Bioinfomatic Genome analysis of *Karaftohelix adamsi* using Illumina 4000 sequencing

Jong Min Chung, Hee Ju Hwang, Min Kyu Sang, Jie Eun Park, Dae Kwon Song, Jun Yang Jeong, Se Won Kang[1], So Young Park[2], Yeon Soo Han[3], Hong seog Park[4], Jun Sang Lee[5] and Yong Seok Lee

Department of Life Science and Biotechnology, College of Natural Sciences, Soonchunhyang University, Asan, Chungcheongnam-do, 31538, Republic of Korea
[1]Biological Resource Center, Korea Research Institute of Bioscience and Biotechnology, Jeongeup, Jeonbuk 56212, Korea
[2]Biodiversity Research Team, Nakdonggang National Institute of Biological Resources, Sangju, Gyeongsangbuk-do, 37242, Republic of Korea
[3]Department of Applied Biology, Institute of Environmentally-Friendly Agriculture (IEFA), College of Agriculture and Life Sciences, Chonnam National University, Gwangju 61186, Korea
[4]Research Institute, GnC BIO Co., LTD., 621-6 Banseok-dong, Yuseong-gu, Daejeon, 34069, Korea
[5]Institute for Basic Sciences, College of Natural Sciences, Soonchunhyang University, Asan, Chungnam, 31538, Korea

Abstract

Karaftohelix adamsi is distributed only in the Ulleung island and corresponds to Eupulmonata in Bradybaenidae. Currently, it is enlisted as an endangered wild species class II under the Red data list of National Institute of Biological Resources, South Korea, and there is no genetic resources of this species, and it is urgent to secure genetic resources for the preservation and reservation of this species in the future. Not only the species information also genus information which the species belongs is unlisted in the NCBI taxonomy browser. On this, To genomic studies to ensure bio-sovereign right and characterization of species are urgently needed. So, we report the pilot genomic study of *K. adamsi* using the Illumina Hiseq 4000 platform. DNA was extracted from the whole body of *K. adamsi*, and whole-genome sequencing generated a total of 301,367,732 raw reads (37,972,334,232 bp). Clean reads of 283,613,488 reads (35,005,636,440 bp) was obtained through Trim galore program, which removes the order of the adaptor, low quality, and Deconseq program, which excludes unrelated sequences such as virus and bacteria. The obtained sequences were assembled through platanus program, a total of 14,253 contigs (N50, 176 bp), of which more than 1 kb was found to be 111 (188,246 bp). In addition, 4,424 (N50, 476 bp) of the scaffold were analyzed, of which 201 (N50, 508,459 bp) were found to be more than 1 kb of scaffold. We believe that these data will provide primary genetic data to secure the bio-sovereign also understanding ecological and genetic characteristics indices of species faced on threatened to extinction. Furthermore, It will provide a guideline for the analysis strategy of further genome study.

Materials and Methods

Scientific Name

Karaftohelix adamsi

Classfication

Mollusca;
Gastropoda;
Eupulmonata;
Bradybaenidae;

Family Bradybaenidae 유전자원	
Nucleotide	65,524
Protein	1,565
Genome	4
Gene	146

Genus Karaftohelix 유전자원	
Nucleotide	71
Protein	10

Figure 1. Current status of genetic resources of *K. adamsi.*

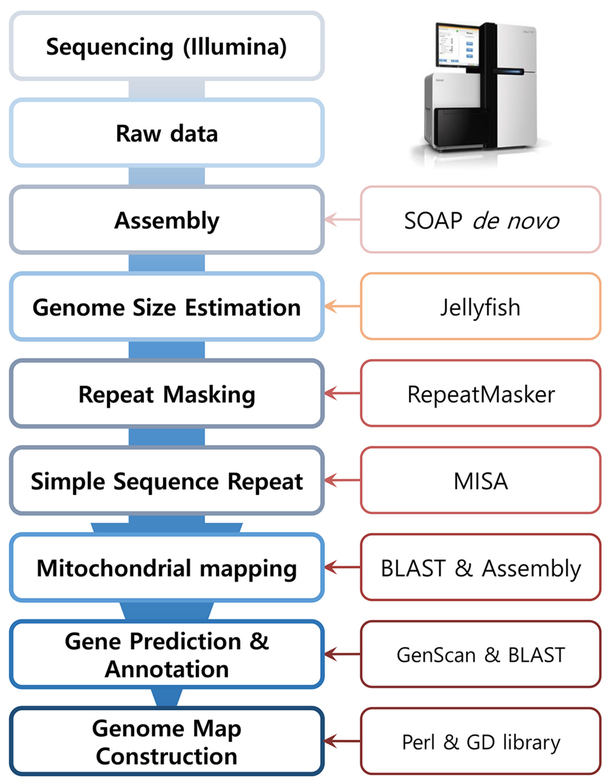


Figure 2. Schematic diagram of *K. adamsi* Genome.

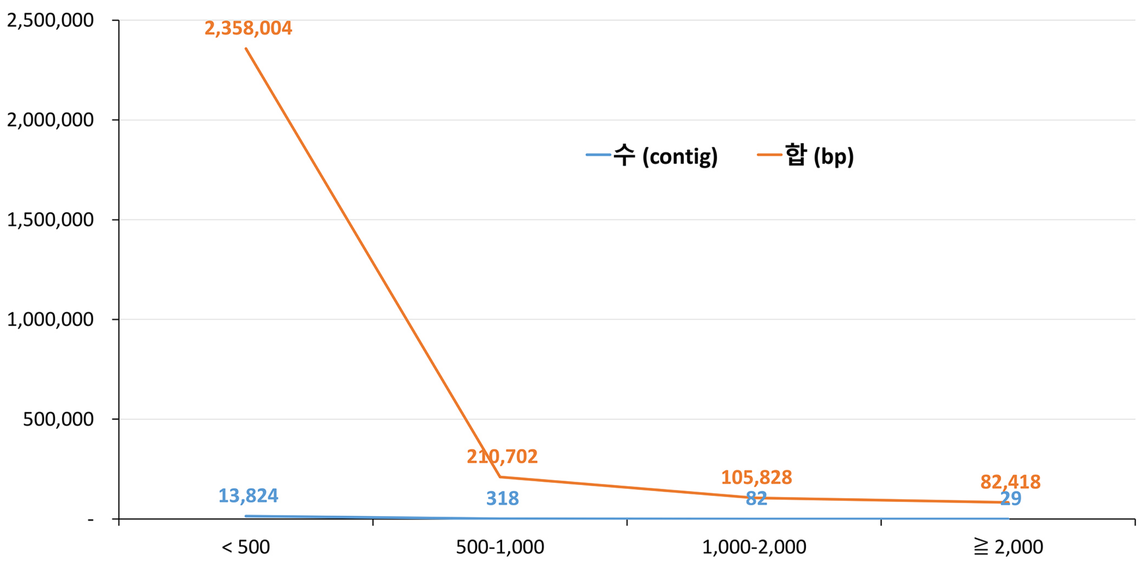
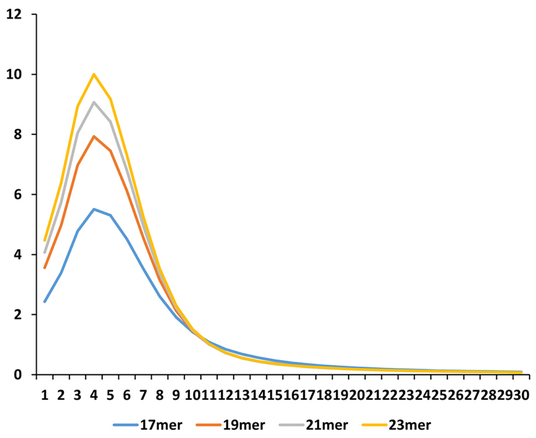


Fig. 3. Length distribution of *K. adamsi* assembled contigs.



K-mer (K-mer size)	Estimated depth (K-mer Coverage depth)	Genome coverage depth (Coverage depth)	Estimated genome Size
17	4	4.63	3,578,869,767
19	4	4.72	3,508,695,850
21	4	4.82	3,438,521,933
23	4	4.91	3,368,348,016

Table 1.Summary of Genome size estimation with Jellyfish program. The k-mer values were 17, 19, 21 and 23 mer for each analysis. Estimation Depth was equal to 4. Genome cover depth (Coverage Depth) and Estimation generation size were identified as 17 mer (4.63), 19 mer (4.72), 21 mer (4.82), 23 mer (4.91). Its average size is about 3.47Gb. The graph shows the peak value by the predicted graph of ddepth, shown by the kmer value.

Result and Discusion

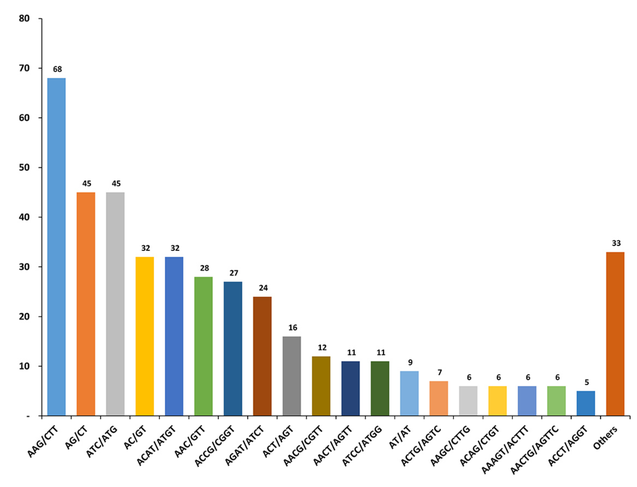
Table 3. The assembly results of the *K. adamsi* sequences using Platanus software.

	Region1	Region2	Total
raw data			
Number Of Reads	150,683,866	150,683,866	301,367,732
Number Of Bases	18,986,167,116	17,577,886,746	37,972,334,232
pair reads			
Number Of Reads	146,876,995	146,876,995	293,753,990
Number Of Bases	18,174,251,900	18,028,711,776	36,202,963,676
single reads			
Number Of Reads	3,343,923	1,347,801	4,691,724
Number Of Bases	368,061,399	107,708,474	475,769,873
clean reads			
Number Of Reads	141,754,534	141,858,954	283,613,488
Number Of Bases	17,566,300,879	17,439,335,561	35,005,636,440

Assembly Results					
Scaffold Metrics	**> 100 bp**	**> 500 bp**	**> 1000 bp**	**> 2000 bp**	
Number Of Scaffold	4,424	366	201	96	
Number Of Bases	1,264,363	624,342	508,459	367,940	
Avg. Scaffold Size	285	1,705	2,529	3,832	
N50 Scaffold Size	476	2,465	3,194	4,050	
N80 Scaffold Size	141	1,052	1,548	2,622	
N90 Scaffold Size	129	741	1,239	2,375	
Largest Scaffold Size	11,741	11,741	11,741	11,741	
Contigs	**> 1 bp**	**> 100 bp**	**> 500 bp**	**> 1000 bp**	**> 2000 bp**
Number Of Contigs	14,253	14,253	429	111	29
Number Of Bases	2,756,952	2,756,952	398,948	188,246	82,418
Avg. Contig Size	193	193	929	1,695	2,842
N50 Contig Size	176	176	935	1,623	2,798
N80 Contig Size	132	132	618	1,203	2,258
N90 Contig Size	126	126	551	1,101	2,108
largest Contig Size	5,471	5,471	5,471	5,471	5,471

Table 4. Summary of simple sequence repeat (SSR) types based on the number of repeat units.

Repeats	Motif length					Total
	Di	Tri	Tetra	Penta	Hexa	
4	-	-	84	12	1	97
5	-	89	28	2	-	119
6	38	36	20	-	-	94
7	11	15	8	-	-	34
8	4	5	5	-	-	14
9	5	2	3	-	-	10
10	6	5	3	-	-	14
11	3	3	-	-	-	6
			~			
30	-	-	1	-	-	1
Total	86	165	152	23	3	429



The SSRs obtained included the most abundant Tri-nucleotide repeats (165), followed by Tetra- (152), Di- (86), Penta- (23) and Hexa- (3). The summary of SSRs based on the number of repeat units has been depicted in Table 4. The five tandem repeats (119) were the predominant, followed by four (97) and six (94) tandem reiterations. An analysis of the frequency distribution of SSRs based on motif sequence types is shown in Table 4. *K. adamsi* Genome is rich in AAG/CTT (68), followed by AG/CT (45), and ATC/ATG (45) repeats. The identification of SSRs can serve as an important lead to genetic improvement programme and for the quantification of genetic diversity within and among populations of this endangered species.

This study supported by National Research Foundation of Korea (NRF-2017R1D1A3B06034971), National Institute of Biological Resources (NIBR201503202)