Journal of Genetics, Vol. 97, No. 4, September 2018, pp. 995–999
https://doi.org/10.1007/s12041-018-0967-z

RESEARCH NOTE

De novo sequencing of the Antarctic krill (*Euphausia superba*) transcriptome to identify functional genes and molecular markers

CHUNYAN MA[1], HONGYU MA[1,2], GUODONG XU[1], CHUNLEI FENG[1], LINGBO MA[1]* and LUMIN WANG[1]*

[1]*East China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Shanghai 200090, People's Republic of China*
[2]*Guangdong Provincial Key Laboratory of Marine Biology, Shantou University, Shantou 515063, People's Republic of China*
*For correspondence. E-mail: Lingbo Ma, malingbo@vip.sina.com; Lumin Wang, lmwang@eastfishery.ac.cn.

Received 5 September 2017; revised 9 November 2017; accepted 6 December 2017; published online 7 August 2018

Abstract. To provide massive genetic resources for the Antarctic krill (*Euphausia superba*), we sequenced and analysed the transcriptome by using high-throughput Illumina paired-end sequencing technology. A total of 77.1 million clean reads representing ~11.0 Gb data were generated. The average length of these reads was 142 bp. *De novo* assembly yielded 125,211 transcripts with a N50 of 690 bp. Further analysis produced 106,250 unigenes, of which 31,683 were annotated based on protein homology searches against protein databases. Gene ontology analysis showed that ion binding, organic substance, metabolic process, and cell part were the most abundantly used terms in molecular function, biological process and cellular component categories, respectively. In addition, 3067 unigenes were mapped onto 311 signal pathways by the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis. Finally, 15,224 simple sequence repeats were identified from 13,535 transcripts, and 103,593 single-nucleotide polymorphisms were found from 21.6% of total transcripts. These genetic resources obtained in this study forms a good foundation for investigating gene function, and evaluating population genetic diversity for this important Southern Ocean fisheries resource, *E. superba*.

Keywords. transcriptome resource; *de novo* assembly; gene annotation; molecular markers; *Euphausia superba*.

Introduction

Antarctic krill (*Euphausia superba*), a pelagic crustacean species, is one of the most essential ecological and fishery resources in the Southern Ocean. It is a plankton and swarming species, and has a key position in oceanic food webs by serving as direct link between primary producers and apex predators. *E. superba* has a circumpolar distribution and always keeps in high density ranging from 10,000 to 30,000 individuals per m^2 (Candeias *et al.* 2014). As a keystone species, the first major study on *E. superba* was conducted since the Discovery Expeditions (1901–1904) (Jia *et al.* 2014). Until now, studies on *E. superba* were mainly focussed on life history (Jia *et al.* 2014), physiological and metabolic activity (Meyer 2012; Auerswald *et al.* 2015), and variations of biomass resources (Atkinson *et al.* 2004; Fielding *et al.* 2014; Shi *et al.* 2014). As sequencing the whole genome in *E. superba* was considered to be difficult (Jeffery 2012), transcriptome sequencing was thought to be an alternative way for mining of genetic resources (Ma *et al.* 2014; Koh *et al.* 2015). Currently, transcriptome resources have been isolated in *E. superba* using 454 next-generation sequencing technology (Clark *et al.* 2011; Martins *et al.* 2015; Meyer *et al.* 2015). Recently, an online, open resource of *E. superba* transcriptome database, Superba SE, was described by Hunt *et al.* (2017) and KrillDB was developed (Sales *et al.* 2017) for the purpose of free access to annotation information for users. However, the availability of molecular data concerning function genes, microsatellites and single-nucleotide polymorphism (SNP) in *E. superba* is still limited, which has severely hampered a better understanding of the genomic and genetic biology of *E. superba*.

Here, we sequenced and analysed the transcriptome of *E. superba* in great detail by using high-throughput Illumina paired-end sequencing technology, and massively

explored function genes, microsatellite markers and SNP loci. This work forms a good foundation for investigation of molecular mechanisms, population genetic diversity and conservation genetics in *E. superba*.

Materials and methods

The specimens of *E. superba* were collected from Southern Ocean (63°6′S, 58°45′W) in February 2013. Total RNA was isolated from six whole animals separately, and then pooled together for mRNA purification using magnetic beads with oligo (dT). After fragmentation into short pieces, mRNAs were reverse transcribed into double-strand cDNA. The cDNA library was finally constructed after end repairing, dA tailing, adapter ligation and products enrichment. RNA-seq was carried out using paired-end sequencing (2×150 bp) on an Illumina Hiseq 3000 platform.

Raw reads were first trimmed and removed adapters and low-quality sequences by using Trim Galore software, then they were analysed and quality controlled using Fast QC software. A *de novo* assembly of the clean and high-quality reads was conducted using the Trinity platform. The efficiency of the assembly was estimated by transcripts length ranges, N50 and the total number of transcripts. The cut-off for minimum transcript length was set at every 100 bp length ranging from 200 to 1000 bp. Unigene was defined as the longest transcript among multiple transcript isoforms or paralogues. The open reading frame (ORF) or coding sequence (CDS) of unigene was predicted using Trans Decoder program by Trinity software. The CDS was translated into amino acids to obtain potential protein sequences of the unigene. Further, unigenes were annotated by searching against UniProt, nonredundant (Nr), Nt, Interpro and clusters of orthologous groups for eukaryotic complete genomes (KOG) databases using BLASTX algorithm. Meanwhile, gene ontology (GO) analysis (level 2) was performed to annotate and classify unigenes by BLASTX against UniProt database. Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis was carried out to classify unigenes into specific signal pathways.

Finally, microsatellite loci were detected from transcripts using MISA software, and SNPs loci were identified by remapping the clean reads back to transcripts using Bowtie software with a minimum variation depth of 10-fold.

Results and discussion

In this study, the high-throughput transcriptome sequencing of *E. superba* generated ~77.9 million raw reads, representing 11.8 Gb data. The raw data were submitted to the NCBI Short Read Archive database under the accession number SRR3089571. After removing low-quality reads and trimming adapters, 77.1 million clean reads corresponding to 11.0 Gb data remained, with an average length of 142 bp. The total high-quality data generated in this study was much higher than those reported in recently published references using 454 sequencing technology (Clark *et al.* 2011; Meyer *et al.* 2015). Compared with the two databases of *E. superba* (Hunt *et al.* 2017; Sales *et al.* 2017), the total high-quality data generated in this study was higher than that by Hunt *et al.* (6.98 million) and lower than Sales *et al*. (368 million). Moreover, the average length of the reads in this study was higher than those of Hunt *et al.* (2017). The GC rate of clean reads was 38.8%, which is much lower than those detected in *Vaccinium macrocarpon* (~47%, Sun *et al.* 2015) and in *Calotropis gigantea* (~43%, Muriira *et al.* 2015). A total of 125,211 transcripts were produced by the *de novo* assembly, with a mean length of 568 bp and a N50 of 690 bp. The average length and N50 were both lower than those in other studies (Hunt *et al.* 2017; Sales *et al.* 2017). The correlation between mean length / N50 of transcripts and the minimum cut-off of transcript length is shown in table 1. The transcripts were clustered and generated 106,250 unigenes with an average length and N50 being 528 and 596 bp, respectively. The prediction analysis

Table 1. Efficiency of *de novo* assembly for transcripts in *E. superba*.

Minimum transcript length cut-off	N50	Mean length	Total number	Total length	Fraction of total length (%)
200	690	568	125,211	71,143,610	100
300	897	736	81,270	59,810,093	84
400	1143	955	52,137	49,813,966	70
500	1340	1153	37,602	43,358,526	61
600	1497	1328	29,193	38,770,847	55
700	1631	1483	23,792	35,282,449	50
800	1756	1629	19,865	32,355,699	45
900	1868	1766	16,892	29,836,839	42
1000	1974	1894	14,614	27,679,139	39

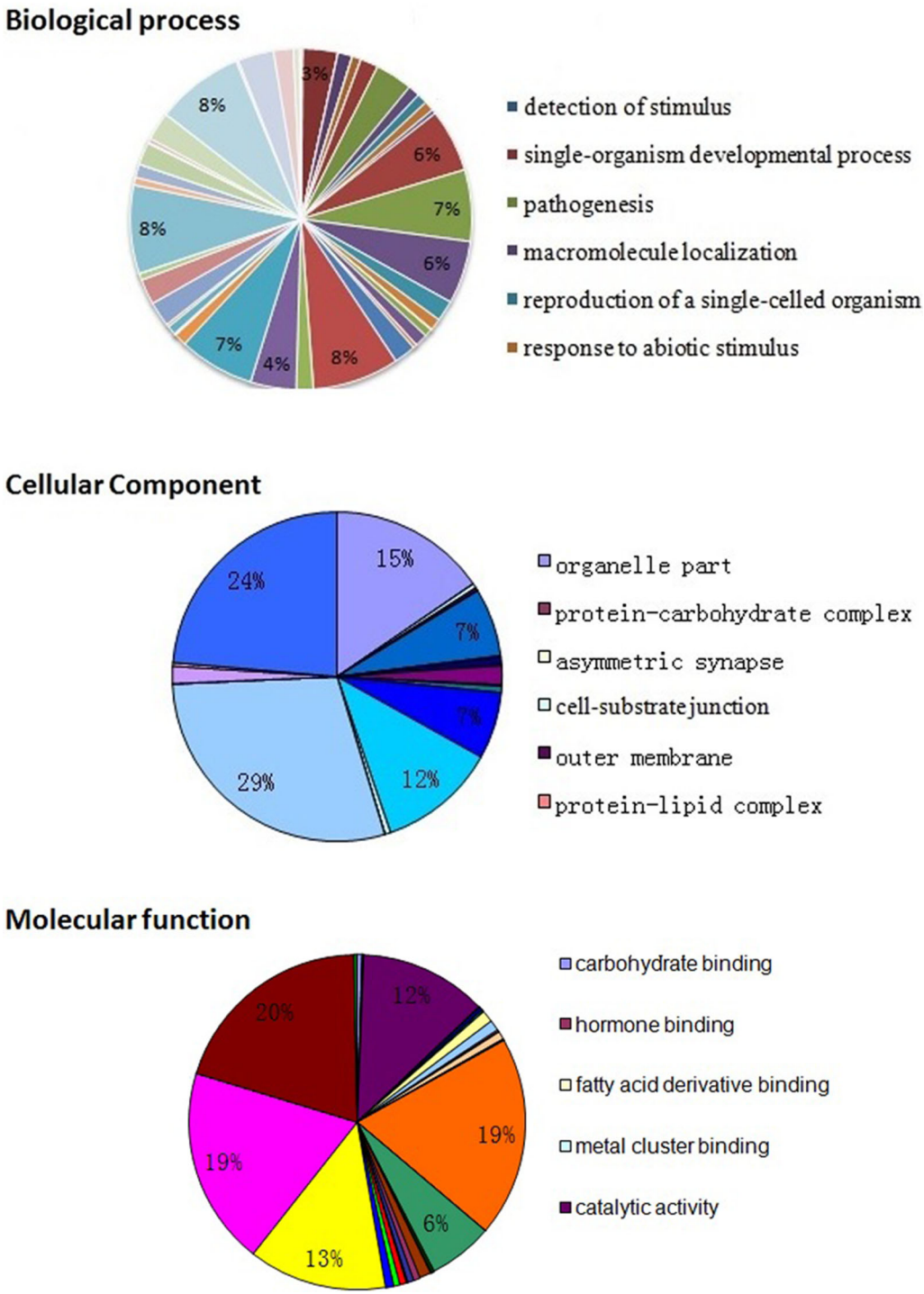

Figure 1. GO classification of the unigenes in *E. superba*.

for unigene ORF showed that 4267 unigenes have complete protein sequences, 2345 unigenes have incomplete protein sequences but containing start codons, and 5974 unigenes have incomplete protein sequences but containing end codons.

To identify and understand as many function genes as possible, seven protein databases including UniProt, Nr, Nt, GO, Interpro, KOG and KEGG were employed to annotate unigenes. Among 106,250 unigenes, 31,683 (29.8%) were successfully annotated by aligning against the above-mentioned protein databases. The number of annotated unigenes in this study was much higher than that in the study on *E. superba* (Clark *et al.* 2011; Martins *et al.* 2015; Meyer *et al.* 2015) but lower than that obtained

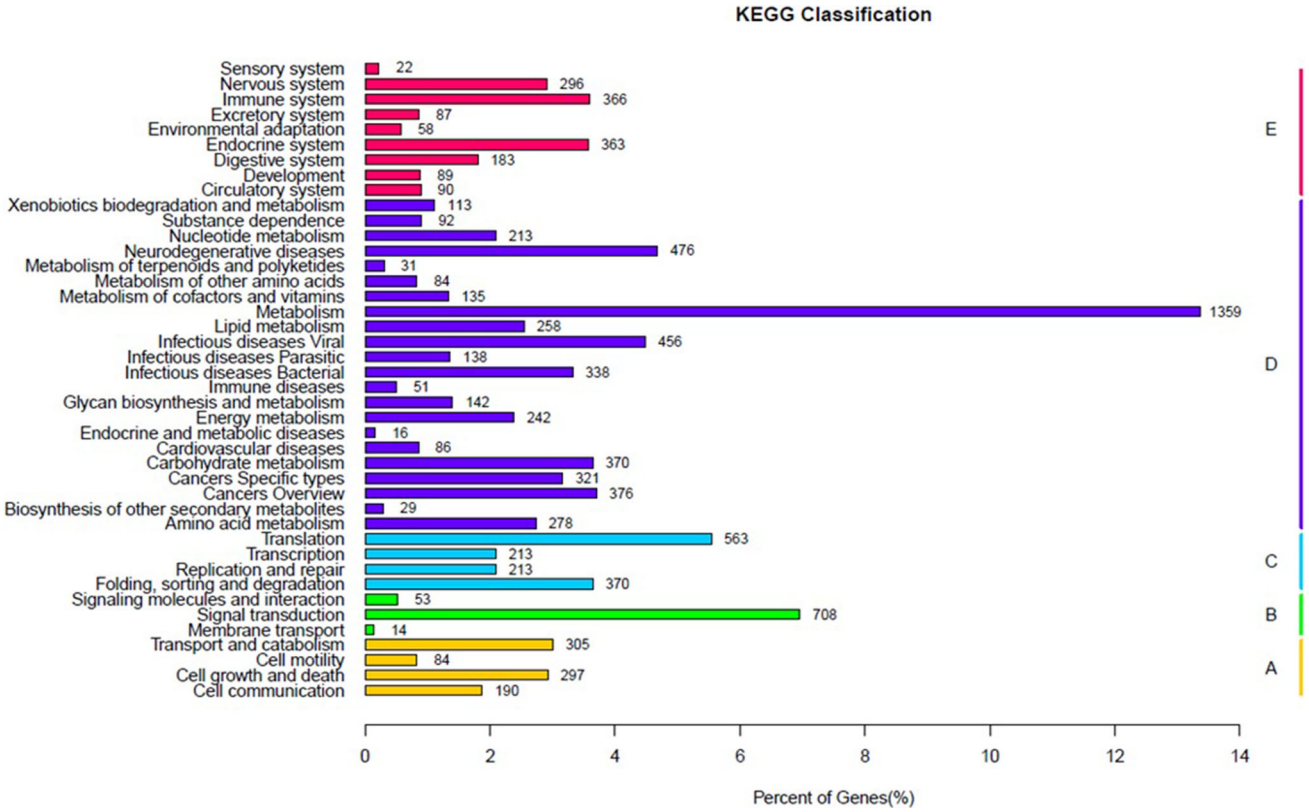

Figure 2. KEGG pathway analysis of unigenes in *E. superba*.

in the two recent studies (Hunt *et al.* 2017; Sales *et al.* 2017). The annotation rate in this study is similar to that observed in *Patinopecten yessoensis* (27.9%, Hou *et al.* 2011), higher than that found in *Carassius auratus* (17.4%, Liao *et al.* 2013), but lower than that detected in *Apostichopus japonicas* (39.1%, Du *et al.* 2012). A total of 25,333 unigenes had significant matches in UniProt protein database, 29,138 had significant matches in Nr database, and 7088 showed significant hits to target proteins in Nt database.

GO annotation of unigenes showed that a total of 25,029 (23.6%) unigenes could be assigned to at least one term (level 2), which was lower than the number annotated in the former study (Hunt *et al.* 2017) (figure 1). Totally, 14,893 unigenes were assigned in biological process, of which organic substance metabolic process was the largest term with 19,369 unigenes. A total of 23,018 unigenes were classified in cellular component process of which the cell part was the dominant term with 22,564 unigenes. Molecular function process contained 21,091 unigenes of which 15,387 were present in the biggest term ion binding. Moreover, a total of 16,482 unigenes were annotated from Interpro database with 14,462 functional domains. Zinc-finger C2H2 type was found to be the most predominant conserved domain, followed by ankyrin repeat domain. Further, based on predicted protein sequence homology, 12,748 unigenes were classified into 25 KOG categories. The most-dominant group was 'general function prediction only' ($n = 4986$), followed by 'transcription' ($n = 3846$), 'signal transduction mechanisms' ($n = 2119$), and 'posttranslational modification, protein turnover, chaperones' ($n = 1429$). 'Nuclear structure' ($n = 28$) was the smallest group in this study. KEGG analysis can further help to understand specific biology processes, gene functions and interactions at the transcriptome level. In this study, a total of 3067 unigenes were classified into 311 KEGG pathways. The most-represented pathways were metabolism, signal transduction, translation and neurodegenerative diseases, and they contained 1359, 708, 563 and 476 unigenes, respectively (figure 2).

Transcriptome has been considered as an important resource for rapid and effective discovery of genetic markers (Feng *et al.* 2014; Sun *et al.* 2015). In this study, a total of 15,224 microsatellites (including 8358 mononucleotides) were identified from 13,535 transcripts. Microsatellite repeat motifs ranged from 1 to 6 bp, and contained 5734 perfect loci and 1132 compound loci. There were 1477 transcripts which contained more than one microsatellite. Further investigation indicated that among perfect loci, dinucleotide repeats were the most abundant ($n = 3789$), followed by trinucleotide repeats ($n = 1568$), and tetranucleotide repeats ($n = 257$). Further, 103,593 potential SNPs were identified from 27,008 (21.6%) transcripts, with a mean density of one SNP every 2634 bp. The number of SNPs identified in this study was much more than

that reported in previous studies on the same krill species (17,776 SNPs, Clark *et al.* 2011). These microsatellite and SNP loci located inside or aside functional genes would be a very useful genetic tool for investigation of the population genetic structure and conservation genetics in *E. superba*.

In conclusion, we sequenced the Antarctic krill (*E. superba*) transcriptome by using high-throughput Illumina paired-end sequencing technology that provided massive genetic resources, including 106,250 unigenes, 15,224 microsatellite markers and 103,593 potential SNPs. This work forms a good foundation for functional gene analysis, population genetic diversity evaluation, and conservation genetics studies in *E. superba* and related krills.

Acknowledgements

This work was supported by the National Science and Technology Support Plan (no. 2013BAD13B03), the National Programme for Support of Top-notch Young Professionals, and the National Natural Science Foundation of China (no. 41406190).

References

Atkinson A., Siegel V., Pakhomov E. and Rothery P. 2004 Long-term decline in krill stock and increase in salps within the Southern Ocean. *Nature* **432**, 100–103.

Auerswald L., Meyer B., Teschke M., Hagen W. and Kawaguchi S. 2015 Physiological response of adult Antarctic krill, *Euphausia superba*, to long-term starvation. *Polar Biol.* **38**, 763–780.

Candeias R., Teixeira S., Duarte C. M. and Pearson G. A. 2014 Characterization of polymorphic microsatellite loci in the Antarctic krill *Euphausia superba*. *BMC Res. Notes* **7**, 73.

Clark M. S., Thorne M. A. S., Toullec J. Y., Meng Y., Guan L. L., Peck L. S. *et al.* 2011 Antarctic krill 454 pyrosequencing reveals chaperone and stress transcriptome. *PLoS One* **6**, e15919.

Du H., Bao Z., Hou R., Wang S., Su H., Yan J. *et al.* 2012 Transcriptome sequencing and characterization for the sea cucumber *Apostichopus japonicus* (Selenka, 1867). *PLoS One* **7**, e33311.

Feng N., Ma H., Ma C., Xu Z., Li S., Jiang W. *et al.* 2014 Characterization of 40 single nucleotide polymorphism (SNP) via T_m-shift assay in the mud crab (*Scylla paramamosain*). *Mol. Biol. Rep.* **41**, 5467–5471.

Fielding S., Watkins J. L., Trathan P. N., Enderlein P., Waluda C. M., Stowasser G. *et al.* 2014 Inter annual variability in Antarctic krill (*Euphausia superba*) density at South Georgia, Southern Ocean: 1997–2013. *ICES J. Mar. Sci.* **71**, 2578–2588.

Hou R., Bao Z., Wang S., Su H., Li Y., Du H. *et al.* 2011 Transcriptome sequencing and de novo analysis for yesso scallop (*Patinopecten yessoensis*) using 454 GS FLX. *PLoS One* **6**, e21560.

Hunt B. J., Özkaya Ö., Davies N. J., Davies N. J., Gaten E., Seear P. *et al.* 2017 The *Euphausia superba* transcriptome database, SuperbaSE: an online, open resource for researchers. *Ecol. Evol.* **7**, 6060–6077.

Jeffery N. W. 2012 The first genome size estimates for six species of krill (Malacostraca, Euphausiidae): large genomes at the north and south poles. *Polar Biol.* **35**, 959–962.

Jia Z., Virtue P., Swadling K. M. and Kawaguchi S. 2014 A photographic documentation of the development of Antarctic krill (*Euphausia superba*) from egg to early juvenile. *Polar Biol.* **37**, 165–179.

Koh H. Y., Lee J. H., Han S. J., Park H., Shin S. C. and Lee S. G. 2015 A transcriptomic analysis of the response of the arctic pteropod *Limacina helicina* to carbon dioxide-driven seawater acidification. *Polar Biol.* **38**, 1727–1740.

Liao X., Cheng L., Xu P., Lu G., Wachholtz M., Sun X. *et al.* 2013 Transcriptome analysis of crucian carp (*Carassius auratus*), an important aquaculture and hypoxia-tolerant species. *PLoS One* **8**, e62308.

Ma H. Y., Ma C. Y., Li S. J., Jiang W., Li X. C., Liu Y. X. *et al.* 2014 Transcriptome analysis of the mud crab (*Scylla paramamosain*) by 454 deep sequencing: assembly, annotation, and marker discovery. *PLoS One* **9**, e102668.

Martins M. J. F., Lago-Leston A., Anjos A., Duarte C. M., Agusti S., Serrao E. A. *et al.* 2015 A transcriptome resource for Antarctic krill (*Euphausia superba* Dana) exposed to short-term stress. *Mar. Genomics* **23**, 45–47.

Meyer B. 2012 The overwintering of Antarctic krill, *Euphausia superb*, from an ecophysiological perspective. *Polar Biol.* **35**, 15–37.

Meyer B., Martini P., Biscontin A., de Pitta C., Romualdi C., Teschke M. *et al.* 2015 Pyrosequencing and *de novo* assembly of Antarctic krill (*Euphausia superba*) transcriptome to study the adaptability of krill to climate-induced environmental changes. *Mol. Ecol. Resour.* **15**, 1460–1471.

Muriira N. G., Xu W., Muchugi A., Xu J. and Liu A. 2015 *De novo* sequencing and assembly analysis of transcriptome in the Sodom apple (*Calotropis gigantea*). *BMC Genomics* **16**, 723.

Sales G., Deagle B. E., Calura E., Martini P., Biscontin A., De Pittà C. *et al.* 2017 KrillDB: a de novo transcriptome database for the Antarctic krill (*Euphausia superba*). *PLoS One* **12**, e0171908.

Shi Y., Sun S., Li C. and Tao Z. 2014 Population distribution, structure and growth condition of Antarctic krill (*Euphausia superb* Dana) during the austral summer in the Southern Ocean. *Adv. Polar Sci.* **25**, 183–191.

Sun H., Liu Y., Gai Y., Geng J., Chen L., Liu H. *et al.* 2015 *De novo* sequencing and analysis of the cranberry fruit transcriptome to identify putative genes involved in flavonoid biosynthesis, transport and regulation. *BMC Genomics* **16**, 652.

Corresponding editor: SILVIA GARAGNA