

Research Article

Cite this article: Wang A, Cheng J, Xin Q, Sha Z, Hui M (2023). A first genome survey sequencing of alvinocaridid shrimp *Shinkaicaris leurokolos* in deep-sea hydrothermal vent environment. *Journal of the Marine Biological Association of the United Kingdom* **103**, e65, 1–8. <https://doi.org/10.1017/S0025315423000504>

Received: 24 April 2023

Revised: 24 June 2023

Accepted: 28 June 2023

Keywords:

genome survey; hydrothermal vent; microsatellite DNA; mitogenome; *Shinkaicaris leurokolos*

Corresponding authors:


Zhongli Sha;

Email: shazl@qdio.ac.cn;

Min Hui;

Email: minhui@qdio.ac.cn

A first genome survey sequencing of alvinocaridid shrimp *Shinkaicaris leurokolos* in deep-sea hydrothermal vent environment

Aiyang Wang^{1,2,3,4}, Jiao Cheng^{1,2,3}, Qian Xin^{4,5}, Zhongli Sha^{1,2,3,4}
and Min Hui^{1,2,3} 

¹Department of Marine Organism Taxonomy & Phylogeny, Institute of Oceanology, Chinese Academy of Sciences, Qingdao 266071, China; ²Laoshan Laboratory, Qingdao 266237, China; ³Shandong Province Key Laboratory of Experimental Marine Biology, Institute of Oceanology, Chinese Academy of Sciences, Qingdao 266071, China; ⁴University of Chinese Academy of Sciences, Beijing 100049, China and ⁵The Affiliated Qingdao Central Hospital of Qingdao University, The Second Affiliated Hospital of Medical College of Qingdao University, Qingdao 266042, China

Abstract

The alvinocaridid shrimp *Shinkaicaris leurokolos* Kikuchi and Hashimoto, 2000, is an evolutionarily important deep-sea species in hydrothermal vents of north-western Pacific. A genome survey of *S. leurokolos* was carried out in order to provide a foundation for its whole-genome sequencing. A total of 599 Gb high-quality sequence data were obtained in the study, representing approximately 118× coverage of the *S. leurokolos* genome. According to the 17-mer distribution frequency, the estimated genome size was 5.08 Gb, and its heterozygosity ratio and percentage of repeated sequences were 2.85 and 87.03%, respectively, showing a complex genome. The final scaffold assembly accounted for a total size of 9.53 Gb (32,796,062 scaffolds, N50 = 597 bp). Repetitive elements nearly constituted 45% of the nuclear genome, among which the most ubiquitous were long interspersed nuclear elements, DNA transposons and long-terminal repeat elements. A total of 12,121,553 genomic simple sequence repeats were identified, with the most frequent repeat motif being di-nucleotide (70.27%), followed by tri-nucleotide and tetra-nucleotide. From the genome survey sequences, the mitochondrial genome of *S. leurokolos* was also constructed and 71 single nucleotide polymorphisms were identified by comparison with previous published reference. This is the first report of *de novo* whole-genome sequencing and assembly of *S. leurokolos*. These newly developed genomic data contribute to a better understanding of genomic characteristics of shrimps from deep-sea chemosynthetic ecosystems, and provides valuable resources for further molecular marker development.

Introduction

Deep-sea hydrothermal vent ecosystems are unique and extreme among marine environments, characterized by high pressure, high temperature (up to 390°C), low oxygen and high levels of toxins (hydrogen sulphide, methane and various heavy metals) (Van Dover, 2000). In such harsh environments, however, there exists lush biological community sustained by chemosynthetic primary production from free-living and symbiotic microbes (Dubilier *et al.*, 2008).

The shrimp *Shinkaicaris leurokolos* Kikuchi and Hashimoto, 2000, is one of the representative species of the Okinawa Trough hydrothermal vent area in the Northwest Pacific Ocean (Watanabe and Kojima, 2015). This species is specifically distributed in the area very close to the vent that can even contact the hydrothermal fluid instantaneously (Yahagi *et al.*, 2015), which is expected to have high thermal resistance and anti-chemical toxicity ability. It offers a biological model for uncovering the mechanisms of animals' adaptation to extreme deep-sea hydrothermal vent environments. Genomic data, especially whole genome map, are essential for clarifying this issue at molecular level.

The genomes of decapods are challenging to assemble due to their large size and complexity (Yuan *et al.*, 2017). Thus far, no whole-genome map of deep-sea decapods has been reported. For *S. leurokolos*, only mitochondrial genome and transcriptome have been sequenced and assembled in order to study the origin, evolution and adaptation of this species (Sun *et al.*, 2018a; Wang *et al.*, 2022a). The lack of genetic and genomic data on *S. leurokolos* greatly restricts the decipherment of its adaptation to extreme environments. Therefore, it highlights the importance of obtaining the whole-genome sequence of this typical vent shrimp, and before this, knowledge of genome size and characteristics is a necessary prerequisite.

Genome survey sequencing (GSS) using next-generation sequencing is currently an important and cost-effective approach to evaluate genome information such as genome size, GC content, heterozygosity and repeat content, as well as developing molecular markers (Li *et al.*, 2019; Baeza, 2020, 2021; Baeza *et al.*, 2022; Choi *et al.*, 2021). In the present study, we aimed to estimate the genomic characteristics of *S. leurokolos* through GSS, identify repetitive elements in the nuclear genome and assemble a complete mitochondrial genome. These data



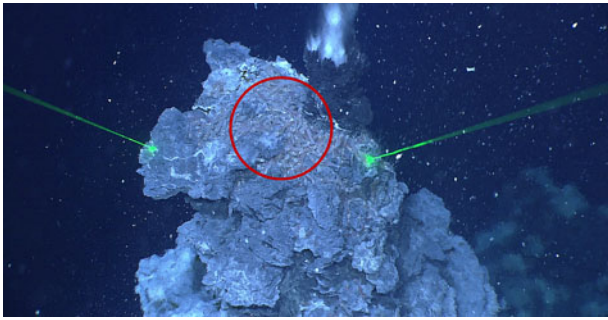


Figure 1. Swarms of *S. leurokolos* individuals (marked in the red circle) are crowded along the hydrothermal vent of Iheya North.

are expected to provide basic information on the *S. leurokolos* genome and serve as a framework for subsequent whole-genome map construction.

Materials and methods

Sample collection

Shrimps of *S. leurokolos* (Figure 1) were collected at Iheya North hydrothermal vent in the Okinawa Trough (126°53.80'E, 27°47.46'N, depth 970 m) during the cruise by the scientific research vessel (RV) KEXUE in July 2018. Species-level morphological identification abided by the main points of Komai and Segonzac (2005). Once aboard, specimens were immediately frozen in liquid nitrogen and stored at -80°C until DNA extraction. One specimen of *S. leurokolos* was subsequently subjected to genome sequencing.

DNA extraction, library construction and sequencing

Total genomic DNA was extracted from the muscle tissue using a DNeasy tissue kit (Qiagen, Beijing, China) according to the manufacturer's protocol. The quality and purity of the DNA were detected with NanoDrop and 1% agarose gel electrophoresis. After DNA extraction and detection, high-quality DNA was fragmented using ultrasonic crusher. The sequencing library with an insert size 300–350 bp was constructed with VAHTS Universal DNA Library Prep Kit for Illumina V3 following the manufacturer's recommendations. Paired-end sequencing was conducted using DNBSEQ-T7 platform (MGI Tech Co., Ltd. in Shenzhen, China) by Wuhan Onemore-tech Co., Ltd.

Sequence quality control and genome assembly

The quality control of raw data was performed using the FastQC v0.11.9 (Andrews, 2010) and Trimmomatic v0.39 (Bolger *et al.*,

2014) based on the four criteria: (1) removing the A-tail and adaptors, (2) deleting the low-quality reads with N content more than 10%, (3) filtering the reads with base quality less than 10 and (4) discarding duplicated reads. Then the clean data were submitted to the Sequence Read Archive (SRA) database (<http://www.ncbi.nlm.nih.gov/sra/>), and were available under the accession number PRJNA926015. Genome size, heterozygosity and repeat content of *S. leurokolos* were estimated based on a *K*-mer method by Jellyfish and GenomeScope with parameters of 17-mer, 21-mer, 27-mer and 31-mer (Marçais and Kingsford, 2011; Vurture *et al.*, 2017). Based on clean data, the draft genome of *S. leurokolos* was *de novo* assembled using SOAPdenovo2 (Luo *et al.*, 2012) with *K*-mer = 41 and *K*-mer = 63.

Genomic repetitive elements and microsatellite identification

In the present study, two methods were used for the discovery, annotation and quantification of the repetitive elements from the draft genome of *S. leurokolos*. First, repetitive elements were *de novo* annotated using the RepeatModeler v2.0.3 (Flynn *et al.*, 2020) and LTR_FINDER v1.0.2 (Xu and Wang, 2007). Second, repetitive sequences were identified by RepeatMasker v4.0.9 (Tempel, 2012) and RepeatProteinMask v4.1.0 (a component of the RepeatMasker application) with the Repbase database. The Perl script MISA (<http://pgrc.ipk-gatersleben.de/misa/misa.html>) was used to identify SSRs in the draft genome of *S. leurokolos*, and search parameters were set as minimum of 6, 5, 5, 5 and 5 repeats for detecting di-, tri-, tetra-, penta- and hexanucleotide motifs, respectively.

Mitochondrial genome assembly and SNP identification

The mitochondrial genome of *S. leurokolos* was *de novo* assembled with Novoplasty v4.3.1 (Dierckxsens *et al.*, 2016) using the published COI sequence of *S. leurokolos* (GenBank accession no. MH398102) as seed sequence. GapCloser v1.12 was used to fill in the missing regions to acquire the complete circular mitochondrial genome. The mitochondrial genome was annotated using the automatic annotators of mitochondrial genes online, Geseq (Tillich *et al.*, 2017) and the MITOS 2 Web server with the invertebrate genetic codes (Donath *et al.*, 2019), followed by strictly manual check.

To identify variation in *S. leurokolos* mitochondrial genome, single nucleotide polymorphisms (SNPs) recovery was performed. The previously published *S. leurokolos* mitochondrial genome (GenBank accession no. MF627741) was set as a reference. Alignment between the two mitochondrial genome sequences was performed using the software MEGA v7.00 (Kumar *et al.*, 2016). The varied sites were supposed to be candidate SNP markers.

Table 1. Summary information for the *S. leurokolos* genome sequencing and genome assembly

Genome sequencing						
Raw base	Raw read number	Clean base	Clean read number	Q20	Q30	GC content
639.75 Gb	4,264,994,284	599.63 Gb	3,905,243,920	96.28%	91.18%	37.60%
Assembled draft genome						
	Total length	Total number	Maximum length	N50 length	GC content	
Contig	8,647,086,942 bp	42,142,373	30,932 bp	227 bp	36.12%	
Scaffold	9,527,856,577 bp	32,796,062	69,344 bp	597 bp	36.12%	

Q20: the ratio of data with accuracy above 99% in total data. Q30: the ratio of data with accuracy above 99.90% in total data

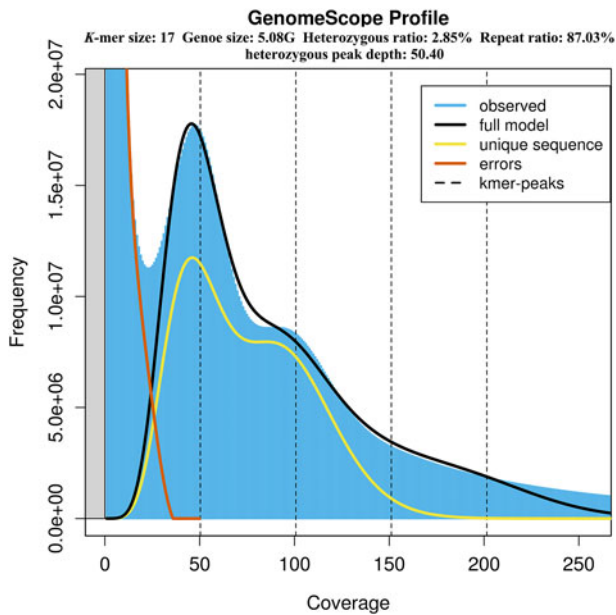


Figure 2. *K*-mer ($K=17$) analysis for estimation of the genome size of *S. leurokolos*. The *x*-axis represents coverage, and the *y*-axis represents the frequency at each depth.

Results and discussion

Sequencing and quality evaluation

A total of 639.75 Gb raw reads were generated for *S. leurokolos*. After filtering and correction, a total of 599.63 Gb clean reads were derived (Table 1). The Q20 and Q30 values of the sequencing data were 96.28 and 91.18%, respectively (Table 1). It has been specified that Q20 and Q30 values should be at least 90 and 85% (Li *et al.*, 2019). Therefore, the sequencing data of *S. leurokolos* genome show extreme precision in the present study. GC content is an important factor in many experiments and bioinformatic analysis, especially for next-generation sequencing where the sequenced DNA has gone through multiple rounds of PCR amplification. High or low GC content will reduce sequencing coverage and cause sequencing bias (Bentley *et al.*, 2008; Aird *et al.*, 2011; Cheung *et al.*, 2011). In this study, GC content of *S. leurokolos* sequences was 37.6% showing a mid GC content (30–47%) (Shangguan *et al.*, 2013). Overall, these results indicate high-quality sequencing data obtained for *S. leurokolos*.

K-mer analysis and genome size estimation

The genome size, heterozygosity and repetitive ratio of *S. leurokolos* were evaluated using *K*-mer distribution analysis, and the 17-mer yielded the highest model fit (Figure 2 and Table 2). *K*-mer analysis revealed a unique bimodal profile with a high heterozygous peak around 50 \times coverage and a lower homozygous peak around 100 \times coverage (Figure 2). By calculation, the genome

size of *S. leurokolos* was estimated to be 5.08 Gb (Table 2). Flow cytometry is another method for the prediction of genome size. Previous study for other four alvinocaridid shrimps based on flow cytometry reveals that genome sizes range from 10,160 Mp in *Rimicaris exoculata* to 13,050 Mp in *Chorocaris chacei* (Bonnivard *et al.*, 2009), displaying a large genome size in the family Alvinocarididae. It seems that the genome size of *S. leurokolos* is much smaller than those of other alvinocaridid shrimps, or its genome size has been underestimated by GSS. The significant discordance of genome size revealed by GSS and flow cytometry has been also detected in other decapods, such as crayfish *Procambarus clarkia*, showing larger genome size by flow cytometry analysis than that revealed by GSS (Shi *et al.*, 2018). However, muscle instead of haemolymph cell has been used in the flow cytometry analysis for alvinocaridid shrimps (Bonnivard *et al.*, 2009), which might be due to the difficulty in collecting living shrimp samples from deep sea. It may influence the quality of cell suspension preparation and in turn affect the precision of genome size estimation. On the other hand, the high heterozygosity and repeat ratio characteristics of *S. leurokolos* genome as shown below might bring biased results in genome size estimation by affecting the *K*-mer depth distribution (Shi *et al.*, 2018). In brief, GSS and flow cytometry should be combined to estimate genome sizes of deep-sea species with large and complex genome, and the genome size of *S. leurokolos* might be larger than 5.08 Gb.

According to the *K*-mer distribution, an extremely high heterozygosity 2.85% was detected in *S. leurokolos* genome (Figure 2 and Table 2). It has been suggested that genome assembly will be difficult if the heterozygosity rate exceeds 0.5%, and it is even more difficult if it exceeds 1% (Marçais and Kingsford, 2011). The repeat ratio of *S. leurokolos* genomic sequences was also high (87.03%) (Figure 2 and Table 2). The high heterozygosity rate and repeat ratio have been also revealed in other decapods, such as *Litopenaeus vannamei*, *Penaeus chinensis* and *P. monodon* (Zhang *et al.*, 2019; Van Quyen *et al.*, 2020; Uengwetwanit *et al.*, 2021; Yuan *et al.*, 2021b; Wang *et al.*, 2022b), and difficulties in genome assembly seem to be common problem in decapods due to high heterozygosity and repeat ratio (Yuan *et al.*, 2021a).

Genome de novo assembly

To assemble the draft genome of *S. leurokolos*, two *K*-mer values, 41 and 63 bp were selected. Unfortunately, too much computer memory was required and the assembly task could not be completed when using the 41 bp *K*-mer value. A complete assembly using 63 bp *K*-mer value was obtained (Table 1). Finally, our efforts recovered a total of 9,527,856,577 bp scaffolds with the scaffold N50 value of 597 bp, and the maximum scaffold was 69,344 bp in length (Table 1). It is apparent that the size of draft genome assembly is almost twice as large as the estimated genome size based on 17-mer analysis. The most plausible explanation for the genome assembly size deviation may be that the presence of a large number of repetitive elements (87.03%) and high heterozygosity (2.85%) of *S. leurokolos* genome might induce the assembly has multiple copies of the same genomic region and

Table 2. Statistics of the estimated *S. leurokolos* genome size and other characteristics

<i>K</i> -mer size	<i>K</i> = 17	<i>K</i> = 21	<i>K</i> = 27	<i>K</i> = 31
Genome size (bp)	5,081,929,970	5,410,446,571	5,338,429,994	5,275,380,535
Heterozygosity (%)	2.85	3.92	3.85	3.74
Repeat ratio (%)	87.03	72.62	68.37	66.41
Model Fit	95.51%	94.98%	95.17%	95.35%

Table 3. Statistics of repetitive sequence annotation in the *S. leurokolos* draft genome assembly

Type	Repeat size (bp)	% of genome
Trf	1,103,995,050	11.59
Repeatmasker	954,688,962	10.02
Proteinmask	637,882,358	6.69
De novo	2,740,517,835	28.76
Total	4,250,866,696	44.62

even contained two divergent haplotypes (Pflug *et al.*, 2020; Hu *et al.*, 2022; Wyngaard *et al.*, 2022). The average GC content of *S. leurokolos* assembled genome was about 36.12%. To further evaluate the data of our assembly, we compared it to previously reported genome survey data of decapods. The scaffold N50 of *S. leurokolos* is much shorter than that of Pacific white shrimp *L. vannamei* (1343 bp) (Yu *et al.*, 2015) and red swamp crayfish *P. clarkia* (1426 bp) (Shi *et al.*, 2018). The inherent defects of second-generation sequencing technology in read length and high complexity of the large genome of *S. leurokolos* itself should be the main reasons for the poor assembly. We hold the opinion that the large and complex genome of *S. leurokolos* represents typical challenges faced by all alvinocaridid shrimp genomes, which partly explains why genomic resources for alvinocaridid shrimps are so limited compared to those of many other deep-sea organisms. Hence, developing new assemblers and bioinformatics tools and using combination of short- and long-read sequencing technologies (i.e. PacBio, Oxford Nanopore Technologies, ONT) are expected to solve these challenges for assembling a high-quality genome. The current GSS data could serve as a reference for subsequent whole-genome sequencing project of *S. leurokolos*.

Genomic repetitive elements annotation

Repetitive sequences, especially transposable elements (TEs), are known to be an evolutionary precursor of many genes, a driving force in the evolution of epigenetic regulation and an important factor in genomic stability maintenance and evolution (Jurka *et al.*, 2007). In total, 4250 Mb repetitive elements were identified in *S. leurokolos* draft genome, accounting for 44.62% of the assembled genome (Table 3). Combining the results from RepeatMasker and RepeatProteinMask analyses, our results revealed that among these repetitive sequences, 38.92% (3708 Mb) were TEs, but 16.49% could not be classified within the

TEs (Table 4). Long interspersed nuclear elements (LINEs) were the most common among the TEs, accounting for 10.45%, followed by DNA transposons (6.09%) and long-terminal repeat elements (LTRs) (4.79%) (Table 4). These repetitive elements, including LINEs, DNA and LTRs, also take up a large proportion of genomes in many other decapod crustaceans (Baeza, 2020; Tang *et al.*, 2020; Chak *et al.*, 2021; Uengwetwanit *et al.*, 2021). However, it has been suggested that the 'unclassified' TEs with a large proportion may contain species-specific variants of known repetitive elements, and we should be cautious when comparing these datasets directly with those of other species (Murgarella *et al.*, 2016).

Microsatellite analysis

It is widely recognized that as a most popular and versatile genetic marker, SSRs are widely used for the genetic characterization of populations due to their abundance in genome, high polymorphism and co-dominant nature (Abdul-Muneer, 2014). In the assembled scaffolds, a total of 12,121,553 microsatellite motifs were identified in *S. leurokolos* (Table 5). Among them, the di-nucleotide was the most abundant, accounting for 70.27% of the total SSRs, which was followed by tri- (25.54%), tetra- (3.33%), penta- (0.50%) and hexa- (3.36%) nucleotide SSRs (Table 6). Our finding shows that both di-nucleotide and tri-nucleotide SSRs are numerous, and the number of repetitions is inversely proportional to the length of repetitions. This result is consistent with those in other crustaceans, such as kuruma prawn *Marsupenaeus japonicus* (Lu *et al.*, 2017), Japanese mantis shrimp *Oratosquilla oratoria* (Cheng *et al.*, 2018) and Antarctic krill *Euphausia superba* (Huang *et al.*, 2020). It has been proposed that longer repeats have downward mutation bias and short persistence times (Harr and Schlotterer, 2000), and therefore, less SSRs with longer repeat units exist in genomes.

Mitochondrial genome and candidate molecular marker identification

Mitochondria are essential organelles that generate most chemical energy to power the cell's biochemical reactions. There is evidence that mitochondrial DNA plays a role in many aspects of biological life history, such as lifespan, fertility, resistance to starvation, altitude adaptation and regulation of temperature (Ballard and Melvin, 2010). It is therefore of significant importance to investigate the mitochondrial genome of *S. leurokolos* inhabiting deep-sea chemosynthetic ecosystems. In this study, we assembled a

Table 4. Statistics of TEs in the *S. leurokolos* draft genome assembly

Type	Repbse TEs		TE protiens		De novo		Combined TEs	
	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome
DNA	411,979,995	4.32	64,425,980	0.68	175,675,369	1.84	580,295,271	6.09
LINE	239,868,220	2.52	384,621,873	4.04	733,379,750	7.7	995,240,891	10.45
SINE	2,112,324	0.02	0	0	19,580,364	0.21	20,866,053	0.22
LTR	171,474,588	1.8	188,870,490	1.98	270,656,796	2.84	456,205,558	4.79
Other	352,246	0	9171	0	0	0	361,417	0
Unknown	12,546,876	0.13	185,093	0	1,559,607,035	16.37	1,571,485,576	16.49
Total	954,688,962	10.02	637,882,358	6.69	2,740,517,835	28.76	3,708,475,620	38.92

RepBase TEs and TE proteins were obtained based on the RepBase library using RepeatMasker and RepeatProteinMask, respectively. *De novo* repeat prediction was performed using RepeatMasker against the *de novo* repeat library of *S. leurokolos*, which was constructed by the programs LTR_FINDER and RepeatModeler. Combined TEs were the union of the three methods.

Table 5. Statistics of SSR distribution in the *S. leurokolos* draft genome assembly

	Di-	Tri-	Tetra-	Penta-	Hexa-
SSR number	8,517,638	3,096,155	403,448	60,908	43,404
Percentage	70.27%	25.54%	3.33%	0.50%	0.36%

Table 6. Organization of the *S. leurokolos* mitogenome

Gene	Positions		Length	Codons		Strand
	Start	Stop		Start/Stop	ovl/nc	
cox1	1	1542	1542	ATT/TAA	-4	+
tRNA-Leu (UAA)	1537	1601	65		-6	+
cox2	1605	2315	711	ATG/TAA	3	+
tRNA-Lys (UUU)	2296	2365	70		-20	+
tRNA-Asp (GUC)	2373	2438	66		7	+
ATP8	2439	2597	159	ATT/TAA	0	+
ATP6	2591	3265	675	ATG/TAA	-7	+
cox3	3265	4053	789	ATG/TAG	-1	+
tRNA-Gly (UCC)	4056	4121	66		2	+
nad3	4121	4474	354	ATT/TAA	-1	+
tRNA-Ala (UGC)	4476	4538	63		1	+
tRNA-Arg (UCG)	4537	4602	66		-2	+
tRNA-Asn (GUU)	4604	4668	65		1	+
tRNA-Ser (UCU)	4669	4736	68		0	+
tRNA-Glu (UUC)	4736	4807	72		-1	+
tRNA-Phe (GAA)	4807	4875	69		-1	-
nad5	4875	6603	1729	GTG/TAT	-1	-
tRNA-His (GUG)	6603	6668	66		-1	-
nad4	6668	8006	1339	ATG/GAT	-1	-
nad4 l	8000	8299	300	ATG/TAA	7	-
tRNA-Thr (UGU)	8305	8370	66		5	+
tRNA-Pro (UGG)	8370	8435	66		-1	-
nad6	8437	8952	516	ATC/TCA	1	+
cytb	8952	10,088	1137	ATA/TAG	-1	+
tRNA-Ser (UGA)	10,087	10,158	72		-2	+
nad1	10,176	11,117	942	ATT/TAG	17	-
tRNA-Leu (UAG)	11,168	11,235	68		50	-
rrnL	11,235	12,543	1309		-1	-
tRNA-Val (UAC)	12,542	12,609	68		-2	-
rrnS	12,634	13,455	822		24	-
Control region	13,456	14,481	1026		0	+
tRNA-Ile (GAU)	14,482	14,548	67		0	+
tRNA-Gln (UUG)	14,566	14,633	68		17	-
tRNA-Met (CAU)	14,644	14,708	65		10	+
nad2	14,709	15,704	996	ATG/TAA	0	+
tRNA-Trp (UCA)	15,707	15,781	75		2	+
tRNA-Cys (GCA)	15,780	15,846	67		-2	-
tRNA-Tyr (GUA)	15,845	4	66		-2	-

Table 7. Summary of SNPs in *S. leurokolos* mitochondrial genome

Gene	Transition	Transversion	Mutation rates, %	Amino acid change
cox1	10	1	0.71	I → M
nad2	7		0.70	A → T
cytb	4		0.35	V → I
nad1	3		0.32	F → L
nad4	8	1	0.67	
nad5	5		0.29	
cox2	2		0.28	
ATP6	2		0.30	
nad3	2		0.56	
nad6	2		0.39	
ATP8	1		0.63	
tRNA-Ala	1		1.59	
rrnL	1		0.08	
tRNA-Trp	1		1.33	
tRNA-Cys	1		1.49	
Control region	16	3	1.58	

15,906 bp long complete mitochondrial genome (GenBank accession no. OQ622002) of *S. leurokolos* from the GSS data. It consisted of 13 protein-coding genes (PCGs), 2 ribosomal RNA genes (*rrnS* and *rrnL*), 22 transfer (tRNA) genes and a non-coding hypervariable control region (1026 bp) between *rrnS* and tRNA-Ile, showing the typical alvinocaridid shrimp mitogenome arrangement model (Table 6). Most of the PCGs and tRNA genes were encoded on the positive strand. Gene overlaps in 19 gene junctions (a total of 57 bp in length) and intergenic spaces in 14 gene junctions (ranging from 1 to 50 bp) were also observed (Table 6).

Moreover, mitochondrial DNA fragments have been proved to be efficient molecular markers in phylogenetic and population genetic analysis. In order to identify candidate markers, we aligned the mitochondrial genome assembled in this study with the previous reported *S. leurokolos* mitochondrial genome (Sun *et al.*, 2018a). By comparison, 3 indels (all located in the control region) and 71 SNPs were detected. The SNPs included 66 transitions and 5 transversions: 47 in PCGs, 3 in tRNAs, 1 in rRNAs and 19 in non-coding regions. Of the 47 SNPs in PCGs, only four mutations were non-synonymous substitutions (Table 7), which occurred in *cox1*, *nad2*, *cytb* and *nad1* (Table 7). It is a general observation in molecular evolution that functional importance and substitution rate are negatively correlated (Sun *et al.*, 2010). This means that the more functionally important genes (or genetic regions) evolve more slowly due to their important effects or strong functional constraints (Kimura, 1983; Yang, 2006). In addition, the relatively high substitution rates observed in tRNA-Ala (1.59%), control region (1.58%), tRNA-Cys (1.49%) and tRNA-Trp (1.33%) may indicate relatively low functional constraints in these regions.

To date, population genetic and phylogenetic studies for alvinocaridid shrimps are mainly based on mitochondrial *cox1*, 12S rDNA and 16S rDNA genes (Yahagi *et al.*, 2015; Sun *et al.*, 2018b). In this study, *cox1*, *nad2*, *nad4* and control region show high mutation rate, and the sequences are long enough for primer design. Hence, these mitochondrial genes can be selected as

candidate markers for population genetic studies for *S. leurokolos*. However, it requires further validation by amplification and sequencing in more individuals.

Conclusions

In summary, this study developed and surveyed the first reference genome for *S. leurokolos*, an alvinocaridid shrimp from Iheya North hydrothermal vent. It represents the first genome survey for crustaceans from deep-sea chemosynthetic ecosystem. The results showed that the genome of *S. leurokolos* was extremely complex, with large genome size, extremely high heterozygosity and repeat ratio. The patterns of genome nuclear repetitive elements were investigated, and a large number of SSRs were detected. The mitochondrial genome of *S. leurokolos* was also assembled, and candidate molecular markers for population genetic study were proposed. These datasets enrich genetic resources of deep-sea life, and are expected to facilitate further studies on the evolutionary biology of alvinocaridid shrimps, as well as the construction of a high-quality genome map of the deep-sea vent *S. leurokolos*.

Data

The clean data of the genome survey sequencing were openly available in NCBI SRA databank under the accession number PRJNA926015. The authors confirm that the other data supporting the findings of this study are available within the article.

Acknowledgements. The samples were collected by RV KEXUE. The authors wish to thank the crews for their help during collection of samples.

Author contributions. M. H. and Z. S. formulated the research question and designed the study. M. H. collected the specimen. Q. X. extracted DNA of the specimen. A. W. and M.H. carried out the study, analysed the data, interpreted the findings and wrote the article. J. C. and Z. S. also interpreted the findings and revised the article.

Financial support. This work was funded by the Science and Technology Innovation Project of Laoshan Laboratory (LSKJ202203104), the National Science Foundation for Distinguished Young Scholars (42025603) and the Strategic Priority Research Program of Chinese Academy of Sciences (XDB42000000).

Competing interests. None.

Ethical standards. No regulated invertebrate was involved in this study.

References

- Abdul-Muneer PM (2014) Application of microsatellite markers in conservation genetics and fisheries management: Recent advances in population structure analysis and conservation strategies. *Genetics Research International* **2014**, 691759.
- Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C and Gnirke A (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology* **12**, R18.
- Andrews S (2010) *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Cambridge, United Kingdom: Babraham Bioinformatics, Babraham Institute, Available at <https://www.bioinformatics.babraham.ac.uk/projects/fastqc>
- Baeza JA (2020) Genome survey sequencing of the Caribbean spiny lobster *Panulirus argus*: Genome size, nuclear rRNA operon, repetitive elements, and microsatellite discovery. *PeerJ* **8**, e10554.
- Baeza JA (2021) A first genomic portrait of the Florida stone crab *Menippe mercenaria*: Genome size, mitochondrial chromosome, and repetitive elements. *Marine Genomics* **57**, 100821.
- Baeza JA, Baker AM and Liu H (2022) Genome survey sequencing of the long-legged spiny lobster *Panulirus longipes* (A. Milne-Edwards, 1868) (Decapoda: Achelata: Palinuridae): Improved mitochondrial genome

- annotation, nuclear repetitive elements classification, and SSR marker discovery. *Journal of Crustacean Biology* **42**, ruac006.
- Ballard JWO and Melvin RG** (2010) Linking the mitochondrial genotype to the organismal phenotype. *Molecular Ecology* **19**, 1523–1539.
- Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL and Bignell HR** (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59.
- Bolger AM, Lohse M and Usadel B** (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120.
- Bonnivard E, Catrice O, Ravaux J, Brown SC and Higuier D** (2009) Survey of genome size in 28 hydrothermal vent species covering 10 families. *Genome* **52**, 524–536.
- Chak STC, Harris SE, Hultgren KM, Jeffery NW and Rubenstein DR** (2021) Eusociality in snapping shrimps is associated with larger genomes and an accumulation of transposable elements. *Proceedings of the National Academy of Sciences of the USA* **118**, e2025051118.
- Cheng J, Zhang N and Sha Z** (2018) Isolation and characterization of microsatellite markers for exploring introgressive hybridization between the *Oratosquilla oratoria* complex. *Molecular Biology Reports* **45**, 1499–1505.
- Cheung MS, Down TA, Latorre I and Ahringer J** (2011) Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Research* **39**, e103.
- Choi E, Kim SH, Lee SJ, Jo E, Kim J, Kim JH, Parker SJ, Chi YM and Park H** (2021) A first genome survey and genomic SSR marker analysis of *Trematopus loennbergii* Regan, 1913. *Animals* **11**, 3186.
- Dierckxens N, Mardulyn P and Smits G** (2016) NOVOPlasty: De novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research* **45**, e18.
- Donath A, Jühling F, Al-Arab M, Bernhart SH, Reinhardt F, Stadler PF, Middendorf M and Bernt M** (2019) Improved annotation of protein-coding genes boundaries in metazoan mitochondrial genomes. *Nucleic Acids Research* **47**, 10543–10552.
- Dubilier N, Bergin C and Lott C** (2008) Symbiotic diversity in marine animals: The art of harnessing chemosynthesis. *Nature Reviews Microbiology* **6**, 725–740.
- Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C and Smit AF** (2020) RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the USA* **117**, 9451–9457.
- Harr B and Schlötterer C** (2000) Long microsatellite alleles in *Drosophila melanogaster* have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation. *Genetics* **155**, 1213–1220.
- Hu G, Feng J, Xiang X, Wang J, Salojärvi J, Liu C, Wu Z, Zhang J, Liang X, Jiang Z, Liu W, Ou L, Li J, Fan G, Mai Y, Chen C, Zhang X, Zheng J, Zhang Y, Peng H, Yao L, Wai CM, Luo X, Fu J, Tang H, Lan T, Lai B, Sun J, Wei Y, Li H, Chen J, Huang X, Yan Q, Liu X, McHale LK, Rolling W, Guyot R, Sankoff D, Zheng C, Albert V.A, Ming R, Chen H, Xia R and Li J** (2022) Two divergent haplotypes from a highly heterozygous lychee genome suggest independent domestication events for early and late-maturing cultivars. *Nature Genetics* **54**, 73–83.
- Huang Y, Bian C, Liu Z, Wang L, Xue C, Huang H, Yi Y, You X, Song W, Mao X, Song L and Shi Q** (2020) The first genome survey of the Antarctic krill (*Euphausia superba*) provides a valuable genetic resource for polar biomedical research. *Marine Drugs* **18**, 185.
- Jurka J, Kapitonov VV, Kohany O and Jurka MV** (2007) Repetitive sequences in complex genomes: Structure and evolution. *Annual Review of Genomics and Human Genetics* **8**, 241–259.
- Kimura M** (1983) *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.
- Komai T and Segonzac M** (2005) A revision of the genus *Alvinocaris* Williams and Chace (Crustacea: Decapoda: Caridea: Alvinocarididae), with descriptions of a new genus and a new species of *Alvinocaris*. *Journal of Natural History* **39**, 1111–1175.
- Kumar S, Stecher G and Tamura K** (2016) MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* **33**, 1870–1874.
- Li G, Song L, Jin C, Li M, Gong S and Wang Y** (2019) Genome survey and SSR analysis of *Apocynum venetum*. *Bioscience Reports* **39**, BSR20190146.
- Lu X, Luan S, Kong J, Hu L, Mao Y and Zhong S** (2017) Genome-wide mining, characterization, and development of microsatellite markers in *Marsipenaenus japonicus* by genome survey sequencing. *Chinese Journal of Oceanology and Limnology* **35**, 203–214.
- Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Zhu X, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW and Wang J** (2012) SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 18.
- Marçais G and Kingsford C** (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770.
- Murgarella M, Puiu D, Novoa B, Figueras A, Posada D and Canchaya C** (2016) A first insight into the genome of the filter-feeder mussel *Mytilus galloprovincialis*. *PLoS ONE* **11**, e0151561.
- Pflug JM, Holmes VR, Burrus C, Johnston JS and Maddison DR** (2020) Measuring genome sizes using read-depth, k-mers, and flow cytometry: Methodological comparisons in Beetles (Coleoptera). *G3 Genes Genomes Genetics* **10**, 3047–3060.
- Shangguan L, Han J, Kayesh E, Sun X, Zhang C, Pervaiz T, Wen X and Fang J** (2013) Evaluation of genome sequencing quality in selected plant species using expressed sequence tags. *PLoS ONE* **8**, e69890.
- Shi L, Yi S and Li Y** (2018) Genome survey sequencing of red swamp crayfish *Procambarus clarkii*. *Molecular Biology Reports* **45**, 799–806.
- Sun S, Hui M, Wang M and Sha Z** (2018a) The complete mitochondrial genome of the alvinocaridid shrimp *Shinkaicaris leurokolos* (Decapoda, Caridea): Insight into the mitochondrial genetic basis of deep-sea hydrothermal vent adaptation in the shrimp. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics* **25**, 42–52.
- Sun XJ, Li Q and Kong LF** (2010) Comparative mitochondrial genomics within sea cucumber (*Apostichopus japonicus*): Provide new insights into relationships among color variants. *Aquaculture* **309**, 280–285.
- Sun S, Sha Z and Wang Y** (2018b) Phylogenetic position of Alvinocarididae (Crustacea: Decapoda: Caridea): New insights into the origin and evolutionary history of the hydrothermal vent alvinocaridid shrimps. *Deep-Sea Research Part I: Oceanographic Research Papers* **141**, 93–105.
- Tang B, Wang Z, Liu Q, Zhang H, Jiang S, Li X, Wang Z, Sun Y, Sha Z, Jiang H, Wu X, Ren Y, Li H, Xuan F, Ge B, Jiang W, She S, Sun H, Qiu Q, Wang W, Wang Q, Qiu G, Zhang D and Li Y** (2020) High-quality genome assembly of *Eriocheir japonica sinensis* reveals its unique genome evolution. *Frontiers in Genetics* **11**, 535.
- Tempel S** (2012) Using and understanding RepeatMasker. In Bigot Y (ed.), *Mobile Genetic Elements: Protocols and Genomic Applications*. Totowa, NJ: Humana Press, 29–51.
- Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R and Greiner S** (2017) GeSeq-versatile and accurate annotation of organelle genomes. *Nucleic Acids Research* **45**, W6–W11.
- Uengwetwanit P, Pootakham W, Nookaew I, Sonthirod C, Angthong P, Sittikankaew K, Rungrassamee W, Arayamethakorn S, Wongsurawat T, Jenjaroenpun P, Sangrakru D, Leelatanawit R, Khudet J, Koehorst JJ, Schaap PJ, Martins dos Santos V, Tangy F and Karoonuthaisiri N** (2021) A chromosome-level assembly of the black tiger shrimp (*Penaeus monodon*) genome facilitates the identification of growth-associated genes. *Molecular Ecology Resources* **21**, 1620–1640.
- Van Dover CL** (2000) *The Ecology of Deep-sea Hydrothermal Vents*. Princeton: Princeton University Press.
- Van Quyen D, Gan HM, Lee YP, Nguyen DD, Nguyen TH, Tran XT, Nguyen VS, Khang DD and Austin CM** (2020) Improved genomic resources for the black tiger prawn (*Penaeus monodon*). *Marine Genomics* **52**, 100751.
- Vurtture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J and Schatz MC** (2017) GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204.
- Wang Q, Ren X, Liu P, Li J, Lv J, Wang J, Zhang H, Wei W, Zhou Y, He Y and Li J** (2022b) Improved genome assembly of Chinese shrimp (*Fenneropenaeus chinensis*) suggests adaptation to the environment during evolution and domestication. *Molecular Ecology Resources* **22**, 334–344.
- Wang A, Sha Z and Hui M** (2022a) Full-length transcriptome comparison provides novel insights into the molecular basis of adaptation to different ecological niches of the deep-sea hydrothermal vent in alvinocaridid shrimps. *Diversity* **14**, 371.
- Watanabe H and Kojima S** (2015) Vent fauna in the Okinawa Trough. In Ishibashi J, Okino K and Sunamura M (eds), *Subseafloor Biosphere Linked to Hydrothermal Systems: TAIGA Concept*. Tokyo, Japan: Springer, 449–459.

- Wyngaard GA, Skern-Mauritzen R, Malde K, Prendergast R and Peruzzi S (2022) The salmon louse genome may be much larger than sequencing suggests. *Scientific Reports* **12**, 6616.
- Xu Z and Wang H (2007) LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35**, W265–W268.
- Yahagi T, Watanabe H, Ishibashi J and Kojima S (2015) Genetic population structure of four hydrothermal vent shrimp species (Alvinocarididae) in the Okinawa Trough, Northwest Pacific. *Marine Ecology Progress Series* **529**, 159–169.
- Yang Z (2006) *Computational Molecular Evolution*. Oxford: Oxford University Press.
- Yu Y, Zhang X, Yuan J, Li F, Chen X, Zhao Y, Huang L, Zheng H and Xiang J (2015) Genome survey and high-density genetic map construction provide genomic and genetic resources for the Pacific White Shrimp *Litopenaeus vannamei*. *Scientific Reports* **5**, 15612.
- Yuan J, Gao Y, Zhang X, Wei J, Liu C, Li F and Xiang J (2017) Genome sequences of marine shrimp *Exopalaemon carinicauda* Holthuis provide insights into genome size evolution of caridea. *Marine Drugs* **15**, 213.
- Yuan J, Zhang X, Li F and Xiang J (2021a) Genome sequencing and assembly strategies and a comparative analysis of the genomic characteristics in Penaeid shrimp species. *Frontiers in Genetics* **12**, 658619.
- Yuan J, Zhang X, Wang M, Sun Y, Liu C, Li S, Yu Y, Gao Y, Liu F, Zhang X, Kong J, Fan G, Zhang C, Feng L, Xiang J and Li F (2021b) Simple sequence repeats drive genome plasticity and promote adaptive evolution in Penaeid shrimp. *Communications Biology* **4**, 186.
- Zhang X, Yuan J, Sun Y, Li S, Gao Y, Yu Y, Liu C, Wang Q, Lv X, Zhang X, Ma KY, Wang X, Lin W, Wang L, Zhu X, Zhang C, Zhang J, Jin S, Yu K, Kong J, Xu P, Chen J, Zhang H, Sorgeloos P, Sagi A, Alcivar-Warren A, Liu Z, Wang L, Ruan J, Chu KH, Liu B, Li F and Xiang J (2019) Penaeid shrimp genome provides insights into benthic adaptation and frequent molting. *Nature Communications* **10**, 356.