Korean J. Malacol. 39(1): 1-4 2023
https://doi.org/10.9710/kjm.2023.39.1.1

PANM 데이터베이스 업데이트 (Version 5.1) : 연체동물 전사체 서열 분석의 정확도 증진을 위한 오염되어진 곰팡이 유전자 서열 검색기능 추가

송대권[1,2], 상민규[1,2], 박지은[1,2], 정준양[3], 홍찬의[3], 김용태[3], 신현준[3], 류자미[3], 이혁[3], Hongray Howrelia Patnaik[1], Bharat Bhusan Patnaik[1,6], 조용훈[3], 박소영[4], 강세원[5], 이용석[1,2,3]

[1]한국자생동물자원활용 융복합연구소, [2]생명자원 바이오빅데이터 분석 및 활용 연구지원센터, [3]순천향대학교 자연과학대학 생명과학과, [4]국립낙동강생물자원관, [5]한국생명공학원구원, [6]PG Department of Biosciences and Biotechnology, Fakir Mohan University

An update of PANM database- version 5.1 for filtering the contaminating fungal gene sequences from molluscan transcriptome data

Dae Kwon Song[1,2], Min Kyu Sang[1,2], Jie Eun Park[1,2], Jun Yang Jeong[3], Chan-Eui Hong[3], Yong Tae Kim[3], Hyeon Jun Shin[3], Ziwei Liu[3], Hyeok Lee[3], Hongray Howrelia Patnaik[1], Bharat Bhusan Patnaik[1,6], Yong Hun Jo[3], So Young Park[4], Se Won Kang[5] and Yong Seok Lee[1,2,3]

[1]Korea Native Animal Resources Utilization Convergence Research Institute (KNAR), Soonchunhyang University, Asan, Chungnam, South Korea
[2]Research Support Center for Bio-Bigdata Analysis and Utilization of Biological Resources, Soonchunhyang University, Asan, Chungnam, South Korea
[3]Department of Biology, College of Natural Sciences, Soonchunhyang University, Asan, Chungnam 31538, Korea
[4]Biodiversity Research Team, Animal & Plant Research Department, Nakdonggang National Institute of Biological Resources, Sangju, Gyeongbuk, South Korea
[5]Biological Resource Center (BRC), Korea Research Institute of Bioscience and Biotechnology (KRIBB), Jeongeup, Jeonbuk, South Korea
[6]PG Department of Biosciences and Biotechnology, Fakir Mohan University, Balasore-756089, Odisha, India

ABSTRACT

The PANM database (Protostome DB) is a public repository of protein sequences available from the Protostomia group (includes Arthropoda, Mollusca, and Nematoda). The latest version of PANM DB v5.0 (released in 2022) contained 21,276,123 protein sequences that comprised 4% of the total NCBI nr protein data. In this study, an update of PANM DB, i.e., version 5.1, is presented that could accurately analyze the large-scale transcriptome data of molluscs for the contaminating fungal gene sequences. This version can filter out the fungal genes, thereby restricting the annotation to molluscs-only sequences. Using the database, we confirmed 1,589,546 amino acid sequences from 32 fungal species, which can be essentially filtered from the unigenes of 20 species of endangered molluscs. In general, the updated version of PANM DB is expected to enhance the accuracy of bioinformatics analyses of invertebrate NGS data, providing a valuable resource for researchers. PANM version 5.1 can be downloaded for free at https://panm.sch.ac.kr/ for local BLAST analysis.

keywords: NGS, Mollusks, Fungal sequences

Received: March 16, 2023; Revised: March 24, 2023; Accepted: March 29, 2023
Corresponding author: Yong Seok Lee
Tel: +82 (41) 530-3040, e-mail: yslee@sch.ac.kr
1225-3480/24833

서 론

차세대염기서열분석 (Next Generation Sequencing: NGS) 기술의 발전으로 염기서열 해독 비용이 저렴해지고 분석 시간이 짧아지면서 유전체 및 전사체 염기서열 데이터를 대량으로 생산할 수 있게 되었다 (Metzker, 2010; van Dijk *et al*., 2014). 이러한 기술 발전으로 인해 국제적으로 진행되던 유전체 프로젝트가 국내 컨소시엄으로 전환되었으며, 미생물, 곰팡

이 등 유전자 서열이 짧은 생물들의 경우 실험실 단위에서 연구 수행이 가능해졌다 (Morozova and Marra, 2008; Yang *et al*., 2009; Bang *et al*., 2010). NGS 기술의 활용도가 높아지면서 이전보다 더 많은 연구자들이 다양한 분야에서 NGS를 활용하기 시작했으며, NCBI Sequence Read Archive (SRA) 에 등록되어지는 데이터량은 NCBI GenBank에 등록되어지는 데이터량에 비해 빠른 속도로 증가하고 있다 (https://www.ncbi.nlm.nih.gov/genbank/statistics/). Long read sequencer의 발전으로 WGS (Whole Genome Shotgun) 데이터의 길이 (Read length) 와 데이터의 양이 증가하면서 NCBI nr database를 활용한 NCBI BLAST (Basic Local Alignment Search Tools) 분석을 통한 annotation의 정확성을 높이는 것이 더욱 중요해지고 있지만, 대량의 염기서열 데이터를 NCBI nr database를 활용한 BLAST 분석은 매우 오랜 시간이 소요된다는 단점이 있다 (Altschul *et al*., 1990; McGinnis and Madden, 2004).

본 연구진은 분석시간의 단축과 annotation의 정확성을 높이기 위해 2004년 연체동물 전용 BLAST 서버를 구축하였으며 (Lee *et al*., 2004), 절지동물을 비롯한 무척추동물의 유전자원에 대해 적절한 분석을 수행하기 위해 동일한 인터페이스를 사용하여 2015년 PANM (Arthropoda, Nematoda, Mollusca) database version 1.0 (https://panm.sch.ac.kr)을 구축하였다 (Kang *et al*., 2015). 유전정보 데이터의 지속적인 증가에 따라 PANM database는 2016년 9월 version 2.0으로 업데이트 되었으며 (Kang *et al*., 2016), 2019년 3월 아미노산 서열을 추가한 version 3.0 업데이트가 수행되었다 (Kang *et al*., 2019). 이후 2021년 2월 추가적으로 업데이트된 NCBI database의 Arthropoda, Nematoda, Mollusca의 아미노산 서열을 PANM database에 추가하는 것으로 version 4.0 업데이트가 수행되었으며 (Sang *et al*., 2021), 최근에는 version 4.0 업데이트 이후 NCBI database에서 아미노산 서열의 수가 133.47% 증가함에 따라 최신 유전정보를 추가하는 것으로 2022년 9월 version 5.0 업데이트가 수행되었다 (Jeong *et al*., 2022). 본 연구진이 구축한 PANM database는 해외에서도 활용하고 있음을 확인할 수 있다 (Capt *et al*., 2018).

Davis 등 (1954) 에 의해 이매패류에 감염된 곰팡이가 언급된 후, Zvereva and Vysotskaya (2005) 에 의해 알려진 이매패류에 감염된 32종의 곰팡이의 유전정보를 데이터베이스화한 후, 본 연구진에 의해 발굴된 20종의 연체동물 unigene을 query로 하여 어느 정도의 서열이 곰팡이의 서열로 판단되어질 수 있을지 알아보는 연구를 통해 연체동물의 NGS 분석 전사체 서열에 곰팡이 서열이 혼합되어 있을 수 있는 가능성이 확인되었다 (Song *et al*., 2022). 그러나 SRA에 등록되어져 있는 감염체 관련 서열들은 분석에 활용되기 어렵기때문에, BLAST database (PANM, NCBI nr) 를 활용하여 분석할 시 기생, 공생 및 감염체와 관련된 서열에 대해 annotation이 제대로 잘 되어지지 않는다. 그러므로 unigene 확보 후 곰팡이 서열을 확인할 수 있는 알고리즘을 추가하는 일이 필요하다. 본 연구는 전사체서열이 해독 되어진 연체동물의 unigene 데이터 에서 곰팡이 유래 서열을 필터링하여 annotation 의 정확도를 높이고자 수행되었다.

재료 및 방법

1. 데이터베이스 업데이트

PANM database version 5.0의 업데이트를 위하여 (Song *et al*., 2022) 에서 확보된 Acremonium, Alternaria, Aspregillus, Aureobasidium, Cladosporium, Penicillium, Dichotomopilus, Chaetomium, Xanthiomyces, Myxotrichum 속에 해당하는 32종의 곰팡이에 대한 염기서열과 아미노산 서열을 NCBI database에서 다운로드하였다. 확보한 유전자 서열을 PANM database version 5.0에 추가하여 BLAST가 가능하도록 formatdb program을 사용하여 데이터베이스화 하였다.

2. 웹 인터페이스의 개선

기존에 사용된 PANM database version 5.0이 구축된 웹 서버에 새로 적용한 PANM database version 5.1을 적용하여 BLAST 수행이 가능하도록 인터페이스를 설정하였으며, PANM database version 5.0과 같이 많은 연구자들이 이용할 수 있도록 PANM database version 5.1을 다운로드 할 수 있도록 구축하였다.

결과 및 고찰

NCBI에 등록된 절지동물문, 선형동물문, 연체동물문의 유전자 아미노산 서열들은 1999년 최초로 등록되기 시작하였으며, PANM database version 1.0이 발표된 2015년 6월까지 4,051,323개의 유전자 아미노산 서열이 등록되었으며, 최근에는 더욱 긴 서열의 유전자 아미노산 서열이 등록되고 있다 (Sang *et al*., 2021). 이는 빠른 속도로 발전하고 있는 차세대 염기서열 분석기에 의해 생물자원의 유전자원 연구가 증가하고 있으며 다양하게 연구가 되고 있음을 의미한다.

NCBI nr, PANM ver 5.0 등의 데이터베이스를 활용한 BLAST 분석이 수행 되었을 때, 기생, 공생 및 감염체 관련 서열에 대한 분석 결과를 명확하게 얻을 수 없는데, 이는 SRA에 등록되어져 있는 감염체 서열들이 분석에 활용되지 못하기 때

Table 1. Status of updated Fungi sequences in PANM DB version 5.1

Genus	Nucleotide sequence	Per (%)	Amino acid sequence	Per (%)
Acremonium	1,231	0.24	164	0.01
Alternaria	57,910	11.17	44	0.00
Aspergillus	199,399	38.45	755,150	47.51
Aureobasidium	129,250	24.93	675,075	42.47
Cladosporium	5,606	1.08	1,644	0.10
Penicillium	106,872	20.61	100,500	6.32
Dichotomopilus	174	0.03	16	0.00
Chaetomium	18,068	3.48	56,949	3.58
Xanthiomyces	12	0.00	4	0.00
Myxotrichum	6	0.00	0	0

문인 것으로 사료된다 (Song *et al*., 2022).

이러한 문제점의 해결을 위해 본 연구진에 의해 수행되어 확보된 Acremonium, Alternaria, Aspregillus, Aureobasidium, Cladosporium, Penicillium, Dichotomopilus, Chaetomiumm, Xanthiomyces, Myxotrichum 속에 해당하는 32종의 곰팡이 유전자인 518,528개의 염기서열과 1,589,546개의 아미노산 서열을 다운로드 하여, Version 5.0을 업데이트 하여 PANM database Version 5.1을 구축하였다. Version 5.1의 곰팡이 데이터에서 염기서열은 Aspregillus 속이 전체의 38.45%를 차지하는 199,399개로 가장 많이 확보되었으며, Aureobasidium 속이 129,250개 (24.93%), Penicillium 속이 106,872개 (20.61%), Alternaria 속이 57,910개 (11.17%) 순으로 많이 확보되었다. 아미노산 서열은 Aspregillus 속이 전체의 47.51%를 차지하는 755,150개로 가장 많이 확보되었으며, Aureobasidium 속이 675,075개 (42.47%), Penicillium 속이 100,500개 (6.32%), Chaetomium 속이 56,949개 (3.58%) 의 순으로 많이 확보되었다 (Table. 1).

PANM DB 5.1 업데이트로 인해 감염되어진 곰팡이의 서열이 unigene의 서열에서 걸러내는 annotation 결과를 도출할 수 있었지만 여전히 내재되어 있는 또 다른 감염체의 서열이 있을 수 있다고 판단된다. 그러므로, 연체동물을 포함한 무척추 동물들에 대한 유전자원이 확보되어 질 필요가 있으며, 감염체에 대한 전사체, 유전체 연구도 많이 수행되어야 명확하게 생물정보학적 분석이 가능해질 것으로 사료된다. 그리고 연체동물의 전사체 및 유전체 데이터에서 곰팡이 서열을 필터링할 수 있도록 더 많은 감염체 관련 유전자원을 확보할 필요성이 요구된다. 따라서 연체동물 전사체 및 유전체 서열에서 곰팡이 서열을 필터링할 수 있는 데이터베이스의 지속적인 업데이트가 필요하며, 연체동물 내 기생, 공생 및 감염체에 대한 기초연구가 많이 수행되어 질 필요가 있다고 판단된다.

요 약

업데이트된 PANM database version 5.1은 이전 version과 비교하여 연체동물에 오염된 곰팡이 유전자를 발굴하여 새로운 인터페이스와 데이터베이스를 구축하는 것으로 518,528개의 nucleotide 서열과 1,589,546 개의 아미노산 서열들이 추가되었다.

연체동물의 유전체 및 전사체 데이터에서 반드시 곰팡이 서열을 필터링 할 수 있는 알고리즘을 추가하여 unigene의 퀄리티를 끌어올리는 것이 필요하며, 연체동물과 관련된 기생 및 공생하는 곰팡이 서열 데이터베이스의 작성을 통해, 이를 기반으로 전사체 및 유전체 서열에 대한 생물정보학적 분석의 정확도를 향상시키고 퀄리티가 높은 unigene을 생성하는데 기여할 것으로 사료된다.

또한 PANM database version 5.1은 이전 version과 같이 많은 연구자들이 활용할 수 있도록 오픈하여 데이터를 다운로드할 수 있도록 하였으며, 연체동물을 포함한 무척추동물을 연구하는 연구자들에게 유용하게 사용될 것으로 사료된다.

사 사

본 연구는 교육부에서 지원하는 지역대학 우수과학자 (한국연구재단, NRF-2017R1D1A3B03034971), 중점연구소 (NRF-2021R1A6A1A03039503) 및 국가연구시설장비진흥센터 (2022R1A6C101B794), 순천향대학교 학술연구비의 지원을 받아 수행하였습니다.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**: 403-410.

Bang, I.S., Han, Y.S., Lee, J.S., and Lee, Y.S. (2010) Current Status of Genome Research in Phylum Mollusks. *The Korean Journal of Malacology,* **26**: 317-326.

Capt, C., Renaut, S., Ghiselli, F., Milani, L., Johnson, N.A., Sietman, B.E., Stewart, D.T., and Breton, S. (2018) Deciphering the Link between Doubly Uniparental Inheritance of mtDNA and Sex Determination in Bivalves: Clues from Comparative Transcriptomics. G*enome Biology and Evolution,* **10**: 577-590.

Davis, H.C., Loosanoff, V.L., Weston, W.H., and Martin, C. (1954) A Fungus Disease in Clam and Oyster Larvae. *Science*, **120**: 36-38.

Jeong, J.Y., Park, J.E., Song, D.K., Hong, C.E., Kim, Y.T., Shin, H.J., Liu, Z.W., Sang, M.K., Patnaik, H.H., Patmaik, B.B., Kang, S.W., Park, S.Y., Lee, J.S., Han, Y.S., Park, H.S., and Lee, Y.S. (2022) PANM DB ver 5.0 : An update of the PANM database for invertebrate NGS data analysis. *The Korean Journal of Malacology,* **38**: 125-128.

Kang, S.W., Park, S.Y., Hwang, H.J., Chung, J.M., Sang, M.K., Min, H.R., Park, J.E., Cho, H.C., Patnaik, B.B., and Lee., Y.S. (2019) PANM DB ver 3.0 An update of the bioinformatics database for annotation of large datasets from sequencing of species under Protostomia clade. *The Korean Journal of Malacology,* **35**: 73-75.

Kang, S.W., Park, S.Y., Patnaik, B.B., Hwang, H.J., Chung, J.M., Song, D.K., Park, Y.-S., Lee, J.S., Han, Y.S., Park, H.S., and Lee, Y.S. (2016) The Protostome database (PANM-DB): Version 2.0 release with updated sequences. . *The Korean Journal of Malacology,* **32**: 185-188.

Kang, S.W., Park, S.Y., Patnaik, B.B., Hwang, H.J., Kim, C., Kim, S., Lee, J.S., Han, Y.S., and Lee., Y.S. (2015) Construction of PANM Database (Protostome DB) for rapid annotation of NGS data in Mollusks. *The Korean Journal of Malacology,* **31**: 243-247.

Lee, Y.-S., Jo, Y.-H., Kim, D.-S., Kim, D.-W., M.-Y. Kim, Choi, S.-H., Yon, J.-O., Byun, I.-S., Kang, B.-R., Jeong, K.-H., and Park, H.-S. (2004) Construction of BLAST Server for Mollusk. *The Korean Journal of Malacology,* **20**: 165-169.

McGinnis, S., and Madden, T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.,* **32**: W20-25.

Metzker, M.L. (2010) Sequencing technologies - the next generation. *Nature Reviews Genetics,* **11**: 31-46.

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

Sang, M.K., Park, J.E., Song, D.K., Jeong, J.Y., Hong, C.-E., Kim, Y.T., Hwang, H.J., Kang, S.W., Park, S.Y., Lee, J.S., Han, Y.S., Park, H.S., and Lee, Y.S. (2021) PANM DB ver 4.0 An update of the bioinformatics database for annotation of large datasets from sequencing of species under Invertebrates. *The Korean Journal of Malacology,* **37**: 33.

Song, D.K., Park, H.J., Sang, M.K., Park, J.E., Jeong, J.Y., Hong, C.E., Kim, Y.T., Shin, H.J., Liu, Z.W., Jo, Y.H., Han, Y.S., Lee, Y.S., and Chang, J.S. (2022) Identification of Fungal Gene Sequence Contamination in Transcriptome Sequence Data of Endangered Molluscs using Bioinformatics. *The Korean Journal of Malacology,* **38**: 221-233.

van Dijk, E.L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014) Ten years of next-generation sequencing technology. *Trends in Genetics,* **30**: 418-426.

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. *Science in China Series C: Life Sciences,* **52**: 50-63.

Zvereva, L.V., and Vysotskaya, M.A. (2005) Filamentous Fungi Associated with Bivalve Mollusks from Polluted Biotopes of Ussuriiskii Bay, Sea of Japan. *Russian Journal of Marine Biology*, **31**: 382-385.