Korean J. Malacol. 36(1): 1-6 2020
https://doi.org/10.9710/kjm.2020.36.1.1

[단보, Short communication]

연체동물의 유전체 및 전사체 연구 동향 (2020년)

상민규, 황희주, 정종민, 박지은, 송대권, 정준양, 강세원[1], 한연수[2], 이용석, 박소영[3]

순천향대학교 자연과학대학 생명시스템학과, [1]한국생명공학연구원 생물자원센터,
[2]전남대학교 농업생명과학대학 응용생물학, [3]국립낙동강생물자원관 다양성연구팀

Trends in genome and transcriptome research on mollusks in the world (2020)

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

Department of Life Science and Biotechnology, College of Natural Sciences, Soonchunhyang University, Asan, Chungnam, 31538, Korea
[1]Biological Resource Center, Korea Research Institute of Bioscience and Biotechnology, Jeongeup, Jeonbuk 56212, Korea
[2]Department of Applied Biology, Institute of Environmentally-Friendly Agriculture (IEFA), College of Agriculture and Life Sciences, Chonnam National University, Gwangju 61186, Korea
[3]Biodiversity Research team, Nakdonggang National Institute of Biological Resources, Sangju, Gyeongbuk, 37242, Korea

ABSTRACT

With the help of Next-Generation Sequencing (NGS) technology, comprehensive genomic and transcriptomic studies have been conducted by researchers working on mollusks around the world. Base on March 2020 data at the GOLD database, 237,877 genetic projects have been registered. It shows that these results are about 4 times higher than the data reported in 2015, meaning that the genetic projects have been steadily conducted with various species. Among them, a total of 71 cases were registered for mollusks (37 cases of bivalves, 28 cases of gastropoda, and 6 cases of cephalopoda). The genome project for mollusks has increased by about 30 cases compared to 2015, mainly in the United States and China. Besides, analysis of the genetic resources registered in the NCBI for ten years indicated that the genome projects have quadrupled depending on the type of database. In case of sequence read archive (SRA) database, 18,476 mollusks-related genomic studies (about 34.7 TB) have been registered. About 66 GB of data was registered by 2010, and also about 32,532 GB was registered from 2011 to 2019, meaning that there is a 500-fold increase over the decade. Taken together, it is expected that genomic research on mollusks will have many advantages such as the preemption of genetic resources.

Key words : Mollusks, Genome, NGS, SRA

서 론

차세대염기서열분석기 (Next Generation Sequencer) 의 발달로 전 세계적으로 유전체 (게놈) 프로젝트의 연구 비율이 급속도로 늘고 있다 (Bang *et al*., 2010; Liolios *et al*., 2010). 유전체 분석 비용과 시간은 감소한 반면 분석되어 나오는 데이터의 양이 폭발적으로 증가하여, 국제컨소시엄으로 진행되던 유전체 프로젝트는 국내컨소시엄 단위로 바뀌어 진행되기 시작하였으며, 최근에는 미생물, 곰팡이, 조류 등과 같은 유전체 길이가 비교적 짧은 생물은 실험실 단위에서 진행되고 있다 (Morozova and Marra, 2008; Yang *et al*., 2009). 또한 metagenome, chip-seq, slide-seq 등 새로운 분석 방법이 나타나면서 기존에 사용되던 Genomic sequencing, Transcriptomic sequencing에서 벗어나 다양한 방식의 연구가 진행되기 시작하였다 (Weber *et al*., 2011; Koboldt *et al*., 2013;

Received: March 22, 2020; Revised: March 26, 2020; Accepted: March 31, 2020
Corresponding author: So Young Park
Tel: +82 (54) 530-0832, e-mail: cindysory@nnibr.re.kr
Co-corresponding author: Yong Seok Lee
Tel: +82 (41) 530-3040, e-mail: yslee@sch.ac.kr

Table 1. Comparison of the GOLD database in 2015 and 2020

		2015	2020
	Archaeal	1,115	1,760
	Bacterial	45,925	163,529
	Microbial	4,564	9,320
Eukaryal	Complete	165	19,290
	Incomplete	11,065	78,203
	Permanent draft	952	139,519
	Targeted	823	865
	Total	65,058	237,877

Table 2. Current status of mollusk genomic research registered in NCBI

Database type	2010	2020
Nucleotide	1,276,010	5,192,310
Protein	62,953	775,084
Structure	246	534
Genome	99	467
Gene	1,234	259,569
SRA Experiments	-	18,476

Rodriques *et al*., 2019).

Dr. Fredric Sanger가 고안한 1세대 염기서열 분석 이후 Roche의 GS-FLS, Illumina (Solexa) 의 Genome Analyzer 등을 위시로 한 차세대 염기서열분석기 (Next Generation Sequencer, NGS) 가 개발되며, 2010년 이후 유전체 연구는 급속도로 증가하였다 (Droege and Hill, 2008; Morozova *et al*., 2008). 최근 차세대염기서열분석 장비의 성능을 충족하며, 더 긴 read length를 생산하는 PacBio SMRT, Oxford Nanopore 등의 3세대 장비가 개발, 발전되면서 다시 한번 유전체 연구의 비약적 성장이 이루어지고 있다 (John Eid, 2009; Lu *et al*., 2016).

이러한 결과 237,877개의 유전체 프로젝트가 진행 중이거나 완료되었다 (2020년 3월 기준). 이중 19,186개의 프로젝트가 완료되어 논문으로 출간되었으며, 78,203개의 프로젝트가 진행 중이다. 진행되고 있는 프로젝트 중 고세균 영역이 491개, 박테리아 영역이 19,262개, 진핵생물 영역이 57,748개이며, 그 외의 플라스미드, 바이러스 등의 영역이 702개로 확인되었다. 하지만 공개하지 않고 진행되고 있는 유전체 프로젝트 또한 무시할 수 없을 정도로 많다는 것을 고려하였을 때 집계된 정보보다 훨씬 많은 수의 유전체 연구 프로젝트가 진행되고 있을 것으로 예측된다.

본 논고에서는 앞서 언급한 유전체 프로젝트 중 연체동물을 대상으로 데이터 또는 프로젝트 내용을 공개한 경우로 한정하여 수행한 프로젝트들의 현황을 알아보고자 한다.

본 론

1. 전 세계 유전체 프로젝트 현황

GOLD (Genomes On-Line Database : https://gold.jgi.doe.gov) 데이터베이스에 따르면, 현재 40,976 종에 대한 유전체 프로젝트가 수행되고 있다 (2020년 3월 기준) (Liolios *et al*., 2010). 5년 전인 2015년 12월에 GOLD 데이터베이스에서 조사한 21,799종에 비해 약 2배 증가하였으며, 이 중 진핵생물에 속하는 종은 현재 5,982종, 63,268개의 유전체 프로젝트가 진행중이거나 완료되어 2015년과 비교하여 2,600여 종, 50,000여개의 유전체 프로젝트가 증가한 것을 확인할 수 있다. 또한 permanent draft로 진행되는 유전체 프로젝트가 7배 이상 증가하여, 전체 유전체를 세세하게 분석하는 프로젝트보다는 빠르게 유전체 서열을 확인함으로써 유전자원 확보의 목적성을 가진 프로젝트가 늘어나고 있는 것으로 판단된다 (Table. 1).

2. 연체동물 유전체 프로젝트 현황

Bang, *et al*. (2010) 에 따르면, 2010년 NCBI 등록 연체동물 유전체 연구는 EST, GSS를 포함해 nucleotide 1,276,010건, protein 62,953건, mitochondrial genome을 포함한 Genome 99건, gene 1,234건 으로 확인되었다. 10년 후인 2020년 현재, NCBI에 등록된 연체동물 유전제 연구는 nucleotide 5,192,310건, genome 467건으로 4배이상 증가

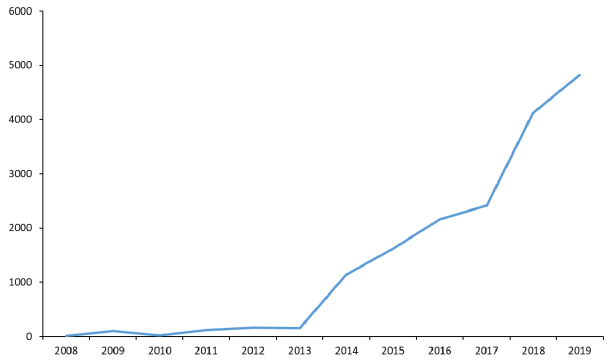

Fig. 1. Annual genomic research registration status of SRA database.

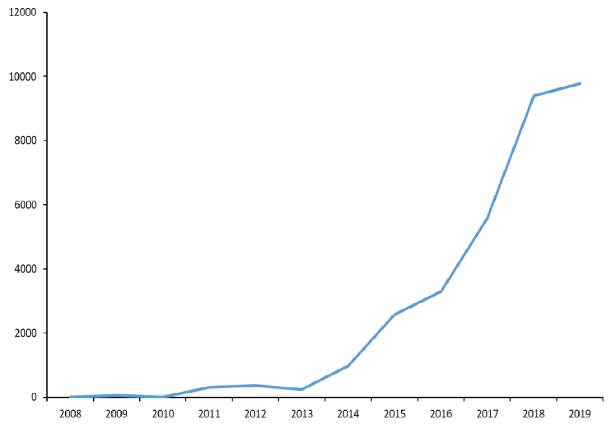

Fig. 2. Annual change of SRA data.

하였으며, 특히 protein 775,084건, gene 259,569건 으로 200 배 이상 증가하였다.

또한 NGS가 등장하며 새로 나타난 데이터베이스인 SRA (Sequence Read Archive) 데이터는 18,476건이 등록되어 있었다 (Table 2). SRA는 NGS의 raw sequence data를 저장하고, 데이터 분석을 통한 재현성을 향상시키며 관련 분야의 새로운 연구를 촉진시키는 것을 목적으로하는 시퀀스 데이터베이스이다 (NCBI). SRA에 등록된 NGS 데이터는 처음 시작된 2008년도부터 2010년도까지 총 115건 (약 66GB) 이었으나, 2011년부터는 1년에 100건 (300GB) 이 넘는 데이터가 등록되기 시작하였다. 이후 2014년부터 등록되는 데이터가 폭발적으로 증가해 1,000건 (약 1TB) 이상의 NGS 데이터가 등록되고 있으며, 최근에는 매년 4,000건 (약 9TB) 이상의 데이터가 등록되고 있다 (Fig. 1, Fig. 2). 본 연구진의 NCBI SRA 데이터베이스 현황 분석에 따르면 SRA에 연체동물의 NGS 데이터를 등록하는 국가는 미국이 33%로 1위를 차지하고 있었고, 그 뒤로 중국 (16%), 독일 (약 9%), 프랑스 (약 7%), 영국 (약 6%) 순 이었다. 우리나라의 경우 255건, 1.1TB (약 1%) 가 등록되어 있어 국내에서도 연체동물의 NGS 연구가 진행되고 있음을 확인할 수 있었다. 이를 상세히 확인하면, 국립수산과학원 (110건), 한국해양과학기술원 극지연구소 (57건), 이화여대 (36건), 서울대 (14건), 순천향대 (14건) 순 이였다. 이 데이터들은 대부분 포스트게놈 다부처 유전체사업 (해양수산부)을 통해 생산되어진 데이터들이며, 순천향대 14건의 경우 환경부 국립생물자원관의 자생동물자원의 유전체 분석연구 사업인 "멸종위기생물 유전자 (체)" 과제의 결과물이다.

SRA Experiments를 통하여 확인한 NGS 연체동물 연구는 *Crassostrea gigas*가 3,200여 건으로 가장 많았으며, *Littorina saxatilis*가 1,400여 건으로 두 번째로 많았다 (Table 3). *C. gigas* 의 경우 굴에 대한 참조유전체서열을 만들기 위하여 Hiseq X 및 PacBio-SMRT2를 사용한 대량의 서열 분석을 영국 Edinberg 대학교 The Roslin Institute 연구그룹이 수행하였으며, 돌연변이 추정을 위하여 상업적으로 가치 있는 생물의 trio sequencing 데이터를 제공하려 하고있다. 또한 6군데 지역에서 서식하는 굴의 Digestive gland, Mantle, Gill, Adductor muscle를 대상으로 한 16S rRNA microbiome을 연구하기 위하여 대량의 NGS 서열 분석을 하는 등 다양한 연구가 수행되고 있음을 확인할 수 있었다. 이러한 연구들을 통하여 단순한 유전체정보의 축적이 아닌 유전육종 관련 연구가 가능하게 되어 추후 생산량을 증대할 수 있게 될 것으로 예상된다. *L. saxatilis* 의 경우, 캐나다 Guelph 대학을 중심으로 double-digest restriction-associated DNA (ddRAD) seq 기법을 이용, foot 조직 대상 대량 SNP 발굴프로젝트가 진행되었으며, (Kess *et al*., 2015), 중국 심천지역에 위치한 Sheffield 대학에서는 일반적인 전사체 연구 및 계통분류학적 연구를 위한 Wave families targeted-capture sequencing 과 hybrid zone에 서식하는 동종생물에 대하여 4만개의 probe 를 이용한 대량의 유전체 프로젝트를 수행한 바 있고, 스웨덴의 Gothenburg 대학에서는 pool-seq 방식을 이용하여 유전다양성 관련 연구를 수행 한 바 있다.

GOLD 데이터베이스에 등록된 연체동물 유전체 프로젝트는 2015년에는 Bivalvia 19건, Gastropoda 17건, Cephalopoda 4건으로 총 40건이었으나, 2020년에는 Bivalvia 37건, Gastropoda 28건, Cephalopoda 6건으로 총 71건의 프로젝트가 진행되어 5년간 31건의 프로젝트가 더 진행된 것을 확인하였다. 진행된 유전체 프로젝트를 살펴보면 2015년과 2020년 모두 *C. gigas*에 대한 연구가 가장 많이 진행되었고, 특히 2020년 *C.gigas*에 대한 프로젝트가 7건이 증가하면서 관련 연구가 활발히 진행됨을 확인할 수 있었다 (Table 4, 5). 국적에 따른 프로젝트 진행 량은 5년전인 2015년과 비슷한 분포를 보였다. 미국과 중국이 가장 많은 프로젝트를 진행하였으며, 2020년 또한 미국 및 중국이 약 10건씩의

Table 3. SRA data status of mollusks in NCBI

Species	SRA Experiments NO.
Crassostrea gigas	3,210
Littorina saxatilis	1,412
Mytilus galloprovincialis	864
Crassostrea virginica	797
Biomphalaria glabrata	715
Ruditapes philippinarum	586
Aplysia californica	502
Ostrea lurida	436
Euprymna scolopes	427
Achatinella mustelina	375
Mizuhopecten yessoensis	355
Mytilus edulis	324
Ancylus fluviatilis	278
Mytilus galloprovincialis × *Mytilus trossulus*	271
Haliotis laevigata	205
Crepidula fornicata	184
Azumapecten farreri	181
Bathymodiolus platifrons	175
Pomacea canaliculata	173
Pecten maximus	163
Leptoxis ampla	161
Haliotis discus hannai	149
Haliotis rufescens	145
Doryteuthis opalescens	143
Laternula elliptica	128
Sinonovacula constricta	115
Limacina antarctica	111
Chlorostoma funebralis	106
Nautilus pompilius	106
Lottia gigantea	105
Other	5,574
Total	18,476

Table 4. GOLD Database of mollusk species project status in 2015

Species	NO.
Crassostrea gigas	3
Azumapecten farreri	2
Biomphalaria glabrata	2
Elysia chlorotica	2
Pinctada fucata	2
Aplysia californica	1
Arctica islandica	1
Argopecten irradians	1
Bankia setacea	1
Cepaea nemoralis	1
Other	24
Total	40

Table 5. GOLD Database of mollusk species project status in 2020

Species	NO.
Crassostrea gigas	10
Azumapecten farreri	2
Bankia setacea	2
Biomphalaria glabrata	2
Crassostrea virginica	2
Elysia chlorotica	2
Mizuhopecten yessoensis	2
Modiolus philippinarum	2
Mytilus galloprovincialis	2
Pinctada fucata	2
Pomacea canaliculata	2
Other	41
Total	71

Table 6. Comparative data of mollusk genome projects within 2015 and 2020 in different countries

Country	2015	2020
USA	22	32
China	8	19
Australia	2	4
Germany	2	3
Japan	2	3
Spain	-	3
U.K	1	2
Brazil	-	1
Canada	-	1
Estonia	-	1
France	1	1
International	1	1
Total	40	71

프로젝트를 추가로 진행하여 가장 많은 프로젝트를 진행하였다 (Table 6).

결 론

GOLD 데이터베이스를 기준으로 하였을 때 그간 진행된 유전체 프로젝트는 5년 전과 비교하여 약 4배 정도 증가하였다. 다만 연체동물의 유전체 프로젝트는 40건에서 71건으로 총 31건 만이 증가하였다.

NCBI 통하여 확인한 전 세계의 연체동물 유전체 연구는 데이터양을 기준으로 하여 10년 전과 비교하였을 때 최소 500배 이상의 데이터가 축적된 것을 확인할 수 있었다.

연체동물의 유전체 및 전사체 분석방법은 단순한 WGS 방식 및 EST 방법으로 이루어졌던 10년전과 달리 다양한 방법으로 수행되고 있었다. 특히 차세대염기서열분석기 즉 NGS 중 Hiseq이 주종을 이루고 있으며 PacBio SMRT를 이용하여 draft 서열이 아닌 참조유전체 서열을 만들려는 시도까지 이루어 지고 있음이 확인되었다.

전세계적으로 진행된 연체동물문의 유전체 프로젝트는 특정 종을 제외하면 다른 생물군에 비하여 미흡하다. 연체동물문이 지구상에서 절지동물문 다음으로 가장 많은 종을 포함하고 있는 분류군임을 고려한다면 500배 이상의 데이터 증가율은 당연한 결과라고 판단된다. 향후 다양한 연체동물에 관한 유전체 및 전사체 연구의 활성화는 유전자원 선점에 있어서도 매우 중

요할 뿐 아니라 분자육종 및 후성유전체 연구 등 새로운 연구 방향의 제시에도 크게 기여할 수 있을 것으로 사료된다.

사 사

본 논문은 환경부의 재원으로 국립낙동강생물자원관 (NNIBR202001107) 과 교육부 (한국연구재단, NRF-2017R1D1A3B06034971) 의 지원으로 수행되었습니다.

REFERENCES

Bang, I., Han, Y., Lee, J., and YS, L. (2010) Current Status of Genome Research in Phylum Mollusks. *Korean J. Malacol*. **26**: 317-326.

Droege, M., and Hill, B. (2008) The Genome Sequencer FLX™ System-Longer reads, more applications, straight forward bioinformatics and more complete data sets. *Journal of Biotechnology,* **136**: 3-10.

John Eid *et al*., (2009) Real-Time DNA Sequencing from Single Polymerase Molecules. *Science,* **323**: 113-138.

Kess, T., Gross, J., Harper, F., and Boulding, E.G. (2015) Low-cost ddRAD method of SNP discovery and genotyping applied to the periwinkleLittorina saxatilis. *Journal of Molluscan Studies*: eyv042.

Koboldt, D.C., Steinberg, K.M., Larson, D.E., Wilson, R.K., and Mardis, E.R. (2013) The next-generation sequencing revolution and its impact on genomics. *Cell,* 1**55**: 27-38.

Liolios, K., Chen, I.M., Mavromatis, K., Tavernarakis, N., Hugenholtz, P., Markowitz, V.M., and Kyrpides, N.C. (2010) The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.,* **38**: D346-354.

Lu, H., Giordano, F., and Ning, Z. (2016) Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics Proteomics Bioinformatics,* **14**: 265-279.

Morozova, O., and Marra, M.A. (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics,* **92**: 255-264.

NCBI, (http://www.ncbi.nlm.nih.gov) The National Center for Biotechnology Information, NIH

Rodriques, S.G., Stickels, R.R., Goeva, A., Martin, C.A., Murray, E., Vanderburg, C.R., Welch, J., Chen, L.M., Chen, F., and Macosko, E.Z. (2019) Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science,* **363**: 1463-1467.

Weber, M., Teeling, H., Huang, S., Waldmann, J., Kassabgy, M., Fuchs, B.M., Klindworth, A., Klockow, C., Wichels, A., Gerdts, G., Amann, R., and Glockner, F.O. (2011) Practical application of self-organizing maps to interrelate biodiversity and functional data in NGS-based metagenomics. *ISME J.,* **5**: 918-928.

Yang, R., Guo, X., Yang, J., Jiang, Y., Pang, B., Chen, C., Yao, Y., Qin, J., and Li, Q. (2009) Genomic research for important pathogenic bacteria in China. *Sci. China C. Life Sci.,* **52**: 50-63.