- Full table

Gene	Status	Sequence	Gene Start	Gene End	Strand	Score	Length	Identity	Fraction	Frameshift events	Best gene	Codons
10011at4827	Single	scaffold_56	43674	44913	-	1244	270	0.6793	0.9853801169590644	0	10011at4827_0	44850_44913_-|44759_44797_-
10014at4827	Single	scaffold_4	649092	650416	+	1347	257	0.6955	1.0	0	10014at4827_9	649093_649293_+|649346_64
10018at4827	Single	scaffold_35	210287	211825	+	1095	297	0.5125	1.0	0	10018at4827_5	210288_210429_+|210482_210
10027at4827	Single	scaffold_5	100934	102940	+	1286	482	0.4455	0.9704433497536946	0	10027at4827_8	100935_101346_+|101408_1015
10032at4827	Single	scaffold_2	441303	442310	-	1291	272	0.8081	0.928125	0	10032at4827_2	442122_442310_-|441913_4420
10036at4827	Single	scaffold_239	7212	8181	-	614	242	0.4327	0.8952095808383234	0	10036at4827_5	7988_8181_-|7213_7924_-
10048at4827	Single	scaffold_16	262345	263164	+	1021	187	0.8097	0.9656652360515021	0	10048at4827_4	262346_262401_+|262485_263
1004at4827	Single	scaffold_19	424396	429224	-	4451	913	0.7502	1.0	0	1004at4827_3	429042_429224_-|428962_428
10053at4827	Single	scaffold_240	20291	21360	+	586	199	0.4813	1.0	0	10053at4827	20292_20531_+|20614_20756_-

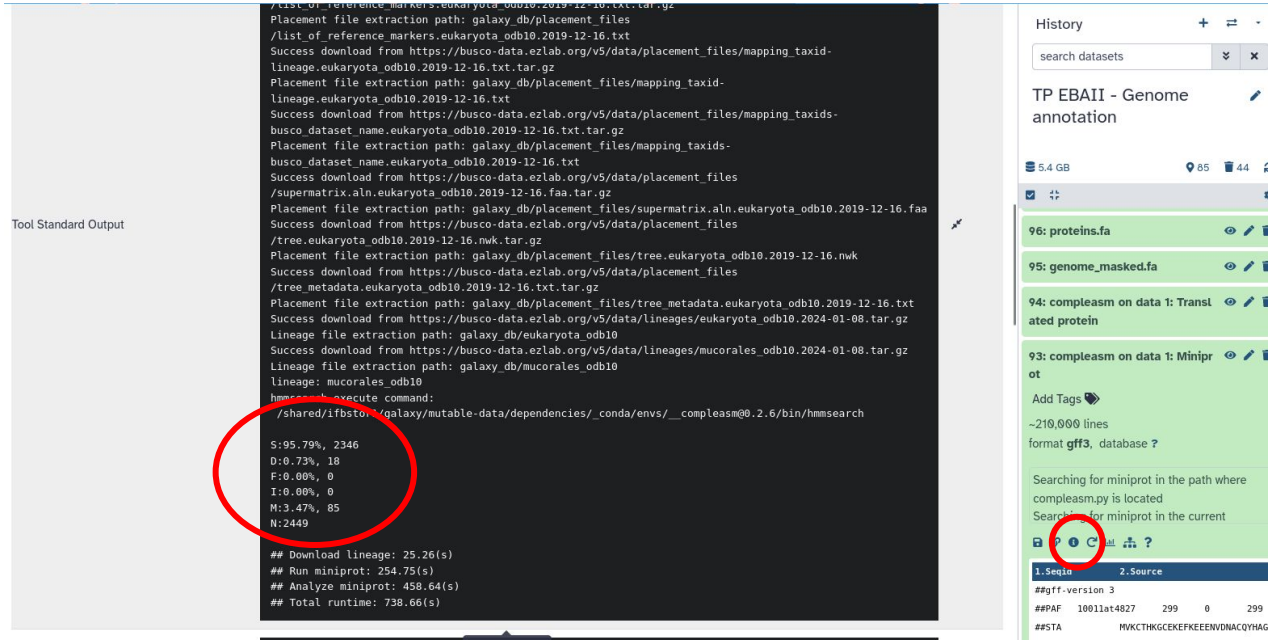
Output

- Full table busco

# Busco id	Status	Sequence	Gene Start	Gene End	Strand	Score	Length	OrthoDB url	Description
# Busco id	Status	Sequence	Gene Start	Gene End	Strand	Score	Length	OrthoDB url	Description
10011at4827	Complete	scaffold_56	43674	44913	-	1244	270	https://v10-1.orthodb.org/?query=10011at4827	Chord-domain-containing protein
10014at4827	Complete	scaffold_4	649092	650416	+	1347	257	https://v10-1.orthodb.org/?query=10014at4827	WD40-repeat-containing domain
10018at4827	Complete	scaffold_35	210287	211825	+	1095	297	https://v10-1.orthodb.org/?query=10018at4827	Dolichyl-diphosphooligosaccharide--protein glycosyltransferase 48kDa subunit
10027at4827	Complete	scaffold_5	100934	102940	+	1286	482	https://v10-1.orthodb.org/?query=10027at4827	Ubiquitin specific protease, conserved site
10032at4827	Complete	scaffold_2	441303	442310	-	1291	272	https://v10-1.orthodb.org/?query=10032at4827	3-methyl-2-oxobutanoate hydroxymethyltransferase
10036at4827	Complete	scaffold_239	7212	8181	-	614	242	https://v10-1.orthodb.org	Leucine-rich repeat

Output

- Short resume dans les logs de l'outil




```
S:95.79%, 2346
D:0.73%, 18
F:0.00%, 0
I:0.00%, 0
M:3.47%, 85
N:2449
```

BUSCO vs Compleasm

- Résultats similaires
  - BUSCO
    - Single : 2308 (94.2%)
    - Duplicated : 36 (1.5%)
    - Fragmented : 8 (0.3%)
    - Missing : 97 (4%)

```
C:95.7%[S:94.2%,D:1.5%],F:0.3%,M:4.0%,n:2449
2344    Complete BUSCOs (C)
2308    Complete and single-copy BUSCOs (S)
36      Complete and duplicated BUSCOs (D)
8       Fragmented BUSCOs (F)
97      Missing BUSCOs (M)
2449    Total BUSCO groups searched
```

  - Compleasm
    - Single : 2346 (95.79%)
    - Duplicated : 18 (0.73%)
    - Fragmented : 0 (0%)
    - Missing : 85 (3.47%)

```
S:95.79%, 2346
D:0.73%, 18
F:0.00%, 0
I:0.00%, 0
M:3.47%, 85
N:2449
```

TP BUSCO

GTN (suite de Funannotate) : https://training.galaxyproject.org/topics/genome-annotation/tutorials/funannotate/tutorial.html#evaluation-with-busco

Paramètres

- Use Augustus or Metaeuk ?
  - Deux prédicteurs de gènes de gènes différents
  - Par défaut → Metaeuk
    - Recherche des gènes codants pour des protéines en fonction de leur similarité avec des protéines de référence ou des profils protéiques
  - Augustus
    - Programmes les plus précis pour l'espèce pour laquelle il est entraîné → Programme ab initio
- Comment choisir ?

Comment choisir ?

- Metaeuk :
  - Approche basée sur des profils de séquences et des alignements multiples
  - Évaluation plus rapide, plus efficace pour traiter des données complexes
- Augustus :
  - Approche basée sur des modèles statistiques (modèles de Markov cachés)
  - Utilise des informations sur les séquences codantes connues et les introns
  - Plus précis en termes de modèles génétiques

Comment choisir ?

- Metaeuk :
  - Approche basée sur des profils de séquences et des alignements multiples
  - Évaluation plus rapide, plus efficace pour traiter des données complexes
- Augustus :
  - Approche basée sur des modèles statistiques (modèles de Markov cachés)
  - Utilise des informations sur les séquences codantes connues et les introns
  - Plus précis en termes de modèles génétiques
- Bilan
  - Metaeuk : adapté pour des analyses rapides et flexibles
  - Augustus : adapté pour des prédictions de gènes précises dansdes génomes eucaryotes bien étudiés

Tips

- Liste des lingées BUSCO : https://github.com/Gaius-Augustus/Augustus/blob/master/docs/RUNNING-AUGUSTUS.md
- Gitlab BUSCO : https://busco.ezlab.org/
- Github Compleasm : https://github.com/huangnengCSU/compleasm