

1
2 **Data note**
3
4
5

6 **Draft genome of the sea cucumber *Apostichopus japonicus* and genetic**
7 **polymorphism among color variants**
8
9

10
11
12
13
14
15
16
17 Jihoon Jo^{a#}, Jooseong Oh^{a#}, Hyun-Gwan Lee^b, Hyun-Hee Hong^a, Sung-Gwon Lee^a,
18 Seongmin Cheon^a, Elizabeth M. A. Kern^c, Soyeong Jin^c, Sung-Jin Cho^{d*}, Joong-Ki Park^{c*},
19 and Chungoo Park^{a*}
20
21
22

23
24 ^a School of Biological Sciences and Technology, Chonnam National University,
25 Gwangju 61186, Republic of Korea
26

27
28 ^b Marine Ecological Disturbing and Harmful Organisms Research Center, Department
29 of Oceanography, Chonnam National University, Gwangju 61186, Republic of Korea
30

31 ^c Division of EcoScience, Ewha Womans University, Seoul 03760, Republic of Korea.
32

33 ^d Department of Biology, College of Natural Sciences, Chungbuk National University,
34 Cheongju, Chungbuk 28644, Republic of Korea
35

36
37 [#] These authors equally contributed this work.
38
39
40
41
42
43
44
45
46
47

48 *Corresponding Authors.
49

50 E-mail addresses:

51 Chungoo Park, chungoo@jnu.ac.kr. Tel: +82-62-530-1913. Fax: +82-62-530-2199
52

53 Joong-Ki Park, jpark@ewha.ac.kr. Tel: +82-2-3277-5948. Fax: +82-2-3277-2385.
54

55 Sung-Jin Cho, sjchobio@chungbuk.ac.kr. Tel: +82-43-261-2294. Fax: +82-43-260-2298.
56
57
58
59
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Abstract

Background: The Japanese sea cucumber (*Apostichopus japonicus* Selenka 1867) is an economically important species as a source of seafood and ingredient in traditional medicine. It is mainly found off the coasts of northeast Asia. Recently, substantial exploitation and widespread biotic disease in *A. japonicus* have generated increasing conservation concern. However, the genomic knowledge base and resources available for researchers to use in managing this natural resource and to establish genetically based breeding systems for sea cucumber aquaculture are still in a nascent stage.

Findings: A total of 312 gigabases (Gb) of raw sequences were generated using the Illumina HiSeq 2000 platform and assembled to a final size of **0.66 Gb** which is about **80.5 %** of the estimated genome size (0.82 Gb). We observed nucleotide-level heterozygosity within the assembled genome to be 0.986 %. The resulting draft genome assembly comprising 132,607 scaffolds with an N50 value of 10.5 kb contains a total of 21,771 predicted protein-coding genes. We identified 6.6 – 14.5 million heterozygous SNPs in the assembled genome of the three natural color variants (green, red, and black), resulting in an estimated nucleotide diversity of 0.00146.

Conclusions: We report the first draft genome of *A. japonicus* and provide a general overview of the genetic variation in the three major color variants of *A. japonicus*. These data will help provide a comprehensive view of the genetic, physiological, and evolutionary relationships among color variants in *A. japonicus*, and will be invaluable resources for sea cucumber genomic research.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Keywords: Sea cucumber genome, *Apostichopus japonicus*, Color variants, Genetic variation, Population genomics

Data description

Background information on *A. japonicus*

The class Holothuroidea (also known as sea cucumbers) belongs to the phylum Echinodermata and comprises approximately 1,250 recorded species worldwide, including some species that are of commercial and medical value [1, 2]. *Apostichopus japonicus* Selenka 1867 is one of the well-known, commercially important sea cucumber species and occurs in the northwestern Pacific coast including China, Japan, Korea and the Far Eastern seas. This species exhibits a wide array of dorsal/ventral color variants (in particular green, red, and black; Fig 1), which differ in their biological and morphological attributes (e.g., shape of ossicle, habitat preference, spawning period, and polian vesicles) [1, 3]. **The red variant is found on rock pebbles and gravel substrate and has higher salinity and temperature tolerance than the other color variants [4, 5]. Green and black variants are found on sandy and muddy bottoms at shallower depths, and the green variant has greater plasticity in thermotolerance than the red variant [6, 7].**

Recently, overexploitation and the prevalence of biotic disease (viral infections) in sea cucumber aquaculture have generated increasing conservation concern [8, 9]. However, the genomic knowledge base and resources available to researchers for use in managing this natural resource or establishing genetically based breeding systems are still in a nascent stage [10].

Sample collection and genomic DNA extraction

Specimens of the three color *A. japonicus* variants (green, red, and black) were collected from same geographical location (GPS data: 34.1 N, 127.18E, Geomun-do,

1
2 Yeosu, Republic of Korea). Genomic DNA of each color variant was extracted manually
3
4 from body wall tissues of single male specimens. Briefly, we ground the tissues to fine
5
6 powder using mortar and pestle with liquid nitrogen freezing. Tissue powders were
7
8 digested for 1 hour at 65 °C in CTAB (Cetyltrimethylammonium bromide) lysis buffer
9
10 (2% CTAB, 1.4 M NaCl, 20 mM EDTA, 100 mM Tris-HCl, and pH 8.0), followed by
11
12 Phenol/Chloroform extraction and ethanol precipitation.
13
14
15
16
17
18

19 **Sequencing and quality control**

20
21 Using the standard protocol provided by Illumina (San Diego, USA), we
22
23 constructed both short-insert (180 and 400 bp) and long-insert (2 kb) libraries for 2 x
24
25 101 bp paired-end reads, which were sequenced using a HiSeq 2000 instrument. For the
26
27 green color variant, a total of 225 Gb of raw data was generated from all three libraries.
28
29 In the case of the red and black color variants, 40 and 47 Gb of raw reads, respectively,
30
31 were produced by 400 bp short-insert library. The raw reads were preprocessed using
32
33 Trimmomatic v0.33 [11] and Trim Galore [12], in which reads containing adapter
34
35 sequences, poly-N sequences, or low quality bases (below a mean Phred score of 20)
36
37 were removed. To correct for errors in the raw sequences, we used ALLPATHS-LG
38
39 v52488 [13]. Approximately 208, 39, and 42 billion clean reads were obtained for
40
41 green, red, and black color variant samples, respectively (Table 1). The *A. japonicus*
42
43 genome size was estimated to be approximately 0.9 Gb based on k-mer measurement
44
45 (Fig 2), which is fully consistent with genome size measured by flow cytometry (~ 0.82
46
47 Gb) [14]. Based on this estimation, the clean sequence reads correspond to about 356-
48
49 fold coverage of the *A. japonicus* genome.
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Assembly

For whole-genome assembly, we used reads only from green color variant libraries and employed Platanus v1.2.4 [15], which is well suited for high-throughput short reads and heterozygous diploid genomes. Briefly, error corrected paired-end (insert size: 180 bp and 400 bp) reads were input for contig assembly. Next, all cleaned paired-end (insert size: 180 bp and 400 bp) and mate-paired (insert size: two 2 kb samples) reads were mapped onto the contigs for scaffold building and were utilized for gap filling (any nucleotide represented by “N” in scaffolds). After gap filling by Platanus, the gaps that still remained in the resulting scaffolds were closed using GapCloser (a module of SOAPdenovo2 [16]). The final genome assembly was 0.66 Gb in total length, which is about 80.5 % of the estimated genome size by flow cytometry (0.82 Gb) [14], and is composed of 132,607 scaffolds and un scaffolded contigs (≥ 1 kb) with an N50 value of 10.5 kb (Table 2). We assessed the completeness of the assembly using CEGMA v2.4.010312 [17] and BUSCO v1.22 [18]. 73.4% of the core eukaryotic genes (based on the 248 core essential genes) and 60.7% of the metazoan single-copy orthologs (based on the 843 genes), respectively were identifiable in the genome. Because assembling highly heterozygous genomes is a major challenge in *de novo* genome sequencing, we further sought to explore whether there are other assemblers that could produce better genome assembly statistics. We applied two popular genome assemblers, SOAPdenovo2 2.04-r240 [16] and ALLPATHS-LG v52488 [13], and as expected [15], the Platanus assembler was superior to the others (Table S1).

Annotation

To identify genomic repeat elements in the *A. japonicus* genome assembly, we

1
2 ran RepeatMasker (version 4.0.6) [19] using the Repbase TE library (release 20150807)
3
4 [20] and the *de novo* repeat library constructed by RepeatModeler (version 1.0.8) [21].
5
6 Approximately 27.2% of the *A. japonicus* genome was identified as interspersed
7
8 repeats.
9
10

11 Protein-coding genes were predicted using four steps. First, *ab initio* gene
12 prediction was performed with trained AUGUSTUS v3.2.1 [22] using hints from
13 splicing alignment of transcripts to the repeat-masked assembled genome with BLAT
14 [23] and PASA v2.0.2 [24]. To obtain high quality spliced alignments of expressed
15 transcript sequences for the AUGUSTUS training set, we collected RNA-seq data from
16 our previous [25] (from body wall tissue of adult stage specimens) and other
17 transcriptome (from embryo, larva, and juvenile stages [developmental-stage specific];
18 from gonads, intestines, respiratory trees, and coelomic fluid of adults [tissue-specific])
19 [26] studies, and assembled reads from the RNA-seq dataset using Trinity v2.1.1 [27].
20
21 Second, for homology-based gene prediction, homologous proteins in other species
22 (from UniProt [28]) were mapped to the repeat-masked assembled genome using
23 tBLASTn [29] with an *E*-value $\leq 1 \times 10^{-5}$. The aligned sequences were predicted using
24 GeneWise v2.4.0 [30] to search for precise spliced alignment and gene structures. Third,
25 for homology-based gene prediction with transcriptome evidence, existing RNA-seq
26 reads [23, 25] were mapped to the repeat-masked assembled genome using TopHat
27 v2.1.0 [31], and gene models were built using Cufflinks v2.2.1 [32]. Finally, the
28 resulting gene sets from each approach were integrated into a comprehensive and non-
29 redundant consensus gene set. We predicted a total of 21,771 (≥ 50 amino acids) genes
30 in the assembled *A. japonicus* genome including 101,776 exons (average 4.67 exons per
31 gene), and an average gene size of 5,402 nucleotides (average transcript size of 982
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2 nucleotides) (Table. 2).
3
4
5
6

7 **Genetic polymorphism among natural color variants**

8
9 To provide a general overview of the total genetic variation in the species, we
10
11 realigned reads from the green color variant to the assembled genome using BWA
12
13 v0.7.13 [33]. Picard v1.141 (<http://picard.sourceforge.net/>) was used to mark and
14
15 remove duplicates. Before SNP and small indel calling, we realigned reads with indels
16
17 using GATK RealignerTargetCreator and IndelRealigner v3.5 [34] to avoid
18
19 misalignment around indels. Next, GATK Haplotypecaller was used to call SNPs and
20
21 indels from the resulting sequences. In this study, we observed nucleotide-level
22
23 heterozygosity within the assembled genome to be 0.986 %; namely, we identified a
24
25 total of 6,550,122 SNPs at the assembled genome, for a heterozygous SNP rate of
26
27 0.00986 per site. This high rate of nucleotide polymorphism is not uncommon in marine
28
29 invertebrates and also has been found in the sea urchin genome (~1%; at least one SNP
30
31 per 100 bases) [35], which belongs to the same phylum.
32
33
34
35
36
37
38

39 To measure nucleotide diversity in *A. japonicus*, the aforementioned analyses
40
41 were repeated for red and black color variants separately, and VCFtools v0.1.14 [36]
42
43 with sliding window analysis (bin 10 kb, step 1 kb) was used to calculate nucleotide
44
45 diversity. We identified 6.6 – 14.5 million heterozygous SNPs (1.7 – 3.7 million small
46
47 indels) in the assembled genome from the three natural color variants (Table 3),
48
49 resulting in an estimated nucleotide diversity of 0.00146.
50
51
52

53 In summary, we report the first draft genome of *A. japonicus* and provide a
54
55 general overview of the genetic variation in its three color variants (green, red, and
56
57 black). These data will help elucidate the genetic, physiological, and evolutionary
58
59
60
61

1
2 relationships among different color variants in *A. japonicus* and will be invaluable
3
4 resources for sea cucumber genomic research.
5
6

7 8 9 **Availability of supporting data**

10
11 The raw dataset of all *Apostichopus japonicus* genome libraries and the
12
13 assembly was deposited in the NCBI database with BioProject accession number
14
15 PRJNA335936, SRA accession number SRP082485, and GenBank accession number
16
17 MODV00000000. The additional dataset associated with genome annotation was
18
19 deposited in GigaScience Database (GigaDB). The RNA-seq datasets used in this study
20
21 were downloaded from the ENA database with accession number PRJEB12167 and the
22
23 NCBI database with SRA accession number SRA046386.
24
25
26
27
28
29
30

31 **Abbreviations**

32
33 bp: base pairs; kb: kilobases; Gb: Gigabases; TE: Transposable element; RNA-seq:
34
35 High-throughput messenger RNA sequencing; SNP: Single nucleotide polymorphism;
36
37 Indel: Insertion and deletion.
38
39
40
41
42

43 **Competing interests**

44
45 The authors declare that they have no competing interests.
46
47
48
49
50

51 **Authors' contributions**

52
53 CP designed the study; CP, JKP, SJC contributed to the project coordination; JJ, HGL,
54
55 HHH, and SJ collected the samples and extracted the genomic DNA; CP, JO, SGL, and
56
57 SC conducted the genome analyses; CP, JKP, JJ and EK wrote the paper; All authors
58
59
60
61
62
63
64
65

1
2 read and approved the final manuscript.
3
4
5
6

7 **Acknowledgements**

8
9 This work was supported by research grants from the Marine Biotechnology Program
10 (PJT200620, Genome Analysis of Marine Organisms and Development of Functional
11 Applications) funded by the Ministry of Oceans and Fisheries of the Republic of Korea
12 to CP, JKP, SJC and from the Basic Science Research Program through the National
13 Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future
14 Planning (NRF-2015R1C1A1A02036896) to CP. This work was also supported by
15 National Research Foundation of Korea (NRF) grant funded by the Korean government
16 (MSIP) (NRF-2015R1A4A1041997) to JKP. This work was carried out with the
17 support of “Cooperative Research Program for Agriculture Science & Technology
18 Development (Signaling regulations and disease mechanisms research on exposure to
19 biological, chemical and environmental hazard substance, PJ01052301)” Rural
20 Development Administration, Republic of Korea to SJC.
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42

43 **Author details**

44
45
46 ^a School of Biological Sciences and Technology, Chonnam National University, Gwangju
47 61186, Republic of Korea. ^b Marine Ecological Disturbing and Harmful Organisms
48 Research Center, Department of Oceanography, Chonnam National University, Gwangju
49 61186, Republic of Korea. ^c Division of EcoScience, Ewha Womans University, Seoul
50 03760, Republic of Korea. ^d Department of Biology, College of Natural Sciences,
51 Chungbuk National University, Cheongju, Chungbuk 28644, Republic of Korea.
52
53
54
55
56
57
58
59
60
61
62
63
64
65

References

1. Choe S, Oshima, Y. On the morphological and ecological differences between two commercial forms, "Green" and "Red", of the Japanese common sea cucumber, *Stichopus japonicus* Selenka. *Nippon Suisan Gakkaishi*. 1961;27:97–105.
2. Kanno M, Kijima, A. Quantitative and qualitative evaluation on the color variation of the Japanese sea cucumber *Stichopus japonicus*. *Suisanzoshoku*. 2002;50:63–9.
3. Hongsheng Yang J-FH, Annie Mercier. *The Sea Cucumber *Apostichopus japonicus*: History, Biology and Aquaculture*. Academic Press; 2015.
4. Yamamoto K, Handa T, Fujimoto K. Differences in tolerance to low-salinity among red, blue and black (color pattern) of the Japanese common sea cucumber, *Apostichopus japonicus* from ventilation in the respiratory tree. *Suisan Zoshoku*. 2003;v. 51:321-26.
5. Yamamoto K, Handa T, Fujimoto K. Effects of Water Temperature on Ventilation of the Japanese Common Sea Cucumber, *Apostichopus japonicus* of Different Color Pattern. *Aquaculture Science*. 2005;53(1):67-74. doi:10.11233/aquaculturesci1953.53.67.
6. Choe S. *Biology of the Japanese common sea cucumber *Stichopus japonicus* Selenka*. Pusan [sic]: Pusan National Univ.; 1963.
7. Dong Y-W, Ji T-T, Meng X-L, Dong S-L, Sun W-M. Difference in Thermotolerance Between Green and Red Color Variants of the Japanese Sea Cucumber, *Apostichopus japonicus* Selenka: Hsp70 and Heat-Hardening Effect. *The Biological Bulletin*. 2010;218(1):87-94. doi:10.1086/BBLv218n1p87.
8. Bordbar S, Anwar F, Saari N. High-value components and bioactives from sea cucumbers for functional foods--a review. *Mar Drugs*. 2011;9(10):1761-805. doi:10.3390/md9101761.
9. Purcell SW. Value, market preferences and trade of Beche-de-mer from Pacific Island sea cucumbers. *PloS one*. 2014;9(4):e95075. doi:10.1371/journal.pone.0095075.
10. Long KA, Nossa CW, Sewell MA, Putnam NH, Ryan JF. Low coverage sequencing of three echinoderm genomes: the brittle star *Ophionereis fasciata*, the sea star *Patiriella regularis*, and the sea cucumber *Australostichopus mollis*. *Gigascience*. 2016;5:20. doi:10.1186/s13742-016-0125-6.
11. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-20. doi:10.1093/bioinformatics/btu170.
12. Krueger F. Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. 2015. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
13. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America*. 2011;108(4):1513-8. doi:10.1073/pnas.1017351108.

- 1
2
3 14. LIU Jin ZX-j, SU Lin, LIU Shi-lin, RU Shao-guo, YANG Hong-sheng. Genome size
4 determination of sea cucumber (*Apostichopus japonicus*). JOURNAL OF FISHERIES OF CHINA.
5 2012;Vol.36, No.5. doi:10.3724/SP.J.1231.2012.27753.
6
7 15. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M et al. Efficient de novo
8 assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome
9 research. 2014;24(8):1384-95. doi:10.1101/gr.170720.113.
10
11 16. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J et al. SOAPdenovo2: an empirically improved
12 memory-efficient short-read de novo assembler. Gigascience. 2012;1(1):18. doi:10.1186/2047-
13 217X-1-18.
14
15 17. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in
16 eukaryotic genomes. Bioinformatics. 2007;23(9):1061-7. doi:10.1093/bioinformatics/btm071.
17
18 18. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing
19 genome assembly and annotation completeness with single-copy orthologs. Bioinformatics.
20 2015;31(19):3210-2. doi:10.1093/bioinformatics/btv351.
21
22 19. Smit AFA, Hubley R, Green P. RepeatMasker 4.0.6. 2015. <http://www.repeatmasker.org/>.
23
24 20. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in
25 eukaryotic genomes. Mob DNA. 2015;6:11. doi:10.1186/s13100-015-0041-9.
26
27 21. Smit A, Hubley R. RepeatModeler Open-1.0. 2015. <http://www.repeatmasker.org/>.
28
29 22. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA
30 alignments to improve de novo gene finding. Bioinformatics. 2008;24(5):637-44.
31 doi:10.1093/bioinformatics/btn013.
32
33 23. Kent WJ. BLAT--the BLAST-like alignment tool. Genome research. 2002;12(4):656-64.
34 doi:10.1101/gr.229202. Article published online before March 2002.
35
36 24. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Jr., Hannick LI et al. Improving the
37 Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic acids
38 research. 2003;31(19):5654-66.
39
40 25. Jo J, Park J, Lee HG, Kern EM, Cheon S, Jin S et al. Comparative transcriptome analysis of
41 three color variants of the sea cucumber *Apostichopus japonicus*. Marine genomics. 2016.
42 doi:10.1016/j.margen.2016.03.009.
43
44 26. Du H, Bao Z, Hou R, Wang S, Su H, Yan J et al. Transcriptome sequencing and
45 characterization for the sea cucumber *Apostichopus japonicus* (Selenka, 1867). PloS one.
46 2012;7(3):e33311. doi:10.1371/journal.pone.0033311.
47
48 27. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I et al. Full-length
49 transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol.
50 2011;29(7):644-52. doi:10.1038/nbt.1883.
51
52 28. UniProt C. UniProt: a hub for protein information. Nucleic acids research. 2015;43(Database
53 issue):D204-12. doi:10.1093/nar/gku989.
54
55
56
57
58
59
60
61
62
63
64
65

- 1
2
3 29. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K et al. BLAST+:
4 architecture and applications. *BMC Bioinformatics*. 2009;10:421. doi:10.1186/1471-2105-10-
5 421.
6
7 30. Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S et al. The EMBL-EBI
8 bioinformatics web and programmatic tools framework. *Nucleic acids research*.
9 2015;43(W1):W580-4. doi:10.1093/nar/gkv279.
10
11 31. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment
12 of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*.
13 2013;14(4):R36. doi:10.1186/gb-2013-14-4-r36.
14
15 32. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR et al. Differential gene and
16 transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature*
17 *protocols*. 2012;7(3):562-78. doi:10.1038/nprot.2012.016.
18
19 33. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
20 *Bioinformatics*. 2009;25(14):1754-60. doi:10.1093/bioinformatics/btp324.
21
22 34. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A et al. The Genome
23 Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.
24 *Genome research*. 2010;20(9):1297-303. doi:10.1101/gr.107524.110.
25
26 35. Sea Urchin Genome Sequencing C, Sodergren E, Weinstock GM, Davidson EH, Cameron
27 RA, Gibbs RA et al. The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science*.
28 2006;314(5801):941-52. doi:10.1126/science.1133609.
29
30 36. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA et al. The variant call
31 format and VCFtools. *Bioinformatics*. 2011;27(15):2156-8. doi:10.1093/bioinformatics/btr330.
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Table 1. Statistics on total reads of the *Apostichopus japonicus* genome.

Variants	Insertion size (bp)	Total reads* (Raw data)	Total reads* (w/o adaptor)	Total reads* (error corrected)	% error corrected
Green	180	498,608,646	474,117,288	466,062,920	1.70
	400	897,432,174	842,766,704	831,964,242	1.28
	2000 (v1)	293,701,464	270,513,434	268,573,812	0.72
	2000 (v2)	538,359,438	496,446,984	493,387,418	0.62
	Total		2,228,101,722	2,083,844,410	2,059,988,392
Red	400	397,799,042	394,984,810	383,734,440	2.85
Black	400	460,597,940	423,543,558	416,007,614	1.78

Note: *The length of each read is 101 bp.

Table 2. Statistics on *Apostichopus japonicus* genome assembly

Statistics	Values
Total assembled bases (bp)	664,375,288
Average length of scaffolds (bp)	5,010
Number of scaffolds	132,607
Number of contigs	197,146
Length of longest scaffold (bp)	131,537
GC content (%)	35.92
Scaffold N50 (bp)	10,488
Contig N50 (bp)	5,525
Number of genes	21,771
Number of exons per gene	4.67
Average exon length (bp)	209
Number of introns per gene	4.21
Average intron length (bp)	1,048

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Table 3. SNP and small indel statistics among three color variants.

Variants	# of heterozygous SNP loci	# of small indel loci
Green	6,550,122	1,662,708
Red	14,509,713	3,681,007
Black	12,627,560	3,198,584

1
2 **Figure legends**
3
4
5

6 **Figure 1. Three color-variants of *Apostichopus japonicus*.** (A) Dorsal view of the
7 three color variants. Left to right: red, green, and black. (B) Ventral view of the three
8 color variants. Left to right: red, green, and black.
9
10

11
12
13 **Figure 2. K-mer distribution of the *Apostichopus japonicus* genome.**
14
15

16
17 **Figure 3. Schematic workflow of *Apostichopus japonicus* genome assembly and**
18 **annotation.** The left side represents the genome assembly and the right side represents
19 the transcriptome assembly that was performed in previous publications. To achieve
20 suitable gene prediction, we integrated these two assembly results.
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

November 03, 2016

Dear colleagues at GigaScience,

Thank you very much for handling our manuscript, “Draft genome of the sea cucumber *Apostichopus japonicus* and genetic polymorphism among color” by Jo et al., which we submitted to GigaScience. The reviewers raised important points that greatly improve our manuscript. Below, we address these points in detail one by one. The reviewers’ comments are in italics. The modifications of the manuscript are shown in red.

Sincerely,

Chungoo Park
Corresponding Author

Editor

Comment 1:

Do you have a picture of the species? We may be able use this to highlight the paper on our homepage, and you may also include a picture of the species as a Figure in your revised manuscript.

Response:

We prepared a picture of the sea cucumber *Apostichopus japonicus* for the GigaScience homepage. We want to use it just as a picture to highlight the paper on the homepage, not as a Figure. Please see the uploaded picture.

Comment 2:

Should you have used any custom scripts or other protocols/materials, please consider providing these via our ftp server.

Response:

In this manuscript, mostly we used well-established bioinformatics tools and pipelines. Because we described all analysis procedures in the manuscript in detail, readers of our manuscript can reproduce all the data without any custom scripts.

Reviewer #1

Comment 1:

The genome assembly is listed as 0.67 Gb. The Assemblathon statistics run on the assembly file indicates that the total scaffold size 0.66 Gb and the number of scaffolds is 132,497 (both numbers are similar to the numbers reported in the manuscript). But there are 102,722 contigs in scaffolds and 164,958 contigs greater than 1kb, also according to those statistics. Which sequences were used to calculate the completeness statistics - was it with or without the unscaffolded contigs? Did it include the small contigs (20,922 total <1kb)?

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Response:

We rechecked the assembly file and it was found that the 0.67 Gb is because 0.664 Gb was uncorrected rounded off to the nearest hundredths. Thus, it has been corrected to 0.66 Gb (page 2 and 6).

Next, the reviewer commented that the number of scaffolds generated by the reviewer was slightly different from assembly statistics we present. In the manuscript, we described that the final genome assembly is composed of 132,607 scaffolds with > 1 kb. We found that there is a typo in the range. Actually, we used all scaffolds that are greater than or equal to 1 kb in length. Thus, the "> 1 kb" has been corrected to "≥ 1 kb" (page 6).

We used all scaffolds (which include unscaffolded contigs if they are ≥ 1 kb in length) with ≥ 1 kb to calculate the completeness statistics.

Comment 2:

Please make it clear the N50 statistics refer to scaffold N50 numbers and also give the contig N50 numbers if the unscaffolded contigs continue to be included in the assembly files.

Response:

We have replaced "N50" with "Scaffold N50" and have added contig N50 with the number of contigs scores in Table 2.

Comment 3:

Also, note if the CEGMA and BUSCO analyses are based upon the assembly as in the files or on the scaffolds only.

Response:

The CEGMA and BUSCO analyses were performed using files we deposited in GigaDB. To be clarified, we have added some words as following (page 6):

"The final genome assembly was 0.66 Gb in total length, which is about 80.5 % of the estimated genome size by flow cytometry (0.82 Gb) [14], and is composed of 132,607 scaffolds and unscaffolded contigs (≥ 1 kb) with an N50 value of 10.5 kb (Table 2). We assessed the completeness of the assembly using CEGMA v2.4.010312 [17] and BUSCO v1.22 [18]."

Comment 4:

The genome assembly should be submitted to the International Nucleotide Sequence Databases (DDBJ/NCBI/EMBL).

Response:

We have deposited our genome assembly in NCBI and have now added the GenBank accession number MODV000000000 in manuscript (page 9).

Note that the deposit is confirmed (because we have had the accession number) and now under processing to release our data in public.

WGS [New submission](#)

Note: To find submissions started before Feb. 3, 2014, go to the [previous version](#) of the WGS submission wizard.

ATTN: to fix or update a recent submission whose status is Queued, Processed-error or Processing, please use the **FIX** button on the existing submission

- or [email your request](#) to have the FIX button enabled for that submission.

Be sure to include the Submission ID and the reason that you need to send new files.

Do not create a new submission to fix or update an existing submission whose status is Queued, Processed-error or Processing!

[Short description and brief instructions](#)

1 submission

Submission	Title	Group	Status	Updated
SUB2033434	Draft genome assembly of <i>Apostichopus japonicus</i>		WGS: Processing MODV00000000 Aj_scaffolds_non_masked_N.fsa	Oct 31

Filter / Search

From date: 2016-10-31 To date: 2016-11-02 Status: Sort by: desc

Query: Search Clear

Reviewer #2

Comment 1:

A small description (a few sentences) referencing what is known of the genetics/biology of the color morphs used in the project would be useful to the reader.

Response:

We have added the following to the Data description (page 4):

“The red variant is found on rock pebbles and gravel substrate and has higher salinity and temperature tolerance than the other color variants [4, 5]. Green and black variants are found on sandy and muddy bottoms at shallower depths, and the green variant has greater plasticity in thermotolerance than the red variant [6, 7].”

Comment 2:

The authors created other assemblies using different methods that appear less complete (Table S1) but did they analyze these using their core eukaryotic and metazoan gene sets (or with other approaches) to assess potential non-overlap from the alternative assemblies. Could they contribute unique sequence?

Response:

We assessed the completeness of the assemblies from two different methods. In the genome assemblies from SOAPdenovo and ALLPATHS-LG, only 50 % and 24.6 % of the core eukaryotic genes (based on the 248 core essential genes using CEGMA) were detected, respectively (vs. 73.4% in our final assembly). Using BUSCO, 21.9 % (28.7 %) and 12.1 % (17.2%) of the metazoan (eukaryotoa) single-copy orthologs in the genome assemblies from SOAPdenovo and ALLPATHS-LG, respectively were identified (vs. 60.7% in our final assembly). We have added these results in Table S1.

SoapDenovo						ALLPATH-LG							
CEGMA						CEGMA							
#	Statistics of the completeness of the genome based on 248 CEGs					#	Statistics of the completeness of the genome based on 248 CEGs						
	#Prots	%Completeness	-	#Total	Average	%Ortho		#Prots	%Completeness	-	#Total	Average	%Ortho
Complete	68	27.42	-	112	1.65	38.24	Complete	41	16.53	-	65	1.59	34.15
Group 1	14	21.21	-	22	1.57	35.71	Group 1	9	13.64	-	16	1.78	44.44
Group 2	18	32.14	-	31	1.72	33.33	Group 2	12	21.43	-	18	1.50	25.00
Group 3	17	27.87	-	20	1.18	17.65	Group 3	8	13.11	-	11	1.38	37.50
Group 4	19	29.23	-	39	2.05	63.16	Group 4	12	18.46	-	20	1.67	33.33
Partial	124	50.00	-	251	2.02	55.65	Partial	61	24.60	-	121	1.98	59.02
Group 1	24	36.36	-	41	1.71	50.00	Group 1	13	19.70	-	23	1.77	53.85
Group 2	31	55.36	-	75	2.42	58.06	Group 2	15	26.79	-	33	2.20	66.67
Group 3	31	50.82	-	54	1.74	48.39	Group 3	18	29.51	-	30	1.67	50.00
Group 4	38	58.46	-	81	2.13	63.16	Group 4	15	23.08	-	35	2.33	66.67
BUSCO (MetazoaDB)						BUSCO (MetazoaDB)							
#Summarized BUSCO benchmarking for file: SC.fa						#Summarized BUSCO benchmarking for file:SC.fa							
#BUSCO was run in mode: genome						#BUSCO was run in mode: genome							
Summarized benchmarks in BUSCO notation:						Summarized benchmarks in BUSCO notation:							
C:0%[D:0%],F:0%,M:0%,n:843						C:0%[D:0%],F:0%,M:0%,n:843							
88	Complete BUSCOs					51	Complete BUSCOs						
77	Complete and single-copy BUSCOs					40	Complete and single-copy BUSCOs						
11	Complete and duplicated BUSCOs					11	Complete and duplicated BUSCOs						
97	Fragmented BUSCOs					51	Fragmented BUSCOs						
658	Missing BUSCOs					741	Missing BUSCOs						
843	Total BUSCO groups searched					843	Total BUSCO groups searched						
BUSCO (EukaryotaDB)						BUSCO (EukaryotaDB)							
#Summarized BUSCO benchmarking for file: SC.fa						#Summarized BUSCO benchmarking for file: SC.fa							
#BUSCO was run in mode: genome						#BUSCO was run in mode: genome							
Summarized benchmarks in BUSCO notation:						Summarized benchmarks in BUSCO notation:							
C:0%[D:0%],F:0%,M:0%,n:429						C:0%[D:0%],F:0%,M:0%,n:429							
47	Complete BUSCOs					25	Complete BUSCOs						
37	Complete and single-copy BUSCOs					18	Complete and single-copy BUSCOs						
10	Complete and duplicated BUSCOs					7	Complete and duplicated BUSCOs						
76	Fragmented BUSCOs					49	Fragmented BUSCOs						
306	Missing BUSCOs					355	Missing BUSCOs						
429	Total BUSCO groups searched					429	Total BUSCO groups searched						

Figure. Results of CEGMA and BUSCO using the assemblies from two different methods

Next, to assess to extent to which unique sequences from alternative assemblies have existed, we aligned all scaffold sequences from alternative assemblies with our final assembled genome sequences. Using $e\text{-value} < 10^{-10}$, 97.7% and 99.1% of scaffolds in SOAPdenovo and ALLPATHS-LG assemblies, respectively, were similar to those in our final assembly.

In summary, our final assembly was superior to the others, consistent with our main argument in the manuscript.

Comment 3:

What transcriptome resources were used in the annotation process and what tissues/developmental stages did these come from?

Response:

We have added the following to the Data description (page 7):

“we collected RNA-seq data from our previous [25] (from body wall tissue of adult stage specimens) and other transcriptome (from embryo, larva, and juvenile stages [developmental-stage specific]; from gonads, intestines, respiratory trees, and coelomic fluid of adults [tissue-specific]) [26] studies,”

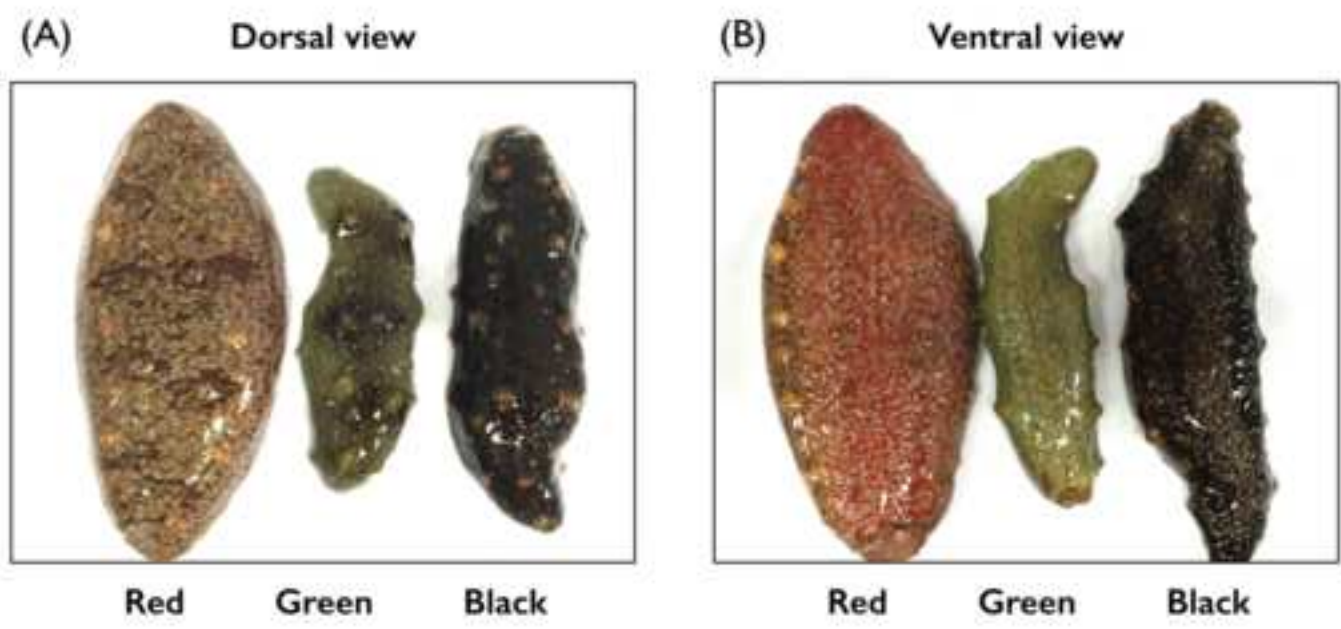


Figure 1

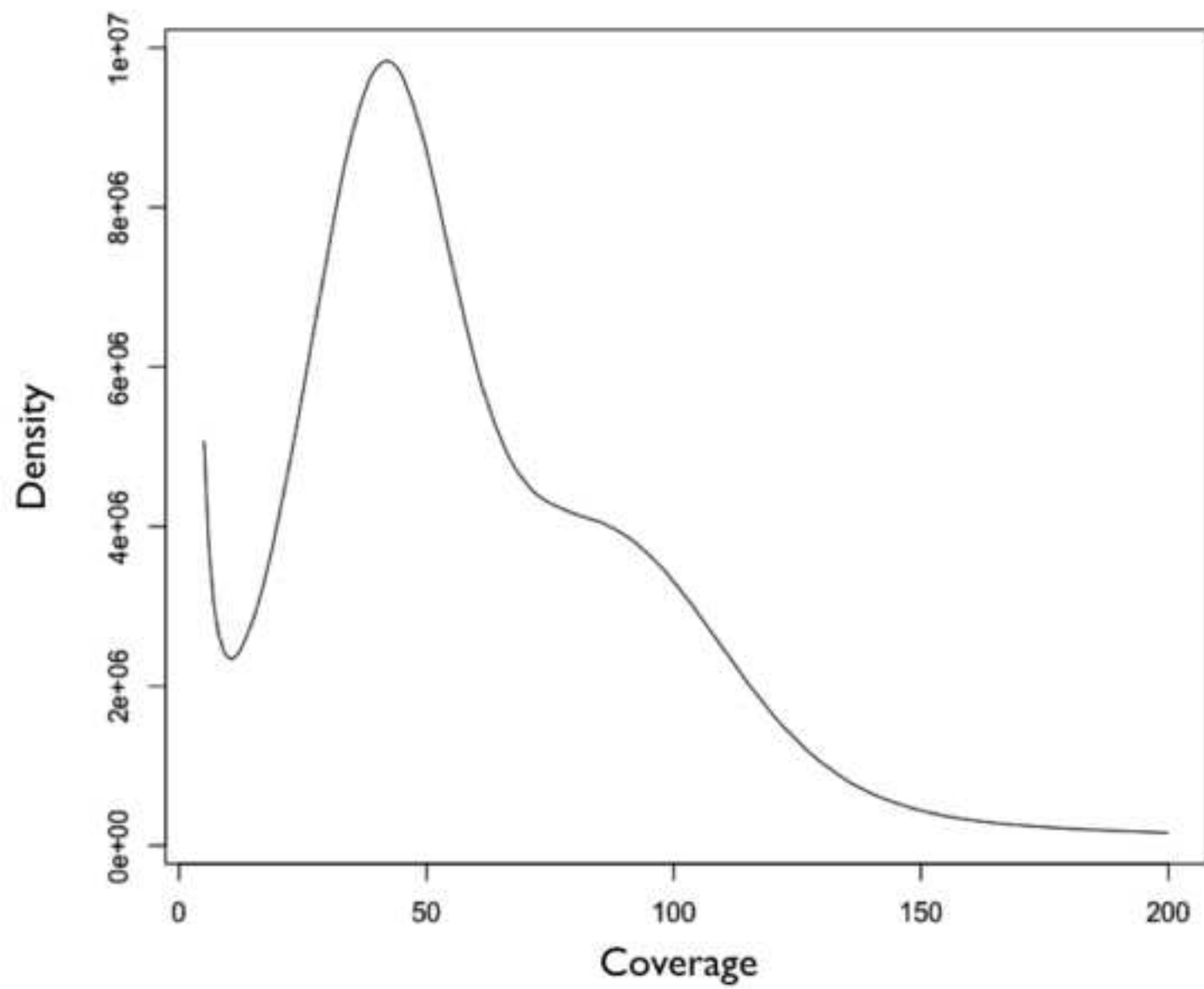


Figure 2

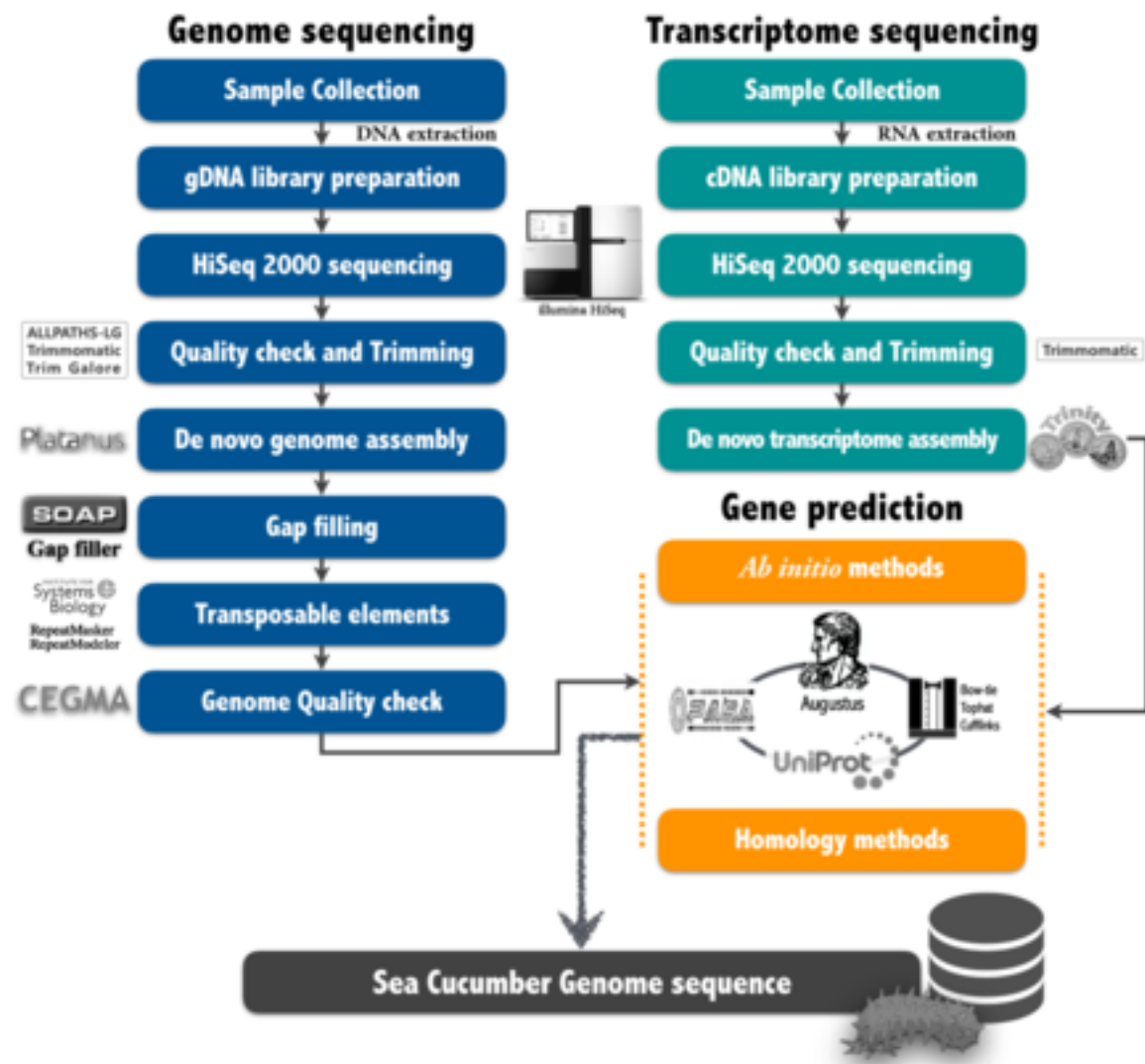



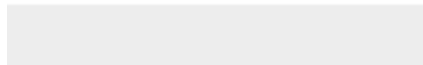
Figure 3



Click here to access/download
Supplementary Material
Table S1.docx



Click here to access/download
Supplementary Material
Picture1.tiff



Sep. 1, 2016.

Dear Editor:

Please consider our manuscript entitled “**Draft genome of the sea cucumber *Apostichopus japonicus* and genetic polymorphism among color variants**” for consideration of publication as a Data note in *Gigascience*. *Apostichopus japonicus* is one of the well-known, commercially important sea cucumber species and occurs in the northwestern Pacific coast including China, Japan, Korea and the Far Eastern seas. This species exhibits a wide array of dorsal/ventral color variants (in particular green, red, and black), which differ in their biological and morphological attributes. Recently, overexploitation and the prevalence of biotic disease in sea cucumber aquaculture have generated increasing conservation concern. However, the genomic knowledge base and resources available to researchers for use in managing this natural resource or establishing genetically based breeding systems are still in a nascent stage.

We believe that our work is suitable for *Gigascience* for the following two reasons. **First**, We report the first draft genome of *A. japonicus*. A total of 312 gigabases (Gb) of raw sequences were generated using the Illumina HiSeq 2000 platform and assembled to a final size of 0.67 Gb which is about 81.7 % of the estimated genome size (0.82 Gb). We observed nucleotide-level heterozygosity within the assembled genome to be 0.986 %. The resulting draft genome assembly comprising 132,607 scaffolds with an N50 value of 10.5 kb contains a total of 21,771 predicted protein-coding genes. **Second**, we provide a general overview of the genetic variation in the three major color variants of *A. japonicus*. We identified 6.6 – 14.5 million heterozygous SNPs in the assembled genome of the three natural color variants (green, red, and black), resulting in an estimated nucleotide diversity of 0.00146. For the above reasons, we expect that our paper will help provide a comprehensive view of the genetic, physiological, and evolutionary relationships among color variants in *A. japonicus*, and will be invaluable resources for sea cucumber genomic research.

Thank you very much for considering our manuscript.

Sincerely,

Chungoo Park
Assistant Professor of School of Biological Sciences and Technology
Chonnam National University
Gwangju, Republic of Korea, 500-757
Phone: +82-62-530-1913
Email: chungoo@jnu.ac.kr