*Bioinformatics Institute* (2022), **xx**:xx 1–5

**REPORT**

BIOINFORMATICS INSTITUTE

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

Vsevolod Zvezdin[1] and Vadim Smirnov[2]

[1]Chemistry and Biology Department, Grenoble Alpes University, Grenoble.
[2]Institute for Theoretical and Mathematical Physics, Moscow State University, Moscow.

**Abstract**
UV radiation is one of the factors that might lead to DNA damage. However, certain organisms show high genome stability even when exposed to it, employing damage prevention and DNA reparation systems. In that regard, Tardigrades, also known as water bears or moss piglets, present unique study opportunities. In this article, we shed some light on genetic peculiarities of *Ramazzottius varieornatus* that make them particularly resistant to radiation.

## Contents

**1 Introduction** **1**
**2 Methods** **2**
2.1 Laboratory Journal . . . 2
2.2 Structural annotation . . . 2
2.3 Physical localisation . . . 2
2.4 Localization prediction . . . 2
2.5 Homologs search . . . 3
2.6 Pfam prediction . . . 3
**3 Results** **3**
**4 Discussion** **3**
**5 Appendix** **4**

## 1. Introduction

The tardigrades are extremophile animals that are able to survive at both water boiling and near-zero temperatures, under the pressure of ocean depths or in a vacuum. Some tardigrade species, such as *Ramazzottius varieornatus* possess yet other intriguing traits. One such trait is the ability to sustain high doses of UV radiation which are deadly for most eukaryotic organisms.[1]

The reasons behind it were debated for a time. It was considered [2] that *R. varieornatus* could utilize the dehydration of their body, another adaptation common to tardigrades, resulting in notably less cellular damage from the radiation in the first place. However, later it was shown that even in a normal, non-dehydrated physiological state, with UV easily penetrating tissues and cells, *R. varieornatus* still can show much higher resilience than other animals. It was suggested then that this species is able to somehow repair the UV-mediated damage (to DNA in particular) in addition to avoiding some of it.

There is still a question as to how they handle this task, however. It is known that some eukaryotes, especially the more primitive ones, are able to borrow DNA from other organisms in a process of horizontal gene transfer[3]. One such example is *Bdelloidea*, a class of wormlike animals in whose genomes a large-scale HGT has happened, with sequences brought from other organisms comprising

tens of percent of genome coding parts. Mobile DNA elements may bring gene sequences within them as well, which a recipient organism may then utilize for their needs. Therefore, a theory appeared, proposing that in tardigrades some proteins with unique traits evolved in a similar fashion from the sequences transferred from other organisms. One research showed that a particular tardigrade species, *Hypsibius dujardini*, has around 1/6 of its genome to derive from recently transferred sequences[4]. However, the article was soon famously disproved[5], noting that such findings were the result of sample contamination.

Consequently, it was theorized that tardigrades might have developed their stress-tolerance mechanisms as a result of a long evolution, and some of the proteins responsible for them bear only little similarity to their homologs in distant organisms, essentially being taxon-specific[2]. This means they can not be reliably mapped to other taxa genomes. As such, they need to be identified *de novo* using gene-prediction assays. In this study, we combine the results of BLAST homolog search, domain prediction, and cellular localization data to analyze the pool of predicted genes and determine those most likely responsible for the unique adaptability of *R. varieornatus*.

## 2. Methods

### *2.1. Laboratory Journal*

The journal is available at https://github.com/Fulmenius/BI-Fall-2022-Practice-IV/blob/main/SmirnovLabJournalIV.md.

### *2.2. Structural annotation*

We analyzed the list of genes predicted in a *R. varieornatus* YOKOZUNA-1 strain genome assembly (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). The AUGUSTUS tool prediction results in .gff format can be procured from https://drive.google.com/file/d/1wBxf6cDgu22NbjAOgTe-8b3Zx60hNKY0/view?usp=drive_web. The perl script http://augustus.gobics.de/binaries/scripts/getAnnoFasta.pl was run on that file with default parameters to extract protein records in FASTA format.

### *2.3. Physical localisation*

To determine proteins of interest, we made use of an output of tandem mass-spectrometry on *R. varieornatus* cells, the set of peptides corresponding to chromatin fraction (can be downloaded from https://disk.yandex.ru/d/xJqQMGX77Xueqg). For the alignment, we used the Blastp program (Command Line Version, 2.5.0). The database was made from a reference file (with makeblastdb tool) and the proteins were aligned to it with output in "tabular" format. Unique protein names were procured from alignment and then used to retrieve their sequence from the reference proteome using samtools faidx tool (see our lab journal).

### *2.4. Localization prediction*

The proteins' most probable localization was obtained by assessing their N-terminal sequence in WoLF PSORT. TargetP 2.0 was similarly used as a complementary source of information to eliminate signals associated with transfer to mitochondria or outside the cell. The proteins that scored most in WoLF PSORT for nuclear localization were taken for further analysis.

### *2.5. Homologs search*

The supposed nuclear proteins were aligned with BlastP (Web-version) against UniProtKB/Swiss-Prot database with *R. varieornatus* excluded from the organism pool. For those of them that didn't produce any results, the Expect threshold value was raised to 0.2, and the run was performed again.

### *2.6. Pfam prediction*

For all nuclear proteins, their motifs were also predicted using the HMMSCAN tool and PFAM database as the reference.

## 3. Results

The initial multi-FASTA file with all proteins contained 16435 records. The alignment of the peptides to the list resulted in 117 hits, with several of them aligning multiple times hinting at gene duplication; filtering for unique records returned 34 distinct proteins. WoLF PSORT annotation attributed nuclear localization to 12 of them. These 12 proteins were chosen for BLAST and HMMSCAN search. The compilation of these results is presented in tables 1 and 2 (in the appendix). 4 proteins are unique in that they don't have any hits in the BLAST database as well as no findings in HMMSCAN report. Their sequences can be found in the table 3 (in the appendix).

## 4. Discussion

In this report, we identify 4 proteins that are possibly localized in the nucleus, and at the same time can not be identified by BLAST search or domain function prediction tools. These results tell us that such sequences do not occur in higher animals, although there is still a possibility for them to exist in other organisms, closer to tardigrades, yet still not found. We suggest that some or all of them can be related to a DNA damage repair and/or prevention.

Most known organisms possess proteins that detect strand breaks, mismatches, and other types of DNA damage and recruit reparases. It is a vital function for survival, with mutations occurring constantly even under normal conditions. It can be assumed that the protein(s) in question either restores impaired DNA more effectively than proteins of similar function in other species or doesn't allow it to be damaged, being quite unique in the latter case.

These proteins annotation can be completed by experiments that might prove their actual localization in the cell, such as FISH or other microscopic techniques, and the ability to bind DNA, which is assessed with electrophoretic mobility shift assay on the purified protein. Similarly, using electrophoresis, it is easy to determine the PI of a protein to see if it is able to bind DNA at all (PI should be significantly above 7 - in cells, such basic proteins are positively charged and are able to interact with negatively charged DNA strand). It is also possible to visualise DNA-protein complexes with techniques such as X-ray crystallography. Finally, the survivability and DNA damage levels should be checked in knockout animals.

*Disclaimer: now in 2022 we know that at least one of these proteins, labeled as g14472.t1 in our reports, indeed has a role in DNA protection. A Dsup protein is involved in packing DNA in such a manner that the access of harmful hydroxyl radicals, generated from UV-radiation treatment, is prohibited. As a result, the same radiation levels are notably less deadly for this species.*

## 5. Appendix

Table 1. Reference nuclear proteins with their corresponding annotation in BLAST.

| Gene Ref № | Annotation | Coverage | E-value | Ident. % | Accession |
|---|---|---|---|---|---|
| g11960.t1 | RecName: Full=E3 ubiquitin-protein ligase BRE1B [Rattus norvegicus] | 96% | 6,00E-98 | 26.96% | Q8CJB9.1 |
| g14472.t1 | ---------- | | | | |
| g15484.t1 | RecName: Full=Vacuolar protein sorting-associated protein 51 homolog [Danio rerio] | 78% | 0 | 45.03% | Q155U0.1 |
| g16318.t1 | RecName: Full=Eukaryotic translation initiation factor 3 subunit A [Xenopus laevis] | 40% | 4,00E-08 | 36.11% | A2VD00.1 |
| g16368.t1 | RecName: Full=Eukaryotic translation initiation factor 3 subunit A [Xenopus tropicalis] | 35% | 1,00E-05 | 39.29% | A4II09.1 |
| g5927.t1 | RecName: Full=Glucosamine 6-phosphate N-acetyltransferase [Caenorhabditis elegans] | 14% | 1,00E-18 | 38.64% | Q17427.1 |
| g7861.t1 | RecName: Full=Sucrose nonfermenting protein 2-like 1 [Rattus norvegicus] | 99% | 2,00E-71 | 37.21% | B4F769.1 |
| g8100.t1 | RecName: Full=Inositol monophosphatase 3 [Danio rerio] | 22% | 3,00E-46 | 36.04% | Q2YDR3.1 |
| g8312.t1 | RecName: Full=Vacuolar protein sorting-associated protein 41 homolog [Mus musculus] | 84% | 0 | 40.84% | Q5KU39.1 |
| g10513.t1 | ---------- | | | | |
| g10514.t1 | ---------- | | | | |
| g.11806.t1 | ---------- | | | | |

Table 2. Reference nuclear proteins with their HMMSCAN prediction results and localization. "Other" in TargetP means any sequences not associated with transport through mitochondrial or cellular membranes.

| Gene Ref N | HMM search | E-value of HMM | WOLF PSORT | TargetP |
|---|---|---|---|---|
| g11960.t1 | zf-C3HC4; Zinc finger, C3HC4 type (RING finger) | 4.20E-05 | Nuclear | Other |
| g14472.t1 | -- | -- | Nuclear | Other |
| g15484.t1 | Vps51/Vps67 | 1.30E-23 | Nuclear | Other |
| g16318.t1 | -- | -- | Nuclear | Other |
| g16368.t1 | -- | -- | Nuclear | Other |
| g5927.t1 | -- | -- | Nuclear | Other |
| g7861.t1 | SNF2-related domain | 1.20E-28 | Nuclear | Other |
| g8100.t1 | Inositol_P; Inositol monophosphatase family | 1.90E-37 | Nuclear | Other |
| g8312.t1 | Clathrin; Region in Clathrin and VPS | 5.40E-23 | Nuclear | Other |
| g10513.t1 | -- | -- | Nuclear | Other |
| g10514.t1 | -- | -- | Nuclear | Other |
| g.11806.t1 | -- | -- | Nuclear | Other |

Table 3. Aminoacid sequences of unique *R. varieornatus* proteins.

| | |
|---|---|
| g10513.t1 | MSTTTSSSSSNKDKDSTDTYVRSADNTSGSSNTTGSSSTNKGGKSSSSDYDSDKTTTSSSYGTGGQNTSSYGQ<br>SGQGGQHNTSTSSSYGQSGQGGQSGQHSSSSYGQSGQHSSGQSGQHSSGQSGQQSGQHGSHEVQQKLK<br>EVGNLLQKAGHLLQDLQGSASDFSSSSSYSQRNQGGNYSGPYGGSDSQFHQYGSSGMGGGSYGQSSSYGQ<br>GSSGYGQGSSSYGQQSGRHQDSDRFGGSSSFGGQSSGQYGGHSQGGYGGQQGGYGGSSGQNYGPYXXX<br>XXXXXXXXXXRGMGGGYGFSQDSGRQGGMGMSGGMGGADRYGGFGGPNRPDMSYGQQYGGRSDRW |
| g10514.t1 | MSYNRTEYRSDSDRHDDDKQGGWRSWFGLGGNKDDNNDRDRGYSYTNTTTYRGDDSNRYGFSGDRMSG<br>PSGYSGAISGGYSGSRGGYGYTTSDDYNRGNTNYSRSNDNSNYNRDNSSRGGDRDVYYRQETRTYPTTGYT<br>GSSNYTSGXXXXXXXXXXXXXXXXXXXXXXXXXXPSYGNTDRDYTVSXXXXXXXXXXXXXXXXXXXXXGYGFSG<br>DRDTSFRNTSFNTTGDRDTFNRGSYSGGYATSGNQGGFSTTSGMGHPGNYSSSYTSPSGTYGQSSNYSYNR<br>NY |
| g.11806.t1 | MCFLKYGGSGRRPCLLNTFKDTYRHLEYPYGGFDSQSHQYEPSGMGGGSYGQSSSYGQGSSGYGQGSSSYG<br>QQSGRPQDSDRFGGSSSFGGQPSGQYGGHSQGGYGGQQGGYGGSNGQSYGPYGSSSNQQSSMESNRG<br>MDGSYGFSQDSGRQGGMGMPDFVGGSDRYGGFEGPNRPDMSYGQQNGGRIGDRWHHCANFNPL |
| g14472.t1 | MASTHQSSTEPSSTGKSEETKKDASQGSGQDSKNVTVTKGTGSSATSAAIVKTGGSQGKDSSTTAGSSSTQG<br>QKFSTTPTDPKTFSSDQKEKSKSPAKEVPSGGDSKSQGDTKSQSDAKSSGQSQGQSKDSGKSSSDSSKSHSVI<br>GAVKDVVAGAKDVAGKAVEDAPSIMHTAVDAVKNAATTVKDVASSAASTVAEKVVDAYHSVVGDKTDDKK<br>EGEHSGDKKDDSKAGSGSGQGGDNKKSEGETSGQAESSSGNEGAAPAKGRGRGRPPAAAKGVAKGAAKG<br>AAASKGAKSGAESSKGGEQSSGDIEMADASSKGGSDQRDSAATVGEGGASGSEGGAKKGRGRGAGKKADA<br>GDTSAEPPRRSSRLTSSGTGAGSAPAAAKGGAKRAASSSSTPSNAKKQATGGAGKAAATKATAAKSAASKAP<br>QNGAGAKKKGGKAGGRKRK |

## References


[1] Survival in extreme environments – on the current knowledge of adaptations in tardigrades N. Møbjerg, K. A. Halberg, A. Jørgensen, D. Persson, M. Bjørn, H. Ramløv, R. M. Kristensen, 2011

[2] Extremotolerant tardigrade genome and improved radiotolerance of human cultured cells by tardigrade-unique protein, Takuma Hashimoto1 et al., 2016

[3] Horizontal gene transfer in eukaryotes: not if, but how much? J Van Etten, D Bhattacharya - Trends in Genetics, 2020 - Elsevier

[4] Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade, Boothby TC, et al. (2015) Proc Natl Acad Sci USA, 112(52):15976–15981

[5] No evidence for extensive horizontal gene transfer in the genome of the tardigrade Hypsibius dujardini, Georgios Koutsovoulosa et al., 2016