BUSCO results for 89-1A

lihui xiang

2025-04-01

For this section, we have already obtained the genome sequence of *Fusarium oxysporum* strain Fov891A using the PacBio sequencing platform. The lab has access to the assembled genome file (`Fov891A.contigs.fasta`), the corresponding gene annotation file (`augustus_Fov891A.gff`), and the initial predicted protein file (`au_Fov891A.proteins.fa`). To ensure the quality of protein predictions, I used the genome + GFF combination to extract high-quality protein sequences and corresponding CDS (coding sequences) using the tool `gffread`. This step was performed on our university's high-performance computing (HPC) cluster.

Step 1: Extracting High-Quality Protein and CDS Sequences

For this project, we obtained the genome sequence of *Fusarium oxysporum* strain Fov891A using the PacBio sequencing platform. The lab already had the assembled genome file (`Fov891A.contigs.fasta`), the gene annotation file (`augustus_Fov891A.gff`), and the initial predicted protein file (`au_Fov891A.proteins.fa`).

To improve the quality of the predicted protein sequences, I used the genome and GFF files to extract high-confidence protein sequences and corresponding CDS (coding DNA sequences) using the tool `gffread`. This step was performed on our university's high-performance computing (HPC) cluster.

Code used

```
module load gffread

gffread augustus_Fov891A.gff \
  -g Fov891A.contigs.fasta \
  -y Fov891A_predicted_proteins.faa \
  -x Fov891A_predicted_CDS.fna
```

This process generated two output files: Fov891A_predicted_proteins.faa and Fov891A_predicted_CDS.fna, which are used in the subsequent BUSCO analysis.

Step 2: Assessing Protein Annotation Completeness Using BUSCO

To evaluate the completeness and reliability of the predicted protein sequences, BUSCO (Benchmarking Universal Single-Copy Orthologs) was used. I ran BUSCO in proteins mode with the sordariomycetes_odb10 lineage dataset on the HPC cluster.

Code used

```
nano run_busco_predicted.sh
```

```
#!/bin/bash
#PBS -N busco_predicted
#PBS -o busco_predicted_output.log
#PBS -e busco_predicted_error.log
#PBS -l nodes=1:ppn=4
#PBS -l walltime=01:00:00
#PBS -l mem=8gb
#PBS -V
#PBS -q medium

cd "$PBS_O_WORKDIR"

module load busco/5.4.3

busco -i Fov891A_predicted_proteins.faa \
      -o busco_predicted_output \
      -l sordariomycetes_odb10 \
      -m proteins \
      --cpu 4 \
      -f
```

```
qsub run_busco_predicted.sh
```

Step 3: Reviewing BUSCO Output

After the BUSCO job finishes, I checked the summary results using the command below:

```
cat busco_predicted_output/short_summary*.txt
```

BUSCO Output

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

C:98.8%[S:97.9%,D:0.9%], F:0.4%, M:0.8%, n:3817
3772    Complete BUSCOs (C)
3738    Complete and single-copy BUSCOs (S)
34      Complete and duplicated BUSCOs (D)
16      Fragmented BUSCOs (F)
29      Missing BUSCOs (M)
3817    Total BUSCO groups searched
```

Conclusion

The BUSCO results show that 98.8% of the expected universal single-copy orthologs were found to be complete in the predicted protein sequences (97.9% single-copy, 0.9% duplicated), while only 0.4% were fragmented and 0.8% were missing.

This high completeness score indicates that the genome assembly and gene annotation for F. oxysporum Fov891A are of high quality and are suitable for downstream analyses, including functional annotation, effector prediction, and comparative genomics.

Step 4: BUSCO Result Visualization

In this step, we use the BUSCO output to visualize the completeness of the predicted protein sequences. A pie chart is generated in R to provide a more intuitive and easy-to-read summary of the results.

```
library(ggplot2)
library(dplyr)

# BUSCO data
busco <- data.frame(
  Category = c("Complete and single-copy", "Complete and duplicated", "Fragmented", "Missing"),
  Count = c(3738, 34, 16, 29)
)

# Custom color
custom_colors <- c(
  "Complete and duplicated"  = "#E76F51",
  "Complete and single-copy" = "#A9BBA9",
  "Fragmented"               = "#2A9D8F",
  "Missing"                  = "#9D4EDD"
)

# Plot
ggplot(busco, aes(x = "", y = Count, fill = Category)) +
  geom_col(width = 1, color = "white") +
  coord_polar(theta = "y") +
  scale_fill_manual(values = custom_colors) +
  theme_void() +
  theme(
    legend.position = "right",
    legend.title = element_blank(),
    legend.text = element_text(size = 10)
  )
```

```
# Generate a summary table
library(ggplot2)
library(dplyr)
library(knitr)

# Add percentage column
busco$Percentage <- round(busco$Count / sum(busco$Count) * 100, 1)

# Display the table
kable(busco, caption = "Summary of BUSCO Result Counts and Percentages")
```

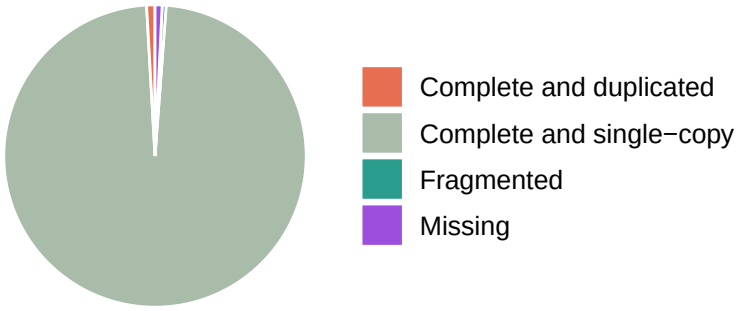

Figure 1: BUSCO Protein Completeness (Fov891A)

Table 1: Summary of BUSCO Result Counts and Percentages

Category	Count	Percentage
Complete and single-copy	3738	97.9
Complete and duplicated	34	0.9
Fragmented	16	0.4
Missing	29	0.8