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

Evaluation of RNA-Seq assemblies of *Matricaria chamomilla* for the definition of a workflow in the construction of de Novo transcriptomes

Maggio Julián F.[1], **García Laura E.**[2,3], ***Costa Tártara, Sabrina M.**[1,4]

[1] Departamento de Ciencias Básicas, Universidad Nacional de Luján. Av. Constitución y Ruta Nac. N° 5 (s/n), Luján (6700), Buenos Aires. [2]IBAM-UNCuyo, CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas). Alte Brown 500, Chacras de Coria, Luján de Cuyo, Mendoza. [3] Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Cuyo. Padre Jorge Contreras 1300, Mendoza (5502). [4] CONICET. *scosta@unlu.edu.ar

INTRODUCTION

An organism's transcriptome represents the portion of its genome expressed at a specific moment under particular conditions and can be constructed using RNA sequences (RNA-Seq) (1). This approach helps to generate primary gene expression information of a species, like some medicinal plants, as non-model species. Various tools exist to assemble short RNA reads in consensus sequences (assemblers) and generate de Novo transcriptomes (without a reference genome). The assemblers can use de Bruijn Graphs (dBG) and Overlap Layout Consensus (OLP) as assembling algorithms (2).

Although there is a base workflow to construct a transcriptome, the parameters chosen to use the different tools are only sometimes reported (3). The first step in this workflow is the library's quality control and post-cleaning if required. Each nucleotide generated by the sequencing platform has a quality value called Phred (Q), which represents a probability of error in its determination. Quality control of the library involves an inspection of the reads based on the Q value for nucleotides and their proportion in each read, which can lead to a correction (4). The quality of post-assembling products (several contigs or transcriptomes) can also be assessed from other perspectives, like metrics regarding the length of each long sequence consensus (contig), the integrity of the transcriptome (the total of contigs) and the magnitude of transformations made in raw read to construct contigs.

For this work, we used *Matricaria chamomilla* (Chamomile) as a non-model organism for which there was no published reference genome at the moment of this analysis. This fact prompted the construction of the de Novo transcriptome.

OBJECTIVE

This work aimed to compare the quality of different versions of the Manzanilla transcriptome constructed by choosing different parameters combinations to control the library's quality and three assemblers tools based on their algorithm implemented.

MATERIALS AND METHODS

The RNA-Seq library of Chamomile was generated in 2014 by a Roche 454 sequencing platform. Fastq files have a size of 262,4 Mb containing 300.719 reads and 119.197.062 nucleotides. The length range among reads was 24 (minimum) - 1308 (maximum) nucleotides. A total of 12 treatments were arranged to generate different de Novo transcriptomes (figure 1). Pre-assembly cleaning was performed with NGS QC Toolkit (5) under four combinations of Q values: first, the nucleotide Q value was set in 30 and 35, and a proportion of the read with that value as minimal set in 35%, 70% and 45% (30-35 and 35-45), one treatment (30-70-80), in which reads were trimmed to get a minimal length of 80 nucleotides, and a control consisted in the raw library. The assembling was performed with three different tools based on their base algorithm: NEWBLER (OLC) (6), SOAPDeNovo-Trans (dBG) (7) and Trinity (dBG) (8). In order to analyze the post-assembling quality control, we implemented rnaQUAST (9) to calculate some transcriptome metrics related to the continuity of the sequence without needing to align to a reference, DETONATE (10) and specifically DETONATE's RSEM-EVAL score which allows comparing the fidelity of assemblies in contrast to reads used as input. For the transcriptomes integrity analysis, we used BUSCO (11), a tool that evaluates the presence of specific orthologous genes considered universal for different taxonomic levels.

RESULTS

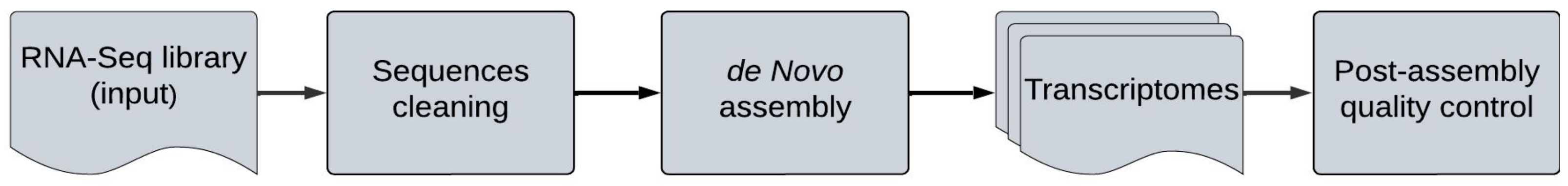
- The rnaQUAST metrics selected were: *Total transcripts*, *Transcripts greater than 500 bp*, *Transcripts greater than 1000 bp, Average transcript length, Total transcriptome length* and *N50*. The pattern of the results observed for each assembler was the same for the four quality combinations: the control and 30-35 quality set recorded the highest values, followed by the 35-45 quality set and last on the 30-70-80. The rnaQUAST indicates longer consensus sequences and greater N50 values for transcriptomes assembled by Trinity than other assemblers.
- Using the DETONATE software, the RSEM-EVAL score was obtained for each transcriptome. Generally, the same pattern was observed for each assembler: the lowest values were for the control and 30-35, with similar values. The highest score was for the treatment 30-70-80, followed by 35-45. Regarding assembler differences, Trinity presented higher scores, followed by NEWBLER and the lower values results for SOAP-DeNovoTrans (figure 2).
- We used the viridiplanate_odb10 database to carry out BUSCO. The results showed a pattern within the assemblers; for the different pre-assembly cleanings, the less demanding ones resulted in higher proportions of Complete and Single-copy (S) BUSCO's genes categories, in contrast with more cleaning treatments for which this proportion decreased. Complete and duplicated (D) and Fragmented (F) BUSCO's genes categories, the proportions were similar for all treatments. Among assemblers, Trinity generates transcriptomes with the highest number of S and D genes, followed by NEWBLER and SOAP-DeNovoTrans. The assembler that generated the greater number of F genes was SOAP-DeNovoTrans, followed by Trinity and NEWBLER (figure 3).

ACKNOWLEDGMENTS

This work takes part in the final grade Thesis of Lic. Julián F. Maggio in current year. The data of the RNA-Seq *M. chamomilla* library was obtained in a framework of a scientific, technical and academic cooperation agreement between INTA (Instituto Nacional de Tecnología Agropecuaria) and Universidad Nacional de Luján through Instituto de Recursos Biológicos and Departamento de Tecnología (RESHCS-LUJ: 241-16) and National funds from INTA projects (PNBIO and PNHFA).

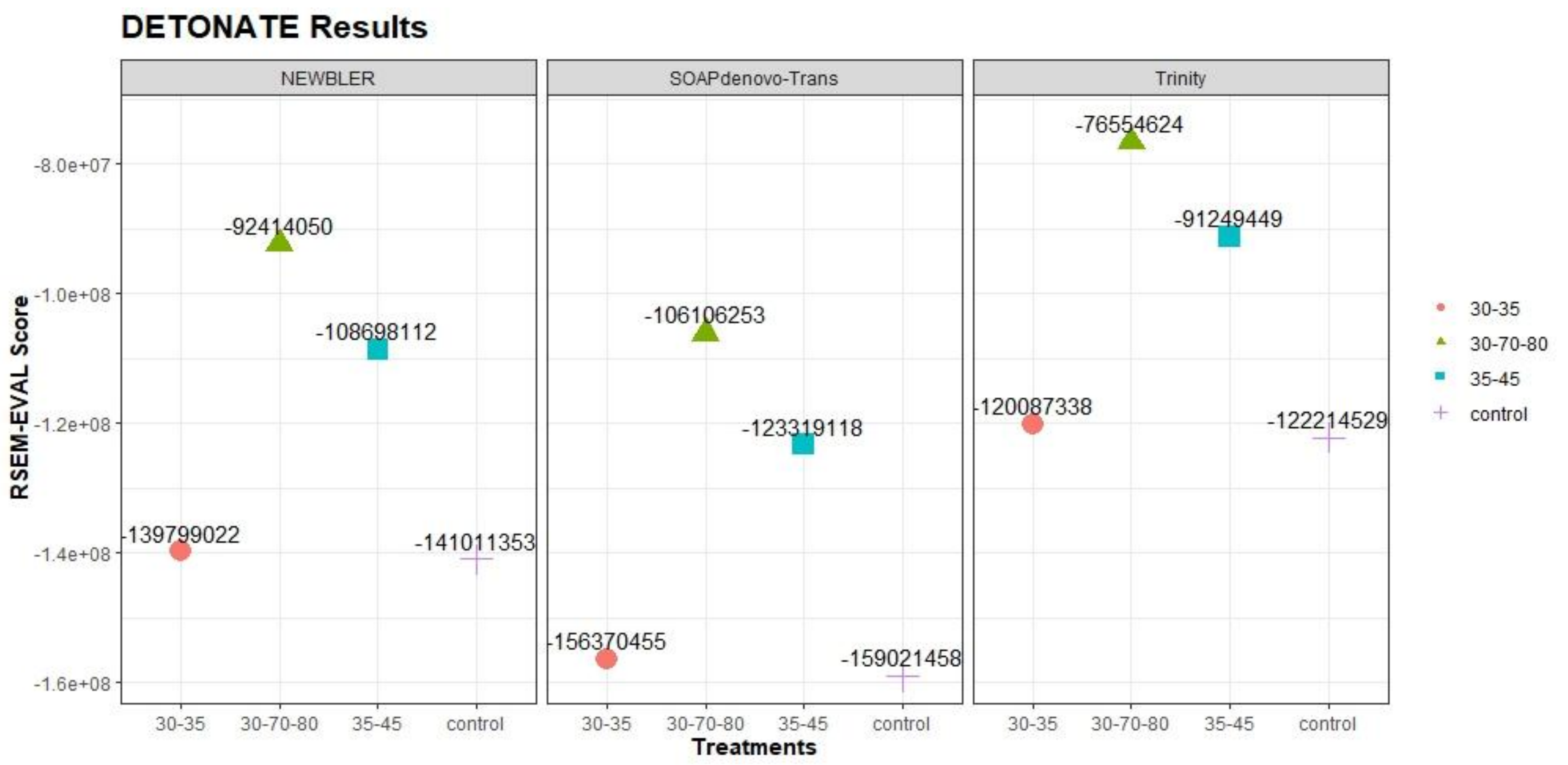
The authors thank the Center for Research, Teaching and Extension in TICs (CIDETIC, http:/cidetic.unlu.edu.ar/) in the Universidad Nacional de Luján for providing human and computational resources necessary for this project.

Figure 1: Proposed workflow.



RNA-Seq library (input) → Sequences cleaning → *de Novo* assembly → Transcriptomes → Post-assembly quality control

Figure 2: Graphic representation of DETONATE results.



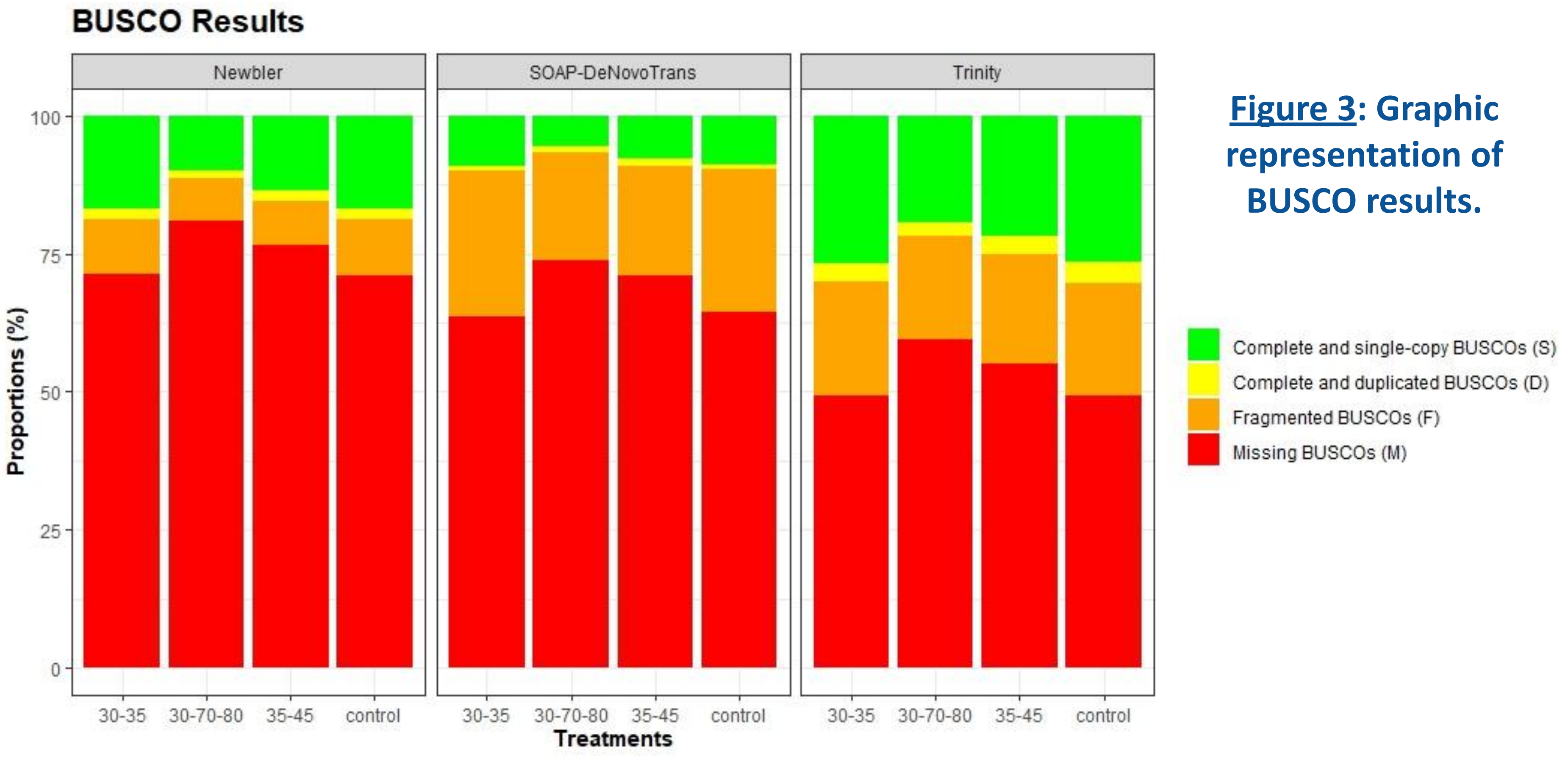


Figure 3: Graphic representation of BUSCO results.

CONCLUSIONS

Despite the small size of the library, it was a helpful input to test different pre-assembly treatments and assemblers. The output analysis with different tools allows for analyzing sequence parameters, the transformation of the raw library at the time of assembly and the integrity of the transcriptome. Trinity (dBG) proved to be the most effective tool for assembling the library, surpassing even NEWBLER (OLC), the assembler recommended by Roche.

Not all cleaning treatment effects are the same for the assemblers. It is likely necessary to adapt the quality control treatment prior to assembly to the assembler that will be used.

Finally, it is essential to report the parameters used in all stages of the de Novo transcriptome assembly to ensure reproducibility since it was demonstrated that different results can be reached.

REFERENCES

1- Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. PLoS Computational Biology, 13(5). https://doi.org/10.1371/journal.pcbi.1005457

2- Miller, J. R., Koren, S., & Sutton, G. (2011). Genome Sequencing Algorithms. 95(6), 315–327. https://doi.org/10.1016/j.ygeno.2010.03.001.Assembly

3- Simoneau, J., Dumontier, S., Gosselin, R., & Scott, M. S. (2021). Current RNA-seq methodology reporting limits reproducibility. Briefings in Bioinformatics, 22(1), 140–145. https://doi.org/10.1093/bib/bbz124

4- Raghavan, V., Kraft, L., Mesny, F., & Rigerte, L. (2022). A simple guide to de novo transcriptome assembly and annotation. Briefings in Bioinformatics, 23(2). https://doi.org/10.1093/bib/bbab563

5- Patel, R. K., & Jain, M. (2012). NGS QC toolkit: A toolkit for quality control of next generation sequencing data. PLoS ONE, 7(2). https://doi.org/10.1371/journal.pone.0030619

6- Roche. (2013). 454 Sequencing System Software Manual Version 2.9. 454 Life Sciences. Signal Processing, June, 14–33.

7- Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T. W., Li, Y., Xu, X., Wong, G. K. S., & Wang, J. (2014). SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. Bioinformatics, 30(12), 1660–1666. https://doi.org/10.1093/bioinformatics/btu077

8- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology, 29(7), 644–652. https://doi.org/10.1038/nbt.1883

9- Bushmanova, E., Antipov, D., Lapidus, A., Suvorov, V., & Prjibelski, A. D. (2016). RnaQUAST: A quality assessment tool for de novo transcriptome assemblies. Bioinformatics, 32(14), 2210–2212. https://doi.org/10.1093/bioinformatics/btw218

10- Li, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J. A., Stewart, R., & Dewey, C. N. (2014). Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biology, 15(12). https://doi.org/10.1186/s13059-014-0553-5

11- Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Molecular Biology and Evolution, 38(10), 4647–4654. https://doi.org/10.1093/molbev/msab199