KFAS 한국수산과학회지 Korean Journal of Fisheries and Aquatic Sciences
한수지 54(4), 489-497, 2021

Original Article
Korean J Fish Aquat Sci 54(4),489-497,2021

전복류(Genus *Haliotis*)의 분류를 위한 단일염기변이 기반 기계학습분석

노은수* · 김주원 · 김동균

국립수산과학원 양식산업연구부 생명공학과

Machine Learning SNP for Classification of Korean Abalone Species (Genus *Haliotis*)

Eun Soo Noh*, Ju-Won Kim and Dong-Gyun Kim

Biotechnology Research Division, National Institute of Fisheries Science, Busan 46083, Korea

Climate change is affecting the evolutionary trajectories of individual species and ecological communities, partly through the creation of new species groups. As population shift geographically and temporally as a result of climate change, reproductive interactions between previously isolated species are inevitable and it could potentially lead to invasion, speciation, or even extinction. Four species of abalone, genus *Haliotis* are present along the Korean coastline and these species are important for commercial and fisheries resources management. In this study, genetic markers for fisheries resources management were discovered based on genomic information, as part of the management of endemic species in response to climate change. Two thousand one hundred and sixty one single nucleotide polymorphisms (SNPs) were discovered using genotyping-by-sequencing (GBS) method. Forty-one SNPs were selected based on their features for species classification. Machine learning analysis using these SNPs makes it possible to differentiate four *Haliotis* species and hybrids. In conclusion, the proposed machine learning method has potentials for species classification of the genus *Haliotis*. Our results will provide valuable data for biodiversity conservation and management of abalone population in Korea.

Keywords: Machine learning, Single nucleotide polymorphism, Abalone, Fisheries resource management

서 론

1930년대에 생물학자들은 인위적 교란이 서로 분리된 두 종 간의 교잡을 초래할 수 있음을 인식하였으며(Wiegand, 1935; Riley, 1938), 이후 생물의 서식지 이동 현상은 교잡종 발생 확률을 높일 수 있음을 언급하였다. 이러한 현상을 설명하기 위하여 처음으로 서식지의 혼성화(hybridization of the habitat)라는 문구가 활용이 되었다(Anderson, 1948). 이후 생물 교잡에 대한 연구는 지속적으로 발전하였으며, 연구를 통해 의도적으로 교잡화된 생물은 일부 농업, 양식업 등에서 매우 중요한 부분을 차지하게 되었다. 하지만 기후변화로 인한 비의도적인 표층수온상승은 해양생물의 서식지 대이동이라는 생태계의 큰 변화를 통한 광범위한 서식지의 혼성화를 야기시키게 되었다(Amanda, 2014). 이러한 현상으로 인해 이종간의 자연적인 교잡의 기회가 증가되는 등 생물다양성보전 관점에서는 큰 문제점으로 인식되고 있으며(Rhymer and Simberloff, 1996; Allendorf et al., 2001), 생물 종 보존을 위한 대책마련이 시급한 상황이다. 생물 종의 식별은 분류학 연구의 기초이며 생태학 및 진화 연구를 포함한 생물학적 연구의 중요한 요소라 할 수 있다(Asaad et al., 2017). 생물 다양성 연구, 멸종 위기에 처한 개체군의 모니터링, 그리고 수산자원관리를 통한 양식 및 수산업 연구는 정확한 종 식별을 기초로 하고 있다(Schwartz et al., 2007; Breed et al., 2019). 생물 종은 일반적으로 형태학적 및 분자생물학적인 방법을 통해 식별할 수 있으며, 특히 분자생물학적 식별 방법은 형태학적 분류의 한계를 보완하여 생물 종의 식별 정확도를 향상시키는데 도움이 되고 있다(Candek and Kuntner, 2014). 생물 종의 식별을 위해 활용되는 주요 분자생물학적 방법 중 하나인 단일염기변이(single nucleotide polymorphism,

***Corresponding author:** Tel: +82. 51. 720. 2452 Fax: +82. 51. 720. 2456
E-mail address: laperm@korea.kr

https://doi.org/10.5657/KFAS.2021.0489
Korean J Fish Aquat Sci 54(4), 489-497, August 2021
Received 23 June 2021; **Revised** 23 July 2021; **Accepted** 2 August 2021
저자 직위: 노은수(연구사), 김주원(연구사), 김동균(연구사)

SNP)는 동일한 종 또는 다른 생물종의 구성원간의 단일염기 변이가 다를 때 발생하는 유전적 변이이다. 동일한 위치에서의 이러한 유전자의 변이는 동물 및 식물 모두에서 효과적인 유전적 마커로 사용이 될 수 있으며, 개체 및 집단의 구성을 밝히는 것이 가능하다(Leache and Oaks, 2017). 또한, 법의학적 연구 및 생물학적 진화 등 다양한 분야에서 SNP 분석 기술의 적용이 가능하여 활발한 연구가 수행되고 있다(Canturk et al., 2014; Leache and Oaks, 2017).

최근 차세대 시퀀싱 기술이 급속도로 발전하고 고도화됨에 따라 생물로부터 유전체정보를 보다 효율적으로 생산할 수 있게 되었다. 특히 genotyping-by-sequencing (GBS)는 차세대 시퀀싱 기술을 바탕으로 새롭게 발전하고 있는 분석법 중 하나이다(Narum et al., 2013; Sonah et al., 2013). GBS 분석은 제한효소(restriction enzyme, RE)를 사용하여 유전체 서열에서 잘리는 영역 주변의 서열만을 시퀀싱하게 되므로, 전장유전체 분석을 수행하지 않고도 넓은 범위의 유전체정보를 고루 확보할 수 있다(Scheben et al., 2017). 또한 적절한 제한효소를 선택함으로써 유전체 서열 내 반복적인 영역을 피할 수 있으며, 동시에 원하는 영역의 염기서열을 확보하는 것이 가능하다(Davey et al., 2011). 이러한 GBS 분석은 식물 종에서 SNP 유전형 분석을 위해 개발되었으나, 최근 수산생물을 포함한 다양한 생물 종을 대상으로 빠르게 확산되고 있다(Narum et al., 2013).

기계학습분석은 컴퓨터과학에서 가장 빠르게 성장하고 있는 분야로서, 차세대 시퀀싱 기술로부터 확보되는 대규모의 데이터세트에 숨겨진 복잡한 상관관계를 도출하는데 도움이 된다(Jordan and Mitchell, 2015). 이는 기존 데이터를 사용하여 모델을 생성하고 예측으로 이어지는 패턴 인식 및 분류를 용이하게 하는 다양한 알고리즘 세트를 사용한다. 기계학습 알고리즘은 크게 지도 학습(supervised learning)과 비지도 학습(unsupervised learning)으로 나눌 수 있다. 지도 학습은 잘 분류된 객체를 훈련하여 새로운 객체의 분류를 예측하는 방법이며, 비지도 학습은 훈련 과정 없이 제공된 객체를 분류하는 방법이다(Ang et al., 2015). 이러한 기계학습분석은 코딩영역 인식, 바이오마커 식별, 질병유전자 탐색 등 다양한 생물학적 분야에서 활용되고 있다(Swan et al., 2013; Han et al., 2016).

본 연구에서 우리는 차세대 시퀀싱(next generation sequencing, NGS) 기술과 기계학습(machine learning, ML) 기반의 접근 방법을 이용하여 국내 서식하는 전복의 종 보존을 위한 연구를 수행하고자 하였다. 전복은 원시복족목(Vetigastropoda), 전복과(Haliotidae)에 속하는 복족류로 전세계적으로 62종 24아종이 보고되어 있으며, 국내에는 한류성인 북방전복(*H. discus hannai*)과 난류성인 둥근전복(*H. discus discus*), 말전복(*H. gigantea*), 왕전복(*H. madaka*), 오분자기(*H. diversicolor supertexta*), 마대오분자기(*S. diversicolor diversicolor*)가 서식하는 것으로 알려져 있다(Kim et al., 2020). 북방전복은 우리나라의 주요 고부가가치 생물자원으로 육종연구 및 자원조성을 위한 방류사업에 활용되고 있다(Park et al., 2013). 또한, 둥근전복은 우리나라의 제주도 일대와 독도 지역에 주로 서식하는 것으로 알려져 있으며, 북방전복과는 아종관계로 형태적으로 뚜렷한 차이는 없으나 서식처에서 차이를 나타내고 있다(Ino, 1953). 인위적인 방법에 의한 북방전복과 둥근전복의 교잡화 가능 여부는 이미 보고된 바 있으며(Lee et al., 2016), 기후변화로 인한 이들 종의 서식지 혼성화는 자연적인 교잡을 야기시킬 수 있기 때문에 생물 종 보존을 위한 기반 마련이 필수적이라 할 수 있다.

재료 및 방법

시료 준비

전복 유전체의 분석을 위하여 형태학적 특징을 기반으로 분류된 북방전복 24미, 둥근전복 29미, 왕전복 9미, 말전복 18미를 수집하였으며, 각 시료로부터 근육조직을 채취하였다(Table 1). Genomic DNA의 추출은 TNES-Urea buffer (8 M urea, 10 mM Tris-HCL pH 7.5, 125 mM NaCl, 10 mM EDTA, 1% SDS)를 사용하였으며, 채취된 근육조직을 포함하는 480 μL의 TNES-Urea buffer에 20 μL의 Proteinase K (20 mg/mL)를 첨가한 후 56°C에서 근육조직이 완전히 분해될 때까지 반응시켰다. 이후 500 μL의 phenol:chloroform:isoamylalcohol (25:24:1)을 처리하여 단백질을 제거하였으며, 이후 2배의 99.9% ethanol과 0.1배의 3 M sodium acetate를 이용하여 genomic DNA를 정제하였다.

유전체 정보 대량생산 및 전처리

추출된 genomic DNA는 synergy HTX multi-mode reader (Biotek, Winooski, VT, USA)와 Quant-iT picoGreen dsDNA assay kit (Molecular Probes, Eugene, OR, USA)를 이용하여 20 ng/μL의 농도로 정량하여 분석에 사용하였다. 200 ng의 Genomic DNA는 adapter ligation을 위해 8 U의 high-fidelity Pst I (New England BioLabs, Ipswich, MA, USA) 제한효소(RE)를 이용하여 37°C에서 2시간 동안 분해(digestion) 후, 20분 동안 65°C로 열처리하여 제한효소를 비활성화 하였다. 이후 QIAquick PCR purification kit (Qiagen, Valencia, CA, USA)을 사용하여 제한효소 처리된 DNA를 정제하였으며, accupower Pfu PCR premix (Bioneer, Daejeon, Korea) 및 25 pmol의 illumina barcode adapter를 이용하여 연결(ligation) 과정을 거쳤다. PCR 산물은 QIAquick PCR purification kit를 사용하여 정제하여 single end GBS library를 구축하였다. 구축된 GBS library는 bioAnalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA)를 사용하여 품질 평가 후 Illumina NextSeq 500 (Illumina, San Diego, CA, USA) 플랫폼을 이용하여 염기서열 분석을 수행하였다. 확보된 데이터는 barcode sequence를 이용하여 demultiplexing (GBSX v 1.3) 후 (Herten et al., 2015),

adapter sequence 제거 및 sequencing quality trimming (trimmomatic v0.33)을 수행하였다(Bolger et al., 2014).

유전체 정보 분석

처리 과정을 거친 각 시료별 clean reads의 공통 염기서열을 분석하기 위해 국립수산과학원에서 해독한 북방전복의 표준유전체 정보(PRJNA3174)를 이용하여 mapping (Bowtie v2.2.5)을 수행하였다(Langmead and Salzberg, 2012; Nam et al., 2017). 이후 mapping 된 유전체 정보로부터 종 식별을 위한 SNP 정보를 탐색하였다. GATK의 variant filtering module을 사용하여 기본 filtering (phred quality score<30, FS>200)을 수행하였고, 선별된 SNP의 신뢰성을 높이기 위해 좀 더 엄격한 조건(read depth<100, phred quality score<50, FS>100)으로 추가 filtering을 수행하였다(McKenna et al., 2010).

유전체 정보 기반 전복류 집단 분석

유전체 정보 비교 분석을 통해 탐색된 SNP 정보를 기반으로 주성분분석(principle component analysis, PCA)을 수행하여 전복류의 집단 분석을 수행하였다. 분석에는 R software v3.4.4이 사용되었으며, 데이터의 변환 및 주성분분석을 위해 dgsfmt package와 SNPRelate package가 이용되었다(Zheng et al., 2012).

전복 품종 구분을 위한 SNP 마커 선별

주성분분석을 통해 분류된 전복류 집단을 기준으로 기계학습분석 프로그램인 Weka software v3.8.3을 이용하여 품종 분류를 위한 SNP 선별을 수행하였다. Weka software의 attribute selected classifier filter를 이용하였으며, RF 알고리즘으로 분류하고자 하는 집단을 효율적으로 선별할 수 있는 SNP 정보를 상위 순서대로 선발하였다. 또한 선발된 SNP는 다양한 통계학적 알고리즘(Bayesian Network, Sequential Minimal Optimization, K-Nearest Neighbor, C4.5, RandomForest)을 대상으로 분류 분석을 수행하여 최적의 마커를 선정하는 방법으로 수행하였다.

유전체 정보 기반 전복 품종 분류

기계학습분석을 이용한 전복 품종 분류에는 국내 연안에서 확보된 전복 총 666개체가 사용되었다. 유전자형 분석은 Fluidigm® SNP Type™ assay 방법이 이용되었으며, 최종 선별된 총 41개의 마커를 대상으로 유전자형(genotype) 분석을 수행하였다. SNP calling 후 확보된 유전자정보는 테스트데이터(test set)로 하며, 주성분분석을 통해 분류된 품종의 유전체정보는 훈련

Table 1. The sample information of 4 *Haliotis* sp.

Sample	area	sampling date	no. of ind.
Haliotis discus hannai	Genetics and breeding research center (NIFS)	2017.05.	2
	Gangwon je-do	2017.09.11	17
	Uljin	2017.07.18	5
Haliotis discus discus	Gyeonju (provided by FIRA)	2017.07.18	1
	Yangyang (provided by FIRA)	2017.07.18	6
	Uljin (provided by FIRA)	2017.07.18	1
	Japan	2016, 2018	7
	Gangwon Je-do	2017.09.11	13
	Jeju Chuja-do	2016.06.30	1
Haliotis madaka	Yeosu Back-do	2017.12.19	2
	Jeju	2016.06.30	7
Haliotis gigantea	Jeju	2016.06.30	10
	Jeju (provided by FIRA)	2017.07.18	8

Table 2. Summary of sequencing information from 4 *Haliotis* sp.

Sample	Cluster			Clean reads (Mbases)
	Raw	Pass-Filter	Demultiplexing	
Haliotis discus hannai	994,247,224	445,098,478	125,441,081	18,937
Haliotis discus discus			110,334,837	16,661
Haliotis madaka			104,454,953	15,773
Haliotis gigantea			104,897,607	15,840

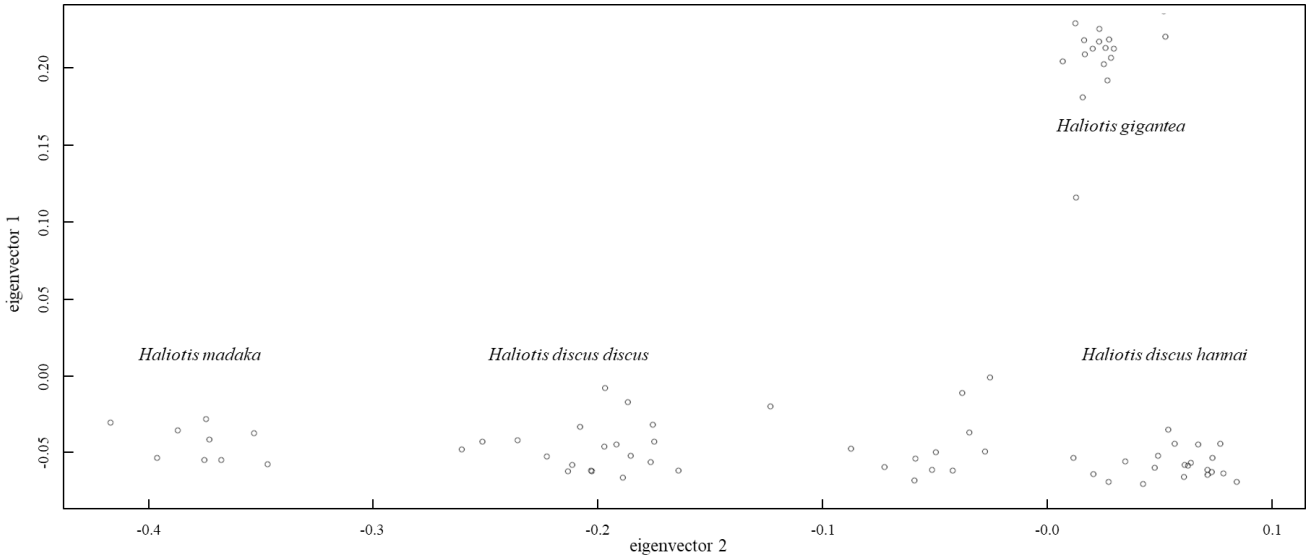

Fig. 1. Distinguish 5 groups *Haliotis* sp. by principal component analysis using 33,125 SNPs.

데이터(training set)로 하는 지도학습으로 품종 분류가 수행되었다.

결 과

유전체 정보 분석

GBS 분석을 통해 80개체의 전복류로부터 최초 994,247,224개의 raw cluster가 생산이 되었으며, pass-filter 후 445,098,478개의 데이터 생산이 가능한 cluster를 확보하였다. 확보된 cluster는 demultiplexing 및 sequencing trimming을 통해 최종적으로 북방전복 18,937 Mbases, 둥근전복 16,661 Mbases, 왕전복 15,773 Mbases 그리고 말전복으로부터 15,840 Mbases의 clean reads를 확보하였다(Table 2). 확보된 clean read를 북방전복 표준유전체에 mapping한 결과 북방전복은 2,995,114 bp (coverage 15.9%), 둥근전복은 3,081,879 bp (coverage 16.4%), 왕전복은 2,688,664 bp (coverage 14.3%) 그리고 말전복은 2,652,455 bp (coverage 14.1%)가 확인되었다(Table 3).

이후 mapping된 유전체정보로부터 Haplotype caller를 이용하여 전복 개체간 차이를 나타내는 유전자 변이정보 45,599개를 확보하였으며, 이중 SNP는 42,770개, InDel은 2,829개로 확인되었다. Variant filtering module을 이용한 기본 조건에서의 filtering 결과 SNP는 33,125개(77.45%), InDel은 2,475개(87.49%)가 선별되었으며, 모든 시료에서 genotype이 확보되어야 품종 구분에 활용될 수 있는 점을 감안하여 InDel을 제외한 33.125개 SNP만을 활용하여 주성분분석을 수행하였다. 또한 추가로 수행한 보다 엄격한 조건에서의 filtering 결과 최종적으로 2,161개(5.05%)의 SNP를 확보되었으며, 이는 기계학습 분석을 위한 품종분류에 활용이 되었다.

Table 3. Results of mapping against *Haliotis discus hannai* genome sequences

Sample	Reference match (bp)	Avg. depth	Coverage (%)
Haliotis discus hannai	2,995,114	61X	15.9
Haliotis discus discus	3,081,879	63X	16.4
Haliotis madaka	2,688,664	53X	14.3
Haliotis gigantea	2,652,455	57X	14.1

유전체 정보 기반 전복류 집단 분석

주성분분석 결과 전복류 80개체는 크게 5개의 집단으로 분류가 되었다. 북방전복, 둥근전복, 왕전복, 말전복과 이외 북방전복과 둥근전복 사이에 넓은 범위의 하나의 집단이 구성되었다. 분석 결과 초기 형태학적 분류와는 달리 일부 개체가 다른 품종에 속하는 것으로 확인되었으며, 북방전복 21개체, 둥근전복 20개체, 왕전복 9개체, 말전복 18개체, 그리고 북방전복과 둥근전복의 유전자를 일부 공유하고 있는 교잡종 추정 12개체로 분석 시료의 집단을 재분류 하였다(Fig. 1).

전복 품종 구분을 위한 SNP 마커 선별

기계학습을 이용한 품종 구분 마커 선별은 총 3단계의 과정을 포함하는 것으로 설계되었다. 먼저 순종 추정집단과 교잡종 추정집단을 선별하는 단계와 이후 순종 추정 집단으로부터 왕전복과 말전복을 선별하는 단계, 그리고 최종적으로 북방전복과 둥근전복을 선별하는 단계로 구분하였다.

Weka software로부터 2,161개의 SNP 정보를 대상으로 순

Table 4. Information of single nucleotide polymorphism for classification of 4 *Haliotis* sp.

	NO	CHR	Pos	Minor	Major	Sequence
Step 1 Pure/Hybrid	SNP01	HDSC00014	850987	A	C	TCACATATTA[A/C]ACAAAAACTG
	SNP02	HDSC00113	599426	T	C	TTGCCGTCAC[T/C]AGCTGTTCTG
	SNP03	HDSC00229	76162	A	C	GACGTCTTCT[A/C]TGCATGAGAC
	SNP04	HDSC00570	148315	T	C	GACTGTAACA[T/C]CTGGTAACGT
	SNP05	HDSC00651	51634	T	C	TTACAGAAGT[T/C]CTGATGACCC
	SNP06	HDSC01049	330465	T	C	TAGGAATAGT[T/C]AGCAGAAGTC
	SNP07	HDSC01168	67729	G	A	ATCGCTTCCA[G/A]AGTGGGAGAG
	SNP08	HDSC01473	137382	G	A	TGTCAAAATA[G/A]CTGCAGTCTT
	SNP09	HDSC01912	135699	C	T	CTCCGTTGTC[C/T]GCCAAACAAC
	SNP10	HDSC01933	226880	T	G	TAACCACTCG[T/G]GCTGTGCTCC
	SNP11	HDSC02821	38709	A	C	AGACATGTGT[A/C]GGCTCTTTCT
	SNP12	HDSC02862	81189	T	A	CAGGCGTGGA[T/A]CCTAACCTGT
	SNP13	HDSC03817	351918	A	G	AACGCTCAAA[A/G]TACATGCTTT
	SNP14	HDSC07287	32876	T	G	ATCATTGCGA[T/G]ACGTAGATGG
Step 2 *H. madaka*/*H. gigantea*	SNP15	HDSC00002	1160146	A	C	CAAAATAACC[A/C]AGCACTGTCG
	SNP16	HDSC00709	1469	G	A	TTCCTCATGG[G/A]GAGATCTGCA
	SNP17	HDSC00982	253667	G	T	GTAATTATCA[G/T]TGCAGCAGTA
	SNP18	HDSC01155	204003	C	G	TGTTGTGTTT[C/G]CAAGTTCGTT
	SNP19	HDSC01168	67723	A	T	CGAACGATCG[A/T]TCCAAAAGTG
	SNP20	HDSC01354	183450	A	G	CCTGCAGCCA[A/G]GCGCAAGTCT
	SNP21	HDSC01791	112542	T	C	CCTTCTTGTA[T/C]GCGGAAGGTA
	SNP22	HDSC02015	37957	G	A	TAACAAATGA[G/A]CAGGTAGATT
	SNP23	HDSC04358	59727	G	T	GTCAGGTTGC[G/T]TTGAGCTGCA
	SNP24	HDSC07320	48229	A	C	AATTTTCAAA[A/C]CAGTTTACCT
Step 3 *H. discus hannai*/ *H. discus discus*	SNP25	HDSC00057	378679	A	G	CTCTAGTGCC[A/G]ACCATGAGTA
	SNP26	HDSC00077	502497	A	G	CCTGAAGAAT[A/G]CCTGGCAGAA
	SNP27	HDSC00153	154876	T	G	TTTTCTTAAT[T/G]AAACCTACAC
	SNP28	HDSC00422	567280	T	C	AGTGGTATAC[T/C]TTACGTTACG
	SNP29	HDSC00570	148333	G	T	ACGTTGTCAT[G/T]TTTTGTCAAA
	SNP30	HDSC00597	342923	A	G	AGAGTCTCAT[A/G]TTCCGGAGGT
	SNP31	HDSC00673	92444	T	G	ATCCCTTGAC[T/G]AGCCTGCAGC
	SNP32	HDSC01025	199924	T	C	CTGTGGCATT[T/C]GATGACGTAC
	SNP33	HDSC01049	330471	T	A	TAGTTCAGCA[T/A]AGTCTGCAGA
	SNP34	HDSC01157	47321	T	C	GCTGCAGTCA[T/C]ATCTCCGTCA
	SNP35	HDSC01157	47336	T	G	TCCGTCACAC[T/G]CCACACCTTC
	SNP36	HDSC01168	67723	A	T	CGAACGATCG[A/T]TCCAAAAGTG
	SNP37	HDSC01385	90419	T	A	ATTCTAAACA[T/A]GCAGCTTGTT
	SNP38	HDSC01456	194409	C	T	GACATGTACA[C/T]GTCAGAAAGC
	SNP39	HDSC01516	161205	C	T	CTGCAGTCGT[C/T]GGTCAGATCT
	SNP40	HDSC02165	45581	C	T	TTTTTGTTCC[C/T]ATTTCACTGC
	SNP41	HDSC03211	198206	T	C	GTCGCTGGAC[T/C]AACTCTGCAG

종 추정집단(68개체)와 교잡종 추정집단(12개체)의 분류를 위한 최적의 마커 선별이 수행되었다. Random Forest (RF) 알고리즘을 기반으로 한 feature selection을 통해 최적의 마커 14개를 선별하였으며(Table 4), sequential minimal optimization (SMO) 알고리즘에서 96.25%의 정확도로 이들을 분류할 수 있음을 확인하였다. 순종 추정집단의 분류 정확도는 100% (true positive 68, false positive 0), 그리고 교잡종 추정 집단의 분류 정확도는 75% (false negative 3, true negative 9)로 확인되었다(Table 5).

왕전복과 말전복의 분류는 교잡종 추정집단이 제외된 왕전복 집단(3개체)와 말전복 집단(17개체) 그리고 북방전복과 둥근전복을 포함한 기타 집단(46개체)으로 데이터를 구성하여 기계학습분석에 사용하였다. 위와 동일한 방법으로 feature selection 결과 총 10개의 마커를 선별하였으며(Table 4), Bayesian network (BayesNet) 알고리즘에서 100%의 정확도로 왕전복(true positive 3, false positive 0)과 말전복(false negative 0, ture negative 17)의 분류가 가능하였다(Table 5).

북방전복과 둥근전복의 분류을 위해 북방전복 집단(40개체), 둥근전복 집단(6개체), 그리고 왕전복과 말전복을 포함한 기타 집단(20개체)으로 데이터를 구성하여 기계학습분석을 수행하였다. Feature selection을 통해 총 17개의 마커를 선별하였으며(Table 4), BayesNet 알고리즘에서 80.49%의 분류 정확도를 갖는 것으로 확인되었다. 북방전복의 분류 정확도는 90.05% (true positive 19, false positive 2), 그리고 둥근전복의 분류 정확도는 70.0% (false negative 6, true negative 14)를 갖는 것으로 나타났다(Table 5).

유전체 정보 기반 전복 품종 분류

국내 연안에서 확보된 666개체의 전복 품종 분류에는 교잡종 추정 그룹 분류에는 SMO 알고리즘, 왕전복과 말전복의 분류 및 북방전복과 둥군전복의 분류에는 BayesNet 알고리즘이 사용되었으며, 분석 결과는 두 방법으로 해석되었다. 먼저 순종 추정 그룹과 교잡종 추정 그룹의 분류에 있어서 정확도가 50% 이상(default)인 경우와 고품질의 순종확보를 위한 90% 이상의 정확도를 기준으로 분류가 수행되었다. 분석 결과 분류 정확도를 50%를 기준으로 하였을 경우 교잡종으로 예측되는 개체가 187개체(28.08%)로 확인되었으며, 북방전복은 41개체(6.16%), 둥근전복은 314개체(47.15%), 왕전복은 46개체(6.91%), 그리고 말전복은 78개체(11.71%)로 분류되었다. 또한, 정확도를 90%로 상향하였을 경우 교잡종으로 예측되는 개체는 427개체(64.11%), 북방전복은 24개체(3.60%), 둥근전복은 161개체(24.17%), 왕전복은 15개체(2.25%), 그리고 말전복은 39개체(5.86%)로 분류되었다(Table 6).

고 찰

Table 5. Accuracy of machine learning by algorithms

	Algorithm		BayesNet		SMO		KNN		C4.5		RF	
Step 1	TP	FP	68	0	68	0	68	0	67	1	67	1
	FN	TN	7	5	3	9	8	4	10	2	5	7
	Accuracy (%)		91.25		96.25		90.00		86.25		92.5	
Step 2	TP	FP	9	0	8	1	8	1	9	0	8	1
	FN	TN	0	18	0	18	0	18	1	17	0	18
	Accuracy (%)		100.00		98.53		98.53		98.53		98.53	
Step 3	TP	FP	19	2	19	2	17	4	17	4	17	4
	FN	TN	6	14	7	13	7	13	10	10	6	14
	Accuracy (%)		80.49		78.05		73.17		65.85		75.61	

SMO, sequential minimal optimization; KNN, k-nearest neighbor; RF, random forest.

Table 6. The species identification of abalones using machine learning analysis (accuracy 50% and 90%)

Accuracy (%)		Hybrid	*Haliotis discus hannai*	*H. discus discus*	*H. gigantea*	*H. madaka*
50	Gangwon Goseong (2018) (N=300)	57	22	173	16	32
	Gangwon Goseong (2019) (N=286)	65	18	138	26	39
	Ulleung-do (2019) (N=80)	65	1	3	4	7
90	Gangwon Goseong (2018) (N=300)	165	12	100	4	19
	Gangwon Goseong (2019) (N=286)	185	11	61	10	19
	Ulleung-do (2019) (N=80)	77	1	0	1	1

생물 종의 분류를 통한 자원의 모니터링 및 평가 등은 생물다양성 구성요소를 확인하기 위한 필수적인 요소이다(Appeltans et al., 2012). 현재까지 국내 전복류의 분류를 대상으로 수행된 선행 연구들로는 형태학적 기반 분류(Lee et al., 2014), 미토콘드리아 DNA의 염기서열 분석 및 이를 간편화한 PCR 분석법과 핵 DNA의 단편일렬반복(short tandem repeat, STR)을 이용한 방법 등이 이루어졌다(Seo et al., 2016; Dong et al., 2018). 최근에는 형태학적 분류의 경우 환경적인 요인에 의해 변화될 수 있는 점을 고려하여 유전학적 분류가 주로 사용이 되고 있다.

본 연구에서는 차세대 시퀀싱 분석법 중 하나인 GBS 방법을 활용해 형태학적 특징을 기반으로 선 분류된 둥근전복 속 4종의 전복 유전체 정보를 확보하였다. 확보된 유전체 정보는 비교유전체 분석을 통해 단일염기변이 정보를 선별하였고, 이를 기반으로 주성분분석을 수행한 결과 기존 형태학적 분류와는 상이한 결과가 확인되었다. 이러한 현상은 아종관계에 있는 북방전복과 둥근전복에서 나타났으며, 일부 개체에서 상호간의 유전적 교류가 이루어지고 있는 것으로 추정되었다.

이들 두 종간의 교잡종은 신품종개발 연구를 통해 인위적인 환경에서 이루어짐이 보고된 바 있으나 자연상태에서의 교잡종은 현재까지는 보고되지 않았다(Lee et al., 2016). 한류성인 북방전복의 생식주기는 1월에서 6월까지이며, 난류성인 둥근전복은 3월에서 10월까지로 일부 시기가 교차하기는 하나 서로의 서식지에 따른 차이로 자연상태에서의 교잡종의 발생 확률은 낮을 것으로 예측된다(Park et al., 2014; Kim et al., 2015). 하지만 급속화되는 기후변화에 따른 해수온도 상승은 전반적인 수서동물의 생식특성 및 서식지의 변화를 야기시키고 있으므로, 서식지의 이동 및 공유에 따른 자연상태에서의 교잡종 발생을 우려하지 않을 수 없다.

자연적으로 발생하는 교잡종의 경우 넓게는 진화의 한 과정으로 볼 수 있으나, 생물다양성보전의 관점에서는 이와 반대된다(Barton, 2001). 전세계적으로 기후변화로 인한 생물 종의 감소를 예방하고 생물다양성을 보전하기 위한 노력이 계속되고 있으며, 특히 생물 자원과 생태계 환경의 지속적인 관찰, 수집, 분석을 통한 변화를 예측하고 대응하는 것이 가장 중요한 방법으로 알려져 있다(Muhlfeld et al., 2014). 그러므로 교잡종의 발생을 관찰하고 순종의 보호, 관리 방안을 마련하기 위해서는 유전자 마커를 개발하고 이를 활용한 지속적인 모니터링이 필수적이라 할 수 있다. 최근 국내에서는 미토콘드리아 cytochrome oxidase subunit I 유전자분석을 통한 붕어와 떡붕어의 교잡종 연구 그리고 초위성체(microsatellite) 마커를 활용한 참돔과 감성돔에 대한 연구가 수행된 바 있다(Kang et al., 2014; Kang et al., 2015).

대용량 유전체 정보에 의한 품종 분류는 표현형과 밀접하게 연관된 유전자 마커의 선별을 위해 기계학습분석을 통한 feature selection 방법의 활용이 요구된다(Jordan and Mitchell, 2015). 최초 2,161개의 SNP 정보를 기반으로 전복 품종 분류 및 예측을 위해 두 개의 서로 다른 기계학습모델이 구성되었다. 이러한 모델은 RF-SMO 및 RF-BayesNet으로 두 모델 모두 RF를 feature selection 알고리즘으로 사용하여 관련 SNP를 식별하였다. 그 후 선택된 SNP는 classification step에서 SMO 및 BayesNet 알고리즘을 통해 분류된다. RF는 효율적인 의사결정 트리를 구축하기 위해 각 기능의 예측 중요성을 측정하는 앙상블 특성을 갖추어 견고성 측면에서 다른 feature selection 방법을 능가하는 것으로 알려져 있다(Sylvester et al., 2017).

전복류의 분류를 위해 총 41개의 SNP 마커가 선별이 되었으며, 분류 단계별로 각각 순종과 교잡종의 예측에는 14개, 왕전복과 말전복의 예측에는 10개, 그리고 북방전복과 둥근전복의 예측에는 보다 많은 17개의 마커가 사용이 된다. 개별 집단의 예측 정확도에서 교잡종 추정 집단은 75%의 정확도를 나타내었으며, 북방전복은 90.05% 그리고 둥근전복은 70.0%의 값을 나타내었다. 예측 정확도를 높이기 위한 노력으로는 유전자 마커의 추가 선별과 기계학습분석의 훈련데이터로 사용될 수 있는 참조집단을 추가 확보하는 방법이 활용될 수 있다. 마커를 추가 선별하는 경우 기존에 분석된 유전체 데이터 외에 추가적인 분석이 요구될 수 있으며, 선별된 마커와의 상호연계성 등을 고려하여야 한다. 일반적으로는 선별된 마커를 활용하여 정확하게 분류가 이루어진 집단으로부터 유전자형분석을 통해 훈련데이터를 추가 보완하는 방법이 정확도를 향상시키기 위한 방법으로 활용되고 있다(Mitchell, 1999).

선별된 마커를 이용한 전복 품종 분류에는 북방전복이 주로 서식하는 것으로 알려진 울릉도와 강원도 고성지에서 확보된 자연산 전복을 대상으로 하였다. 북방전복은 우리나라의 중요 양식 대상종으로 자연산의 경우 양식계통의 유전학적 다양성 유지 및 자원복원을 위한 방류사업 등에 활용되고 있다. 확보된 666개체의 전복으로부터 유전자형 정보를 분석하고 기계학습분석을 통해 품종을 분류한 결과 분류 정확도 기준에 따라 187개체(28.08%)에서 426개체(63.96%)가 교잡종으로 예측이 되었다. 또한, 순종으로 예측되는 개체 중에서는 대부분이 둥근전복으로 확인되었다. 일부 지역에 국한된 전복을 대상으로 분석이 수행되어 우리나라 전체를 대표할 수는 없으나 분석에 사용된 전복이 주로 북방전복의 주요 서식지에서 확보되었음을 감안할 때 분석 전복의 절반 이상이 교잡종으로 예측이 되었다는 점은 교잡화로 인한 생태계 건강성의 악화가 급속도로 진행되고 있음을 알려준다.

유전체정보를 기반으로 한 생물 종의 분류는 유전학적 분류방법 중에서도 가장 높은 정확도를 나타낸다고 할 수 있다. 본 연구에서 개발된 전복류 품종 분류를 위한 SNP 마커는 정확한 분류가 이루어진 훈련데이터를 추가 확보함으로써 활용성을 보다 높일 수 있을 것으로 사료된다. 또한, 생물다양성보전을 위한 모니터링 사업과 해역특성을 고려한 방류사업 등에 적용하여 고유종에 대한 보호, 보존 및 지속적, 체계적 자원 관리 체계의 구축에 적극 활용될 수 있을 것이다.

사 사

본 연구는 2021년도 (R2021024)의 지원으로 수행되었으며, 연구비 지원에 감사드립니다.

References

Allendorf FW, Leary RF, Spruell P and Wenberg JK. 2001. The problems with hybrids: setting conservation guidelines. Trend Ecol Evol 16, 613-622. https://doi.org/10.1016/S0169-5347(01)02290-X.

Amanda JC. 2014. Hybridization in a warmer world. Ecol Evol 4, 2019-2031. https://doi.org/10.1002/ece3.1052.

Anderson E. 1948. Hybridization of the habitat. Evolution 2, 1-9. https://doi.org/10.2307/2405610.

Ang JC, Mirzal A, Haron H and Hamed HNA. 2015. Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection. IEEE-ACM Trans. Comput Biol Bioinform 13, 971-989. https://doi.org/10.1109/TCBB.2015.2478454.

Appeltans W, Ahyong ST, Anderson G, Angel MV, Artois T, Bailly N, Bamber R, Barber A, Bartsch I, Berta A, Blazewicz-Paszkowycz M, Bock P, Boxshall G, Boyko CB, Brandao SN, Bray RA, Bruce NL, Cairns SD and Costello MJ. 2012. The magnitude of global marine species diversity. Curr Biol 22, 2189-2202. https://doi.org/10.1016/j.cub.2012.09.036.

Asaad I, Lundquist CJ, Erdmann MV and Costello MJ. 2017. Ecological criteria to identify areas for biodiversity conservation. Biol Conserv 213, 309-316. https://doi.org/10.1016/j.biocon.2016.10.007.

Barton NH. 2001. The role of hybridization in evolution. Mol Ecol 10, 551-568. https://doi.org/10.1046/j.1365-294x.2001.01216.x.

Bolger AM, Lohse M and Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-2120. https://doi.org/10.1093/bioinformatics/btu170.

Breed MF, Harrison PA, Blyth C, Byrne M, Gaget V, Gellie NJC, Groom SVC, Hodgson R, Mills JG, Prowse TAA, Steane DA and Mohr JJ. 2019. The potential of genomics for restoring ecosystems and biodiversity. Nat Rev Genet 20, 615-628. https://doi.org/10.1038/s41576-019-0152-0.

Candek K and Kuntner M. 2014. DNA barcoding gap: reliable species identification over morphological and geographical scales. Mol Ecol Resour 15, 268-277. https://doi.org/10.1111/1755-0998.12304.

Canturk KM, Emre R, Kinoglu K, Baspinar B, Sahin F and Ozen M. 2014. Current status of the use of single-nucleotide polymorphisms in forensic practices. Genet Test Mol Biomark 18, 455-460. https://doi.org/10.1089/gtmb.2013.0466.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM and Blaxter ML. 2011 Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12, 499-510. https://doi.org/10.1038/nrg3012.

Dong CH, Lee MN, Kang JH, Park JY, Nam BH, Noh JK, Kim PY, Cho YA and Kim EM. 2018. Development of a rapid and simple method for identification of *Haliotis gigantean* using species-specific PCR. Korean J Malacol 34, 51-58. https://doi.org/10.9710/kjm.2018.34.1.51.

Han S, Liang Y, Li Y and Du W. 2016. Long noncoding RNA identification: Comparing machine learning based tools for long noncoding transcripts discrimination. Biomed Res Int 2016, 8496165. https://doi.org/10.1155/2016/8496165.

Herten K, Hestand MS, Vermeesch JR and Van Houdt JK. 2015. GBSX: a toolkit for experimental design and demultiplexing genotyping by sequencing expreriments. BMC Bioinformatics 16, 73. https://doi.org/10.1186/s12859-015-0514-3.

Ino T. 1953. Biological studies on the propagation of Japanese abalone (genus *Haliotis*). Bull Tokai reg Fish Res Lab 5, 1-102.

Jordan MI and Mitchell TM. 2015. Machine learning: Trends, perspectives, and prospects. Science 349, 255-260. https://doi.org/10.1126/science.aaa8415.

Kang JH, Noh ES, Lim JH, Han HK, Kim BS and Lim SK. 2014. Genetic differentiation of *Carassius auratus* and *C. cuvieri* by the cytochrome c oxidase I gene analysis. J Aquac Res Development 5, 1-4. https://doi.org/10.4172/2155-9546.1000231.

Kang JH, Yang SG, Kim EM, Noh ES, Kim DG, Kim BS and Choi TJ. 2015. Possibility of natural hybridization between red seabream *Pagrus major* and blackhead seabream *Acanthopagrus schlegeli*. J Life Sci 25, 16-20. http://doi.org/10.5352/JLS.2015.25.1.16.

Kim HJ, Kim HJ, Kim YS and Lee JS. 2020. Microstructural differentiation of the oocyte in the abalone *Haliotis discus hannai*. Korean J Fish Aquat Sci 53, 90-97. https://doi.org/10.5657/KFAS.2020.0090.

Kim JW, Lee BW, Kang JC, Min EY, Won SH, Lim HG, Kang SW, Jeon MA and Lee JS. 2015. Reproductive cycle of the abalone, *Haliotis discus discus* collected from Jeju island of Korea. Korean J Malacol 31, 21-26. https://doi.org/10.9710/kjm.2015.31.1.21.

Langmead B and Salzberg SL. 2012. Fast gapped-read alignment with bowtie 2. Nat Methods 9, 357-359. https://doi.org/10.1038/nmeth.1923.

Leache AD and Oaks JR. 2017. The utility of single nucleotide polymorphism (SNP) data in phylogenetics. Ann Ecol Evol Syst 48, 69-84. https://doi.org/10.1146/annurev-ecolsys-110316-022645.

Lee JK, Seo YB, Kim GD and Lim HK. 2016. Molecular and physiological aspects of breeding program for development of hybrids between abalones distributed in the coast of Korea. J Life Sci 26, 1218-1223. https://doi.org/10.5352/JLS.2016.26.10.1218.

Lee JS, Won SH, Kim SK, Lim HK and Lee JS. 2014. Classification and description of genus *Hordotis* (Gastropoda: Vestigastropoda) from Korea. Korean J Malacol 30, 79-86. https://doi.org/10.9710/KJM.2014.30.1.79.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M and DePristo MA. 2010. The genome analysis toolkit: a map reduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297-1303. https://doi.org/10.1101/gr.107524.110.

Mitchell TM. 1999. Machine learning and data mining. Commun ACM 42, 30-36. https://doi.org/10.1145/319382.319388.

Muhlfeld CC, Kovach RP, Jones LA, Chokhachy RA, Boyer MC, Leary RF, Lowe WH, Luikart G and Allendorf FW. 2014. Invasive hybridization in a threatened species is accelerated by climate change. Nat Clim Chang 4, 620-624. https://doi.org/10.1038/nclimate2252.

Nam BH, Kwak WR, Kim YO, Kim DG, Kong HJ, Kim WJ, Kang JH, Park JY, An CM, Moon JY, Park CJ, Yu JW, Yoon J, Seo MS, Kim KD, Kim DK, Lee SB, Sung SS, Lee C, Shin YH, Jung MH, Kang BC, Shin GH, Ka SJ, Anolles KS, Cho SA and Kim HB. 2017. Genome sequence of Pacific abalone *Haliotis discus hannai*: The first draft genome in family Haliotidae. Gigascience 6, 1-8. https://doi.org/10.1093/gigascience/gix014.

Narum SR, Buerkel CA, Davey JW, Miller MR and Hohenlohe PA. 2013. Genotyping-by-sequencing in ecological and conservation genomics. Mol Ecol 22, 2841-2847. https://doi.org/10.1111/mec.12350.

Park CJ, Nam WS, Lee JH, Noh JK, Kim HC, Park JW, Hwang IJ and Kim SY. 2013. Analysis of genetic divergence according to each mitochondrial DNA region of *Haliotis discus hannai*. Korean J Malacol 29, 335-341. https://doi.org/10.9710/kjm.2013.29.4.335.

Park MW, Kim HJ, Kim BH, Som M, Choi JS and Lee J. 2014. Reproductive cycle of the abalone *Haliotis discus hannai* collected from Jindo of Korea. Korean J Malacol 30, 243-248. https://doi.org/10.9710/kjm.2014.30.3.243.

Rhymer JM and Simberloff D. 1996. Extinction by hybridization and introgression. Annu Rev Ecol Syst 27, 83-109. https://doi.org/10.1146/annurev.ecolsys.27.1.83.

Riley HP. 1938. A character analysis of colonies of *Iris fulva*, *Iris hexagona var. giganticaerulea* and natural hybrids. Am J Bot 25, 727-738. https://doi.org/10.2307/2436599.

Scheben A, Batley J and Edwards D. 2017. Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. Plant Biotechnol J 15, 149-161. https://doi.org/10.1111/pbi.12645.

Schwartz MK, Luikart G and Waples RS. 2007. Genetic monitoring as a promising tool for conservation and management. Trends Ecol Evol 22, 25-33. https://doi.org/10.1016/j.tree.2006.08.009.

Seo YB, Kang SC, Choi SS, Lee JK, Jeong TH, Lim HK and Kim GD. 2016. Phylogenetic study of genus *Haliotis* in Korea by cytochrome c oxidase subunit 1 and RAPD analysis. J Life Sci 26, 406-413. https://doi.org/10.5352/JLS.2016.26.4.406.

Sonah H. Bastien M. Iquira E, Tardivel A, Legare G, Boyle B, Normandeau E, Laroche J, Larose S, Jean M and Belzile F. 2013. An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PLoS One 8, e54603. https://doi.org/10.1371/journal.pone.0054603.

Swan AL, Mobasheri A, Allaway D, Liddell S and Bacardit J. 2013. Application of machine learning to proteomics data: Classification and biomarker identification in postgenomics biology. OMICS 17, 595-610. https://doi.org/10.1089/omi.2013.0017.

Sylvester EVA, Bentzen P, Bradbury IR, Clement M, Pearce J, Horne J and Beiko RG. 2017. Applications of random forest feature selection for fine-scale genetic population assignment. Evol Appl 11, 153-165. https://doi.org/10.1111/eva.12524.

Wiegand KM. 1935. A taxonomist's experience with hybrids in the wild. Science 81, 161-166. https://doi.org/10.1126/science.81.2094.161.

Zheng X, Levine D, Shen J, Gogarten SM, Laurie C and Weir BS. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28, 3326-3328. https://doi.org/10.1093/bioinformatics/bts606.