De novo assembly of the genome of the Shortfin scad, *Decapterus macrosoma* (Perciformes, Carangidae): constructing a genomic resource for galunggong biology

Zae-Zae A. Aguinaldo[1], Ricardo P. Babaran[2], and Arturo O. Lluisma[1]

[1] The Marine Science Institute, University of the Philippines, Diliman, Quezon City, Philippines

[2] Institute of Marine Fisheries and Oceanology, College of Fisheries and Ocean Sciences, University of the Philippines Visayas, Miag-ao, Iloilo, Philippines

ABSTRACT

Locally known as “Galunggong”, *Decapterus* species (round scads) are small pelagic fishes that belong to the family Carangidae and are widely distributed in the Indo-Pacific oceans. These species are among the most economically important fishes in Philippine fisheries, both municipal and commercial, in terms of volume and value. Genetic approaches and tools can play an important role in elucidating the impact of high level of fishing effort on, and hence in the development of effective management strategies for, local populations of these species but currently such tools are lacking.

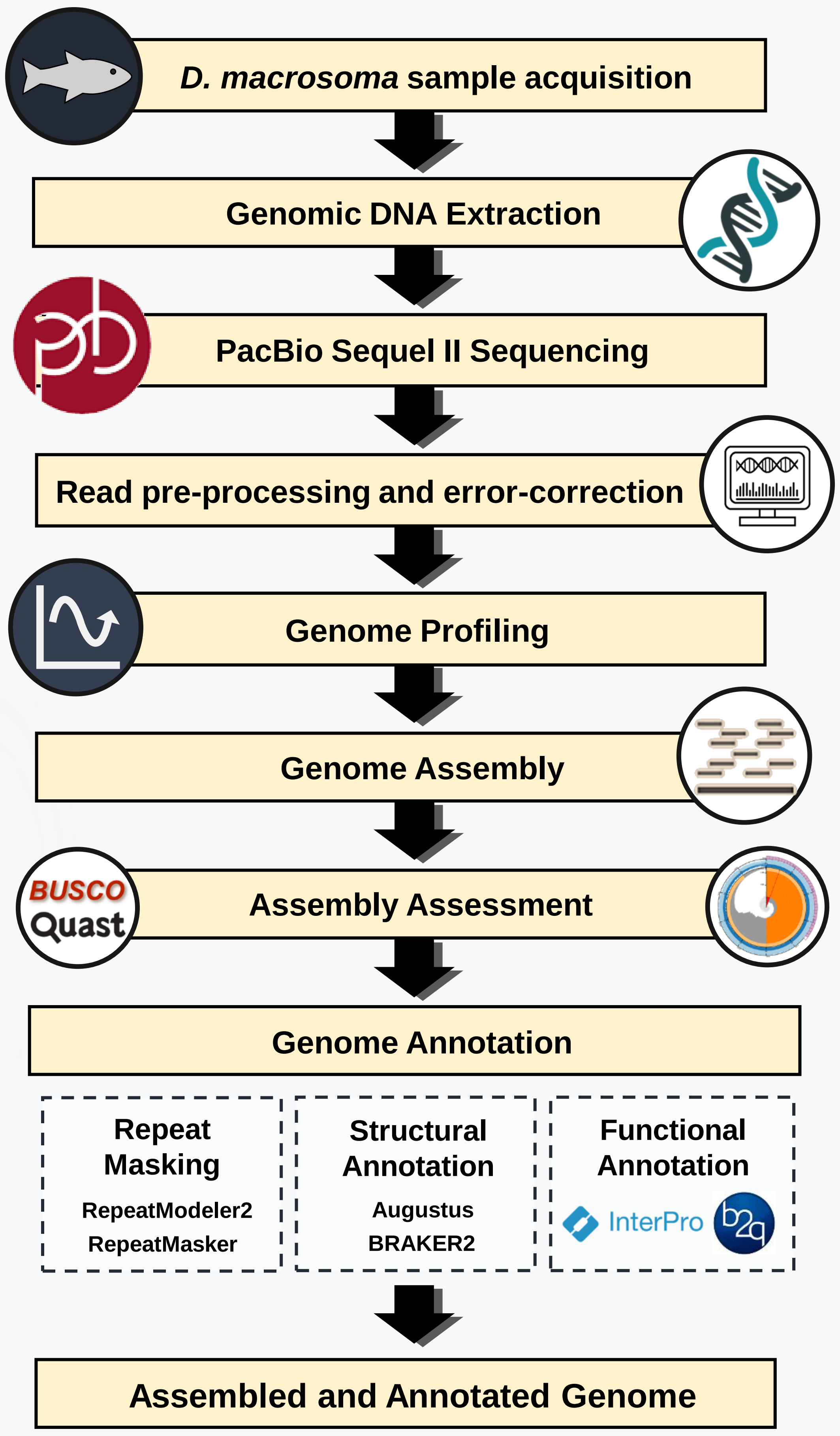
In this study, we generated and annotated a high-quality draft genome assembly for the shortfin scad, *D. macrosoma*, using long-read sequencing (PacBio HiFi reads). The total assembly size of the genome was 724 Mb, close to the expected size, comprising just 974 contigs with an N50 of 4.36 Mb; the longest contig was 18.9 Mb. BUSCO analysis indicated that the draft genome assembly was 97.8% complete with 96.5% of the single-copy orthologs in the Actinopterygii library profile. The assembled genome was also characterized by having a relatively small number of repetitive elements. The generated genome assembly is the first high-quality genome assembly to be reported for the genus *Decapterus* and will serve as a valuable resource for population genomics studies and for the development of fisheries management strategies for these species.

Keywords: *Decapterus*, PacBio Sequencing, genome, fisheries management

INTRODUCTION

The round scads account for a significant fraction of the Philippines’ fisheries production and the management of their fisheries is therefore crucial. Genetic tools have been routinely used to generate information relevant to fisheries management. In particular, the generation of genomic resources represents a key step towards the development of sound and sustainable fisheries management strategies. However, despite the economic importance of round scads to the country’s fisheries production, genomic resources for these species are still limited. Here, we sequenced, assembled, and annotated the genome of *D. macrosoma* using the PacBio HiFi sequencing platform.

MATERIALS & METHODS



D. macrosoma sample acquisition

Genomic DNA Extraction

PacBio Sequel II Sequencing

Read pre-processing and error-correction

Genome Profiling

Genome Assembly

Assembly Assessment

Genome Annotation

Repeat Masking: RepeatModeler2, RepeatMasker

Structural Annotation: Augustus, BRAKER2

Functional Annotation: InterPro, b2g

Assembled and Annotated Genome

RESULTS

A. Genome Profiling (K-mer distribution and estimation of genome properties)

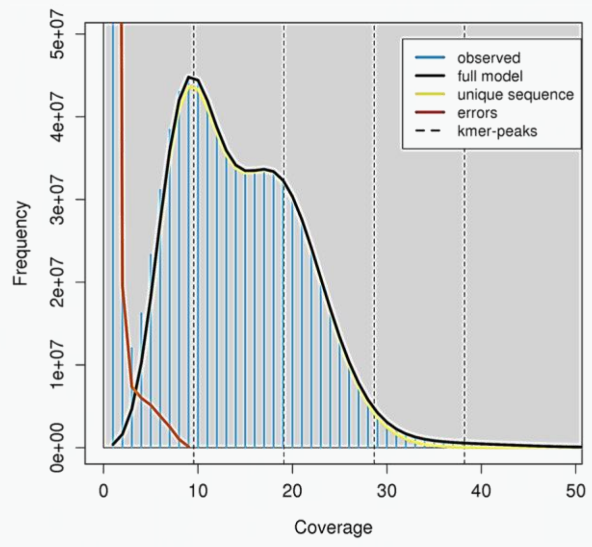


Figure 1. K-mer distribution analysis of PacBio HiFi reads using GenomeScope based on k value 21. K-mer occurrences (x axis) were plotted against their frequencies (y axis).

Table 1. Properties of *D. macrosoma* genome inferred from PacBio HiFi reads using the GenomeScope software.

Metric	*D. macrosoma* genome
Estimated Genome Size	644,508,705 bp
Genome unique length	531,662,443 bp (82.5%)
Genome repeat length	112,846,263 bp (17.5%)
Heterozygosity	1.82%
Duplication Rate	0.14%
Read Error Rate	0.19%

B. *De Novo* Genome Assembly and Quality Assessment

Table 2. Statistics of *D. macrosoma* genome assembly as evaluated from the Quast software

Assembly Statistics	*D. macrosoma* genome
Total length	724,210,758 bp
Number of contigs	974
Largest contig	18,981,966 bp
GC (%)	42.36
N50	4,366,527 bp

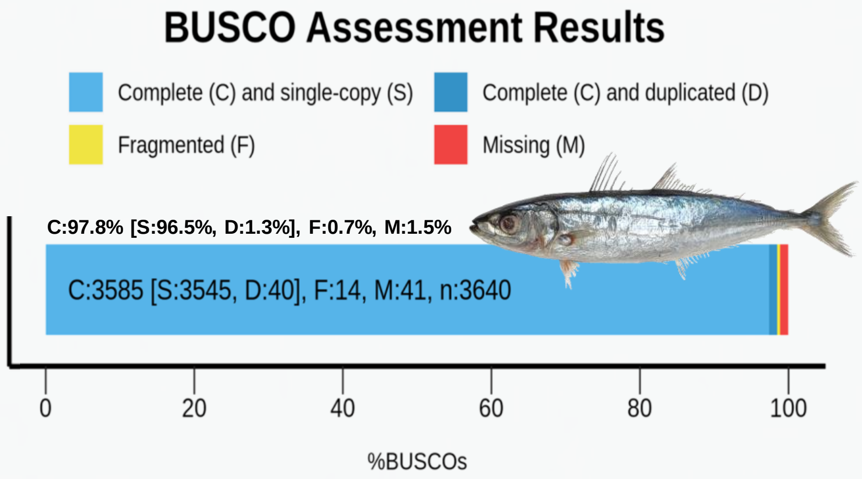


Figure 1. BUSCO evaluation of the *D. macrosoma* genes compared with the Actinopterygii gene set.

C. Genome Features (Repeats and Gene structures)

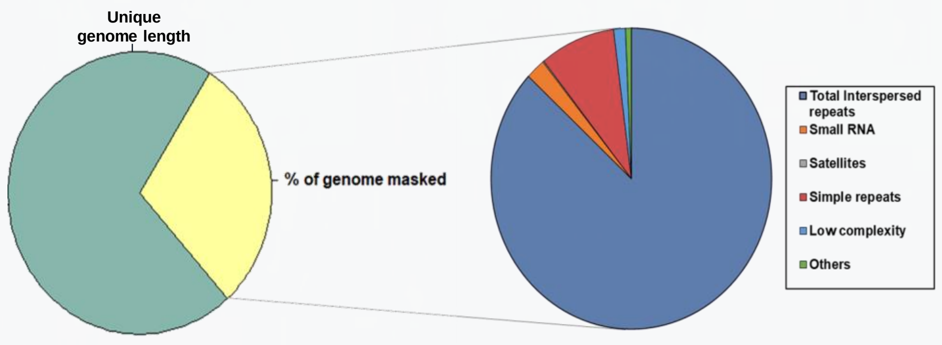


Figure 4. Repetitive elements in the assembled *D. macrosoma* genome. The repeats accounted for 29.50% of the assembled genome. The most abundant are the total interspersed repeats comprising 25.64% of the total repeats, followed by simple repeats (2.57%), small RNA (0.66%), low complexity elements (0.40%), and satellite repeats (0.04%).

Table 2. Structural annotation of *D. macrosoma* genome assembly

Gene Feature	Number	Total size (kb)
Protein-coding genes	37,358	247,323,986
Exons	268,212	47,570,271
Introns	230,854	199,753,715
Overlapping genes	0	-
Mean gene length	-	6,620
Mean exon length	-	177
Mean intron length	-	865

CONCLUSIONS

We present the first high-quality draft genome assembly for the shortfin scad, *D. macrosoma.* The total assembly size of the genome was 724 Mb, consisting of 974 contigs with an N50 length of 4.36 Mb and longest contig length of 18.9 Mb. The assembly was 97.8% complete with 96.5% of the single-copy orthologs in the Actinopterygii library profile. Further, the genome is comprised of 29.50% repetitive elements and was annotated with 37,358 protein-coding genes. The functional annotation of the draft genome is currently in progress.

REFERENCES

- Bruna, T., Hoff, K.J., Lomsadze, A., Stanke, M., & Borodovsky, M. 2021. BRAKER2: Automatic Eukaryotic Genome Annotation with GeneMark-EP+ and AUGUSTUS Supported by a Protein Database. NAR Genomics and Bioinformatics 3(1):lqaa108, doi: 10.1093/nargab/lqaa108.
- Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., & Li, H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods, https://doi.org/10.1038/s41592-020-01056-5.
- 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.
- Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. 2017. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics 33(14):2202-2204. doi: 10.1093/bioinformatics/btx153.

ACKNOWLEDGEMENTS

This study is part of the project entitled “Development of microsatellite and SNP markers for population structure in commercially important Carangid species in Philippine waters” under the program “Application of molecular genetics and otolith as tools for fisheries management” funded by the University of the Philippines System Emerging Inter-Disciplinary Research Grant (EIDR-C08-011-R).