Journal of Life Science **2010** Vol. 20. No. 3. 411~415 ©JLS / ISSN 1225-9918

Construction of Web-Based Database for Anisakis Research

Yong Seok Lee, Moon Ki Baek, Yong-Hun Jo[1], Se Won Kang, Jae Bong Lee, Yeon Soo Han[1], Hee-Jae Cha[2], Hak Sun Yu[3] and Mee Sun Ock[2]*

Department of Parasitology, Inje University College of Medicine, Busan 614-735, Korea
[1]*Department of Agricultural Biology, College of Agriculture and Life Science, Chonnam National University, Gwangju 500-757, Korea*
[2]*Department of Parasitology, Pusan National University College of Medicine, Busan 602-739, Korea*
[3]*Department of Parasitology and Genetics, Kosin University College of medicine, Busan 602-703, Korea*

Received December 17, 2009 / Accepted December 21, 2009

Anisakis simplex is one of the parasitic nematodes, and has a complex life cycle in crustaceans, fish, squid or whale. When people eat under-processed or raw fish, it causes anisakidosis and also plays a critical role in inducing serious allergic reactions in humans. However, no web-based database on *A. simplex* at the level of DNA or protein has been so far reported. In this context, we constructed a web-based database for Anisakis research. To build up the web-based database for Anisakis research, we proceeded with the following measures: First, sequences of order Ascaridida were downloaded and translated into the multifasta format which was stored as database for stand-alone BLAST. Second, all of the nucleotide and EST sequences were clustered and assembled. And EST sequences were translated into amino acid sequences for Nuclear Localization Signal prediction. In addition, we added the vector, *E. coli*, and repeat sequences into the database to confirm a potential contamination. The web-based database gave us several advantages. Only data that agrees with the nucleotide sequences directly related with the order Ascaridida can be found and retrieved when searching BLAST. It is also very convenient to confirm contamination when making the cDNA or genomic library from Anisakis. Furthermore, BLAST results on the Anisakis sequence information can be quickly accessed. Taken together, the Web-based database on *A. simplex* will be valuable in developing species specific PCR markers and in studying SNP in *A. simplex*-related researches in the future.

Key words : Web-based database, Anisakis, interface, sequence information

서 론

Genome project는 세계적으로 총 6,283개가 진행 중 이거나 완성되었다(2009년 12월 09일 기준). 그 중 1,150 개 생물 종에 대한 게놈프로젝트는 이미 완성되어 논문으로 출판되었으며 원생동물에서 3,568 종류, 고세균류에서 111 종류, 그리고 진핵생물에서 1,253 종류의 genome project가 현재 진행 중에 있다[7]. 이와 같이 사람, 마우스, 랫트, 복어 및 많은 미생물 등 게놈프로젝트가 끝난 생물들의 경우에는 게놈정보를 바탕으로 유전자 및 아미노산 정보가 밝혀져 있어, 연구자들이 분자생물학적 연구를 수행할 경우 NCBI (National Center for Biotechnology Information), Ensembl, UCSC genome browser 등 전문데이터베이스 등을 통해 매우 손쉽게 정보를 얻을 수 있는 이점이 있다[6,16,22]. 연구자 층이 두터운 생물군의 경우 VectorBase 및 Plasmodb처럼 컨소시엄 형태로 데이터베이스가 운용되고 있는 경우도 있지만 아직 연구자 층이 두텁지 않은 생물군의 경우에는 개인 연구자들이 독립적으로 데이터베이스를 구축하여 운용하는 일도 종종 있다. 대표적인 예로 연체동물서열 데이터베이스, 가시아메바, 동양달팽이 EST 데이터베이스(unpublished) 등이 있다[12,15,19].

본 데이터베이스의 주제인 고래회충(*Anisakis simplex*)은 고래, 돌고래, 물개 등의 소화관에 성충이 기생하는 바다포유류 기생충이다. 성충의 분변으로 배출된 충란은 유충으로 성장하여 중간숙주인 바다갑각류에 먹히게 되고 운반숙주인 해산어류나 오징어, 한치 등의 두족류에서 제3기 유충으로 성장한다. 사람은 감염된 물고기나 두족류를 날것으로 또는 덜 익혀 섭취하여 감염된다[10]. 고래회충증(Anisakidosis)은 대부분 우리나라, 일본 등의 생선회를 즐겨먹는 아시아 지역과 유럽에 국한되어 보고되고 있으며 연구자의 층도 두텁지 않았다 [2,9,11,17]. 그러나 인체 감염된 고래회충 유충은 급성복통을 일으키거나 위장관 계통에 호산구성 육아종을 형성할 뿐 만 아니라 알레르기 반응을 유도하는 것으로 밝혀져 최근에는 식품 알레르기의 한 원인으로도 주목을 받고 있다. 최근 기생충을 이용한 위생가설 관련 및 선천성면역 관련 연구의 중요 대상 기생생물로 재조명을 받고 있으며 EST 연구도 수행되어진 바 있어 NCBI Genbank를 통해 고래회충 EST 서열을 다운로드 하여 사용할 수 있게 되어있다[4,14,23]. 하지만 현재 EST

*Corresponding author
Tel : +82-51-990-6424, Fax : +82-51-990-3081
E-mail : sunnyock@kosin.ac.kr

단순 서열정보를 받을 수는 있으나 그 서열의 annotation 정보 및 클론번호 등은 입력 되어 있지 않기 때문에 고래회충의 연구에 이러한 정보를 활용하기가 매우 어렵다.

본 연구에서는 이미 밝혀진 고래회충의 EST 정보 및 근연 카테고리에 속한 생물들의 유전정보만을 모아 단독으로 BLAST 가 가능하고 필요한 유전자의 서열 및 annotation 정보를 받을 수 있는 웹인터페이스를 구축하여 앞으로 고래회충을 대상으로 한 다양한 분자생물학적 연구에 도움이 되고자 하였다.

재료 및 방법

서버구축 및 환경설정

사용된 서버는 Intel Server Platform ZSS130 (Samsung)에 Xeon 3.2 GHz cpu 시스템을 사용하였으며, 운영체제(operating system)는 Cent OS를 사용하였다. 운영체제 설치 후 Apache, PHP, Mysql 연동 시스템을 구축하였으며, 서버의 설정에서 웹 접속 사용자가 cgi (common gate interface)를 사용할 수 있도록 환경설정을 한 후 WebBLAST 패키지를 설치하였다.

BLAST 용 데이터베이스 구축

NCBI 에 등록되어 있는 회충목(Order Ascaridida) 관련 게놈정보(미토콘드리아 게놈정보), 유전자서열 정보, 아미노산 서열정보를 taxonomy browser와 연계하여 모두 다운 받은 후, 멀티파스타 형태의 정보로 만든 후 NCBI에서 제공하는 formatdb 프로그램을 사용하여 BLAST 용 데이터베이스로 만들었으며 부가적으로 실험 후 데이터 확인 시 필요한 벡터서열, *E. coli* 서열, 반복서열 등을 모두 데이터베이스에 포함하여, 실험 데이터를 검증할 때 용이하도록 하였다.

SNP 연구의 기초가 될 EST 및 NT core 서열들의 clustering 및 assembly

종내 또는 종간 SNPs 연구를 위하여 NCBI taxonomy browser를 활용하여 genbank에 등록된 Anisakis EST (총 398 개) 및 NT (총 475 개) 서열들을 다운로드 한 후 vector sequence database 와 cross_match 프로그램을 사용하여 vector sequence를 제거하였다. 정리된 서열은 BLAST 및 cap3 소프트웨어를 엔진으로 한 TGICL package (TIGR tools)를 이용하여 clustering 및 assembly 를 수행하였다[1,8,18].

유전자의 기능을 예측하기 위하여 clusters of orthologous groups for eukaryotic complete genomes (KOG) 분석을 시행하였다. 실험군과 대조군의 서열을 KOG 데이터베이스에 local BLAST (blastx, $E < e^{-10}$) 검색을 통하여 각각의 유전자의 기능을 예측하였으며[21], NLS 영역을 예측하기 위해 EST 및 핵산서열들을 모두 Genscan [3] 및 EMBOSS package [20]의 sixpack을 사용하여 아미노산 서열로 생성한 후 predictNLS 프로그램을 엔진으로 하는 perl script를 활용하여 NLS를 포함하는 서열들을 추출하였다[5,13].

Web interface 구축

Blast 메뉴에서는 핵산, 아미노산, EST 3개의 소메뉴로 구성하였다. 회충목에서 현재까지 밝혀진 핵산 서열, 아미노산 서열, EST 서열을 대상으로 각각 BLAST 가 가능하도록 구성하였으며 query 및 데이터베이스가 허용하는 한 blastp, blastn, blastx, tblastn, tblastx 모두 수행 가능하도록 하였다. Vector, *E. coli*, Repeat 서열을 따로 검색 할 수 있도록 하였으며, multi DB 메뉴를 만들어 라이브러리 확인(insert size 측정) 등을 할 때 용이하도록 하였다. 또한 perl script를 기반으로 한 검색엔진을 설치하여 Annotation이 되어진 nisakis EST 서열정보를 종 이름, 유전자 이름 및 NCBI accession number 등을 query 로 하여 찾을 수 있도록 하였다. 그리고 primer3 등 실험 시 부가적으로 필요한 웹용 프로그램을 설치하여 연구자들이 편리하게 이용하도록 하였다.

결과 및 고찰

로컬서버에 다운로드 된 Anisakis의 EST (총 398 개) 및 NT (총 475 개) 서열들을 vector sequence database와 cross_match 프로그램을 사용하여 vector sequence를 제거한 결과 62개의 서열들이 벡터서열 제거 후 100 bp 이하였다. 정리된 811개의 서열은 BLAST 및 cap3 소프트웨어를 엔진으로 한 TGICL package (TIGR tools)를 이용하여 clustering 및 assembly를 수행한 결과 총 90개의 cluster에서 100개의 contig 그리고 235개의 singleton이 생성되었다. 이러한 clustering결과 및 assembly 결과의 활용도를 높이기 위해 웹기반의 인터페이스를 구축하였다.

Home menu 에서는 전반적인 data processing 방법을 도식화하여 보여주고 있으며, clustering res. 메뉴에서는 clustering 결과에 의해 만들어진 contig의 서열, align 결과, cluster를 이루는 각 read 의 서열을 웹상에서 볼 수 있도록 하였다. contig 서열을 NCBI nr database에 BLASTx로 homology test 한 결과 중 best hit에 해당하는 annotation 및 source (species)를 바로 볼 수 있으며 전체 BLAST 결과를 볼 수 있는 링크를 구성하였다. 또한 결과는 Query ID, Align results, Cluster를 이루는 각 read 의 서열, Annotation, Source (종별) 등으로 sorting 된 결과를 보여 줄 수 있도록 하였다(Fig. 1).

Blast 메뉴에서는 핵산, 아미노산, EST 3개의 소메뉴로 구성하였다. 회충목에서 현재까지 밝혀진 핵산 서열, 아미노산 서열, EST 서열 및 미토콘드리아 게놈(Mitochondrial Genome), 그리고 선충강(Class Nematoda)의 아미노산, 핵산, EST와 고래회충의 핵산, EST, 아미노산과 *A. simplex* 미토콘드리아의 게놈, 핵산, 아미노산과 저자 등이 분석한 Anisakis EST를 포함하여 총 15종류의 데이터베이스를 대상으로 BLAST 가 가능

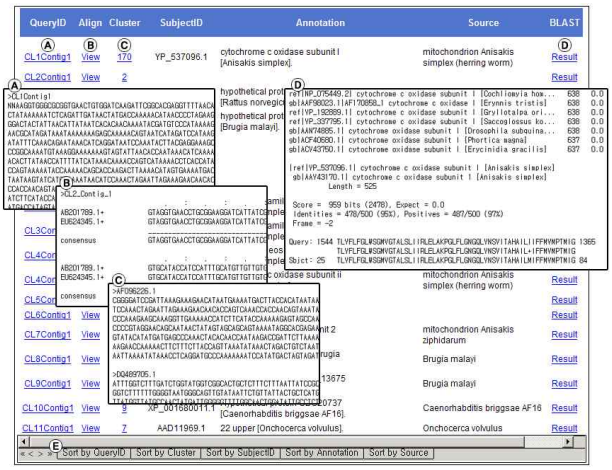

Fig. 1. The annotation results by clustering and assembly of nucleotide and EST sequences downloaded through NCBI taxonomy browser. (A) Contig sequences, (B) Alignment results, (C) Sequences contained in the contig, (D) Blast results, (E) Tab menu for sorting

하도록 구축하였다. BLAST는 query 및 데이터베이스의 관계가 허용하는 한 blastp, blastn, blastx, tblastn, tblastx 프로그램의 수행이 모두 가능하도록 구축하였다. 또한 Vector, *E. coli*, Repeat 서열을 따로 검색 할 수 있도록 구축하여, multi DB 메뉴를 만들어 라이브러리 확인(삽입체 크기 측정) 등을 할 때 용이하도록 하였다(Fig. 2).

Search 메뉴에서는 자체적으로 구축된 cgi와 perl script를 이용한 검색엔진을 달아놓아 유전자 이름, accession number, 종 이름 등을 query로 하여 데이터베이스 내부의 모든 정보를 검색할 수 있도록 구축하였다(Fig. 3). 또한 웹페이지 자체에 Primer3 엔진을 탑재하여 필요한 서열의 시발체를 데이터베이스 내부에서 제작 할 수 있도록 하였다.

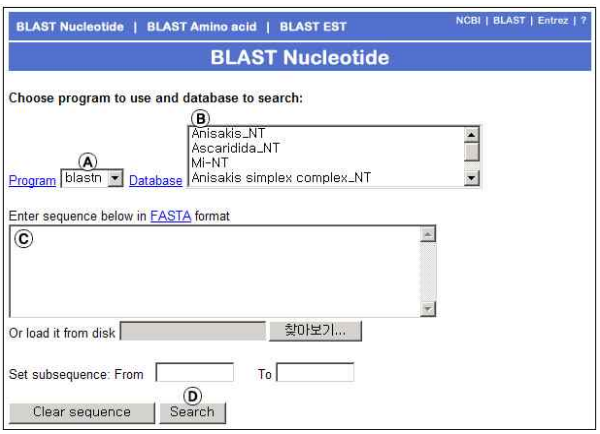

Fig. 2. The screen of searching the BLAST through constructed database. (A) Selection of blast program, (B) Selection of blast database, (C) Insertion of query sequences, (D) Start to search

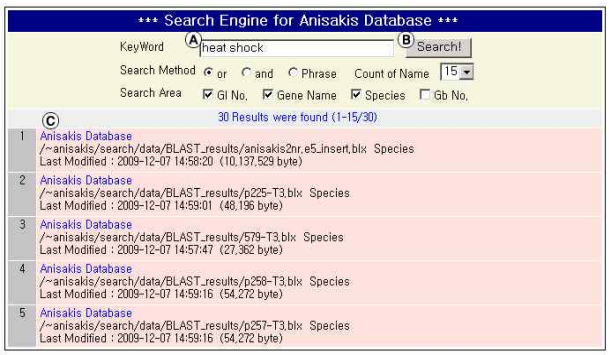

Fig. 3. Installation of search engine within constructed Anisakis database. (A) Insertion of keywords, (B) Start to search, (C) Results of searching

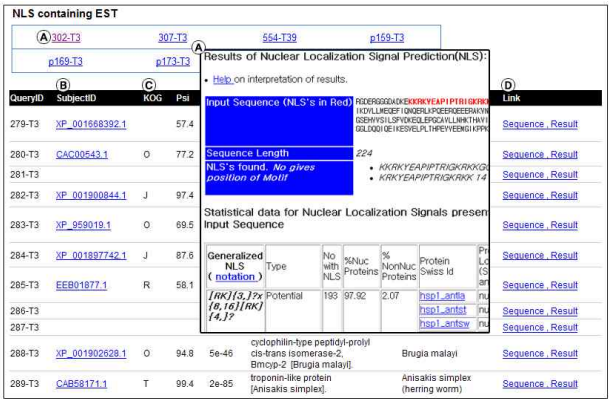

Fig. 4. The interface showing the annotation results of EST through the BLAST against NCBI nr database constructed in the local server. (A) Results of Nuclear Localization Signal Prediction (NLS) (B) The NCBI link of annotated genes, (C) KOG results, (D) EST sequences and total results of blast

NLS 영역을 예측하기 위해 EST 및 핵산 서열들을 모두 아미노산으로 변환 한 후 predictNLS 프로그램을 엔진으로 하는 perl script를 활용하여 NLS를 포함하는 서열들을 추출한 결과 8개의 서열이 확인되어 Annotated EST 메뉴에 추가하였다. KOG 데이터베이스에 local BLAST (blastx, $E<e^{-10}$) 검색을 통하여 각각의 유전자의 기능을 예측 한 결과 총 388개의 서열 중 262개의 서열에 대해 KOG 결과를 도출 할 수 있었으며, 이러한 모든 결과들 또한 Annotated EST 메뉴에 추가하였다.

본 웹데이터베이스 서버의 구축을 통해 고래회충 및 회충목의 염기서열과 일치하는 서열을 자체 BLAST를 통해 매우 빠른 속도로 추출 할 수 있었으며, repeat elements, *E. coli*, vector 등의 서열들과 동시에 BLAST를 시행할 수 있어 cDNA 또는 genomic DNA 라이브러리를 구축할 때 라이브러리의 오염, 삽입체의 길이 등의 상태를 쉽게 확인 할 수 있었다. 또한 Clustering Res. 인터페이스를 통해 SNPs 발굴이 용이하게 되었으며 자체 구축된 primer3를 통해 실험용 시발체를 제작할

수 있게 되었다. 더 나아가 저자 등이 기 구축한 cDNA library의 활용을 annotated EST 을 통해 극대화 시킬 수 있어 차후 수행되어질 고래회충 관련 분자생물학적 연구에 도움이 될 것으로 기대된다.

감사의 글

본 연구는 고신대학교 의과대학 학술연구비의 지원에 의해 수행되었음(2008).

References

1. Altschul, S. F., W. Gish, W. Miller, E. W. Meyers, and D. J. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410.
2. Bouree, P., A. Paugam, and J. C. Petithory. 1995. Anisakidosis: report of 25 cases and review of the literature. *Comp. Immunol. Microbiol. Infect Dis.* **18**, 75-84.
3. Burge, C. and S. Karlin. 1997. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268**, 78-94.
4. Choi, S. J., J. C. Lee, M. J. Kim, G. Y. Hur, S. Y. Shin, and H. S. Park. 2009. The clinical characteristics of Anisakis allergy in Korea. *Korean J. Intern. Med.* **24**, 160-163.
5. Cokol, M., R. Nair, and B. Rost. 2000. Finding nuclear localization signals. *EMBO Rep.* **1**, 411-415.
6. Ensembl. http://www.ensembl.org
7. Genomes online Database V 3.0. http://www.genomesonline.org
8. Huang, X. and A. Madan. 1999. CAP3: A DNA sequence assembly program. *Genome Res.* **9**, 868-877.
9. Ishikura, H. and K. Kikuchi. 1990. Intestinal anisakiasis in Japan. Infected fish, sero-immunology, diagnosis, and prevention. pp. 129-143, Springer Verlag. Tokyo
10. Ishikura, H., K. Kikuchi, K. Nagasawa, T. Ooiwa, H. Takamiya, N. Sato, and K. Sugane. 1993. Anisakidae and anisakidosis. *Prog. Clin. Parasitol.* **3**, 43-102.
11. Lee, E. J., Y. C. Kim, H. G. Jeong, and O. J. Lee. 2009. The mucosal changes and influencing factors in upper gastrointestinal anisakiasis: analysis of 141 cases. *Korean J. Gastroenterol* **53**, 90-97.
12. *Nesiohelix samarangae* EST database. http://edunabi.com/~nsdb
13. Lee, Y. S., S. W. Kang, Y. H. Jo, H. C. Gwak, S. H. Chae, S. H. Choi, I. Y. Ahn, H. S. Park, Y. S. Han, and W. G. Kho. 2006. Bioinformatic analysis of NLS (Nuclear Localization Signals)-containing proteins from MOLLUSKS. *The Korean Journal of Malacology* **22**, 109-113.
14. Moneo, I., M. L. Caballero, R. Rodriguez-Perez, A. I. Rodriguez-Mahillo, and M. Gonzalez-Munoz. 2007. Sensitization to the fish parasite *Anisakis simplex*: clinical and laboratory aspects. *Parasitol Res.* **101**, 1051-1055.
15. Moon, E. K., J. O. Kim, Y. H. Xuan, Y. S. Yun, S. W. Kang, Y. S. Lee, T. I. Ahn, Y. C. Hong, D. I. Chung, and H. H. Kong. 2009. Construction of EST database for comparative gene studies of acanthamoeba. *Korean J. Parasitol.* **47**, 103-107.
16. NCBI (National Center for Biotechnology Information). http://www.ncbi.nlm.nih.gov/
17. Pampiglione, S., F. Rivasi, M. Criscuolo, A. De Benedittis, A. Gentile, S. Russo, M. Testini, and M. Villan. 2002. Human anisakiasis in Italy: a report of eleven new cases. *Pathol. Res. Pract.* **198**, 429-434.
18. Pertea, G., X. Huang, F. Liang, V. Antonescu, R. Sultana, S. Karamycheva, Y. Lee, J. White, F. Cheung, B. Parvizi, J. Tsai, and J. Quackenbush. 2003. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* **19**, 651-652.
19. PlasmoDB. http://plasmodb.org
20. Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* **16**, 276-277.
21. Tatusov, R., N. Fedorova, J. Jackson, A. Jacobs, B. Kiryutin, E. Koonin, D. Krylov, R. Mazumder, S. Mekhedov, A. Nikolskaya, B. S. Rao, S. Smirnov, A. Sverdlov, S. Vasudevan, Y. Wolf, J. Yin, and D. Natale. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41.
22. UCSC Genome Browser. http://genome.ucsc.edu
23. Yu, H. S., S. K. Park, K. H. Lee, S. J. Lee, S. H. Choi, M. S. Ock, and H. J. Jeong. 2007. *Anisakis simplex*: analysis of expressed sequence tags (ESTs) of third-stage larva. *Exp. Parasitol.* **117**, 51-56.

초록 : 고래회충 연구를 위한 웹기반 데이터베이스 구축

이용석 · 백문기 · 조용훈[1] · 강세원 · 이재봉 · 한연수[1] · 차희재[2] · 유학선[3] · 옥미선[2]*
(인제대학교 의과대학 기생충학교실, [1]전남대학교 농업생명과학대학 식물생명공학부, [2]고신대학교 의과대학 기생충학.유전학교실, [3]부산대학교 의학대학원 기생충학교실)

본 연구에서는Anisakis 연구를 위하여 웹을 기반으로 하는 데이터베이스를 리눅스 Cent OS 시스템이 설치된 Xeon 3.2 GHz cpu의 인텔 서버플랫폼 ZSS130 (삼성) 서버에 구축하였다. 운영체제를 설치한 후에 common gate interface (cgi) 기반의 웹서버(http://www.anisakis.org)를 구축하고 NCBI에서 제공하는 WebBLAST 프로그램을 설치하였다. Anisakis 연구를 위한 웹기반 데이터베이스를 다음과 같은 순서로 구축하였다. 우선 회충목에 속하는 각종 서열(염기서열, 아미노산서열, EST 서열, 미토콘드리아 Genome 서열)들을 멀티파스타 형식으로 다운로드 하였다. 다음으로NCBI 에서 제공하는 formatdb 프로그램을 통하여 BLAST 검색이 가능하도록 데이터베이스화 하였으며 모든 염기서열들과 EST 서열들을 TGICL 프로그램을 통하여 clustering 및 assembing을 하였다. 그리고 NLS (Nuclear Localization Signal) 예측을 위해 EST 서열들은 Genscan 프로그램과 Emboss sixpack 프로그램을 사용하여 아미노산으로 변환하였다. 또한 벡터 서열과 *E. coli* 서열, 그리고 반복 서열들을 서버에 구축하여 서열들의 오염을 확인할 수 있게 하였다. 본 웹데이터베이스 서버의 구축을 통해 고래회충 및 회충목의 염기서열과 일치하는 서열을 자체 BLAST 를 통해 매우 빠른 속도로 추출 할 수 있었으며, cDNA나 genomic DNA 라이브러리를 구축할 때 라이브러리의 상태를 쉽게 확인 할 수 있게 되었다. 또한 Clustering Res. 인터페이스를 통해 SNPs 연구 수행 시 매우 쉽게 실험용 시발체를 제작할 수 있으며 기 구축된 cDNA library의 활용을 annotated EST를 통해 극대화 시킬 수 있어 고래회충 관련 분자생물학적 연구에 도움이 될 것으로 기대된다.