RNA sequencing-based identification of candidate hypersensitive transcripts in *Perna viridis*

Srinivasa Raghavan Vasudevan[1,*], Vinaya Kumar Katneni[2], Sudheesh K. Prabhudas[2] and Karthic Krishnan[2]

[1]Madras Regional Station, ICAR-Central Marine Fisheries Research Institute, Chennai 600 028, India
[2]Centre for Bioinformatics, Nutrition Genetics and Biotechnology Division, ICAR-Central Institute of Brackishwater Aquaculture, Chennai 600 028, India

The Asian Green Lipped mussel, *Perna viridis* one of the widely distributed bivalves act as a source of low cost protein providing nutritious meal to the coastal population in the form of well balanced amino acids and micronutrients. The immune system produces antibodies to certain class of proteins present in shellfish meat thereby causing hypersensitive reactions in the body. The next generation integrated transcriptome sequencing approach identifies all the potential allergenic proteins expressed in an animal very effectively. The present study describes the transcriptome of *P. viridis* based on the sequence data generated using five tissues. Transcriptome level candidate allergens and epitopes were observed and identified that might play a role in hypersensitive reaction to shellfish proteins including certain novel candidate allergens like Ran protein and a filamin A like protein. The existence of epitope hotspots in an important protein, arginine kinase was also observed and the unigenes identified would be a valuable resource for conduct of functional studies.

Keywords: Allergen, hypersensitive transcripts, *Perna viridis*, RNA sequencing, unigenes.

THE Asian green-lipped mussel, *Perna viridis,* one of the widely distributed bivalves, contributes significantly to the economy and livelihood of the coastal community. The farming of this mussel provides financial security to small and marginal farmers[1]. Mussels also act as a source of low-cost protein, providing nutritious food to the coastal population in the form of well-balanced amino acids and micronutrients[2,3]. On account of their fast growth, high fecundity, reproductive potential and lower maintenance cost, mussels are one of the dominant candidate species suitable for coastal mariculture[4–8]. The large-scale commercial cultivation of green mussels is being carried out extensively in tropical countries. The scientific farming of green mussels has been carried out by several researchers of ICAR-Central Marine Fisheries Research Institute (CMFRI), Chennai. The technology has been successfully transferred and adopted on a larger scale by the coastal communities[9–11]. The farming of green mussels gained momentum with well-developed marketing facilities, storage and subsequent value addition of their products[12,13]. The ICAR-CMFRI has also successfully commercialized its first nutraceutical product, Cadalmin™ green mussel extract, to combat joint pain and rheumatoid arthritis.

Hypersensitivity to muscle proteins is one of the major concerns in all developed and industrialized countries. Shellfish comprising crustaceans and molluscs are highly nutritious. However, they are also one of the major causes of food allergies among the coastal communities[14], making it mandatory for labelling allergens in European countries. All age groups, including children and adults, irrespective of geographical location, are sensitive to the proteins present in these animals. The immune system produces antibodies against certain proteins in shellfish meat, thereby causing hypersensitive reactions in the body. The recent increase in shellfish consumption has led to the enormous growth and production of crustaceans and molluscs[15]. The drastic change in the consumption of these animals has also led to adverse health issues and abnormal reactions of the immune system, leading to severe anaphylactic reactions[16,17]. The hypersensitivity reaction of shellfish meat does not stop with consumption but also propagates as aerosol particles in the processing industries[18–20]. Both the edible and non-edible tissues of shellfish harbour potential allergens[15,21]. The highly heat-stable, water-soluble and low-molecular-weight tropomyosin appears to be one of the major allergens of shellfish meat, apart from arginine kinase, actin, troponin-C and sarcoplasmic-binding protein[14,22–24]. Due to the similarity of amino acid sequences of tropomyosin, an individual's hypersensitive reaction to one species of crustacean/mollusc cross-reacts with other species of shellfish and other non-dietary invertebrates[25,26]. The identification of potential allergens and their subsequent characterization traditionally require large amounts of base material. The next-generation integrated transcriptome sequencing approach effectively identifies all the potential allergenic proteins expressed in an animal. During the present study on the *de novo* assembly of transcripts of *P. viridis*, observations were made on the transcriptome-level candidate allergens and epitopes based on the sequence data generated using different tissues, which might play a role in shellfish allergy and also help understand the species-specific nature of allergy.

Adult healthy, wild *P. viridis* in the size range 110–220 g were collected from the second largest brackishwater lagoon, viz. Pulicat Lake, Tamil Nadu, India. Tissues (adductor muscle, digestive gland, gills, gonad and mantle) were dissected, and the samples were snap-frozen in liquid nitrogen and stored at –80°C until further processing.

The total RNA was extracted from all the tissues of green mussels using TRIzol (Qiagen RNeasy mini kit, USA), according to the manufacturer's instructions. The integrity and concentration of RNA were checked and assessed using Nanodrop2000 (Thermo Scientific, USA), Qubit (Thermo

*For correspondence. (e-mail: vetvsr@gmail.com)

Scientific) and Tapestation (Agilent, USA). mRNA isolation, fragmentation and priming were carried out using 500 ng of total RNA. The primed RNA was used for both first and second-strand synthesis. The double-stranded cDNA was purified using JetSeq beads (Bioline 68031) and ligated to Illumina multiplex barcode adapters according to NEBNext® Ultra™ II directional RNA library preparation protocol, followed by second strand excision utilizing USER enzyme at 37°C for 15 min. The adapter-ligated fragments were enriched by purification with JetSeq beads followed by 12 cycles of indexing. The sequencing libraries were quantified and checked for concentration using the Qubit fluorometer (Thermo Fisher Scientific, USA). Analysis of the distribution of fragment size was carried out on an Agilent 2200 Tapestation. The sequencing libraries were prepared with Illumina Compatible NEBNext Ultra II directional RNA library prep kit (New England Biolabs, USA). The libraries were sequenced following the manufacturer's instructions on the Illumina HiSeq X Ten sequencer (Illumina, USA) and NextSeq 550 for 150 cycles. In total, 137 million reads of 150 bp length in paired-end mode consisting of 40.7 Gb data were generated on Illumina HiSeq4000.

The quality of the raw, paired-end reads was examined using FastQC v0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The reads were trimmed for poor-quality bases with Trimmomatic v0.39 using the following parameters[27]: LEADING:3, TRAILING:22, SLIDINGWINDOW:4:22 and MINLEN:100. The good-quality reads were then assembled to transcripts *de novo* using Trinity v2.12.0 (ref. 28). The assembly contained 390,415 transcripts with an N50 length of 1443 bases. About 86.38% of paired-end reads could be aligned back to the transcripts with bowtie2 v2.3.4.3 (ref. 29), thus validating the quality of the transcript assembly. The candidate coding regions in the assembled transcripts were predicted with TransDecoder v5.5.0 (https://github.com/TransDecoder) to identify 71,315 transcripts. CDHIT v4.6 was used to cluster similar transcripts (>95% similarity)[30], which resulted in a final set of 39,970 non-redundant unigenes (N50: 1632 bases).

The final transcript assembly containing 39,970 unigenes was checked with BUSCO single-copy orthologs (eukaryota_odb10, 2020-09-10, 70 genomes, 255 BUSCOs) in transcriptome mode for completeness assessment. Of the 255 orthologs, 249 (97.65%) were completely present, four (1.57%) were fragmented and two (0.78%) were missing in the transcript assembly of *P. viridis* (Figure 1 *a*). The same transcript assembly was found to have 9.12% missing orthologs when checked with the Mollusca lineage (mollusca_odb10, 2020-08-05, 7 genomes, 5295 BUSCOs). The transcript annotation was performed using OmicsBox v 2.0.24 (ref. 31). Briefly, a blastx search using the Eukaryota subset of the non-redundant protein database of GenBank and GO mapping were performed employing Blast2GO methodology. Next, the transcripts were searched separately for GO terms using InterProScan and EggNOG mapper[32] as implemented in OmicsBox v2.0.36 (ref. 31). Thereafter, the GO terms were merged and final annotations were obtained. About 24,109 (60.32%) transcripts could be annotated and 6867 were left without any blast hits. The five top hit species during blast search were scallops (*Pectan maximus* and *Mizuhopecten yessoensis*), oysters (*Crassostrea gigas* and *Crassostrea virginica*) and mussels (*Mytilus galloprovincialis*) of phylum Mollusca. As evidenced by the GO terms, the cellular protein metabolic process is the major biological process, the metal ion-binding is the major metabolic function, and the intracellular membrane-bound organelle is the major cellular component for the transcripts in the assembly. The majority of the enzyme code classes observed in the *P. viridis* transcript assembly were hydrolases followed by transferases and translocases. For 99.99% of the transcripts, the InterProScan (IPS) search provided functional information about domain/family/repeat/site. The unigenes were also searched for simple sequence repeats (SSRs) using the MISAv1.0 tool. About 1333 SSRs were identified in 1162 unigenes, of which A/T, AG/CT and ATC/ATG were the predominant mono-, di- and tri-nucleotide SSRs respectively, in *P. viridis* unigenes.

Identification of candidate hypersensitive transcript sequences in the transcriptome of *P. viridis* was done using two public data resources, one containing the known hypersensitive protein sequences and the other containing the epitope peptide sequences.

About 1043 hypersensitive sequences were sourced from the allergens database www.allergen.org (accessed on 8 June 2021), the official site for systematic allergen nomenclature approved by the World Health Organization, Geneva and International Union of Immunological Societies, Germany. A blast database was generated with these allergen sequences, and the unigenes of *P. viridis* transcriptome were queried through a blastx search. The unigenes that gave significant blast hits (e-05) having an alignment covering >80 amino acid of query, >90% subject length and >50% similarity were considered as potential candidate allergens. Overall, the study identified 318 unigenes as candidate allergens in *P. viridis* transcript assembly exhibiting similarity to 116 unique allergens. The identified candidate allergens were predominantly airborne and food allergens (Figure 1 *b*). The major shellfish allergens, viz. arginine kinase, myosin light chain, a sarcoplasmic calcium-binding protein, tropomyosin, troponin and triose phosphate isomerase, are among the identified candidate allergens. It was observed that the major GO terms obtained for the candidate allergens were similar to those observed for all unigenes. Whereas, the major enzyme code classes observed for candidate allergens differed from those of unigenes with a particular increase in isomerases. The EF-hand domain, EF-hand domain pair, filamin and EF-hand-binding site were the predominant IPS domain, family, repeat and site respectively, among the candidate allergens.

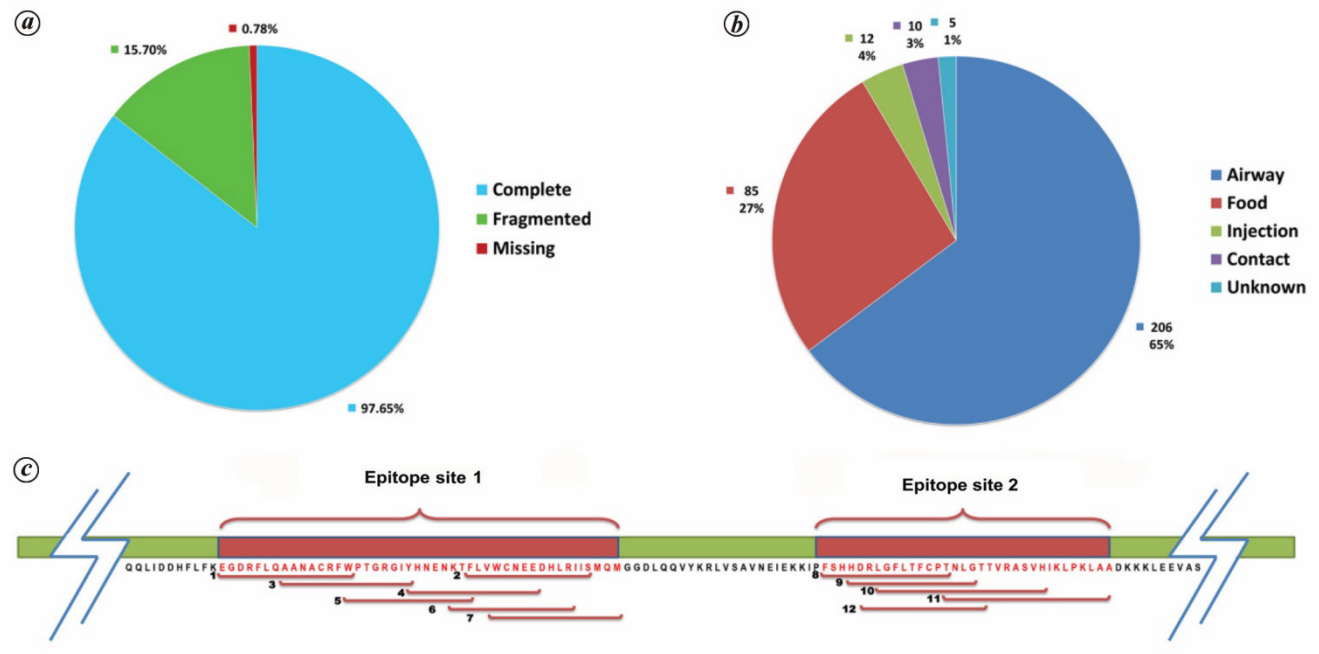

Figure 1. BUSCO analysis, allergen identification and epitope analysis.

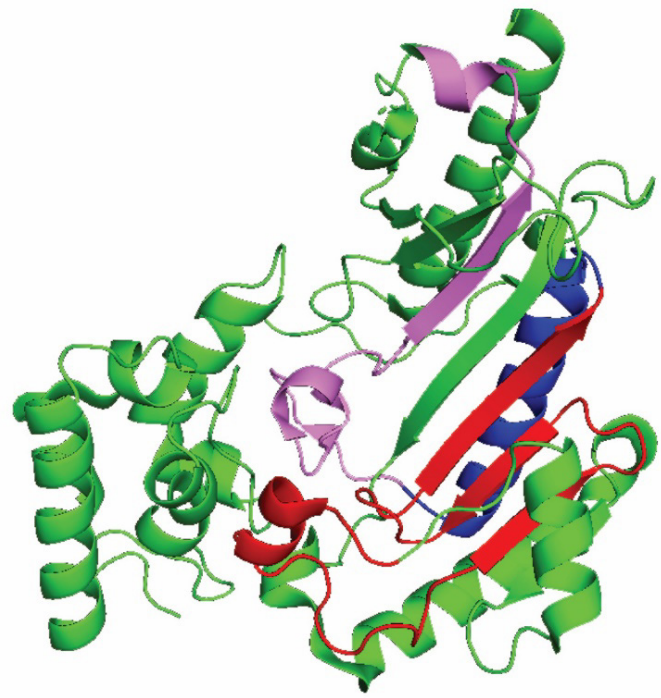

Figure 2. Structure of the arginine kinase of *Daphnia magna* as sourced from the RCSB Protein Data Bank (6KY2). The two epitope hotspots are shown in pink and red respectively which are separated by an α-helix shown in blue colour.

For epitope sequence-based analysis, 21,839 linear peptide epitope sequences associated with allergy were extracted from the Immune Epitope Database and Analysis Resource (https://www.iedb.org/home_v3.php; accessed on 23 July 2021). The epitopes were clustered using CDHIT v4.6 (ref. 30) with a sequence identity threshold of 1.0 to remove redundant entries, which resulted in 14,185 non-redundant epitopes. A tblastn (ref) search was performed for the epitopes against the *P. viridis* unigene sequences to find matching entries. Twenty-two epitopes gave matching hits to

10 different unigenes when a 100% cut-off was imposed for query coverage and sequence identity. These unigenes include well-known allergens like arginine kinase, actin, cyclophilin and tropomyosin, in addition to certain novel candidate allergens like Ran protein and a filamin A-like protein. A particular observation made for arginine kinase is the presence of multiple epitope sites ($n = 12$). Here, we observed the localization of multiple epitope sites at two different regions towards the *C*-terminal of the arginine kinase protein (Figure 1 *c*).

The identified epitope sites were further compared with the structure of homologous arginine kinase of *Daphnia magna* obtained from RCSB Protein Data Bank (accession no. 6KY2) using PyMOL v2.5.2. The two epitope regions were found spanning four of the eight antiparallel β-sheets at the *C*-terminal of the protein and separated by an α-helix (Figure 2). The present study mapped hotspots of epitopes for arginine kinase in *P. viridis*, which may be a factor in deciding the allergenicity potential of allergens. Similar hotspots might exist for different allergens in different species, which needs to be explored.

In this study, we have identified 318 unigenes using the RNA sequence data of a single specimen as a candidate allergen. The publicly available RNA sequence datasets of *P. viridis* available at GenBank were explored to validate the correctness of candidate allergens. Twenty-three SRA datasets were downloaded from GenBank to identify the evidence for the candidate allergens. Briefly, the quality of SRA datasets was assessed using the FastQC tool and quality trimming was performed with Trimmomatic v0.39 (ref. 27) to remove adaptor contamination and retain high-quality reads with an average quality of 25. The predicted allergen sequences were considered reference and the good-quality reads of SRA datasets were mapped to this reference using Burrows–Wheeler Aligner tool (bwa v 0.7.17)[33] to generate a sam file. The generated sam file was converted to a bam file using SAMtools v1.11 (ref. 34), and Qualimap v.2.2.2 (ref. 35) was used to evaluate the mean coverage. A mean coverage >1 was considered as evidence for the presence of candidate allergen in that dataset.

We observed that for 300 out of 318 unigenes, evidence for candidate allergens exists from at least one dataset. For 295 unigenes (92.7%), evidence exists from two more samples than the one used in this study. All the 23 studied samples provided evidence for 131 unigenes. It is to be mentioned that the tissue type, sequence coverage and sequence depth of SRA datasets can influence the validation process. Nevertheless, excluding 18 unigenes, evidence could be established for candidate allergens with the help of 23 public RNA sequence datasets originating from multiple individuals.

The present study describes the transcriptome of *P. viridis* based on the sequence data generated using five tissues. Utilizing the public datasets, transcriptome-level candidate allergens and epitopes were observed and identified that might play a role in the hypersensitive reaction to shellfish proteins. Particularly, we observed the existence of epitope hotspots in an important protein, viz. arginine kinase. The unigenes identified for *P. viridis* would be a valuable resource for future functional studies.

This study focused on the merit of RNA sequence datasets in identifying transcriptome-level candidate allergens, which would further help understand the species-specific nature of allergy. The datasets, supplementary figures and tables generated in this study can be found online. The repository and accessions can be found at: https://www.ncbi.nlm.nih.gov/, Bioproject PRJNA660597; https://doi.org/10.6084/m9.figshare.20296671.v3.

Conflict of interest: The authors declare that they have no conflict of interest.

1. Shinoj, P. *et al.*, Green mussel (*Perna viridis* L.) farming in India: an analysis of major growth milestones, recent decline due to disease incidence, and prospects for revival. *Aquacult. Int*., 2021, **29**, 1813–1828.
2. Khan, M. A. A., Assim, Z. B. and Ismail, N., Population dynamics of the green-lipped mussel, *Perna viridis* from the offshore waters of Naf River coast, Bangladesh. *Chiang Mai J. Sci*., 2010, **37**(2), 344–354.
3. Guo, X., Ford, S. E. and Zhang, F., Molluscan aquaculture in China. *J. Shellfish Res.*, 1999, **18**, 19–31.
4. Rajagopal, S., Venugopalan, V. P., Van der Velde, G. and Jenner, H. A., Greening of the coasts: a review of the *Perna viridis* success story. *Aquat. Ecol*., 2006, **40**, 273–297.
5. Laxmilatha, P., A review of the green mussel, *Perna viridis* fishery of southwest coast of India. *Int. J. Mar. Sci*., 2013, **3**, 408–416.
6. Hickman, R. W., Mussel cultivation. In *The Mussel Mytilus: Ecology, Physiology, Genetics and Culture* (ed. Gosling, E.), Elsevier, Amsterdam, The Netherlands, 1992, pp. 465–511.
7. Tan, K. S. and Ransangan, J., Feeding behaviour of green-lipped mussels, *Perna viridis* in Marudu Bay, Malaysia. *Aquacult. Res*., 2017, **48**(3), 1216–1231.
8. Mohammed, K. S., Mussel farming and its potential in India. In *Advances in Marine and Brackishwater Aquaculture* (ed. Perumal, S.), Springer, New Delhi, 2015, pp. 187–193.
9. Qasim, S. Z., Parulekar, A. H., Harkantra, S. N., Ansari, Z. A. and Nair, A., Aquaculture green mussel *Mytilus viridis* L. cultivation on ropes from floating ropes. *Indian J. Mar. Sci*., 1977, **4**, 189–197.
10. Pillai, V. N. *et al.* (eds), *Bivalve Mariculture in India (Pearl Oyster, Edible Mussel and Oyster): A Success Story in Coastal Ecosystem Development*, Asia-Pacific Association of Agricultural Research Institutions, FAO Regional Office for Asia and the Pacific, Bangkok, Thailand, 2001.
11. Kripa, V. and Mohammed, K. S., Green mussel, *Perna viridis*, farming in Kerala, India – technology diffusion process and socio-economic impacts. *J. World Aquacult. Soc.*, 2008, **39**(5), 612–624.
12. Asokan, P. K., Vipinkumar, V. P., Appukuttan, K. K., Surendranathan, V. G. and Sivadasan, M. P., Mussel culture in backwaters of Kasargod district, Kerala. *Mar. Fish Inf. Ser*., 2001, **169**, 9–11.
13. Mohammed, K. S. *et al.*, Guidance for good mussel farming practices in India based on a case study from Kerala. *Mar. Fish Policy Ser.*, 2019, **10**, 64.
14. Emoto, A., Ishizaki, S. and Shiomi, K., Tropomyosins in gastropods and bivalves: Identification as major allergens and amino acid sequence features. *Food Chem*., 2009, **114**(2), 634–641.
15. Lopata, A. L., Kleine-Tebbe, J. and Kamath, S. D., Allergens and molecular diagnostics of shellfish allergy: part 22 of the series Molecular Allergology. *Allergo J. Int*., 2016, **25**(7), 210–218.

16. Tham, E. H. *et al.*, Epinephrine auto-injector prescriptions as a reflection of the pattern of anaphylaxis in an Asian population. *Allergy Asthma Proc*., 2008, **29**(2), 211–215.
17. Matricardi, P. M. *et al.*, EAACI molecular allergology user's guide. *Pediatr. Allergy Immunol*., 2016, **27**(23), 1–250.
18. Jeebhay, M. F., Robins, T. G., Lehrer, S. B. and Lopata, A. L., Occupational seafood allergy: a review. *Occup. Environ. Med*., 2001, **58**(9), 553–562.
19. Bønløkke, J. H., Gautrin, D., Sigsgaard, T., Lehrer, S. B., Maghni, K. and Cartier, A., Snow crab allergy and asthma among Greenlandic workers – a pilot study. *Int. J. Circumpolar Health*, 2012, **71**, 19126.
20. Kamath, S. D., Thomassen, M. R., Saptarshi, S. R., Nguyen, H. M., Aasmoe, L., Bang, B. E. and Lopata, A. L., Molecular and immunological approaches in quantifying the air-borne food allergen tropomyosin in crab processing facilities. *Int. J. Hyg. Environ. Health*, 2014, **217**(7), 740–750.
21. Sun, S. and Lopata, A., The role of shellfish proteases in allergic diseases and inflammation. *Curr. Allergy Clin. Immunol*., 2010, **23**, 174–179.
22. Reese, G., Ayuso, R. and Lehrer, S. B., Tropomyosin: an invertebrate pan-allergen. *Int. Arch. Allergy Immunol*., 1999, **119**(4), 247–258.
23. Gámez, C. *et al.*, Tropomyosin IgE-positive results are a good predictor of shrimp allergy. *Allergy*, 2011, **66**(10), 1375–1383.
24. Pascal, M. *et al.*, Molecular diagnosis of shrimp allergy: efficiency of several allergens to predict clinical reactivity. *J. Allergy Clin. Immunol*., 2015, **3**(4), 521–529.
25. Torres Borrego, J., Martínez Cuevas, J. F. and Tejero García, J., Reactividad cruzada entre pescados y mariscos [Cross reactivity between fish and shellfish]. *Allergol. Immunopathol*., 2003, **31**(3), 146–151.
26. Zhang, Y., Matsuo, H. and Morita, E., Cross-reactivity among shrimp, crab and scallops in a patient with a seafood allergy. *J. Dermatol*., 2006, **33**(3), 174–177.
27. Bolger, A. M., Marc, L. and Bjoern, U., Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 2014, **30**(15), 2114–2120.
28. Grabherr, M. G. *et al.*, Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnol*., 2011, **29**(7), 644–652.
29. Langmead, B. and Salzberg, S. L., Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 2012, **9**(4), 357–359.
30. Li, W. and Godzik, A., Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 2006, **22**, 1658–1659.
31. OmicsBox – Bioinformatics made easy. BioBam Bioinformatics (version 2.0.24). 3 March 2019; www.biobam.com/omicsbox
32. Huerta-Cepas, J. *et al.*, eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res*., 2019, **47**(D1), D309–D314.
33. Li, H. and Durbin, R., Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatic*s, 2009, **25**, 1754–1760.
34. Li, H. *et al.*, 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 2009, **25**(16), 2078–2079.
35. Okonechnikov, K., Conesa, A. and García-Alcalde, F., Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, 2016, **32**(2), 292–294.

ACKNOWLEDGEMENTS. We thank the Director, ICAR-Central Marine Fisheries Research Institute (ICAR-CMFRI), Chennai for providing the necessary facilities for this study. We also thank the Director, ICAR-Central Institute of Brackishwater Aquaculture, Chennai for support. The laboratory assistance rendered by T. Balaraman (ICAR-CMFRI) is acknowledged. This study received institutional grant from ICAR-CMFRI.

Received 20 April 2023; revised accepted 26 August 2023

doi: 10.18520/cs/v125/i9/1008-1012