Korean J. Malacol. 31(3): 243-247 2015
http://dx.doi.org/10.9710/kjm.2015.31.3.243

Construction of PANM Database (Protostome DB) for rapid annotation of NGS data in Mollusks

Se Won Kang[1#], So Young Park[1#], Bharat Bhusan Patnaik[1,5], Hee Ju Hwang[1], Changmu Kim[2], Soonok Kim[2], Jun Sang Lee[3], Yeon Soo Han[4] and Yong Seok Lee[1]

[1]Department of Life Science and Biotechnology, College of Natural Sciences, Soonchunhyang University, Asan, Chungnam 31538, Korea
[2]National Institute of Biological Resources, Incheon 22689, Korea
[3]Institute of Environmental Research, Kangwon National University, Chuncheon, Gangwon 24341, Korea
[4]College of Agriculture and Life Science, Chonnam National University, Gwangju 61186, Korea
[5]Trident School of Biotech Sciences, Trident Academy of Creative Technology (TACT), Bhubaneswar-751024, Odisha, India

ABSTRACT

A stand-alone BLAST server is available that provides a convenient and amenable platform for the analysis of molluscan sequence information especially the EST sequences generated by traditional sequencing methods. However, it is found that the server has limitations in the annotation of molluscan sequences generated using next-generation sequencing (NGS) platforms due to inconsistencies in molluscan sequence available at NCBI. We constructed a web-based interface for a new stand-alone BLAST, called PANM-DB (Protostome DB) for the analysis of molluscan NGS data. The PANM-DB includes the amino acid sequences from the protostome groups-Arthropoda, Nematoda, and Mollusca downloaded from GenBank with the NCBI taxonomy Browser. The sequences were translated into multi-FASTA format and stored in the database by using the formatdb program at NCBI. PANM-DB contains 6% of NCBInr database sequences (as of 24-06-2015), and for an input of 10,000 RNA-seq sequences the processing speed was 15 times faster by using PANM-DB when compared with NCBInr DB. It was also noted that PANM-DB show two times more significant hits with diverse annotation profiles as compared with Mollusks DB. Hence, the construction of PANM-DB is a significant step in the annotation of molluscan sequence information obtained from NGS platforms. The PANM-DB is freely downloadable from the web-based interface (Malacological Society of Korea, http://malacol.or/kr/blast) as compressed file system and can run on any compatible operating system.

Keywords: PANM-DB, protostome, mollusks, NCBInr, next-generation sequencing

INTRODUCTION

The field of DNA sequencing, originally developed by Sanger *et al*. (1977) has seen unprecedented growth, with the evolution of chain-terminated 'Sanger DNA sequencing' method with fluorescent-labelling and capillary electrophoresis into automated sequencing instruments. Notwithstanding the method development, Sanger sequencing using first-generation technology showed limitations as high running time and costs and low high-throughput capacity. The 'next- generation sequencing' (NGS) platforms including the 454 Genome Sequencer (http://www.454.com), Illumina Genome Analyzer (http://www.illumina.com), and SOLiD Genome Sequencer (http://www.lifetechnologies.com) were capable in improving the sequence read efficiency (-100 fold), necessitating their applications to map transcripts and genes at a global level (Mardis, 2013;

Received: September 20, 2015; Revised: September 24, 2015; Accepted: September 30, 2015
Co-first author : Se Won Kang & So Young Park
Corresponding author : Yong Seok Lee
Tel: +82 (41) 530-3040 e-mail: yslee@sch.ac.kr
1225-3480/24589

Kim *et al*., 2014). To put it into perspective, the most prolific NGS, the short read length (150 bp) Solexa GAIIx/Illumina platform was able to yield 96,000 Mb reads/run at a reagent cost of $0.12 per Mb (Harrison and Kidner, 2011). With NGS sequencing platforms becoming increasingly cost-efficient with a high-throughput technology model, genetic studies on non-model organisms have improved, enabling the availability of data from the taxons of evolutionary significance (Riesgo *et al*., 2012).

With NGS platforms reaching the pinnacle of gene and transcript discovery in research, there has been an extraordinary increase in the datasets, and hence a need for upgradation of data storage, processing power to data output. From the NGS sequencing and assembly processes to the annotation analysis, many datasets in different file formats are shared within researchers that warrants an expansion in hard disk storage (Baker, 2010). Similarly, with an increased emphasis towards the study of non-model species, especially with reference to the *de novo* assembled transcriptomes, leading research laboratories are strategizing towards development of local server stations and databases. The benchmark software tools for analysis of NGS data includes the visualization softwares to explore NGS data alongside the reference genome such as Map View (Bao *et al*., 2009), MGA Viewer (Zhu *et al*., 2013) among others, and the merging annotation platforms such as AmalgamScope (Tsiliki *et al*., 2014) among others.

We are in a continuous pursuit to refine and validate the utilization of mollusks sequence information submitted to National Centre for Biotechnology Information (NCBI) sequence databases for the annotation of transcriptome data from both model and non-model members of the phylum. Earlier, we constructed a local stand-alone BLAST server for the mollusk by translating the downloaded DNA and amino acid sequences from NCBI using the formatdb program (Lee *et al*., 2004). The BLAST server was convenient for quick and contamination-free annotation of large-scale EST data generated through first-generation Sanger sequencers (Jeong *et al*., 2013; Jeong *et al*., 2015). Subsequently, we developed mollusks sequence database (an updated version of stand-alone BLAST server for the mollusk) (http://www.malacol.or.kr/blast) that suggested a modification of the server to improve the significant hits percentage during the annotation of molluscan genes (Kang *et al*., 2014). In this study, we report the construction of a new stand-alone BLAST server for the analysis of mollusks NGS data. The new server comes with added features as it accounts for the amino acid sequences of a significant proportion of protostomia clade (includes Arthropoda, Nematoda and Mollusca). The NGS sequence annotation results using the new stand-alone BLAST server show advances over the NCBInr database in analyzing the assembled protostome transcriptomes.

MATERIALS AND METHODS

1. Database construction

For the construction of a composite protostome database, we downloaded the amino acid sequences of Mollusca, Arthropoda and Nematoda from GenBank with the NCBI taxonomy browser (http://www.ncbi.nlm.nih.gov/taxonomy). The downloaded amino acid sequences from the independent protostome groups were combined and translated into multi-FASTA format. The independent multi-FASTA files of Mollusca, Arthropoda and Nematoda were combined and stored as database by using the formatdb program provided by NCBI. The database was named as PANM-DB (Protostome DB) that accesses a significant fraction of protostome groups including Mollusks.

2. Database verification

For the verification of PANM-DB, a total of 10,000 sequences from RNA-Seq library (average length of 118 bp) were used. The sequences were used as query sequences and used for BLAST analysis using local server against the NCBInr, PANM-DB (Protostome DB), Arthropoda, Nematoda, and Mollusks database, respectively. The local server used for the analysis had Intel Xeon Processor 8-core E5-4620 (2.4 GHz) central processing unit (CPU), 768G random access memory (RAM) system and used a Linux Cent operating

Fig. 1. A web interface for PANM database. Screenshot from MSK (the Malacological Society of Korea, http://malacol.or.kr/blast/aminoacid.html). (**A**) Users can click Aminoacid DB for using PANM DB (**B**) Users can select PANM DB for the blast within web interface (**C**) The web interface provide the entire PANM-DB file as a compressed file 'tar.gz'. (**D**) Users put the own sequences as FASTA format in the drop-box or uploaded own sequences (**E**) and then click the search button will BLAST runs.

system release 6.4.

3. Web-based interface construction

The PANM-DB is interfaced under amino acid DB BLAST of the Malacological Society of Korea (MSK) (http://malacol.or.kr/blast/aminoacid.html) with the provision of downloading the entire application file (as a compressed file 'tar.gz'). This is to facilitate the prompt utilization of the stand-alone server in the researcher's independent local server. The PANM-DB integrated into the MSK web-page is shown in Fig. 1. The amino acid DB menu in the home page provides a scheme for transcriptome data mining in molluscan species. In the menu, under the selection tab, PANM-DB can be accessed for homology testing of the unigene sequences from the transcriptome assembly. The sequences can be entered manually as FASTA format files in the drop-box or uploaded from the disk. The BLASTp/BLASTx results show the query ID, alignment results, subject IDs, query and subject length, annotation and source (species).

RESULTS AND DISCUSSION

PANM-DB was constructed to aid the researcher in deciphering the annotation profile of large transcriptome datasets after assembly and clustering. For PANB-DB to be effectual and consistent, it is imperative that it provides significant quality hits for the transcriptome sequences in a short time compared to the NCBInr database. After the compilation of the Mollusks, Arthropod, and Nematode sequences from GenBank with the NCBI taxonomy Browser and database formatting, we validated the efficiency of PANM-DB by conducting BLAST with a set of 10,000 sample sequences from molluscan species against NCBInr, PANM-DB, and Mollusks DB.

The current status of PANM-DB has been shown in Table 1. A total of 66,387,522 sequences were made available in the NCBInr database as of 06-05-2015. PANM-DB was configured (as on 24-06-2015) by the extraction of a total of 4,051,323 sequences (3,111,849 Arthropoda, 652,125 Nematoda and 287,349 Mollusca sequences) from the NCBI taxonomy browser, and that constitutes of only 6% of the total available sequences in the NCBInr database.

Hence, the arthropod, nematode, and mollusks

Table 1. Status of the available amino acid sequences in PANM-DB

Database	Number of Sequences	Length	Remark
NCBI nr	66,387,522	23,805,201,081	--
PANM-DB	4,051,323	1,522,609,669	6% of NCBInr DB
- Arthropoda	3,111,849	1,196,730,565	77% of PANM-DB
- Nematoda	652,125	226,168,391	16% of PANM-DB
- Mollusks	287,349	99,710,713	7% of PANM-DB

Table 2. A comparative depiction of the speed and quality of RNA-sequence annotation using the developed databases

Input sequences	DB name	Processing time	Significant hit
10,000 sequences	NCBI nr	10 days	5,409
	PANM-DB	16 hours	5,709
	Mollusks DB	1 hours	3,061

amino acid sequences constitute 77%, 16%, and 7% of total PANM-DB available sequences, respectively.

The BLAST results for a total of 10,000 molluscan RNA-seq sequences against the NCBI nr, PANM-DB, and Mollusks DB with an E-value cut-off of $1E^{-5}$ has been shown in Table 2.

We find that for an input of 10,000 sequences, PANM-DB show a superior significant hit and processing time (15 times faster) compared with NCBInr database. It is significant to note that even though Mollusks DB has the quickest processing time of 1 hour, it shows much lesser sequence hit (2 times) compared with either NCBInr database or PANM-DB. The compactness of PANM-DB over NCBInr database could be attributed to the use of target protostome specific sequences registered with NCBI. Moreover, NCBInr is updated regularly with sequences from diverse organisms.

In conclusion, we state to have constructed PANM-DB (Protostome DB) for the annotation of transcriptome sequences from Mollusks, Arthropods and Nematodes that show a superior significant hit with a shorter processing time. It will be a prudent and effective database for quick information on NGS data from protostomes including mollusks.

ACKNOWLEDGEMENTS

This work was supported by the grant entitled "The Genetic and Genomic Evaluation of Indigenous Biological Resources" funded by the National Institute of Biological Resources (NIBR201503202).

REFERENCES

Baker, M. (2010) Next-generation sequencing: adjusting to data overload. *Nature Methods*, **7**: 495-499.

Bao, H., Guo, H., Wang, J., Zhou, R., Lu, X. and Shi, S. (2009) MapView: visualization of short reads alignment on a desktop computer. *Bioinformatics*, **25**: 1554-1555.

Harrison, N. and Kidner, C.A. (2011) Next-generation sequencing and systematics: What can a billion base pairs of DNA sequences data do to you? *Taxon*, **60**: 1552-1566.

Jeong, J.E., Kang, S.W., Hwang, H.J., Chae, S-H., Patnaik, B.B., Han, Y.S., Lee, J.B., Jo, Y.H., Lee, B.L., Seog, D-H. and Lee, Y.S. (2013) Expressed sequence tags (ESTs) analysis of *Tenebrio molitor* larvae. *Entomological Research*, **43**: 168-176.

Jeong, J.E., Patnaik, B.B., Kang, S.W., Hwang, H-J., Park, S.Y., Wang, T.H., Park, E.B., Lee, J.B., Nam, M-M., Jo, Y.H., Han, Y.S., Lee, J-S., Park, H.S. and Lee, Y.S. (2015) Characterization of *Physa acuta* expressed sequence tags and transcript mining following cadmium exposure. *Genes and Genomics*. doi:10.1007/s13258-015-0334-x.

Kang, S.W., Hwang, H.J., Park, S.Y., Wang, T.H., Park, E.B., Lee, T.H., Hwang, U.W., Lee, J.S., Park, H.S., Han, Y.S., Lim, C.E., Kim, S. and Lee, Y.S. (2014) Mollusks sequence database: Version II. *Korean Journal of Malacology*, **30**: 429-431.

Kim, K.M., Park, J-H., Bhattacharya, D. and Yoon, H.S. (2014) Applications of next-generation sequencing to unravelling the evolutionary history of algae. *International Journal of Systematic and Evolutionary Microbiology*, **64**: 333-345.

Lee, Y-S., Jo, Y-H., Kim, D-S., Kim, D-W., Kim, M-Y., 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 Mollusks. *Korean Journal of Malacology*, **20**: 165-169.

Mardis, E.R. (2013) Next-Generation Sequencing Platforms. *Annual Review of Analytical Chemistry*, **6**: 287-303.

Riesgo, A., Andrade, S.C.S., Sharma, P.P., Novo, M., Perez-Porro, A.R., Vahtera, V., Gonzalez, V.L., Kawauchi, G.Y. and Giribet, G. (2012) Comparative description of ten transcriptomes of newly sequenced invertebrates and efficiency estimation of genomic sampling in non-model taxa. *Frontiers in Zoology*, **9**: 33.

Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of National Academy of Sciences, USA*, **74**: 5463-5467.

Tsiliki, G., Tsaramirsis, K. and Kossida, S. (2014) AmalgamScope: Merging annotations data across the Human genome. *BioMed Research International*, Article ID 893501.

Zhu, Z., Niu, B., Chen, J., Wu, S., Sun, S. and Li, W. (2013) MGAviewer: a desktop visualization tool for analysis of metagenomics alignment data. *Bioinformatics*, **29**: 122-123.