XIII Argentine Congress of Bioinformatics and Computational Biology
XIII International Conference of the Iberoamerican Society of Bioinformatics
III Annual Meeting of the Ibero-American Artificial Intelligence Network for Big BioData

A comparative assessment on fungal genome annotation

Andreiu, Erika María Martha[1] ; Chelaliche, Anibal Sebastian[12] ; Benitez, Silvana Florencia[12] ; Zapata, Pedro Dario[12] ; Fonseca, Maria Isabel[12] .

1. *Universidad Nacional de Misiones. Facultad de Ciencias Exactas, Químicas y Naturales. Instituto de Biotecnología Misiones. Laboratorio de Biotecnología Molecular. CP3300 Posadas, Misiones, Argentina. Tel. +54 376 4480200. E-Mail: biotecmol2010@gmail.com*
2. *CONICET. Buenos Aires, Argentina.*

Background

Genome annotation techniques and methodologies used in the assessment of fungal genomes play a crucial role in understanding the genetic background and functionalities of these organisms. An ambitious genome-sequencing program provides a wealth of data on metabolic diversity within the fungal kingdom, thereby enhancing research into medical science, agriculture science, ecology, bioremediation, bioenergy, and the biotechnology industry.

There are various approaches to this complex process, and it is important to evaluate the quality and reliability of the annotation performed on fungal genomes. In this context, *Pleurotus pulmonarius* LBM 105, a fungus isolated from Misiones (Argentina), has demonstrated qualities suitable for environmental biotechnology applications. Therefore, progress in annotating its genome represents a significant advancement in both scientific and technological understanding.

Materials and methods

-The fungal strain used in this work is *Pleurotus pulmonarius* LBM 105, isolated from the subtropical forest of Misiones (Argentina).

-Whole genome sequencing (WGS) was performed using an **Illumina** HiSeq 4000 platform.

-*De novo* assembly of the reads was carried out using the **De Novo Assembly 1.5** tool of **CLC Genomics Workbench 22.0.2.**

-The *de novo* assembly was analyzed with the **BUSCO** software in order to determine the expected gene content. This analysis was performed with the Galaxy server (www.Galaxy.eu) (Galaxy version 2.0.4+galaxy1). The MetaEuk tool was used as a gene search engine and the search lineage (Agaricales) was manually selected. The E-value cutoff value for BLAST searches was set to 0.001.

-The reference annotation was generated with the **Annotate from Reference tool** of **CLC Genomics Workbench 23.0.1.** using the PM_ss13_v1 genome as a reference.

-We use two different masking approaches for the *ab initio* annotations, one using **RepeatModeler** to generate repeat sequences from the genome, and another utilizing the **Dfam** repeat database.

-For *de novo* annotation of the genome, **Augustus** software version 3.4.0 was used.

-The prediction was made using the **Funannotate** prediction annotation tool, available on the Galaxy server. For this, the Funannotate database 2022-01-17-193541 was used. For the BUSCO algorithm, the mushroom option was chosen and within the set of species trained in Augustus for the alignment with BUSCO, the species *Phanerochaete chrysosporium* was chosen.

References

https://www.qiagen.com/us/products/discovery-and-translational-research/next-generation-sequencing/informatics-and-data/analysis-and-visualization/clc-genomics-workbench.

Bourque, G., et al. (2018). *Genome Biology*, *19*(1).

Flynn, J. M., et al. (2020). *Proceedings of the National Academy of Sciences*, 117(17), 9451–9457

Keller, O., et al. (2011). *Bioinformatics*, *27*(6), 757–763.

Krzywinski, M., et al. (2009). *Genome Research*, *19*(9), 1639–1645

Palmer, J. M., & Stajich, J. (2020). *Funannotate v1.8.1: Eukaryotic genome annotation*. Zenodo.

Rasche, H., & Hiltemann, S. (2020). *GigaScience,* 9(6).

Simão, F. A., et al. (2015). *Bioinformatics,* 31(19), 3210-3212.

Stanke, M., & Waack, S. (2003). *Bioinformatics*, *19*(Suppl 2), ii215–ii225.

Stanke, M., et al. (2008). *Bioinformatics*, *24*(5), 637–644.

Results

In this study, we examined various annotation methods applied to the *de novo* assembly of *P. pulmonarius* LBM 105 in order to evaluate their accuracy and completeness. We compared a reference annotation generated with CLC Genomic Workbench to classic *ab initio* annotation using the Augustus software and fungal optimized annotation software called Funnanotate. Additionally, we incorporated two different masking approaches for the *ab initio* annotations - one using RepeatModeler to generate repeat sequences from the genome, and another utilizing the Dfam repeat database. The most complete annotation was achieved by the reference annotation yielding a total of 13560 genes, 79283 CDS and 12693 mRNAs. However, using Augustus we achieved similar results with 13331 genes, 57933 CDS and 13331 mRNA

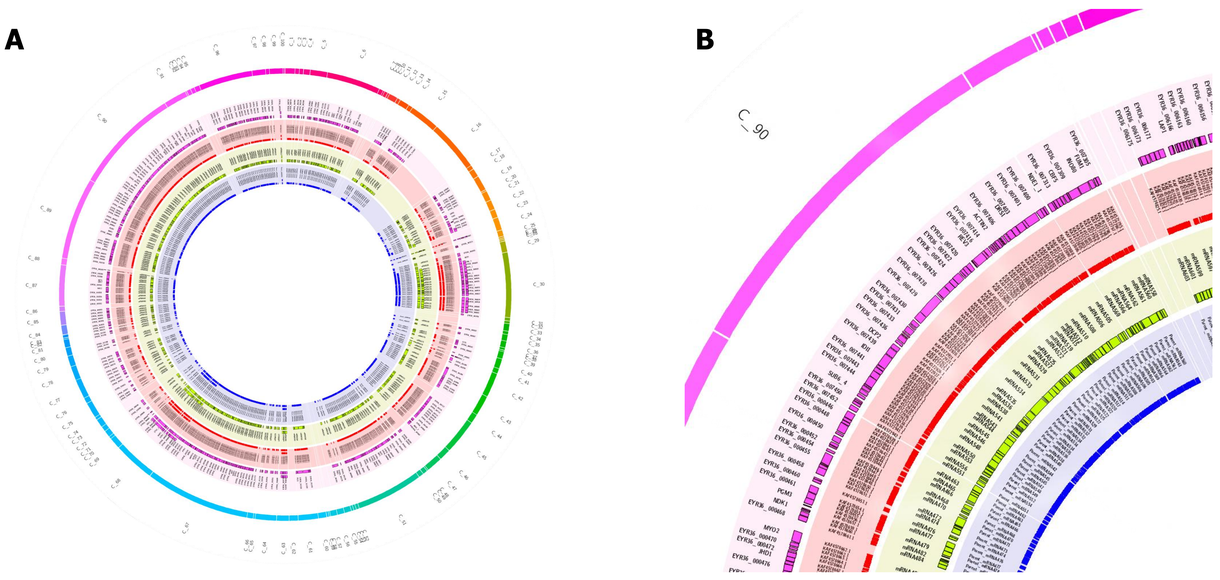

Figure 1. Circos plot that shows the annotation to the reference of the first hundred contigs **(A)** and the annotations corresponding to contig number 90 **(B)**.

Comparison of results from different annotations

Table 1. Results of the three annotations for *Pleurotus pulmonarius* LBM 105

Annotation	CDS	Exon	Gen	RNAm	RNAt
To Reference	79283	79731	13560	12693	448
Augustus (ab initio)	57933	57933	13331	13331	0
Funnanotate	50166	50291	11820	11695	125

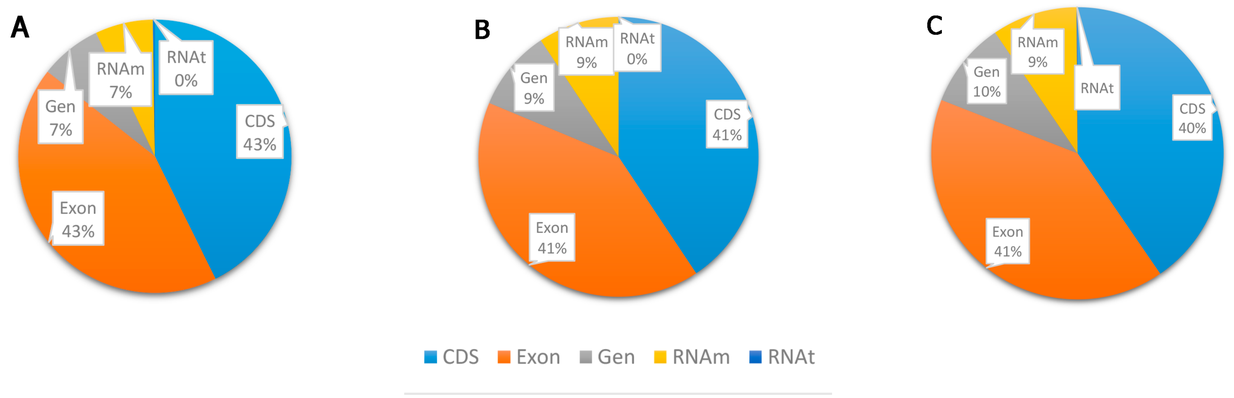

Figure 2. Percentages corresponding to each category (CDS, Exon, Gen, RNAm, RNAt) for the different types of annotation tested for *Pleurotus pulmonarius* LBM 105. **A.** to Reference. **B.** Augustus. **C.** Funnanotate.

Conclusions

This level of completeness suggests that the annotation pipeline used in this study successfully captured a significant portion of the fungal genome, highlighting the importance of using a high quality reference for the genomic studies. However, the ab initio annotation results showed promising results, allowing to better capture the strain-specific characteristics of the genome