Whole genome survey and microsatellite motif identification of *Artemia franciscana*

Euna Jo\(^1\), Seung Jae Lee\(^1\),†, Eunkyung Choi\(^1\), Jinmu Kim\(^1\), Sung Gu Lee\(^2,3\), Jun Hyuck Lee\(^2,3\), Jeong-Hoon Kim\(^2\), Hyun Park\(^1,*\)

\(^1\) Department of Biotechnology, College of Life Sciences and Biotechnology, Korea University, Seoul 02841, Korea

\(^2\) Korea Polar Research Institute (KOPRI), Yeonsu-gu, Incheon 21990, Korea

\(^3\) Department of Polar Sciences, University of Science and Technology, Yeonsu-gu, Incheon 21990, Korea

† These authors contributed equally to this work.

* Corresponding author: **Hyun Park**, Department of Biotechnology, College of Life Sciences and Biotechnology, Korea University, Seoul 02841, Korea. E-mail: hpark@korea.ac.kr
Abstract

*Artemia* is an industrially important genus used in aquaculture as a nutritious diet for fish and as an aquatic model organism for toxicity tests. However, despite the significance of *Artemia*, genomic research remains incomplete and knowledge on its genomic characteristics is insufficient. In particular, *A. franciscana* of North America has been widely used in fisheries of other continents, resulting in invasion to native species. Therefore, studies on population genetics and molecular marker development as well as morphological analyses are required to investigate its population structure and to discriminate closely related species. Here, we used the Illumina Hi-Seq platform to estimate the genomic characteristics of *A. franciscana* through genome survey sequencing. Further, simple sequence repeat (SSR) loci were identified for microsatellite marker development. The predicted genome size was ~867 Mb using K-mer analysis (K=17), and heterozygosity and duplication rates were 0.655% and 0.809%, respectively. A total of 421,467 SSRs were identified from the genome survey assembly, most of which were dinucleotide motifs with a frequency of 77.22%. This study will be a useful basis in genomic and genetic research for *A. franciscana*.

Keywords: *Artemia franciscana*; Genome size; GC content; Genome assembly; SSR
Introduction

The genus *Artemia* (Crustacea: Branchiopoda: Anostraca), known as brine shrimp, is an aquatic invertebrate living mainly in salt lakes. To date, seven species have been assigned to the genus *Artemia* excluding the parthenogenetic populations called *A. parthenogenetica* (1). *Artemia* species are important in aquaculture industry because their dormant cysts are easily hatched, and the nauplii can be used as a nutrient-rich food for fish (2, 3). In addition, they are also widely used as aquatic model organisms for ecotoxicity tests, along with *Daphnia* (4-6). However, despite the significance of *Artemia* in aquaculture, genomic research is still incomplete, and the genomic characteristics are less known, compared to *Daphnia*, because of the relatively large estimated genome size of 0.93 Gb to 2.93 Gb (7-9).

*Artemia franciscana*, a native species in North America, has been extensively used and introduced to other continents for commercial purposes, thereby affecting the local population’s biodiversity (10, 11). Additionally, there were some incorrect identification, especially *A. franciscana* as *A. salina*, in citing the species used as test organisms in literature (12). These problems, including population structure changes and species misidentification, suggest the need for not only morphological analyses but also population genetic studies and additional marker development to discriminate closely related species.

Genome survey sequencing (GSS) using next-generation sequencing (NGS) is a time- and cost-effective way to evaluate genome information, such as genome size, heterozygosity level, and repeat content, and can be used to develop molecular markers (13, 14). Simple sequence repeats (SSRs) or microsatellites are short tandem repeats of 1–6 nucleotides that have been utilized as genetic markers because of their outstanding abundance and high variability (15). In *Artemia*, several microsatellite markers have already been developed for use in population genetic studies (16, 17), but they are limited because they are not based on genome-wide data.

In this study, we aimed to estimate the genomic characteristics of *A. franciscana* through GSS...
and then identify SSRs from GSS for microsatellite marker development. This study would be useful for population genetics and molecular species identification and as a framework for subsequent whole-genome sequencing of *A. franciscana*.

**Materials and methods**

**Materials and DNA extraction**

Nauplii of *A. franciscana*, originating from the Great Salt Lake (Utah, USA), were hatched on commercial cysts (INVE Technologies NV, Dendermonde, Belgium). Cultures were maintained in 30 g/L salt water at 25ºC with aeration. Live *Tetraselmis* sp. was fed to *A. franciscana* during the culture period. One egg-bearing female individual were cultured separately and the progenies were used in the subsequent experiments. Genomic DNA was extracted from whole 5 adults using phenol/chloroform method. The quality and quantity of the DNA were checked using a BioAnalyzer (Agilent Technologies, Santa Clara, CA, USA) and Qubit fluorometer (Invitrogen, Life Technologies, Carlsbad, CA, USA).

**Genome sequencing, assembly, and K-mer analysis**

Genomic DNA was randomly sheared into 350-bp fragments using an ultrasonicator (Covaris, USA). A paired-end DNA library was prepared and sequenced with Illumina HiSeq 2000 platform. To ensure the quality of data, adaptors, poly(N) sequences, and low-quality reads were filtered out, and only clean reads were subjected to K-mer analysis. K-mer analysis was performed using Jellyfish 2.1.4 (18) with a K-value of 17, 19 and 25. Based on the 17-mer distribution, GenomeScope (19) in R version 3.4.4 (20) was used to estimate genome size, heterozygosity rate, and repeat content. The de novo genome assembly was carried out using Maryland Super-Read Celera Assembler (MaSuRCA) version 3.3.4 (21).
SSR detection and primer design

Genome-wide SSR identification was conducted using QDD version 3.1.2 pipeline (22). First, the assembled sequences were used to extract microsatellite sequences with di- to hexanucleotides motifs. Next, sequences were compared using all-against-all BLAST to detect unique singleton sequences. In the last step, primer pairs were designed using two iterative methods for each sequence.

Results and discussion

Genome sequencing data statistics

In the present study, a total of 22.8 Gb of raw data for *A. franciscana* were generated by Illumina paired-end library (Table 1). Quality value (Q) is regarded to be correctly sequenced when Q20 and Q30 values are at least 90% and 85%, respectively (23). The Q20 and Q30 values for this study were 99.3% and 96.0%, respectively (Table 1); hence, the sequencing accuracy of *A. franciscana* was high. Additionally, the GC content of the raw data was 35.8% (Table 1).

Genome size prediction

In this study, Illumina paired-end data was used for K-mer analysis using a K-value of 17. The predicted genome size was approximately 867 Mb (Figure 1). In a previous study, the genome size of *A. franciscana* based on flow cytometry was 930 Mb (9). Our estimation and previous results, measured using different methods, were quite similar, with a difference of 63 Mb. These estimates are smaller than 1.47 Gb in *A. salina* and 2.93 Gb in tetraploid parthenogenetic population (7, 8). In addition, the heterozygosity and duplication rates were calculated to be 0.655% and 0.809%, respectively (Figure 1).
**Genome assembly results**

The results of preliminary genome assembly of *A. franciscana* are shown in Table 2. We obtained 122,231 contigs with a total length of 841,603,395 bp. The maximum and N50 contig lengths were 1,508,123 bp and 14,130 bp, respectively. The GC content of contigs was 35.50% (Table 2). Further assembly generated 46,193 scaffolds with a total length of 938,041,450 bp. The maximum and N50 scaffold lengths were 2,555,521 bp and 67,542 bp, respectively. The GC content of the scaffolds was 31.85% (Table 2).

These genome survey data provide useful information for genomic research of *A. franciscana* and related species. However, further study combined with more advanced NGS technologies using PacBio long read sequencing and high-throughput chromosome conformation capture (Hi-C) method are necessary to improve whole genome sequencing and assembly data. If so, *A. franciscana* could be used in a wider range of research fields (e.g. comparative genomics) as a reference genome.

**SSR loci identification**

From the genome survey assembly of *A. franciscana* with a total length of ~938 Mb, a total of 421,467 repeat motifs were identified. The types of motifs contained 77.22% (325,450) dinucleotide, 20.38% (85,912) trinucleotide, 2.11% (8,877) tetranucleotide, 0.20% (838) pentanucleotide, and 0.09% (390) hexanucleotide (Table 3). The percentage of dinucleotide repeats was the highest, and as the repeat motif length increased, the number of loci decreased, similar to other studies (13, 24, 25). It has been suggested that longer repeats have higher mutation rates, causing instability (26, 27) and shorter persistence times because of their downward mutation bias toward a reduction in repeat number (28).

Of the dinucleotides, the most frequent motif was AT/AT (53.31%), followed by AG/CT (26.66%), AC/GT (17.74%), and CG/CG (2.28%). Of the trinucleotides, the most frequent
motif was AAT/ATT (41.74%), followed by ACT/AGT (33.51%) and AAG/CTT (13.38%). ACG/CGT (0.08%) was the least frequent trinucleotides motif. The most abundant motifs among the tetra-, penta- and hexanucleotides were AAAT/ATTT (24.14%), AATAT/ATATT (14.32%) and AGAGCC/GGCTCT (15.64%), respectively (Table 3).

Overall, the motifs including A or T were more abundant than those including C or G, consistent with the findings of Daphnia pulex genome-wide SSR study (29). These results might be due to the high slippage rate of A/T motifs, addition of 3’ poly(A) tail by retrotransposon elements or transition of methylated C to T residues at CpG sites (14, 30-32). These data for SSRs in A. franciscana will be used as valuable references for the development of microsatellite markers, although further validation studies using various Artemia population are needed.

Conclusion

In this study, the genome of A. franciscana was analyzed and assembled, and SSR loci were identified from the GSS data. The K-mer analysis (K=17) estimated the genome size of A. franciscana to be ~867 Mb, and the heterozygosity and duplication rates were 0.655% and 0.809%, respectively. Genome assembly results showed that contig N50 was 14,130 bp, with a total length of 841,603,395 bp, whereas scaffold N50 was 67,542 bp, with a total length of 938,041,450 bp. A total of 421,467 SSRs were identified, of which dinucleotide motifs were the most abundant and hexanucleotide motifs were the least abundant. This study will be a useful foundation for genomic and genetic studies on A. franciscana.

Acknowledgements: This research was a part of the project titled “Development of potential antibiotic compounds using polar organism resources (15250103, KOPRI Grant PM20030)” and “Ecosystem Structure and Function of Marine Protected Area (MPA) in Antarctica (20170336, KOPRI Grant PM19060)” funded by the Ministry of Oceans and Fisheries, Korea.
Conflict of interest: The authors declare that they have no competing interests.

Data accessibility statement: The *Artemia franciscana* genome project has been registered in NCBI under the BioProject number PRJNA449186. The whole-genome sequence has been deposited in the Sequence Read Archive (SRA) database under accession numbers SRS3156165 and SAMN08892388.

Author contributions: H.P. conceived the study. E.J., S.J.L., E.K.C., J.K., S.G.K., J.H.L., and J.-H.K. performed the genome sequencing, assembly, and analysis. E.J., S.J.L., and H.P. mainly wrote the manuscript. All authors contributed in writing and editing the manuscript, collating supplementary information, and creating the figures.

Abbreviations

GSS, Genome survey sequencing; Hi-C, high-throughput chromosome conformation capture; K-mer, A sequence of k characters in a string; MaSuRCA, Maryland Super-Read Celera Assembler; NGS, next-generation sequencing; SSR, simple sequence repeat

References

1. Asem A, Rastegar-Pouyani N, De Los Ríos-Escalante P. The genus *Artemia* leach, 1819 (Crustacea: Branchiopoda). I. True and false taxonomical descriptions. Latin American Journal of Aquatic Research. 2010;38(3):501-6.
2. Bengtson DA, Léger P, Sorgeloos P. Use of *Artemia* as a food source for aquaculture. *Artemia* Biology. 1991;11:255-85.
3. Kolkovski S, Curnow J, King J. Intensive rearing system for fish larvae research II:
200  Artemia hatching and enriching system. Aquacultural Engineering. 2004;31(3-4):309-17.

201  4.  Nunes BS, Carvalho FD, Guilhermino LM, Van Stappen G. Use of the genus Artemia in ecotoxicity testing. Environmental Pollution. 2006;144(2):453-62.

203  5.  Yu J, Lu Y. Artemia spp. model-a well-established method for rapidly assessing the toxicity on an environmental perspective. Medical Research Archives. 2018;6(2).

205  6.  Libralato G, Prato E, Migliore L, Cicero A, Manfra L. A review of toxicity testing protocols and endpoints with Artemia spp. Ecological Indicators. 2016;69:35-49.

207  7.  Vaughn J. DNA reassociation kinetic analysis of the brine shrimp, Artemia salina. Biochemical and Biophysical Research Communications. 1977;79(2):525-31.

209  8.  Rheinsmith E, Hinegardner R, Bachmann K. Nuclear DNA amounts in Crustacea. Comparative Biochemistry and Physiology Part B: Comparative Biochemistry. 1974;48(3):343-8.

211  9.  De Vos S, Bossier P, Van Stappen G, Vercauteren I, Sorgeloos P, Vuylstekke M. A first AFLP-based genetic linkage map for brine shrimp Artemia franciscana and its application in mapping the sex locus. PLoS One. 2013;8(3):e57585.

213  10.  Muñoz J, Gomez A, Green AJ, Figuerola J, Amat F, Rico C. Phylogeography and local endemism of the native Mediterranean brine shrimp Artemia salina (Branchiopoda: Anostraca). Molecular Ecology. 2008;17(13):3160-77.

215  11.  Naceur HB, Jenhanive ABR, Romdhane MS. New distribution record of the brine shrimp Artemia (Crustacea, Branchiopoda, Anostraca) in Tunisia. Check List. 2009;5(2):281-8.

217  12.  Ruebhart DR, Cock IE, Shaw GR. Brine shrimp bioassay: importance of correct taxonomic identification of Artemia (Anostraca) species. Environmental Toxicology: An International Journal. 2008;23(4):555-60.

219  13.  Xu S-y, Song N, Xiao S-j, Gao T-x. Whole genome survey analysis and microsatellite motif identification of Sebastiscus marmoratus. Bioscience Reports. 2020;40(2).
14. Li J, Li S, Kong L, Wang L, Wei A, Liu Y. Genome survey of *Zanthoxylum bungeanum* and development of genomic-SSR markers in congeneric species. Bioscience Reports. 2020;40(6).

15. Buschiazzo E, Gemmell NJ. The rise, fall and renaissance of microsatellites in eukaryotic genomes. Bioessays. 2006;28(10):1040-50.

16. Muñoz J, Green AJ, Figuerola J, Amat F, Rico C. Characterization of polymorphic microsatellite markers in the brine shrimp *Artemia* (Branchiopoda, Anostraca). Molecular Ecology Resources. 2009;9(2):547-50.

17. Eimanifar A, Asem A, Wang P-Z, Li W, Wink M. Using ISSR genomic fingerprinting to study the genetic differentiation of *Artemia* Leach, 1819 (Crustacea: Anostraca) from Iran and neighbor regions with the focus on the invasive American *Artemia franciscana*. Diversity. 2020;12(4):132.

18. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27(6):764-70.

19. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 2017;33(14):2202-4.

20. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing 2017.

21. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics. 2013;29(21):2669-77.

22. Meglécz E, Pech N, Gilles A, Dubut V, Hingamp P, Trilles A, et al. QDD version 3.1: A user-friendly computer program for microsatellite selection and primer design revisited: Experimental validation of variables determining genotyping success rate. Molecular Ecology Resources. 2014;14(6):1302-13.
23. Li G-q, Song L-x, Jin C-q, Li M, Gong S-p, Wang Y-f. Genome survey and SSR analysis of Apocynum venetum. Bioscience Reports. 2019;39(6).

24. Chen M, Tan Z, Zeng G, Peng J. Comprehensive analysis of simple sequence repeats in pre-miRNAs. Molecular Biology and Evolution. 2010;27(10):2227-32.

25. Katti MV, Ranjekar PK, Gupta VS. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Molecular Biology and Evolution. 2001;18(7):1161-7.

26. Kruglyak S, Durrett RT, Schug MD, Aquadro CF. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proceedings of the National Academy of Sciences. 1998;95(18):10774-8.

27. Wierdl M, Dominska M, Petes TD. Microsatellite instability in yeast: dependence on the length of the microsatellite. Genetics. 1997;146(3):769-79.

28. Harr B, Schlötterer C. Long microsatellite alleles in Drosophila melanogaster have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation. Genetics. 2000;155(3):1213-20.

29. Sung W, Tucker A, Bergeron RD, Lynch M, Thomas WK. Simple sequence repeat variation in the Daphnia pulex genome. BMC genomics. 2010;11(1):691.

30. Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD. The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome research. 2008;18(1):30-8.

31. Tóth G, Gáspári Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Research. 2000;10(7):967-81.

32. Zhou Y, Bizzaro JW, Marx KA. Homopolymer tract length dependent enrichments in functional regions of 27 eukaryotes and their novel dependence on the organism DNA (G+ C)% composition. BMC Genomics. 2004;5(1):95
Table 1. Statistics for the genome survey sequencing data of *Artemia franciscana*

| Lib ID | Raw data (bp)   | Q20 (%) | Q30 (%) | GC content (%) |
|--------|-----------------|---------|---------|----------------|
| PE350  | 22,814,108,862  | 99.3    | 96.0    | 35.8           |
Table 2. Statistics of the assembly in *Artemia franciscana*

|       | Total length (bp) | Total number | Max length (bp) | N50 length (bp) | GC content (%) |
|-------|-------------------|--------------|-----------------|-----------------|----------------|
| Contig| 841,603,395       | 122,231      | 1,508,123       | 14,130          | 35.50          |
| Scaffold | 938,041,450     | 46,193       | 2,555,521       | 67,542          | 31.85          |
Table 3. Distribution pattern of simple sequence repeat motifs

| Repeat motif  | Number of repeats | Total |
|---------------|-------------------|-------|
|               | 5     | 6     | 7     | 8     | 9     | 10    | 11–20 | >20   |       |
| **Dinucleotide (325,450)** |       |       |       |       |       |       |       |       |       |
| AC/GT         | 31,132| 11,605| 6,082 | 3,178 | 2,092 | 1,243 | 2,171 | 244   | 57,747|
| AG/CT         | 35,678| 11,966| 7,985 | 5,895 | 3,480 | 2,493 | 11,440| 7,824 | 86,761|
| AT/AT         | 79,855| 25,592| 14,230| 10,458| 6,305 | 4,932 | 19,250| 12,887| 173,509|
| CG/CG         | 6,515 | 796   | 100   | 22    |       |       |       |       | 7,433 |
| **Trinucleotide (85,912)** |       |       |       |       |       |       |       |       |       |
| AAT/ATT       | 21,685| 7,466 | 3,070 | 1,400 | 510   | 233   | 1,033 | 466   | 35,863|
| ACT/AGT       | 17,268| 6,614 | 2,388 | 1,297 | 385   | 172   | 586   | 78    | 28,788|
| AAG/CTT       | 9,019 | 1,737 | 467   | 126   | 40    | 36    | 59    | 12    | 11,496|
| AAC/GTT       | 3,136 | 807   | 203   | 42    |       |       |       |       | 4,188 |
| ATC/GAT       | 1,617 | 337   | 120   | 57    |       |       |       |       | 2,131 |
| AGG/CCT       | 1,355 | 142   | 24    |       |       |       |       |       | 1,521 |
| ACC/GGT       | 767   | 139   | 24    |       |       |       |       |       | 930   |
| AGC/GCT       | 555   | 51    | 42    | 15    |       |       |       |       | 663   |
| CCG/CGG       | 176   | 83    |       |       |       |       |       |       | 259   |
| ACG/CGT       | 73    |       |       |       |       |       |       |       | 73    |
| **Tetranucleotide (8,877)** |       |       |       |       |       |       |       |       |       |
| AAAT/ATTT     | 1,844 | 233   | 54    | 12    |       |       |       |       | 2,143 |
| AATT/AATT     | 720   | 238   | 12    | 33    |       |       |       |       | 1,003 |
| AATC/GATT     | 746   | 134   | 36    | 9     |       |       |       |       | 925   |
| AAGG/CTTT     | 564   | 156   | 48    | 15    | 12    | 26    |       |       | 821   |
| AGAT/ATCT     | 430   | 252   | 79    | 6     | 15    | 24    |       |       | 821   |
| ACAG/CTGT     | 552   | 175   | 3     | 9     | 26    |       |       |       | 765   |
| AATG/CATT     | 477   | 84    | 30    | 15    |       |       |       |       | 606   |
| AAGT/ACTT     | 344   | 58    | 35    |       |       |       |       |       | 437   |
| AAAC/GTTT     | 302   | 108   | 17    |       |       |       |       |       | 427   |
| ACAT/ATGT     | 202   | 116   | 57    | 18    | 12    |       |       |       | 405   |
| Others         | 380   | 60    | 66    | 6     | 12    |       |       |       | 524   |
| **Pentanucleotide (838)** |       |       |       |       |       |       |       |       |       |
| AATAT/ATATT   | 68    | 40    | 12    |       |       |       |       |       | 120   |
| AAAAT/ATTTT   | 78    | 15    | 12    |       |       |       |       |       | 105   |
| ACTAT/ATAGT   | 42    | 24    | 15    | 9     | 12    |       |       |       | 102   |
| AAATT/AATTT   | 81    | 15    |       |       |       |       |       |       | 96    |
| AATTC/GAATT   | 66    |       |       |       |       |       |       |       | 66    |
| Others         | 259   | 57    | 15    | 6     | 12    |       |       |       | 349   |
| **Hexanucleotide (390)** |       |       |       |       |       |       |       |       |       |

Downloaded from http://portlandpress.com/bioscirep/article-pdf/doi/10.1042/BSR20203868/904827/bsr-2020-3868.pdf by guest on 01 March 2021
| Sequence        | 24 | 37 | 61 | 15 | 20 | 35 | 30 | 27 | 24 | 126 | 36 | 27 | 9 | 15 | 213 |
|-----------------|----|----|----|----|----|----|----|----|----|-----|----|----|---|----|-----|
| AGAGCC/GGCTCT   |    |    |    |    |    |    |    |    |    |     |    |    |   |    |     |
| AATATT/AATATT   | 15 | 20 |    |    |    |    |    |    |    |     |    |    |   |    |     |
| ACAAT/ATTGTT    | 30 |    |    |    |    |    |    |    |    |     |    |    |   |    |     |
| AATCT/AGTATT    | 27 |    |    |    |    |    |    |    |    |     |    |    |   |    |     |
| AATAT/ATATTT    | 24 |    |    |    |    |    |    |    |    |     |    |    |   |    |     |
| Others          | 126| 36 | 27 | 9 | 15 | 213|    |    |    |     |    |    |   |    |     |
| **Total**       | 216,208 | 69,156 | 35,211 | 22,631 | 12,900 | 9,165 | 34,673 | 21,523 | 421,467 |
**Figure legend**

**Figure 1. Distribution of K-mer (K=17).** Blue bars represent the observed K-mer distribution; black line represents the modelled distribution without the error K-mers (red line), up to a maximum K-mer coverage specified in the model (yellow line). Len, estimated total genome length; Uniq, unique portion of the genome (not repetitive); Het, heterozygosity rate; Kcov, mean K-mer coverage for heterozygous bases; Err, error rate; Dup, duplication rate.
GenomeScope Profile

len: 867,121,397bp uniq: 46.4% het: 0.655% kcov: 17.9 err: 0.23% dup: 0.809% k: 17