Bioinformatics
doi.10.1093/bioinformatics/xxxxxx

Genome analysis

"Tardigrades: from genestealers to space marines". Eukaryotic genome analysis.

Vlasenko Daniil[1], Dracheva Kseniia[2]

[1]St. Petersburg State University, School of Mathematics and Mechanics, St. Petersburg, Russia, E-mail: vlasnko.daniil.vl@gmail.com.
[2]Pavlov First State Medical University of St. Petersburg, St. Petersburg, Russia, E-mail: xeniadracheva@gmail.com

Abstract

Tardigrades are microscopic invertebrate animals. They are also called aquatic bears. Their special feature is their ability to survive in extreme environmental conditions. It will be interesting to understand what made tardigrades get this feature. We had the tardigrade genome assembled and then we need to computationally identify proteins that can be associate with DNA protection and/or to effectively DNA repair. We obtained several proteins with nuclear localization that can be suggested for further study. One of them is termed Damage suppressor (Dsup) that is unique to the tardigrade.
Supplementary information: https://github.com/Daniil-Vlasenko/IBBioinformaticsWorkshop/blob/main/Project%204

1 Introduction

Tardigrades, also known as water bears, are tiny aquatic animals having four pairs of legs [1]. They were discovered in the 18th Century with the development of early microscopes and were first described by the German zoologist Goeze in 1773 [1]. They can live in a variety of habitats, such as marine, freshwater, or lima-terrestrial environments. Tardigrades are exceptional among metazoans in their adaptations to the most extreme environments. There are about 1,300 species of tardigrades in the world. They can go up to 30 years without food or water, and live at temperatures as cold as absolute zero or above boiling, at pressures six times that of the ocean's deepest trenches, and in the vacuum of space [2].

At the moment there is a debate about the effect of horizontal gene transfer on the ability of the tardigrades to survive in extreme conditions. According to a study by Boothby T.K. et al [3]. suggest that an large fraction of an animal genome can be derived from foreign sources and this helps them survive in an extremal environment. Opposite Georgios Koutsovoulos et al. detected a low level of horizontal gene transfer [4].

There are lots of different ways to predict coding regions in the genome. Two classes of methods are generally adopted: similarity based searches and ab initio prediction. Sequence similarity search is a conceptually simple approach that is based on finding similarity in gene sequences between ESTs (expressed sequence tags), proteins, or other genomes to the input genome. This approach is based on the assumption that functional regions (exons) are more conserved evolutionarily than nonfunctional regions (intergenic or intronic regions). Once there is similarity between a certain genomic region and an EST, DNA, or protein, the similarity information can be used to infer gene structure or function of that region [5].

ab initio method for the computational identification of genes is to use gene structure as a template to detect genes [5].

Sequence similarity searching, typically with BLAST, is the most widely used, and most reliable, strategy for characterizing newly determined sequences. Sequence similarity searches can identify "homologous" proteins or genes [6].

2 Methods

We used already assembled genome of Ramazzottius varieornatus [7] and it's functional annotation [8] obtained by AUGUSTUS [9]. Then we extracted protein sequences from functional annotation and put them into fasta file with getAnnoFasta.pl script [10].

We extracted chromatin fraction and analyzed the extracted proteins using tandem mass spectrometry [11]. Then we made protein database by BLAST [12] and aligned mass spectrometry results to database to understand which of proteins were most likely from cell nucleus.

We used WoLF PSORT [13] to predict the localization of the proteins that were obtained in the previous step and selected proteins that WoLF PSORT identified as nuclear.

For protein functional properties prediction we used BLAST protein search against the UniProtKB/Swiss-Prot database [14] and HMMER hmmscan search with Pfam protein family [15].

3 Results

AUGUSTUS's result contained 16435 proteins. After aligning mass spectrometry results to database 34 proteins left, 17 of which WoLF PSORT identified as nuclear.

Results of HMMER and BLAST search present in Table 1 and Table 2.

Table 1. HMMER hmmscan search with Pfam protein family.

Accession Number	Annotation	Conditional E-value	Independent E-value
g2203.t1	Glycosyl hydrolases family 31	2.4e-45	4.8e-41
g7861.t1	SNF2-related domain	1.8e-32	1.2e-28
g8100.t1	Inositol monophosphatase family	2.0e-41	1.9e-37
g8312.t1	Region in Clathrin and VPS	2.7e-27	5.4e-23
g11513.t1	Transport protein Trs120 or TRAPPC9, TRAPP II complex subunit	4.9e-14	9.6e-10
g11960.t1	Zinc finger, C3HC4 type (RING finger)	2.1e-09	4.2e-05
g15484.t1	Vps51/Vps67	3.2e-27	1.3e-23

4 Discussion

From proteins with nuclear localization we selected 2 proteins: g11960.t1 and g7861.t1 which had homologues in BLAST. g11960.t1 is a homologue of the E3 ubiquitin-protein ligase. E3 ubiquitin ligases are a group of proteins that transfer ubiquitin from E2 conjugating enzymes to highly specific substrates such as DNA repair proteins [16]. g7861.t1 s a homologue of SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A-like protein 1. ATP-dependent annealing helicase that binds selectively to fork DNA relative to ssDNA or dsDNA and catalyzes the rewinding of the stably unwound DNA. Acts throughout the genome to reanneal stably unwound DNA, performing the opposite reaction of many enzymes, such as helicases and polymerases, that unwind DNA. May play an important role in DNA damage response by acting at stalled replication forks [17].

A number of proteins with nuclear localization did not have holomogs in the analyzed databases. There are g10513t1, g10514.t1, g11806.t1, g14472.t1. Therefore, we can assume that they may be specific to the tardigrades and can be offered for further study.

An article by Takuma Hashimoto et al. [2] was studied the protein Dsup as specific to tardigrades. We also found this protein - g14472.t1. According to our data, it has a nuclear localization and has no similar proteins in other organisms.

Table 2. Results of BLAST protein search against the UniProtKB/Swiss-Prot database.

Accession	E-value	% Ident	% Query coverage	Annotation
g2203.t1	2e-126	35.93%	75%	Q69ZQ1.2
g3428.t1	9e-65	56.60%	91%	Q09510.1
g5927.t1	1e-18	38.64%	14%	Q17427.1
g7861.t1	2e-71	37.21%	99%	B4F769.1
g8100.t1	3e-46	36.04%	22%	Q2YDR3.1
g8312.t1	0.0	40.84%	84%	Q5KU39.1
g11513.t1	7e-83	28.61%	68%	Q32PH0.1
g11960.t1	6e-98	26.96%	96%	Q8CJB9.1
g14472.t1	0.0	100.00%	100%	P0DOW4.1
g15484.t1	0.0	45.03%	78%	Q155U0.1
g16318.t1	4e-08	36.11%	40%	A2VD00.1
g16368.t1	1e-05	39.29%	35%	A4II09.1

References

[1]Møbjerg, Nadia, et al. "Survival in extreme environments–on the current knowledge of adaptations in tardigrades." Acta physiologica 202.3 (2011): 409-420. https://doi.org/10.1111/j.1748-1716.2011.02252.x

[2]Hashimoto, Takuma, et al. "Extremotolerant tardigrade genome and improved radiotolerance of human cultured cells by tardigrade-unique protein." Nature communications 7.1 (2016): 1-14. https://doi.org/10.1038/ncomms12808

[3]Boothby, Thomas C., et al. "Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade." Proceedings of the National Academy of Sciences 112.52 (2015): 15976-15981. https://doi.org/10.1073/pnas.1510461112

[4]Koutsovoulos, Georgios, et al. "No evidence for extensive horizontal gene transfer in the genome of the tardigrade Hypsibius dujardini." Proceedings of the National Academy of Sciences 113.18 (2016): 5053-5058. https://doi.org/10.1073/pnas.1600338113

[5]Wang, Zhuo, Yazhu Chen, and Yixue Li. "A brief review of computational gene prediction methods." Genomics, proteomics & bioinformatics 2.4 (2004): 216-221. https://doi.org/10.1016/S1672-0229(04)02028-5

[6]Pearson, William R. "An introduction to sequence similarity ("homology") searching." Current protocols in bioinformatics 42.1 (2013): 3-1. https://doi.org/10.1002/0471250953.bi0301s42.

[7]Genome; Available from: ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/949/185/GCA_001949185.1_Rvar_4.0/GCA_001949185.1_Rvar_4.0_genomic.fna.gz.

[8]Annotation; Available from: https://drive.google.com/file/d/1wBxf6cDgu22NbjAOgTe-8b3Zx60hNKY0/view?usp=drive_web.

[9]Mario Stanke, Mark Diekhans, Robert Baertsch, David Haussler (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics, 24(5), pages 637–644, doi: 10.1093/bioinformatics/btn013

[10]Script; Available from: http://augustus.gobics.de/binaries/scripts/getAnnoFasta.pl.

[11]Mass spectrometry; Available from: https://disk.yandex.ru/d/xJqQMGX77Xueqg.

[12]Camacho, C., Coulouris, G., Avagyan, V. et al. BLAST+: architecture and applications. BMC Bioinformatics 10, 421 (2009). https://doi.org/10.1186/1471-2105-10-421.

[13]Paul Horton, Keun-Joon Park, Takeshi Obayashi, Naoya Fujita, Hajime Harada, C.J. Adams-Collier, Kenta Nakai, WoLF PSORT: protein localization predictor, Nucleic Acids Research, Volume 35, Issue suppl_2, 1 July 2007, Pages W585–W587, https://doi.org/10.1093/nar/gkm259

[14]Madden T. The BLAST Sequence Analysis Tool. 2002 Oct 9 [Updated 2003 Aug 13]. In: McEntyre J, Ostell J, editors. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2002-. Chapter 16. Available from: http://www.ncbi.nlm.nih.gov/books/NBK21097.

[15]hmmscan :: search sequence(s) against a profile database HMMER 3.3.2 (Nov 2020); http://hmmer.org/ Copyright (C) 2020 Howard Hughes Medical Institute. Freely distributed under the BSD open source license.

[16]Natarajan, Chandramouli, and Kenichi Takeda. "Regulation of various DNA repair pathways by E3 ubiquitin ligases." Journal of Cancer Research and Therapeutics 13.2 (2017): 157. https://doi.org/10.4103/0973-1482.204879

[17]Yusufzai, Timur, and James T. Kadonaga. "HARP is an ATP-driven annealing helicase." Science 322.5902 (2008). https://doi.org/748-750. 10.1126/science.1161233