

## 꼬막(*Tegillarca granosa*)의 유전적 다양성 분석을 위한 드래프트 게놈분석과 마이크로새틀라이트 마커 발굴

김진무<sup>1</sup> · 이승재<sup>1</sup> · 조은아<sup>1,2</sup> · 최은경<sup>1</sup> · 김현진<sup>3</sup> · 이정식<sup>3</sup> · 박현<sup>1\*</sup>

<sup>1</sup>고려대학교 생명공학과

<sup>2</sup>한국해양과학기술원 부속 극지연구소

<sup>3</sup>전남대학교 수산생명의학과

## Genome Survey and Microsatellite Marker Selection of *Tegillarca granosa*

Jinmu Kim<sup>1</sup>, Seung Jae Lee<sup>1</sup>, Euna Jo<sup>1,2</sup>, Eunkyung Choi<sup>1</sup>, Hyeon Jin Kim<sup>3</sup>, Jung Sick Lee<sup>3</sup>, Hyun Park<sup>1\*</sup>

<sup>1</sup>Division of Biotechnology, College of Life Sciences and Biotechnology, Korea University, Seoul 02841, Korea

<sup>2</sup>Unit of Research for Practical Application, Korea Polar Research Institute (KOPRI), Incheon 21990, Korea

<sup>3</sup>Department of Aquaculture Medicine, Chonnam National University, Yeosu 59626, Korea

### Corresponding Author

Hyun Park

Division of Biotechnology, College of Life  
Sciences and Biotechnology, Korea  
University, Seoul 02841, Korea

E-mail : hpark@korea.ac.kr

Received : May 07, 2021

Revised : May 07, 2021

Accepted : May 17, 2021

꼬막 종류 중 하나인 *Tegillarca granosa*는 해양 이매패류로서 한국, 중국, 일본 등의 중요한 수산 자원 중 하나이다. 꼬막의 염색체 수는  $2n=38$ 로 알려져 있지만, 유전체의 크기와 유전 정보에 대해서는 아직 명확하게 알려져 있지 않다. 꼬막의 유전체 크기 예측을 위하여 NGS Illumina HiSeq 플랫폼을 이용하여 얻은 짧은 DNA 서열 정보를 통하여 *in silico* 분석으로 유전체 크기를 분석하였다. 그 결과 꼬막의 유전체 크기는 770.61 Mb로 예측되었다. 이후 MaSuRCA assembler를 통하여 드래프트 게놈 조립 작업을 수행하고, QDD pipeline을 이용하여 SSR (simple sequence repeats) 분석을 수행하였다. 꼬막의 유전체로부터 43,944개의 SSR을 발굴하였으며, 다이-뉴클레오타이드(di-nucleotide) 69.51%, 트라이-뉴클레오타이드(tri-nucleotide) 16.68%, 테트라-뉴클레오타이드(tetra-nucleotide) 12.96%, 펜타-뉴클레오타이드(penta-nucleotide) 0.82% 그리고 헥사-뉴클레오타이드(hexa-nucleotide) 0.03%로 구성되었다. 이후 꼬막의 유전적 다양성 연구에 활용할 수 있는 100개의 마이크로새틀라이트 마커의 프라이머 세트를 선별하였다. 앞으로 이번 연구를 통해서, 꼬막의 집단유전학적 연구와 유전적 다양성을 규명하는데 도움이 될 것이며, 나아가 동종들 간의 원산지 분류를 알아낼 수 있을 것이다.

The blood clam, *Tegillarca granosa*, is economically important in marine bivalve and is used in fisheries industry among western Pacific Ocean Coasts especially in Korea, China, and Japan. The number of chromosomes in the blood clam is known as  $2n=38$ , but the genome size and genetic information of the genome are not still clear. In order to predict the genomic size of the *T. granosa*, the *in-silico* analysis analysed the genomic size using short DNA sequence information obtained using the NGS Illumina HiSeq platform. As a result, the genomic size of *T. granosa* was estimated to be 770.61 Mb. Subsequently, a draft genome assembly was performed through the MaSuRCA assembler, and a simple sequence repeat (SSR) analysis was done by using the QDD pipeline. 43,944 SSRs were detected from the genome of *T. granosa* and 69.51% di-nucleotide, 16.68% trinucleotide, 12.96% tetra-nucleotide, 0.82% penta-nucleotide, and 0.03% hexa-nucleotide were consisted. 100 primer sets that could be used for genetic diversity studies were selected. In the future, this study will help identify the genetic diversity of *T. granosa* and population genetic studies, and further identify the classification of origin between homogenous groups.

**Keywords:** *Tegillarca granosa*(꼬막), Illumina HiSeq platform(일루미나 하이식 플랫폼), MaSuRCA assembler, SSR(마이크로새틀라이트), Homogeneous(동종)

## 서론

꼬막은 연체동물문에서 이매패강에 속하는 종이다. 이 종은 주로 한국, 중국, 태국, 필리핀, 말레이시아를 따라 서태평양 연안의 중서부 지역에 주로 서식하며(Liu et al., 2017), 과거부터 현재까지 한국, 중국, 일본의 서태평양 해안을 중심으로 수산업에 있어서 매우 의미 있는 해양식품의 자원이었다(Liu et al., 2017). 예를 들어, 2010년에는 중국 갯벌 조개류 생산량의 30%에 해당되는 약 13만 톤의 꼬막이 양식되었다(Dong-li, 2011).

따라서, 수산업, 집단 유전학, 성별격리 등을 위한 *Tegillarca granosa*의 유전적 영역을 해결하기 위한 많은 유전학적 연구가 진행되고 있다. 굴, 홍합, 가리비와 같은 유전적 정보와 전사체의 정보가 양식 산업의 중요한 원천이지만, 돌조개과에 있어서는 여전히 제한적이다(Bai et al., 2019). 게다가 이매패류를 포함한 연체동물은 유전적 다양성이 높은 것으로 나타나, 수산업에 대한 집단 유전체학과 성별 분리를 해결하기에 어렵게 만든다(Penalzoza et al., 2014).

본 연구에서는 꼬막의 유전체 크기를 예측하고 드래프트 게놈 어셈블리를 통하여 마이크로새틀라이트 DNA 마커를 발굴하여 집단 유전체 연구와 중간 다양성 연구의 유전자 마커를 제시하고자 한다.

## 재료 및 방법

### 1. DNA 추출과 라이브러리 제작 및 게놈 시퀀싱

*T. granosa*의 샘플은 2019년 9월 대한민국 순천 벌교에서 포획되었다. 꼬막의 게놈 DNA (genomic DNA)는 근육 조직에서 추출했으며 페놀: 클로로포름: 이소아밀-알코올 방법을 이용하여 추출하게 되었다. DNA의 품질과 정량분석을 위해서 BioAnalyzer (Agilent Technologies, CA, USA)와 Qubit fluorometer (Invitrogen, Life Technologies, CA, USA)를 이용하여 확인하였다.

게놈 DNA는 ultrasonicator (Covaris, USA)를 사용하여 무작위로 350 bp 조각으로 절단하였고 paired-end library를 Illumina 제조사의 실험 방법에 따라 제작하였다. DNA 서열분석은 Illumina HiSeq 2000 (Liu et al., 2011)을 이용하여 수행하였다.

### 2. K-mer 분석, 게놈 조립과 마이크로새틀라이트 선별

생산된 염기서열 데이터는 어댑터, 폴리(N) 시퀀스, 낮은 품질의 리드들을 필터링하여 제거한 후, Jellyfish (Marcais and Kingsford, 2011) 외에 GenomeScope (Vurture et al., 2017)를 사용하여 K-mer 17, 19, 25를 값으로 이용하여 *T. granosa*의 게놈 크기를 예측했다. 그 중에서 19 K-mer를 최종 게놈 크기로 예측했다. 드래프트 유

**Table 1.** Statistics for the genome survey sequencing data of *T. granosa*

Raw data (bp)	Total reads	Q20 (%)	Q30 (%)	GC content (%)
87,857,793,048	350,031,048	76.2	73.0	55.4

전체 서열은 MaSuRCA (Maryland Super Read Cabog Assembler) (Zimin et al., 2013)을 이용하여 조립하였으며, 이는 *de Bruijn* graphs 방법과 겹침 기반 조립 전략의 유연성을 기반으로 한다. 콘티그(contig)와 스캐폴드(scaffold) 수준 어셈블리 통계를 계산하기 위해 *assemblathon\_stats.pl* script (<https://github.com/ucdavis-bioinformatics/assemblathon2-analysis>)을 사용했다.

꼬막의 드래프트 유전체를 활용하여 마이크로새틀라이트를 분석하기 위해 QDD 버전 3.1.2 파이프라인(Meglecz et al., 2014)을 사용하여 게놈 내 마이크로새틀라이트 반복단위를 확인했다. 파이프라인에서 설정된 한계로는 모노뉴클레오타이드(mononucleotides) 반복 횟수에서 헥사뉴클레오타이드(hexa-nucleotide) 반복 횟수이다. 기본 파라미터로 진행된 세 가지 QDD 파이프라인 단계를 사용하였다. 세 가지 QDD 단계로는 -config 1(1단계), -make\_cons 0(2단계), -config 1(3단계) 옵션이 추가되었다.

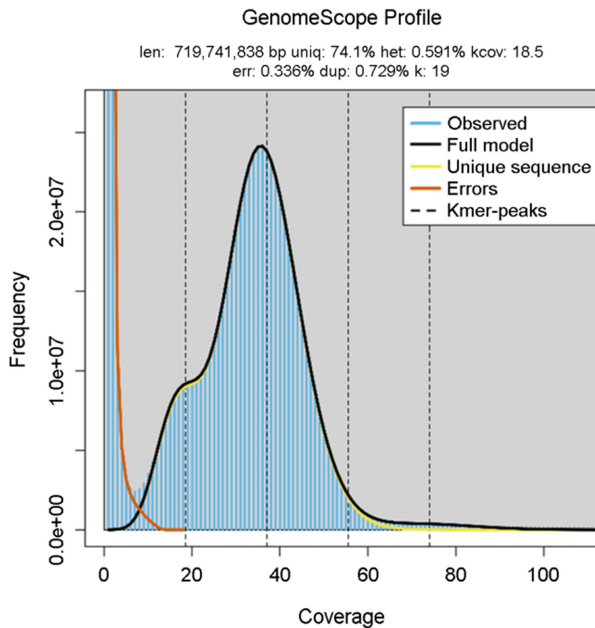
프라이머 쌍을 선별하기 위해서 예측 PCR product 크기로는 120~150 bp와 160~180 bp, 프라이머 길이로는 20~22 bp, 프라이머 용융 온도(Tm)로는 59~60°C, 다이뉴클레오타이드 모티프(dinucleotide motifs) 선택 등의 기준에 따라 Primer3 소프트웨어 (Rozen and Skaletsky, 2000)을 사용해서 선정했다.

## 결과

### 1. 게놈 염기서열 분석 및 K-mer 분석을 이용한 게놈 크기 예측

꼬막의 DNA 염기서열은 Illumina HiSeq 2000 장비를 이용하여 분석하였다. 분석된 꼬막의 DNA 염기서열 데이터 중 전체 리드로는 350,031,048가 확인이 되었고 모두 87.86 Gb의 데이터가 생산되었다. Seqtk (<https://github.com/lh3/seqtk>)를 이용해서 Q20, Q30 그리고 GC 함량이 총 76.2%, 73.0%, 그리고 55.4%인 것으로 확인하였다(Table 1).

*T. granosa* 게놈의 정확한 크기는 보고되지 않았지만, 핵형과 염색체 수(2n=38)는 *T. granosa* (Lu et al., 2008)에 대해서 보고되었다. 따라서, 게놈의 크기를 예측하기 위해 *in silico* 분석 방법 중 하나인 K-mer 분석을 진행했고, 그 결과, K-mer 19를 이용하여 최종적으로 추정되었다. 게놈의 크기는 719,741,838 bp로 확인되었고, 이형성 비율(heterozygosity rate), 중복성 비율(duplication rate)



**Fig. 1.** Distribution of K-mer analysis with GenomeScope (K=19). This represents the overall distribution of k-mer and genome size in *T. granosa*. The blue bar in the figure represents the observed k-mer and the yellow, orange lines in the figure represents unique sequences and errors. The statistics of overall analysis are displayed above the graph.<sup>1</sup>

및 오차율(error rate)은 0.591%, 0.729%, 그리고 0.336%로 확인하였다(Fig. 1).

## 2. 게놈 조립 및 향후 연구를 위한 마이크로새틀라이트 선별

게놈 조립 도구인 MaSuRCA (Zimin et al., 2013)에 의해 꼬막의 게놈을 조립하였으며, 그 결과, 스캐폴드(scaffold)의 총 수는 38,981개이고 가장 긴 스캐폴드는 1,157,674 bp이며 스캐폴드의 총 크기로는 793,914,538 bp이다. 또한, N50 스캐폴드 길이와 GC 함량은 148,364 bp와 27.55%로 확인하였다. 이로써, *T. granosa* 게놈 어셈블리의 전체 크기는 770.66 Mb (Table 2)로 확인하였다.

MaSuRCA에 의해 조립된 드래프트 게놈을 통하여 마이크로새틀라이트를 분석하는데 사용이 되었으며, 유전적 영역을 식별했고 QDD 버전 3.1.2 (Meglecz et al., 2014)을 사용하여 필터링을 통해 총 43,944개의 마이크로새틀라이트를 확인하였다. 검출된 전체

**Table 2.** Statistics of *T. granosa* assembly with MaSuRCA

	MaSuRCA
Number of scaffolds	38,981
Total size of scaffolds	793,914,538
Longest scaffold	1,157,674
Number of scaffolds > 1K nt	38,899 (99.8%)
Number of scaffolds > 10K nt	8,009 (20.5%)
Number of scaffolds > 100K nt	2,423 (6.2%)
Number of scaffolds > 1M nt	1
N50 scaffold length	148,364
L50 scaffold count	1,526
GC Contents (%)	27.55

마이크로새틀라이트는 다이-뉴클레오타이드(di-nucleotide) 69.51% (30,548), 트라이-뉴클레오타이드(tri-nucleotide) 16.68% (7,332), 테트라-뉴클레오타이드(tetra-nucleotide) 12.96% (5,694), 펜타-뉴클레오타이드(penta-nucleotide) 0.82% (360) 그리고 헥사-뉴클레오타이드(hexa-nucleotide) 0.03% (10)이다. 다이-뉴클레오타이드 중 가장 빈번한 모티프는 AT/AT (47.62%)였고, 그 다음이 AC/GT (27.93%)였다. 트라이-뉴클레오타이드 중 가장 빈번한 모티프는 AAT/ATT (70.11%)였으며, AAC/GTT (11.82%)가 그 뒤를 이었다. 테트라, 펜타, 헥사-뉴클레오타이드 중에서 AAAAT/ATTTT (73.56%), AAAAT/ATTTT (44.17%), AAAAAC/GTTTTT (20%), AAAAAAT/ATTTTT (20%)가 가장 풍부한 모티프였다(Table 3).

마이크로새틀라이트를 이용한 꼬막의 기존 연구에 의하면, 중국에 있는 개체에서 나타나는 마이크로새틀라이트의 loci가 말레이시아에 사는 꼬막의 개체에서는 나타나지 않았다는 것을 실험을 통해 확인하였다(Wang et al., 2013). 아직까지 정확한 원인은 밝혀지지 않았지만, 아주 긴 시간에 걸쳐 진화가 이루어지게 되면서 어떤 중요한 유전적 분화가 말레이시아와 중국 꼬막 개체 사이에서 이루어졌을 것이라고 추정한다(Wang et al., 2013). 따라서, 꼬막의 집단 유전학 연구 등을 위한 정확한 마이크로새틀라이트 마커를 선별하기 위해서 차세대 염기서열 분석을 이용해서 더 정확하고 명확하게 이루어져야 한다.

추후, 꼬막의 집단 유전학 연구 및 동종들 간의 원산지 분류를 위해 우리는 100쌍의 마이크로새틀라이트 마커를 Primer 3 소프트웨어(Table 4) 사용하여 선별하였다. 선별한 마이크로새틀라이트 마커는 사용적합성을 검정하기 위해서는 추가적으로 후속 검증과 연구가 필요하다.

<sup>1</sup> The len displayed on the figure means inferred total genome length and the uniq means the percentage of the genome that is unique which are not repetitive. The het means the overall rate of heterozygosity and kcov means the kmer coverage for heterozygous bases. The err and dup means the error rate of the reads and the average rate of read duplications.

**Table 3.** Statistics of Microsatellite Analysis by using QDD version 3.1.2

Repeat Motif	Number of repeats								Total
	5	6	7	8	9	10	11~20	>20	
Di-nucleotide (30,548)									
AC/GT	1,027	483	315	267	235	228	2,644	3,332	8,531
AG/CT	612	330	278	237	198	160	1,324	4,284	7,423
AT/AT	5,266	1,840	1,051	697	452	240	1,262	3,740	14,548
CG/CG	24	10	6	2	2	1	1	–	46
Tri-nucleotide (7,332)									
AAC/GTT	398	204	117	68	25	27	28	–	867
AAG/CTT	113	60	25	25	12	7	25	2	269
AAT/ATT	1,427	924	576	410	300	263	1,152	89	5,141
ACC/GGT	86	33	17	6	5	2	6	2	157
ACG/CGT	8	7	6	1	–	–	–	–	22
ACT/AGT	20	9	7	4	7	8	11	–	66
AGC/GCT	30	13	4	1	–	–	1	–	49
AGG/CCT	19	4	2	1	–	–	–	–	26
ATC/GAT	235	145	103	78	41	46	85	1	734
CCG/CGG	1	–	–	–	–	–	–	–	1
Tetra-nucleotide (5,694)									
AAAT/ATTT	1,301	700	465	347	270	225	867	13	4,188
AAAC/GTTT	149	48	11	6	1	1	3	–	219
AAAG/CTTT	84	50	14	6	2	7	10	–	173
AATC/GATT	75	35	20	15	10	7	10	–	172
ACAG/CTGT	64	24	11	4	–	1	2	–	106
AATT/AATT	57	22	8	3	5	3	3	–	101
AATG/CATT	53	55	30	24	22	19	53	2	258
AACC/GGTT	28	10	4	1	–	–	1	–	44
AGAT/ATCT	25	21	12	14	14	17	72	2	177
ACAT/ATGT	23	8	8	13	8	6	16	–	82
Others	80	42	25	9	7	6	5	–	174
Penta-nucleotide (360)									
AAAAT/ATTTT	52	20	10	21	9	13	34	–	159
AAAAC/GTTTT	39	5	2	3	–	–	2	–	51
AACTC/GAGTT	7	1	1	–	1	–	2	–	12
ACTCC/GGAGT	7	3	1	1	–	–	–	–	12

**Table 3.** Statistics of Microsatellite Analysis by using QDD version 3.1.2 (Continued)

Repeat Motif	Number of repeats								Total
	5	6	7	8	9	10	11~20	>20	
AATCC/GGATT	5	3	–	–	1	–	–	–	9
ACCCC/GGGGT	5	2	–	–	–	–	–	–	7
AAAAG/CTTTT	4	4	–	–	–	–	–	–	8
AAATC/GATT	2	1	–	–	–	–	–	–	3
AACAT/ATGTT	2	2	–	2	3	1	2	–	12
ACGAT/ATCGT	2	1	–	–	–	–	–	–	3
Others	26	10	8	7	5	5	21	2	84
Hexa-nucleotide (10)									
AAAAAC/GTTTTT	2	–	–	–	–	–	–	–	2
AAAAAT/ATTTTT	2	–	–	–	–	–	–	–	2
AAATTC/GAATTT	1	–	–	–	–	–	–	–	1
AACTAG/CTAGTT	1	–	–	–	–	–	–	–	1
AATATC/GATATT	1	–	–	–	–	–	–	–	1
AATCAC/GTGATT	1	–	–	–	–	–	–	–	1
ACATGC/GCATGT	1	–	–	–	–	–	–	–	1
ACCCCC/GGGGGT	1	–	–	–	–	–	–	–	1
Total	11,366	5,129	3,137	2,273	1,635	1,293	7,642	11,469	43,944

**Table 4.** Microsatellite primer set of *Tegillarca granosa*

Primer No.	Motif	PCR product size (bp)	Forward primer (5' → 3')	Tm (°C)	Reverse primer (5' → 3')	Tm (°C)
1	CG (5)	120	ACTGAAGATGATTGGTGGCCT	59.958	ATAAACAACCCGCAACCTGC	59.399
2	AC (9)	120	TGCATAGGACAAGACAACAAGC	59.186	CGAGATGACCCTTGTGACCA	59.998
3	AT (15)	120	ATTCCGACTCGTCAACAGCT	59.396	TGTTTCCGACTGAAGCTCAAGA	59.898
4	AT (5)	120	TCTTCGGAATAAGCTAGGCTGTT	59.549	ATACGTCAGCGATCCGTGTC	59.971
5	AC (7)	120	GTCTGTGTGTGTCGGTCTGT	59.898	TACAGACACACGACACACCC	59.614
6	AC (13)	121	GCCATCTAGACAGCTAAGTGGT	59.567	ACCATCACATACCCATGCCA	59.074
7	AC (7)	121	TGTCTGTGTGTGTCGGTCTG	59.898	TACAGACACACGACACACCC	59.614
8	AT (5)	121	ACACAAGCAGGCTACACTCC	59.965	CACTCTCAAGTGTCCCTCCC	59.389
9	AC (10)	122	TCTATAATTGCGCGCGTGTT	59.338	TCTTGTTGCCGACTCCATGT	59.604
10	AC (15)	122	ACTGGTAAAGTGCGAGTGAA	59.585	TCTGAAGCAACACGCACAC	59.623
11	AT (5)	122	CGCTCTTTATTGCTCGAAGAACA	59.632	AGACAGGAACTTCGTCACTTGA	59.306
12	AG (19)	122	GCATGCTGTCTCTTTCTCCA	59.746	AGAGTAGGCATGTGTGTATGTGT	59.484

**Table 4.** Microsatellite primer set of *Tegillarca granosa* (Continued)

Primer No.	Motif	PCR product size (bp)	Forward primer (5' → 3')	Tm (°C)	Reverse primer (5' → 3')	Tm (°C)
13	AG (8)	122	TGAGCCAGAATTTCACTGAAACC	59.43	ACAGAATTCAGGCCAGGCTT	59.593
14	AC (16)	122	TGAAGAAACATCTTTGGCAAGCA	59.616	ACAGCACCTAGATTGAATCCCA	59.733
15	AC (19)	123	AGGCATTGCTTGTGAGTCGA	59.965	GGCGACTGTAGCTAGTTGGA	59.184
16	AC (21)	123	GGCCATGCTAAACATTTCAAACC	59.316	AGATCCGACTGATTCTTCATGCT	59.613
17	AT (20)	124	TGTCGTGTGCTCGATTTCCA	59.967	GTCGAATCTGTACGGGACCC	59.898
18	AC (13)	124	ACTCAGAAGACCAAACCAGCA	59.51	TGGGTCACTGTGATCTTGCA	59.236
19	AC (19)	125	TCAGTCGGTTTGGTACTTGTGT	59.831	GTTGTCACAAGTGTGCTTTAGCT	59.935
20	AT (7)	126	GCCGGTATAGTTCGCTGTGA	59.898	TGTACGACAGCTGTTACCGG	59.758
21	AC (16)	127	TCTTGGCTTCAAATCAAAGTGGG	59.678	TGGCTGGCTAATCAAGCTGA	59.995
22	AT (8)	127	TGGTAGAGCATCCCACTGGTA	59.992	ACTCTGTGAGTGCTTGAGTGT	59.241
23	AC (13)	127	AGTACTGTCAAGTGTGGCACA	59.631	CTTGAGTCTCTGCCATCCTGA	59.17
24	AC (14)	127	CGAATTAAGTCAACCGAGCGG	59.676	ACGTCATGTCTACAGCCTCG	59.548
25	AT (8)	129	TCATCAAGGAAGTGAGGCGG	59.75	ACCGTCAAGACAGTACTGTTGT	59.57
26	AG (15)	129	CCAACCTCGTCAATGACTCTCT	59.768	GCAGTTGCAATGGGATCTCC	59.257
27	AG (12)	129	ACTCATGTTTCTCCTCCATCCC	59.493	CTGTTGCCATTGACCATCATGA	59.244
28	AC (10)	129	AGCAGCTTACGTCTTGTACG	59.27	AAACAGGGTCTGAAACGAGCT	59.859
29	AC (12)	129	GCTAGCACATAACAGACCATGC	59.711	ATGGCAGTGTGTGTTGTGC	59.897
30	AC (19)	129	TCATGGTGCTGATGCTTGGT	59.961	GCAACAACCAACCTCAAC	59.898
31	AT (10)	129	CACCAGCAAGGATAAATGCATGT	59.868	ACTTGGTTGTTAGGGCTACAGT	59.291
32	AC (12)	129	GCTAGCACATAACAGACCATGC	59.711	ATGGCAGTGTGTGTTGTGC	59.897
33	AG (5)	129	CCTGGAATGCGTAGACAGACA	59.795	TCTCTCACACACGAGCACAA	59.256
34	AT (6)	129	AGTCTCTGGTCATTGCGCA	59.965	GCTAGTTGGATTGCGGTGT	59.46
35	AG (15)	130	ACCAACTCGTCAATGACTCT	59.029	GCAGTTGCAATGGGATCTCC	59.257
36	AT (24)	130	GGTCATCTACCCACAACAGACA	59.698	AAGGGCTATCGCTATACCG	59.396
37	AT (10)	136	AGTGCAGCATAATCCCTGTCT	59.759	GCAACAACCTGTTCTCTGATGTCA	59.439
38	AT (18)	136	ATACGTAACCCGGCACCATC	59.895	AGAGATTTACAACCTGCCAAGT	59.925
39	AG (8)	138	GTTGCCAGTTTCCATGCCAA	59.608	CTTCTTTGTCTGGCACTCTGC	59.465
40	AT (19)	138	ACATGTCATGTGGGCAGACT	59.305	ACCCGGAATTATTCATTCTCGA	59.55
41	AC (6)	139	TCAGCTTGAGGGATTGAGCC	59.746	CTCGCATGAGAGAGCAGGAT	59.325
42	AT (7)	139	TACCTCCCACTCCAGTCTG	59.959	GCATGTATTCTGGCAGTGACAA	59.251
43	AT (5)	139	AGCCCAAGGTCTTTGCATCA	59.887	TGTACAATTGTAGCACCAGGGT	59.626
44	AT (5)	139	AGACATGCATCCGATTACCTCA	59.865	TGGCACATCAACATTAGCAGG	59.507
45	AT (10)	140	TGATACATTGGCCAGTTGTGC	59.769	TGACGCTGATTAGTCACGTGA	59.462
46	AC (12)	140	TCTTCTTTGTGCATGTGTGTGT	59.557	AGTGAAACCATTGACTGTAAGGC	59.179

Table 4. Microsatellite primer set of *Tegillarca granosa* (Continued)

Primer No.	Motif	PCR product size (bp)	Forward primer (5' → 3')	Tm (°C)	Reverse primer (5' → 3')	Tm (°C)
47	AC (14)	140	AAGTTGGTTCTCCCTTCTTCACA	59.798	AACTTGTTACCGCACCACCT	59.818
48	AC (19)	140	GGCTTACGCATTATGTACACCC	59.453	GCACATGACAACCTGAATTAGCCA	59.809
49	AT (8)	140	GCAGTCCACATAGCTAGAGCA	59.589	AGTATTACGGACGTGTTGTGAGA	59.497
50	AT (5)	200	CGGCGCATCAGATAATACG	59.092	CCGGCTTAAGTTTGCTTTGACT	59.707
51	AG (15)	200	CTGCAGTTCTCCTCAACCT	59.676	TGCTGATAGTAAAGGCGCAGT	59.793
52	AG (19)	200	ATTACCGTCGATGCCTGCTT	59.822	TACGGAGCGAGCACTTTACC	59.828
53	AT (9)	200	CGCACCGGTCCAGATTCTAA	59.825	TTCGTGTCGCTCGAACTACA	59.412
54	AT (5)	200	TCCTAGCCACGTGCAACTTT	59.892	AGGTGTTGAGAACCACCTTGT	59.436
55	AT (5)	200	GCCTGGGCTTTAGTCTGGAA	59.672	ACACTCTGGGTAAGGTGTTGA	59.225
56	AT (6)	200	GACAAGTTTGCTGGTGGCTG	59.969	CTGAAATGTGCCTGACGTCC	59.201
57	AG (7)	200	TCAACTGGTAGGGATTGGCC	59.376	AAGCTAGAATAAGGCGGGCA	59.163
58	AT (6)	200	CGAAGAGTCGCAAATTAGTCACC	59.939	ACCTGATGTGGGTCTGTGATT	59.016
59	AT (5)	200	AGCCGCCATGTTCTGAAAGA	59.963	TATGGTGTCCAGTGGTGTGG	59.31
60	AT (6)	200	GGGAATGGTCTACGTCCACTC	59.864	ACACCGACCTTCGAACTTCA	59.252
61	AC (5)	222	ACCGTGACACCACATACACC	59.966	GGTTTGTGTGTACGGTGTG	59.902
62	AT (6)	222	GCAGCCATTTCAGAATGCAC	59.002	TCAATACGAAGAATAGCCCGAGT	59.365
63	AT (5)	222	GGATTGCCAACACATCAGTCA	59.442	CTCTCATGGTTTGGTCTCGGT	59.724
64	AT (5)	222	GGCCATTAGAATGACCAGAACG	59.935	TCCAAATGTTGTGCTTGATCACA	59.364
65	AG (5)	222	AGCATCAAGAACCTAAATTGGGC	59.551	ATGGCAGTGATGTGACAGCA	59.963
66	AT (5)	222	CCACCATCATCAGTCTGCA	59.746	ACCACTGAAGTGTGCTACCT	59.231
67	AG (18)	222	CCTTCTCACTTTACAGGCGTG	59.52	AGCTTGCTCAAGTTGAATGTGCA	59.311
68	AC (13)	226	CACTTGCGCAAACAGTCACT	59.622	CACAGATGATGGTTATTGGTGGC	59.933
69	AT (8)	227	TGGTGGTCACGGAACAAAGT	59.746	CCCAGATTCGAACCTAGCTGA	59.243
70	AT (6)	229	TACGTGCCGCTGTGACTTA	59.685	CCAACGAAGGCTCTATACCCA	59.239
71	AT (5)	229	GGTAACTATGGCGTGGTAGCA	59.863	TCAGATCTACCACGACAGCG	59.267
72	AC (6)	229	CCTTGGAATGGTTGTCTGCC	59.109	TGTGCTCAACCAGTCAAGTCA	59.791
73	AC (15)	229	GCGAACTACATGATGTTAGCCT	59.128	TTGCCCACAAGAGGATTGCT	59.887
74	AT (8)	232	ACACCATCCTTTCACCTCACT	59.557	TCAGGATGAGGATGATGATGTCG	59.742
75	AC (5)	232	GCAATTCCTGTTGCCTCACA	59.659	TGCAAAGGTTCTATTAGTCTCGGA	59.992
76	AC (11)	232	CCACGTGCTGCTTTGGAATT	59.685	TGTCAAGTACATCCGCATCTGA	59.504
77	AT (12)	232	TGACATTTGCCTCTCAGTCTGT	59.63	CCACAATGCCACACACAACA	59.541
78	AT (6)	232	ACAAACTTGCCATTGACCCG	59.327	GCTAGTTTGGAATTGCGGTGT	59.46
79	AC (15)	232	TTGGTGTAAGGGCGGTTC	59.817	GCTCAACAATGACCAAGCAGG	59.475
80	AT (6)	232	ACGGAAGCATCTGATATGCCA	59.582	GGGAGCCGTTAAGTATCTGGG	59.93

**Table 4.** Microsatellite primer set of *Tegillarca granosa* (Continued)

Primer No.	Motif	PCR product size (bp)	Forward primer (5' → 3')	Tm (°C)	Reverse primer (5' → 3')	Tm (°C)
81	AT (5)	234	GGTGGAGGTGGAGCTTGTA	59.309	TGCAGCAATTTGATGTAACGGT	59.444
82	AT (5)	234	TTCGACGACAGTGTGAGTGT	59.263	TGCGCAATAATAAATCTGCTGCT	59.684
83	AT (7)	234	TCGAAGCAGTTAAGCAGGGT	59.315	GTGATGCCTCGTAGTTGTTTCAG	59.876
84	AT (5)	235	TCTGTGGGTGACATTGATTGGT	59.892	AGGAATTCAAAGTGAGCGGAGA	59.697
85	AG (5)	235	AGCTTATTACTTTCCGCTAGCA	59.866	TTCCAGAACAGACTTCCGC	59.965
86	AT (7)	235	ATGCTGCTTCAGGAACGAGA	59.39	CCTACGACGCATATCTACTTGGA	59.499
87	AG (5)	235	GGATTACGGTCAAGCGACAC	59.003	AGGCAAAGACGAGGTTCTGT	59.241
88	AT (11)	235	AAACGAGTTTGACGCCAAC	59.898	TAGCACGTGTCTAGTACAGGG	59.114
89	AC (7)	235	CCGGAATGGTTTGTCTCTGTG	59.774	CACATGCATGGGTATACACACC	59.382
90	AT (5)	235	GCCTACTACCGACACAGTT	59.107	GCATTCCATACAATCGGACACA	59.058
91	AT (7)	236	GCCTGTTGTCTGTCCATGC	59.758	AGTTTCACACCTCACTGACCTC	59.897
92	AG (5)	240	TTGACAGGTGCCTCTATCTG	59.166	ACTATCCACATACAGAGTGAGC	59.361
93	AG (11)	240	TTGCCTCATGTTACCCTCTTGA	59.358	TACTGGTAAATCCTCCATCGGC	59.632
94	AG (11)	240	ATTGCCTCATGTTACCCTCTTGA	59.733	ACTGGTAAATCCTCCATCGGC	59.859
95	AC (10)	240	GCCAGAGCAACAAGGATCATG	59.596	AACAGGCACAGAGGGACTTG	59.891
96	AT (5)	240	AGAATTGCGGGCTAGCTTCA	59.747	AATCCGCCGAGAATGACTC	59.895
97	AC (5)	240	AACAGCTTCATGGAGATGAGGT	59.425	CAGACTTACAAACAGGCAGGC	59.463
98	AC (18)	240	TACAATCCACTGCTTATCCGGG	59.896	AACGATCGTTCCTACTCCGC	59.9
99	AC (17)	240	GTTGGTAACCAAGGTGATCAGC	59.512	AAGTGGGCTCTAGTACAATGCA	59.428
100	AG (15)	240	TCGCCTTATTGCAACACAGC	59.479	AGCCTGAAGTTGCACCAGAT	59.597

## 고찰

*T. granosa*는 특히 한국, 중국, 일본의 서태평양 해안의 중요한 수산 자원이다(Liu et al., 2017). 꼬막이 중요한 수산 자원이기 때문에 질병 또는 양식업에 있어서 도움이 될 수 있는 유전적 영역을 알아내려고 많은 연구와 노력들이 행해지고 있다. 물론, 잔류성 유기 오염 물질(Persistent Organic Pollutants, POPs)과 면역반응에 관한 연구(Su et al., 2017), DCMU와 같은 살충제 농도에 따른 연구(Mohamat-Yusuff et al., 2020) 등 꼬막에 관한 수많은 연구들이 이루어지고 있지만, 이러한 문제점들을 좀 더 쉽게 접근하기 위해서, 전체 게놈을 알아내는 것이 더 도움이 될 수 있다.

꼬막의 전체 게놈을 알아내기 전에, 우리는 K-mer 분석을 이용하여 게놈의 크기를 719,741,838 bp로 추정했으며, 중복성 비율, 이형성 비율 및 오차율은 0.729%, 0.591% 그리고 0.336%로 확인되었다(Fig. 1). *T. granosa*의 게놈 어셈블리 총 크기는 770.66 Mb

였으며, 수집된 게놈 데이터는 QDD 버전 3.1.2 (Meglecz et al., 2014)에 의해 마이크로새틀라이트를 분석하였다. 꼬막의 마이크로새틀라이트는 총 43,944개로 확인하였다. 마이크로새틀라이트 같은 경우 전체 유전체의 상당한 부분을 차지하고 있고, 두 가지 측면에서 모두 코딩 유전자 시퀀스를 능가할 만큼 매우 다형적이다(Katti et al., 2001). 이러한 이유 때문에, 마이크로새틀라이트의 높은 돌연변이 비율은 진화 연구에 있어서 중요하고, 모집단 유전자형 및 마커 지원 사육과 같은 매우 폭넓은 응용 분야에서도 유용하고 유용하다(Mohamat-Yusuff et al., 2020). 또한, 마이크로새틀라이트 분석을 통해 중요한 시퀀스의 특성에 대해서 더 나은 이해를 제공하고 게놈 조직에서 마이크로새틀라이트의 역할에 대한 연구를 촉진한다(Manee et al., 2020).

추가적으로 집단 유전학을 연구하기 위해, Primer 3 소프트웨어(Rozen and Skaletsky, 2000) 이용하여 100쌍의 마이크로새틀라이트 마커를 선별하게 되었다(Table 4). 이를 통해, 유전학 및 진화



연구를 위한 풍부한 마커의 분리 및 개발을 포함하여 앞으로 많은 수산업 분야에 유용하게 도움이 될 것이다. 특히, 가장 빈번하게 일어나는 반복을 밝혀내는 것은 한 종 내의 개체를 구별하거나 심지어 이종 간에도 유용한 라이브러리 기반의 마커 선택에 있어서 필수적인 출발점을 제공하게 될 것이다(Manee et al., 2020). 이러한 데이터 기반이 갖춰지게 된다면, 이 종의 수산업 개발과 보존 계획을 지원하는데 있어서 의미 있을 것이다. 게다가, 마이크로새틀라이트는 그 종의 지리적 분포 전반에 걸친 추가적인 유전적 개체 연구 말고도, 양식활동이나 수산업 분야에 필요한 다른 유전자 분석을 가능하게 할 것이다(Fernández-Pérez et al., 2019). 앞으로의 후속연구를 위해서, 마이크로새틀라이트 마커의 사용 적합성을 검증하기 위해서는 추가적인 연구와 검증이 필요할 것이다.

## 참고문헌

- Bai CM, Xin LS, Rosani U, Wu B, Wang QC, Duan XK, Liu ZH, Wang CM. 2019. "Chromosomal-level assembly of the blood clam, *Scapharca* (Anadara) *broughtonii*, using long sequence reads and Hi-C." *Gigascience* 8.
- Dong-li Z. 2011. Analysis of genetic variation in the fast growth families of *Tegillarca granosa*. *Journal of Fisheries of China* 35: 350-357
- Fernández-Pérez J, Nantón A, Arias-Pérez A, Insua A, Méndez J. 2019. "Fifteen novel microsatellite loci, developed using next-generation sequencing, reveal the lack of genetic structure in *Donax vittatus* from Iberian Peninsula." *Estuarine, Coastal and Shelf Science* 217: 218-225.
- Katti MV, Ranjekar PK, Gupta VS. 2001. "Differential Distribution of Simple Sequence Repeats in Eukaryotic Genome Sequences." *Molecular Biology and Evolution* 18: 1161-1167.
- Liu B, Teng S, Shao Y, Chai X, Xiao G, Fang J, Zhang J, Wang C. 2017. "A Genetic Linkage Map of Blood Clam (*Tegillarca granosa*) Based on Simple Sequence Repeat and Amplified Fragment Length Polymorphism Markers." *Journal of Shellfish Research* 36: 31-40.
- Liu L, Hu N, Wang B, Chen M, Wang J, Tian Z, He Y, Lin D. 2011. "A brief utilization report on the Illumina HiSeq 2000 sequencer." *Mycology* 2: 169-191.
- Lu R, Lin Z, Zhang Y, Chai X, Dong Y, Xiao G, Zhang J-M, Fang J, Hu L-H. 2008. "Comparison on the karyotypes of *Scapharca subcrenata*, *Tegillarca granosa* and *Estellarca olivacea*." *Journal of Shanghai Fisheries University* 17: 625-629.
- Manee MM, Al-Shomrani BM, Al-Fageeh MB. 2020. "Genome-wide characterization of simple sequence repeats in Palmae genomes." *Genes Genomics* 42: 597-608.
- Marçais G, Kingsford C. 2011. "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers." *Bioinformatics* 27: 764-770.
- Meglecz E, Pech N, Gilles A, Dubut V, Hingamp P, Trilles A, Grenier R, Martin JF. 2014. "QDD version 3.1: a user-friendly computer program for microsatellite selection and primer design revisited: experimental validation of variables determining genotyping success rate." *Mol Ecol Resour* 14: 1302-1313.
- Mohamat-Yusuff F, Zulkarnain Z, Anuar NZA, Joni AAM, Kusin FM, Mohamed KN, Zulkeflee Z, Asha'ari ZH, Zulkifli SZ, Arshad A, Ismail A. 2020. "Impact of Diuron contamination on blood cockles (*Tegillarca granosa* Linnaeus, 1758)." *Mar Pollut Bull* 161: 111698.
- Penalzoza C, Bishop SC, Toro J, Houston RD. 2014. RAD Sequencing reveals genome-wide heterozygote deficiency in pair crosses of the Chilean mussel *Mytilus* spp. *Proceedings of the World Congress on Genetics Applied to Livestock Production. Species Breeding: Breeding in Aquaculture Species*: 275.
- Rozen S, Skaletsky H. 2000. "Primer3 on the WWW for general users and for biologist programmers." *Methods Mol Biol* 132: 365-386.
- Su W, Zha S, Wang Y, Shi W, Xiao G, Chai X, Wu H, Liu G. 2017. "Benzo[a]pyrene exposure under future ocean acidification scenarios weakens the immune responses of blood clam, *Tegillarca granosa*." *Fish Shellfish Immunol* 63: 465-470.
- Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. 2017. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 33: 2202-2204.
- Wang Y, Zeng Q, Xu L. 2013. "Population structure of the blood clam (*Tegillarca granosa*) in China based on microsatellite markers." *Genet Mol Res* 12: 892-900.
- Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. "The MaSuRCA genome assembler." *Bioinformatics* 29: 2669-2677.