

Assembly of continuous high-resolution draft genome sequence of *Hemicentrotus pulcherrimus* using long-read sequencing

Tetsushi Komoto¹ | Kazuho Ikeo² | Shunsuke Yaguchi³  |
Takashi Yamamoto^{1,4} | Naoaki Sakamoto^{1,4}  | Akinori Awazu^{1,4} 

¹Graduate School of Integrated Sciences for Life, Hiroshima University, Higashi-Hiroshima, Japan

²Department of Genomics and Evolutionary Biology, National Institute of Genetics, Shizuoka, Japan

³Shimoda Marine Research Center, University of Tsukuba, Shimoda, Japan

⁴Research Center for the Mathematics on Chromatin Live Dynamics, Hiroshima University, Higashi-Hiroshima, Japan

Correspondence

Akinori Awazu and Naoaki Sakamoto, Division of Integrated Sciences for Life, Graduate School of Integrated Sciences for Life, Hiroshima University, Higashi-Hiroshima 739-8526, Japan.

Email: awa@hiroshima-u.ac.jp and naosaka@hiroshima-u.ac.jp

Funding information

Japan Society for the Promotion of Science, Grant/Award Number: 21K06124

Communicating Editor: Hiroshi Wada

Abstract

The update of the draft genome assembly of sea urchin, *Hemicentrotus pulcherrimus*, which is widely studied in East Asia as a model organism of early development, was performed using Oxford nanopore long-read sequencing. The updated assembly provided ~600-Mb genome sequences divided into 2,163 contigs with N50 = 516 kb. BUSCO completeness score and transcriptome model mapping ratio (TMMR) of the present assembly were obtained as 96.5% and 77.8%, respectively. These results were more continuous with higher resolution than those by the previous version of *H. pulcherrimus* draft genome, HpulGenome_v1, where the number of scaffolds = 16,251 with a total of ~100 Mb, N50 = 143 kb, BUSCO completeness score = 86.1%, and TMMR = 55.4%. The obtained genome contained 36,055 gene models that were consistent with those in other echinoderms. Additionally, two tandem repeat sequences of early histone gene locus containing 47 copies and 34 copies of all histone genes, and 185 of the homologous sequences of the interspecifically conserved region of the *Ars* insulator, *ArsInsC*, were obtained. These results provide further advance for genome-wide research of development, gene regulation, and intranuclear structural dynamics of multicellular organisms using *H. pulcherrimus*.

KEYWORDS

ArsInsC, *cis*-regulatory sequences, long-read sequencing, sea urchin genome, repeat sequences

1 | INTRODUCTION

Sea urchins are model organisms commonly used for studying early development and morphogenesis due to their evolutionary position as diverged at the early period of deuterostome evolution. For example, *Hemicentrotus pulcherrimus*, which is one of the typical sea urchins in East Asia, has been widely investigated to reveal developmental systems of multicellular organisms such as intranuclear chromatin structural dynamics, establishment of left-right asymmetric body axis, and

nervous system formation (Matsushita et al., 2017; Takemoto et al., 2016; Yaguchi & Katow, 2003). The gene regulatory network controlling endomesoderm specification in sea urchin embryos was also extensively studied using *Strongylocentrotus purpuratus* (Davidson et al., 2002; Oliveri & Davidson, 2004).

The detailed features of genome sequences of various organisms promote research for a variety of universal and specific activities of entire living systems. Therefore, similar to various organisms, sea urchins draft genomes have also been assembled; for example, *S. purpuratus* in

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *Development, Growth & Differentiation* published by John Wiley & Sons Australia, Ltd on behalf of Japanese Society of Developmental Biologists.

2006 (Sodergren et al., 2006) (the latest version was updated in 2019), *H. pulcherrimus* in 2018 (Kinjo et al., 2018), *Lytechinus variegatus* in 2020 (Davidson et al., 2020), *Temnopleurus reevesii* in 2022 (Kinjo et al., 2022), and *Paracentrotus lividus* in 2023 (Marletaz et al., 2023).

Recent research using the Hi-C method and its derivatives (Lieberman-Aiden et al., 2009), which has progressed rapidly in the last decade, has revealed that the expression of each gene is affected by various *cis*-regulatory elements and their higher-ordered structures such as enhancer–promoter loops, nucleosome exclusive nonlooping insulator sequences (NENLIS) (Matsushima et al., 2019), topologically associated domains (TADs), A/B compartments, etc. (Lieberman-Aiden et al., 2009). To analyze the effects of such *cis*-regulatory structures at various scales ranging from several kb to Mb, a highly contiguous genomic sequence longer than several Mb of the target organism is required.

However, it is difficult to construct such a long contiguous scaffold that contains a transcribed gene region and its *cis*-regulatory elements such as enhancers and insulators with conventional short-read-based genome assembly. For example, the previously reported draft genome sequence of *H. pulcherrimus*, called HpulGenome_v1, did not contain the *Ars* insulator sequence located 2 kb upstream of the arylsulfatase (*Hp-Ars*) gene (Akasaka et al., 1999; Takagi et al., 2012), which has an interspecies-conserved AT-rich core region (ArsInsC) known as a typical NENLIS (Matsushima et al., 2019). Furthermore, the *Hp-Ars* gene-containing scaffold in the HpulGenome_v1 did not include another important regulatory region, the C15 enhancer (Iuchi et al., 1995; Sakamoto et al., 1997). Since repetitive sequences are dispersed throughout the genome, the presence of repeat sequences leads to difficulty in the short-read-based genome assembly. In addition, it was impossible to reconstruct long repeat sequences far exceeding several 100 bp, which frequently appear in multicellular organism genomes; therefore, many gaps had to exist in the scaffold. Thus, although it has been suggested that the *H. pulcherrimus* genome contains long repetitive sequences consisting of 10–100 of tandem sequences of genes encoding histones H1, H2A, H2B, H3, and H4, HpulGenome_v1 does not contain such long repeat regions.

However, the above problems are being overcome by technological advances in long-read sequencing, which yields sequences of 10–100 kb per read. At present, reads obtained by long-read sequencing appear to contain errors 100 times more frequently than those obtained by short-read sequencing (Logsdon et al., 2020; Rang et al., 2018). However, the hybrid assembly method, which assembles short reads from the same organism while correcting errors in long reads, may enable highly accurate genome assembly with long continuity by taking advantage of both (Holley et al., 2021).

Therefore, in this study, the *H. pulcherrimus* draft genome sequence was updated by hybrid assembly using both long reads obtained using the Oxford Nanopore sequencer and short reads used in conventional *H. pulcherrimus* draft genome assembly. In this paper, we compared the updated draft genome sequence with HpulGenome_v1, and provided the annotation of genes by comparison with related species. We also show that the updated draft genome contains the ArsInsC sequence at upstream of the *Hp-Ars* gene and its homologous sequences in the updated genome. Furthermore, we identified two distinct long tandem repeats of early histone genes.

Such improvements of the draft genome sequence will drive further progress of studies for physiological and structural–dynamical features of gene regulatory behaviors in *H. pulcherrimus*.

2 | MATERIALS AND METHODS

2.1 | Sample collection and DNA extraction

Adult males of *H. pulcherrimus* were collected from the sea around Etajima City, Hiroshima Prefecture, Japan, with permission of the Hiroshima Fishery Cooperative, and the genomic DNA (gDNA) was extracted from the sperm of three adult individuals. The *H. pulcherrimus* genomic DNA was extracted by using the Blood & Cell Culture DNA Midi Kit (QIAGEN), and short genomic DNA segments were removed by the Short Read Elimination XL Kit (Circulomics Inc.). To obtain extremely long reads, we followed the protocol described by Logsdon et al. (2020).

2.2 | Sequencing library preparation

Nanopore sequencing libraries were individually prepared from gDNA derived from each sample using the Rapid Sequencing Kit (SQK-RAD004; Oxford Nanopore Technologies [ONT]). Nanopore sequencing were performed on an ONT MinION sequencer using R9.4.1 flow cells and FAST5 files were basecalled to FASTQ files using ONT MinIT (MNT-001). In the following, each of the generated FASTQ formatted reads was named ONT-read.

2.3 | Preprocessing of illumina short reads and ONT read data

The preprocessing (adapter trimming and removal of low-quality reads) of Illumina short reads from gDNA of *H. pulcherrimus* (DRR107784, DRR107785, and DRR107786) and transcriptome (DRR107783) were performed using fastp v0.22.0 (Chen et al., 2018). Error correction of ONT read data derived from each library was individually performed using Ratatosk v0.7.0 (Holley et al., 2021) with preprocessed Illumina short reads (160.47 Gb in total).

2.4 | Genome assembly and polishing

De novo genome assembly for *H. pulcherrimus* was performed with over 5 kb error corrected and merged ONT reads using Raven v1.8.1 (Vaser & Šikić, 2021), Flye v2.8.3-b1695 (Kolmogorov et al., 2019), and Wtdbg2 v2.5 (Ruan & Li, 2020). These draft genome sequences were corrected with each other and a single consensus sequence was generated using MAECI (Lang, 2022). This consensus sequence was polished with Illumina short reads using MAECI.

2.5 | Assessment of draft genome sequences

To evaluate the completeness of the genome assembly, highly conserved genes in metazoan were searched using BUSCO v5.3 (Manni et al., 2021). As an additional analysis, predicted transcriptome models (nonredundant open reading frames [ORFs] derived from transcriptome assembly in HpBase, HpulTranscriptome_nucl.fa) were mapped to each draft genome sequence using hisat2 v2.2.1 (Kim et al., 2019).

2.6 | Gene prediction and gene annotation

The preprocessed Illumina short reads from the transcriptome were mapped to draft genome sequences using hisat2 v2.2.1 (Kim et al., 2019), and BAM formatted files were generated using SAMtools v1.17 (Li et al., 2009). Gene prediction was performed using these BAM files by BRAKER2 v2.1.4 (Brůna et al., 2021). Gene annotation (orthology search) of gene models predicted by BRAKER2 was performed using BLASTP v2.6.0 (Altschul et al., 1990) with protein models in HpBase and Echinobase (*Strongylocentrotus purpuratus* v5.0 and *Lytechinus variegatus* v3.0). In addition, homologous gene models were searched using BLASTP v2.6.0 with SwissProt and NCBI nr database.

3 | RESULTS

3.1 | Genome assembly and completeness of draft genome sequence

Sequencing by the ONT MinION sequencer generated 17.29 million reads with 51.54 Gb of sequenced data. By the error correction and filtering out ONT reads, 2.76 million reads with 35.26 Gb were generated for genome assembly. The draft genome sequence was updated

by assembling and polishing these reads by Raven, Flye, Wtdbg2, and MAECI (Table 1, Table S1). Here, the contig (contig ID: Utg196084) that was most homologous to the mitochondrial genome of *H. pulcherrimus* [NC_023771.1] by BLASTN was removed. The updated draft genome sequence consists of 2,163 contigs with a total length of ~626.4 Mb and N50 length of ~515.7 kb (Table S1), with a smaller number of contigs and more continuously assembled than HpulGenome_v1; HpulGenome_v1 consists of ~16,000 scaffolds (contigs) with total length of ~600 Mb containing a total of ~100 Mb gaps and N50 = 142.6 kb (Table S1).

Analysis of genome completeness of the updated draft genome by BUSCO using the metazoa_odb10 showed a larger value of 96.5% (89.9% single copy and 6.6 duplicated) than that of HpulGenome_v1 (86.1% [84.8% single copy and 1.3 duplicated]). Additionally, the mapping ratio of transcriptome models of the present draft genome also obtained a larger value of 77.78% than that of HpulGenome_v1 (55.35%).

3.2 | Gene prediction based on genome assembly and gene annotation with related species

The analysis by BRAKER2 and transcriptome data (DRR107783) predicted 46,914 genes exist in the updated draft genome sequence. By eliminating the predicted genes that were too short (less than 50 amino acids), contained multiple stop codons, or were encoded on the mitochondrial contig, 46,826 gene models were obtained in the updated draft genome, with 5.85 exons per gene model and an average coding sequence (CDS) length of 1,249.23 bp (Table S2). The inference from the orthology relationships between the present gene models and gene models in other sea urchins through reciprocal BLASTP searches showed that 36,055 gene models (number of reciprocal best hit pair, 20,434 and number of not reciprocal but best hit,

TABLE 1 Summary of genome assembly, BUSCO completeness and gene annotation in HpulGenome_v1 and our draft genome sequence.

Version of genome assembly	Updated draft genome	HpulGenome_v1
Assembly size ^a	626.4 Mb	568.9 Mb
No. contigs ^a	2,163	86,128
N50 contig length ^a	515.7 kb	9.641 kb
No. scaffolds ^a	2,163	16,251
N50 scaffold length ^a	515.7 kb	142.6 kb
N (%) ^a	0	16.41
GC-content (%) ^a	37.12	36.72
BUSCO completeness (%) (metazoa_odb10: 954 genes)	Complete: 96.5	Complete: 86.1
	Duplicated: 6.6	Duplicated: 1.3
	Fragmented: 2.1	Fragmented: 9.9
	Missing: 1.4	Missing: 4.0
Mapping ratio of transcriptome models (%) (20,564 sequences)	75.65 (aligned exactly 1 time)	54.64 (aligned exactly 1 time)
	2.13 (aligned >1 times)	0.71 (aligned >1 times)

^aCalculated from BBtools (Bushnell et al., 2017) (stats.sh).

15,621) were homologous (best hit with $e\text{-value} \leq 1e-10$) to publicly known gene models of *H. pulcherrimus*, *S. purpuratus*, or *L. variegatus* (Tables 2 and S2).

3.3 | Detection of early histone gene loci with long tandem repeats

The *H. pulcherrimus* genome has been suggested to contain two non-allelic early histone gene loci that are the long repetitive sequences consisting of 10–100 of tandem sequences of genes encoding histones (Table S3) (Matsushita et al., 2017). However, such long tandem repeat sequences could not be detected by recent short-read-based genome assemblies; in HpulGenome_v1, BLASTN searches did not find any genomic regions homologous even to single repeat sequences of early histone loci.

In contrast, two contigs containing the regions of long tandem repeats of histone genes were found on the updated draft genome sequence; where one contig (contig ID: Utg198178, with 479,333 bp) involved 47 tandem repeats, and the other contig (contig ID: Utg200276, with 127,611 bp) involved 34 tandem repeats (Table S3).

3.4 | Regulatory sequences of arylsulfatase gene

We have investigated the regulatory mechanism of the transcription of the arylsulfatase (*Hp-Ars*) gene. The promoter (−252 to +38) is the minimum region required for temporal expression (Iuchi et al., 1995; Morokuma et al., 1997). The C15 enhancer required for large enhancement of the expression is present in the first intron (Iuchi et al., 1995; Sakamoto et al., 1997). The polypyrimidine sequence (−2,201 to −1,680) that can form an intramolecular triplex structure (Sakamoto et al., 1996; Yamamoto et al., 1994) and the *Ars* insulator (−2,686 to −2,109) (Akasaka et al., 1999) are present in the upstream region.

In HpulGenome_v1, scaffold989 contains the the *Hp-Ars* gene. A BLASTN search revealed that this scaffold includes all exons, the

promoter, and upstream sequences to the polypyrimidine region. However, it does not include the C15 enhancer in the first intron and the *Ars* insulator in the upstream flanking region, which play important roles in the transcriptional regulation.

On the other hand, in the updated draft genome, the *Hp-Ars* gene-containing contig (contig ID: Utg200732) contains all exons and regulatory sequences including *Ars* insulator and C15 enhancer. Furthermore, the *Hp-Ars* gene has been reported to include various types of repetitive sequences such as two direct repeats (DIR, which are called DIR1 and DIR2, respectively) and an inverted repeat (INV) in the upstream region (Akasaka et al., 1994). All of these repetitive sequences could be found in contig Utg200732. Interestingly, when we searched homologous sequences of the inverted repeat (−475 to −217) with 90–120 bp of the unit length, >90% identity and >90% query coverage among inverted repeats were obtained from IRF v3.08 (Warburton et al., 2004), and nine homologous inverted repeat sequences were found in the updated draft genome (Table S4). Although seven out of these homologous inverted repeats were present in intergenic regions, there was no tendency in their position. Furthermore, eight homologous sequences of the direct repeat DIR1 (−3,440 to −3,109) with two repeat units, >90% identity, and >95% query coverage were found, and 205 homologous sequences of another direct repeat DIR2 (−3,096 to −2,592) with two to five repeat units, >90% identity, and >95% query coverage were found in the updated draft genome (Table S5). These DIR1 and DIR2 homologs tend to exist in the upstream region of the nearest genes (Figure 1).

3.5 | Homologous genome region to *Ars* insulator

The *Hp-Ars* gene has been known to involve the *Ars* insulator, which includes the interspecifically conserved *ArsInsC* sequence (Akasaka et al., 1999; Takagi et al., 2012) (Table S4) that is a typical example of NENLIS (Matsushima et al., 2019). In HpulGenome_v1, the *Hp-Ars* gene was present, but the *ArsInsC* sequence was not present. In contrast, the updated draft genome contained *ArsInsC* from 46,417 to

TABLE 2 Summary of BLASTP searches with protein models of HpBase and other sea urchin.

Subject	Reciprocal best hit pair	Not reciprocal but best hit	No hit
<i>H. pulcherrimus</i>	15,652	16,847	14,327
(HpBase, HpulGenome_v1_prot.fa)	(33.43%)	(35.98%)	(30.60%)
<i>S. purpuratus</i>	16,153	17,190	13,483
(Echinobase, sp5_0_GCF_proteins.fa)	(34.50%)	(36.71%)	(28.79%)
<i>L. variegatus</i>	14,802	16,613	15,411
(Echinobase, Lvar3_0_GCF_proteins.fa)	(31.61%)	(35.48%)	(32.91%)
<i>P. lividus</i> *	14,958	18,388	13,480
	(31.94%)	(39.27%)	(28.70%)
<i>H. pulcherrimus</i> , <i>S. purpuratus</i> , or <i>L. variegatus</i>	20,434	15,621	10,771
	(43.64%)	(33.36%)	(23.00%)

*Proteins fasta file of *Paracentrotus lividus* was created using gffread v0.12.7 (Pertea & Pertea, 2020). The genome and annotation in GTF format of *P. lividus* were downloaded from Zenodo (<https://zenodo.org/record/7459274>).

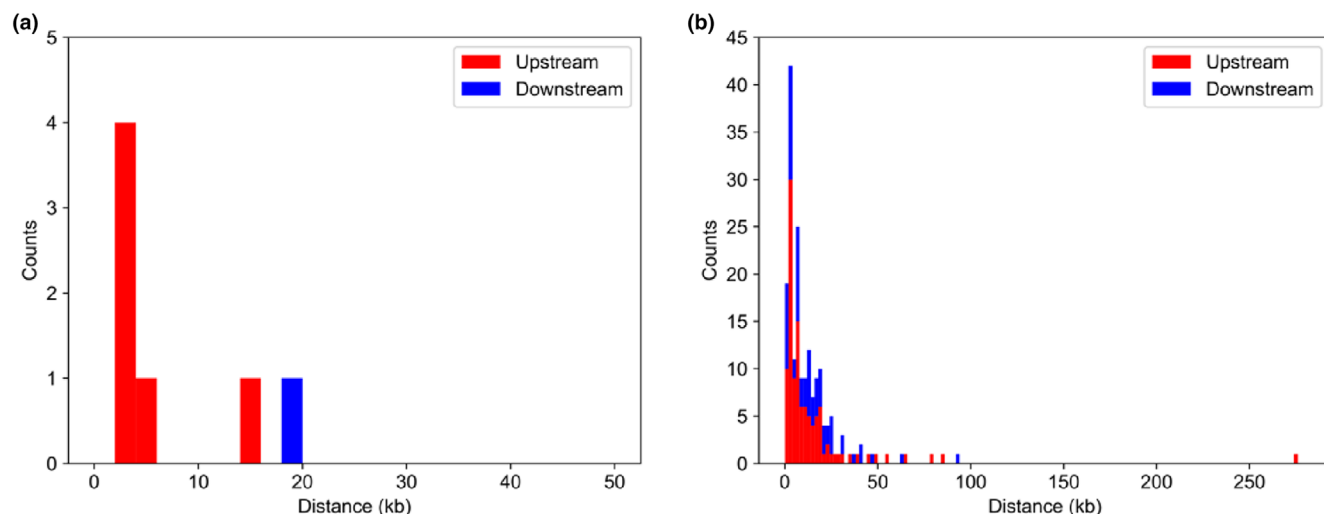


FIGURE 1 Stacked histogram of genomic distances between DIR1 homologs and their nearest gene (a), and that between DIR2 homologs and their nearest gene (b). The height of red and blue bars indicates counts of direct repeats (DIR) homologs at the upstream and downstream region of the nearest gene, respectively. Distances between DIR homolog and its nearest gene was defined as the genomic distance between the centers of them.

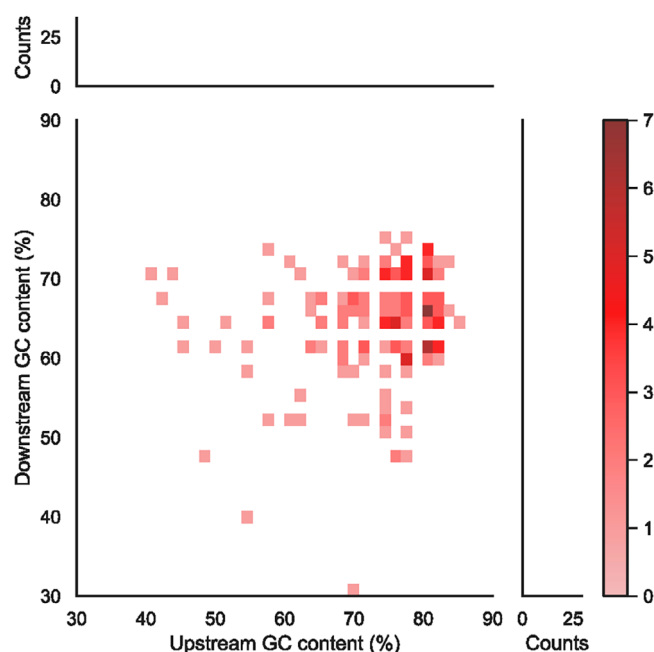


FIGURE 2 Heat map of the number of ArslnsC homologs with each GC content in the upstream 50 bp and downstream 50 bp regions. The histograms represent the projection onto the upstream GC content axis and the downstream GC content axis, respectively.

46,236 (minus strand) in contig ID: Utg200732, which is 2-kb upstream of the *Hp-Ars* gene region. ArslnsC was originally identified by interspecific sequence comparison with 75% identity in the 30 bp of the sliding window and does not have specific motifs except its interspecies-conserved AT-rich sequence (Takagi et al., 2012). Therefore, to find candidate sequences for functional homologs of ArslnsC, nucleotide sequences that have more than 90% identity to ArslnsC were searched. BLASTN analysis found 185 sequences that have

more than 90% identity to ArslnsC (Table S6). As observed in the *Ars* insulator, the 50-bp regions surrounding the ArslnsC homologous sequences tended to have high GC contents (50%–80% GC) (Figures 2 and S1), and 121 homologous sequences (65.4%) contained a G(C)-stretch of more than 8 bp in these GC-rich regions (Figure S1). Thus, the sequence properties of the *Ars* insulator are well conserved in these homologous sequences, suggesting that these homologous sequences may function as insulators in the sea urchin genome.

3.6 | Short tandem repeat

It has been reported that short tandem repeats (STRs) are enriched in gene promoters and variable repeat number is associated with gene expression (Sawaya et al., 2013; Vences et al., 2009). Furthermore, approximately 90% of transcription factors preferentially bind to STRs (Horton et al., 2023). Therefore, we also analyzed the presence of STRs in the updated draft genome using TRF v4.09 (Benson, 1999). According to the definition of microsatellite by Sawaya et al. (2013), uninterrupted consecutively repeated units of one to four nucleotides that are at least 12 bp in length were searched. After searching STRs that obtained from TRF results and matched this definition, a total of 111,553 mononucleotide repeats, 66,665 dinucleotide repeats, 61,983 trinucleotide repeats, and 62,448 tetranucleotide repeats were found in the updated draft genome (Table S7). However, the enrichment of STRs in the proximity of genes or the translation start sites was not observed.

3.7 | Transposable elements prediction

Transposable elements (TEs) were predicted and classified using RepeatModeler v2.0.5 (Smit, AFA, Hubley, R. *RepeatModeler Open-*

1.0. 2008–2015, <http://www.repeatmasker.org>) and RepeatMasker v4.1.5 (Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-4.0*. 2013–2015, <http://www.repeatmasker.org>) for the updated draft genome of *H. pulcherrimus*, and the genomes of *S. purpuratus*, *L. variegatus*, and *P. lividus*, respectively. Most TEs were not annotated as known elements, but the classification of TEs and their proportions indicated that the trends of TEs in the updated draft genome sequence were closer to those in other sea urchins than in HpulGenome_v1 (Table S8).

3.8 | Microsynteny plots

Microsynteny plots were plotted for the five fragment sets of *H. pulcherrimus* contig and *S. purpuratus* scaffold in order of having the most reciprocal BLASTP best hit pairs (Figure S2). While genome-wide synteny plots showed little interchromosomal gene order rearrangement (data not shown), microsynteny plots showed intrachromosomal gene order rearrangement, similar to the results of genome-scale synteny plots in other sea urchins (Marletaz et al., 2023).

4 | DISCUSSION

Long-read sequencing and hybrid assembly using long and short reads were performed to update the *H. pulcherrimus* draft genome with more continuous and higher accuracy than the recently reported HpulGenome_v1. This study resulted in a draft genome of *H. pulcherrimus* with higher values of the metrics by BUSCO and higher mapping rate of the transcriptome model than HpulGenome_v1.

This draft genome, involving 46,826 gene models, is larger than HpulGenome_v1 containing 24,860 model genes. Here, 36,055 of these gene models were annotated based on the analysis with the homologous gene models of related sea urchins.

The present draft genome contained two long repeat sequences, including the tandemly connected early histone genes. This result may drive the progress of genomic and epigenomic analyses to reveal the developmental stage-dependent changes in intranuclear location and interactions of early histone loci (Matsushita et al., 2017).

Additionally, this draft genome also contained Arslnc sequence and more than 100 Arslnc homologous sequences. Furthermore, the number of Arslnc homologous sequences with more than 75% identity in the updated draft genome was 241. In vertebrates, the CTCF-binding sequence is known as a typical insulator sequence responsible for gene expression control. Recently, however, it has been suggested that sea urchin CTCF functions during mitosis rather than interphase (Watanabe et al., 2023). Therefore, Arslnc and its homologous sequences may play an important role in the control mechanism of sea urchin interphase gene expression, and the present draft genome may promote research to clarify the mechanism.

The updated draft genome involved more than 10,000 non-annotated genes. These genes were expected to include

noncoding RNA-derived or *H. pulcherrimus*-specific genes. The annotation of these gene models should be progressed by the comparison among genes in various organisms and the function analysis by experiments in the future. Additionally, a large number of duplicated genes were found by analysis with BUSCO in the present draft genome compared with HpulGenome_v1. The reason for this was not clear, nor whether each duplication is true in *H. pulcherrimus*. Further validation should be a future research direction.

Sawaya et al. (2013) suggested that enrichment of STRs in promoters by calculating distances from transcription start site to STR. However, since gene model estimation based on mapping results of bulk RNA-sequencing did not show the transcription start site and the distribution of the 5'-UTR length in the *H. pulcherrimus* genome, we analyzed the distribution of distances from translation start sites to STRs. Although the enrichment of STRs in the proximity of genes and translation start site was not observed in this study, it is necessary to examine the enrichment of STRs in gene promoters after the estimation of transcription start site from other sequencing methods such as CAGE-sequencing or Iso-sequencing.

In this study, HpulGenome_v1 was updated using Oxford Nanopore long-read sequencing, resulting in a smaller number of contigs with more continuous assembly. Since each contig was expected to contain a complete set of *cis*-regulatory sequences, we believe that this updated draft genome improves the applicability of *H. pulcherrimus* genome information to various genome-wide analyses.

HpulGenome_v1 was assembled using genomic DNA isolated from *H. pulcherrimus* collected in Shimoda, Shizuoka, Japan (Kinjo et al., 2018). However, the sea urchin genome is highly polymorphic (Britten et al., 1978; Sodergren et al., 2006; Yamamoto et al., 2007). For genome-wide analysis of *H. pulcherrimus*, a genomic sequence that can be widely used is required. Therefore, to perform genome-wide analysis reflecting the degree of polymorphism in *H. pulcherrimus*, we used genomic DNA isolated from multiple individuals of *H. pulcherrimus* collected in Hiroshima. After error correction and genome assembly, there were 1,625 mixed bases on the updated draft genome sequence (Table S9), and the draft genome size was restrained to ~600 Mb, although the estimated genome size is ~800 Mb (Kinjo et al., 2018). We expect that excessive redundancy is suppressed, and the updated draft genome represents polymorphism in *H. pulcherrimus*. On the other hand, genes determined to be duplicated in BUSCO and the contigs containing them are likely to be redundant. However, we did not remove these contigs because they may contain important information such as polymorphic regions when editing the genome of *H. pulcherrimus*. In addition, it is difficult to perform quantitative haplotype-resolved analysis based on *k*-mer distribution with raw reads obtained from the Oxford Nanopore sequencer because of the high error rate of sequencing by this sequencer. Thus, we plan to correct the redundancy of draft genome sequences by, for example, using the Hi-C method, in the future.

ACKNOWLEDGMENTS

This work was supported by a Grant-in-Aid for Scientific Research (KAKENHI) (grant number 21K06124) from the Japan Society for the Promotion of Science. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

DATA AVAILABILITY STATEMENT

The sequence data of the updated genome assembly were featured on HpBase (<https://cell-innovation.nig.ac.jp/Hpul/>) with the name HpulGenome_kure_v1. The raw sequencing data of whole genome sequencing was deposited in the DDBJ Sequence Read Archive (DRA) with accession DRA017089, and was also submitted to DDBJ/EMBL/GenBank databases (BioProject accession PRJDB16611).

ORCID

Shunsuke Yaguchi  <https://orcid.org/0000-0002-8326-5762>

Naoaki Sakamoto  <https://orcid.org/0000-0003-0546-4018>

Akinori Awazu  <https://orcid.org/0000-0002-0778-021X>

REFERENCES

- Akasaka, K., Nishimura, A., Takata, K., Mitsunaga, K., Mibuka, F., Ueda, H., ... Shimada, H. (1999). Upstream element of the sea urchin arylsulfatase gene serves as an insulator. *Cellular and Molecular Biology*, 45(5), 555–565. <http://www.ncbi.nlm.nih.gov/pubmed/10512188>
- Akasaka, K., Sakamoto, N., Yamamoto, T., Morokuma, J., Fujikawa, N., Takata, K., ... Shimada, H. (1994). Corrected structure of the 5' flanking region of arylsulfatase gene of the sea urchin, *Hemicentrotus pulcherrimus* 5' flanking sequence/sea urchin/aryl sulfatase gene/G-string. *Development, Growth & Differentiation*, 36(6), 633–636. <https://doi.org/10.1111/j.1440-169X.1994.00633.x>
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
- Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research*, 27(2), 573–580. <https://doi.org/10.1093/nar/27.2.573>
- Britten, R. J., Cetta, A., & Davidson, E. H. (1978). The single-copy DNA sequence polymorphism of the sea urchin *strongylocentrotus purpuratus*. *Cell*, 15(4), 1175–1186. [https://doi.org/10.1016/0092-8674\(78\)90044-2](https://doi.org/10.1016/0092-8674(78)90044-2)
- Brůna, 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), 1–11. <https://doi.org/10.1093/nargab/lqaa108>
- Bushnell, B., Rood, J., & Singer, E. (2017). BBMerge – Accurate paired shotgun read merging via overlap. *PLoS One*, 12(10), 1–15. <https://doi.org/10.1371/journal.pone.0185056>
- Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17), i884–i890. <https://doi.org/10.1093/bioinformatics/bty560>
- Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Caestani, C., Yuh, C. H., ... Bolouri, H. (2002). A genomic regulatory network for development. *Science*, 295(5560), 1669–1678. <https://doi.org/10.1126/science.1069883>
- Davidson, P. L., Guo, H., Wang, L., Berrio, A., Zhang, H., Chang, Y., ... Wray, G. A. (2020). Chromosomal-level genome assembly of the sea urchin *Lytechinus variegatus* substantially improves functional genomic analyses. *Genome Biology and Evolution*, 12(7), 1080–1086. <https://doi.org/10.1093/GBE/EVAA101>
- Holley, G., Beyter, D., Ingimundardottir, H., Møller, P. L., Kristmundsdottir, S., Eggertsson, H. P., & Halldorsson, B. V. (2021). Ratatosk: Hybrid error correction of long reads enables accurate variant calling and assembly. *Genome Biology*, 22(1), 28. <https://doi.org/10.1186/s13059-020-02244-4>
- Horton, C. A., Alexandari, A. M., Hayes, M. G. B., Marklund, E., Schaepe, J. M., Aditham, A. K., Shah, N., Suzuki, P. H., Shrikumar, A., Afek, A., Greenleaf, W. J., Gordân, R., Zeitlinger, J., Kundaje, A., & Fordyce, P. M. (2023). Short tandem repeats bind transcription factors to tune eukaryotic gene expression. *Science*, 381(6664), eadd1250. <https://doi.org/10.1126/science.add1250>
- Iuchi, Y., Morokuma, J., Akasaka, K., & Shimada, H. (1995). Detection and characterization of the cis-element in the first intron of the *Ars* gene in the sea urchin. *Development, Growth and Differentiation*, 37(4), 373–378. <https://doi.org/10.1046/j.1440-169X.1995.t01-3-00003.x>
- Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, 37(8), 907–915. <https://doi.org/10.1038/s41587-019-0201-4>
- Kinjo, S., Kiyomoto, M., Suzuki, H., Yamamoto, T., Ikeo, K., & Yaguchi, S. (2022). TrBase: A genome and transcriptome database of *Temnopleurus reevesii*. *Development Growth and Differentiation*, 64(4), 210–218. <https://doi.org/10.1111/dgd.12780>
- Kinjo, S., Kiyomoto, M., Yamamoto, T., Ikeo, K., & Yaguchi, S. (2018). HpBase: A genome database of a sea urchin, *Hemicentrotus pulcherrimus*. *Development Growth and Differentiation*, 60(3), 174–182. <https://doi.org/10.1111/dgd.12429>
- Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37(5), 540–546. <https://doi.org/10.1038/s41587-019-0072-8>
- Lang, J. (2022). MAECI: A pipeline for generating consensus sequence with nanopore sequencing long-read assembly and error correction. *PLoS One*, 17(5), e0267066. <https://doi.org/10.1371/journal.pone.0267066>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., ... Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950), 289–293. <https://doi.org/10.1126/science.1181369>
- Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, 21(10), 597–614. <https://doi.org/10.1038/s41576-020-0236-x>
- Manni, M., Berkeley, M. R., Seppely, M., & Zdobnov, E. M. (2021). BUSCO: Assessing genomic data quality and beyond. *Current Protocols*, 1(12), e323. <https://doi.org/10.1002/cpz1.323>
- Marletaz, F., Couloux, A., & Poulain, J. (2023). Analysis of the *P. lividus* sea urchin genome highlights contrasting trends of genomic and regulatory evolution in deuterostomes. *Cell Genomics*, 3, 100295. <https://doi.org/10.1016/j.xgen.2023.100295>
- Matsushima, Y., Sakamoto, N., & Awazu, A. (2019). Insulator activities of nucleosome-excluding DNA sequences without bound chromatin looping proteins. *Journal of Physical Chemistry B*, 123(5), 1035–1043. <https://doi.org/10.1021/acs.jpcc.8b10518>
- Matsushita, M., Ochiai, H., Suzuki, K. i. T., Hayashi, S., Yamamoto, T., Awazu, A., & Sakamoto, N. (2017). Dynamic changes in the interchromosomal interaction of early histone gene loci during development of sea urchin. *Journal of Cell Science*, 130(24), 4097–4107. <https://doi.org/10.1242/jcs.206862>
- Morokuma, J., Akasaka, K., Mitsunaga-Nakatsubo, K., & Shimada, H. (1997). A cis-regulatory element within the 5' flanking region of arylsulfatase gene of sea urchin, *Hemicentrotus pulcherrimus*. *Development, Growth and Differentiation*, 39(4), 469–476. <https://doi.org/10.1046/j.1440-169X.1997.t01-3-00008.x>

- Oliveri, P., & Davidson, E. H. (2004). Gene regulatory network controlling embryonic specification in the sea urchin. *Current Opinion in Genetics and Development*, 14(4), 351–360. <https://doi.org/10.1016/j.gde.2004.06.004>
- Pertea, G., & Pertea, M. (2020). GFF utilities: GffRead and GffCompare [version 2; peer review: 3 approved]. *F1000Research*, 9(304), 1–20. <https://f1000research.com/articles/9-304/v2>
- Rang, F. J., Kloosterman, W. P., & de Ridder, J. (2018). From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. *Genome Biology*, 19(1), 90. <https://doi.org/10.1186/s13059-018-1462-9>
- Ruan, J., & Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nature Methods*, 17(2), 155–158. <https://doi.org/10.1038/s41592-019-0669-3>
- Sakamoto, N., Akasaka, K., Mitsunaga-Nakatsubo, K., Takata, K., Nishitani, T., & Shimada, H. (1997). Two isoforms of orthodenticle-related proteins (HpOtx) bind to the enhancer element of sea urchin arylsulfatase gene. *Developmental Biology*, 181(2), 284–295. <https://doi.org/10.1006/dbio.1996.8455>
- Sakamoto, N., Akasaka, K., Yamamoto, T., & Shimada, H. (1996). A triplex DNA structure of the polypyrimidine: Polypurine stretch in the 5' flanking region of the sea urchin arylsulfatase gene. *Zoological Science*, 13(1), 105–109. <https://doi.org/10.2108/zsj.13.105>
- Sawaya, S., Bagshaw, A., Buschiazio, E., Kumar, P., Chowdhury, S., Black, M. A., & Gemmell, N. (2013). Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. *PLoS One*, 8(2), e54710. <https://doi.org/10.1371/journal.pone.0054710>
- Sodergren, E., Weinstock, G. M., Davidson, E. H., Cameron, R. A., Gibbs, R. A., Angerer, R. C., ... Wright, R. (2006). The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science*, 314(5801), 941–952. <https://doi.org/10.1126/science.1133609>
- Takagi, H., Inai, Y., Watanabe, S. I., Tatemoto, S., Yajima, M., Akasaka, K., ... Sakamoto, N. (2012). Nucleosome exclusion from the interspecies-conserved central AT-rich region of the *Ars* insulator. *Journal of Biochemistry*, 151(1), 75–87. <https://doi.org/10.1093/jb/mvr118>
- Takemoto, A., Miyamoto, T., Simono, F., Kurogi, N., Shirae-Kurabayashi, M., Awazu, A., ... Sakamoto, N. (2016). Cilia play a role in breaking left–right symmetry of the sea urchin embryo. *Genes to Cells*, 21(6), 568–578. <https://doi.org/10.1111/gtc.12362>
- Vaser, R., & Šikić, M. (2021). Raven: A de novo genome assembler for long reads. *Nature Computational Science*, 1(5), 332–336. <https://doi.org/10.1038/s43588-021-00073-4>
- Vinces, M. D., Legendre, M., Caldara, M., Hagihara, M., & Verstrepen, K. J. (2009). Unstable tandem repeats in promoters confer transcriptional evolvability. *Science*, 324(5931), 1213–1216. <https://doi.org/10.1126/science.1170097>
- Warburton, P. E., Giordano, J., Cheung, F., Gelfand, Y., & Benson, G. (2004). Inverted repeat structure of the human genome: The X-chromosome contains a preponderance of large, highly homologous inverted repeated that contain testes genes. *Genome Research*, 14(10a), 1861–1869. <https://doi.org/10.1101/gr.2542904>
- Watanabe, K., Fujita, M., Okamoto, K., Yoshioka, H., Moriwaki, M., Tagashira, H., ... Sakamoto, N. (2023). The crucial role of CTCF in mitotic progression during early development of sea urchin. *Development, Growth & Differentiation*, 65(7), 395–407. <https://doi.org/10.1111/dgd.12875>
- Yaguchi, S., & Katow, H. (2003). Expression of tryptophan 5-hydroxylase gene during sea urchin neurogenesis and role of serotonergic nervous system in larval behavior. *Journal of Comparative Neurology*, 466(2), 219–229. <https://doi.org/10.1002/cne.10865>
- Yamamoto, T., Akasaka, K., Irie, S., & Shimada, H. (1994). A long polypyrimidine: Polypurine sequence in 5' flanking region of arylsulfatase gene of sea urchin embryo. *International Journal of Developmental Biology*, 38(2), 337–344.
- Yamamoto, T., Kawamoto, R., Fujii, T., Sakamoto, N., & Shibata, T. (2007). DNA variations within the sea urchin *Otx* gene enhancer. *FEBS Letters*, 581(27), 5234–5240. <https://doi.org/10.1016/j.febslet.2007.10.013>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Komoto, T., Ikeo, K., Yaguchi, S., Yamamoto, T., Sakamoto, N., & Awazu, A. (2024). Assembly of continuous high-resolution draft genome sequence of *Hemicentrotus pulcherrimus* using long-read sequencing. *Development, Growth & Differentiation*, 66(4), 297–304. <https://doi.org/10.1111/dgd.12924>