Project 4. Tardigrades: from genestealers to space marines

Made by Alisa Fedorenko and Maria Uzun

1. Obtaining data. Genome sequence.

For this project we will be using a sequence of the Ramazzottius varieornatus, the YOKOZUNA-1 strain (sequenced in the University of Tokyo and named after the highest rank in professional sumo).

Download tarigrades assembled genome

```
mkdir Project4
cd Project4
mkdir raw_data
cd raw_data
wget ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/949/185/GCA_001949185.1_Rvar_4.0/GCA_001
949185.1_Rvar_4.0_genomic.fna.gz
```

QUAST online results for Tarigrade genome:

Statistics without reference	GCA_001949185.1_Rvar_4.0_geno…
# contigs	200
# contigs (>= 0 bp)	200
# contigs (>= 1000 bp)	200
# contigs (>= 5000 bp)	45
# contigs (>= 10000 bp)	37
# contigs (>= 25000 bp)	29
# contigs (>= 50000 bp)	28
Largest contig	9333084
Total length	55842812
Total length (>= 0 bp)	55842812
Total length (>= 1000 bp)	55842812
Total length (>= 5000 bp)	55527331
Total length (>= 10000 bp)	55468139
Total length (>= 25000 bp)	55346641
Total length (>= 50000 bp)	55307997
N50	4740345
N90	1295620
auN	4838420
L50	4
L90	15
GC (%)	47.51
Per base quality	
# N's per 100 kbp	750.86
# N's	419303

2. Structural annotation.

Download precomputed AUGUSTUS results - protein fasta and gff.

```
grep ">" augustus.whole.aa | wc -l
# 16435
```

We obtained 16435 protein sequences.

A list of peptides that were associated with the DNA (you can download that list here)

3. Physical localization

Local alignment-based search: create a local database from protein fasta file and look it up using your peptide sequence file as a query.

```
# install blast
mamba install -c bioconda blast
cd ..
mkdir blast
cd blast
makeblastdb --in ../raw_data/augustus.whole.aa --dbtype prot  --out ../raw_data/augustus_db
blastp -db ../raw_data/augustus_db --query ../raw_data/peptides.fa -outfmt 6  --out pepdides_blast
```

Totally, it was generated 34 proteins, which had matches with studied peptides.

g10513.t1	g10514.t1	g11320.t1	g11513.t1	g11806.t1	g11960.t1
g15153.t1	g15484.t1	g16318.t1	g16368.t1	g2203.t1	g3428.t1
g5502.t1	g5503.t1	g5510.t1	g5616.t1	g5641.t1	g5927.t1
g12388.t1	g12510.t1	g12562.t1	g1285.t1	g13530.t1	g14472.t1
g3679.t1	g4106.t1	g4970.t1	g5237.t1	g5443.t1	g5467.t1
g702.t1	g7861.t1	g8100.t1	g8312.t1		

```
# install diamond
conda install -c bioconda diamond

diamond makedb --in ../raw_data/augustus.whole.aa  --db ../raw_data/augustus_db_diamomd
diamond blastp -d ../raw_data/augustus_db_diamomd -q ../raw_data/peptides.fa -f 6 -o pepdides_diamond --very-sensitive
grep 'название' augustus.whole.aa
```

8	g4106.t1	100	18	0	0	1	18	222	239	2.90E-05	42
20	g12510.t1	100	18	0	0	1	18	425	442	1.50E-04	39.7
21	g12510.t1	100	22	0	0	1	22	443	464	1.50E-06	46.6
29	g4106.t1	100	18	0	0	1	18	222	239	2.90E-05	42

4. Localization prediction

4a. WoLF PSORT (https://wolfpsort.hgc.jp/)

As the result we found protein localization for 34 proteins from blast.

g702.t1 details extr: 29, plas: 2, lyso: 1

g1285.t1 details extr: 25, plas: 5, mito: 1, lyso: 1

g2203.t1 details plas: 29, nucl: 2, golg: 1

g3428.t1 details mito: 18, cyto: 11, extr: 2, nucl: 1

g3679.t1 details extr: 26, mito: 2, lyso: 2, plas: 1, E.R.: 1

g4106.t1 details E.R.: 14.5, E.R._golg: 9.5, extr: 7, golg: 3.5, lyso: 3, pero: 2, plas: 1, mito: 1 g4970.t1 details plas: 32

g5237.t1 details plas: 24, mito: 8

g5443.t1 details extr: 28, nucl: 3, cyto: 1

g5467.t1 details extr: 27, plas: 4, mito: 1

g5502.t1 details extr: 31, lyso: 1

g5503.t1 details extr: 29, plas: 1, mito: 1, lyso: 1

g5510.t1 details plas: 23, mito: 7, E.R.: 1, golg: 1

g5616.t1 details extr: 31, mito: 1

g5641.t1 details extr: 31, lyso: 1

g5927.t1 details nucl: 30.5, cyto_nucl: 16.5, cyto: 1.5

g7861.t1 details nucl: 16, cyto_nucl: 14, cyto: 8, plas: 5, pero: 1, cysk: 1, golg: 1

g8100.t1 details nucl: 16.5, cyto_nucl: 12.5, cyto: 7.5, plas: 5, extr: 2, E.R.: 1

g8312.t1 details nucl: 15.5, cyto_nucl: 15.5, cyto: 12.5, mito: 2, plas: 1, golg: 1

g10513.t1 details nucl: 20, cyto_nucl: 14.5, cyto: 7, extr: 3, E.R.: 1, golg: 1

g10514.t1 details nucl: 19, cyto_nucl: 15, cyto: 9, extr: 3, mito: 1

g11320.t1 details plas: 24.5, extr_plas: 16, extr: 6.5, lyso: 1

g11513.t1 details cyto: 17, cyto_nucl: 12.8333, cyto_mito: 9.83333, nucl: 7.5, E.R.: 3, mito: 1.5, plas: 1, pero: 1, golg: 1

g11806.t1 details nucl: 18, cyto_nucl: 11.8333, mito: 5, extr: 4, cyto: 3.5, cyto_pero: 2.66667, cysk_plas: 1

g11960.t1 details nucl: 32

g12388.t1 details extr: 25, plas: 4, mito: 2, lyso: 1

g12510.t1 details plas: 29, cyto: 3

g12562.t1 details extr: 30, lyso: 2

g13530.t1 details extr: 13, nucl: 6.5, lyso: 5, cyto_nucl: 4.5, plas: 3, E.R.: 3, cyto: 1.5 g14472.t1 details nucl: 28, plas: 2, cyto: 1, cysk: 1

g15153.t1 details extr: 32

g15484.t1 details nucl: 17.5, cyto_nucl: 15.3333, cyto: 12, cyto_mito: 6.83333, plas: 1, golg: 1 g16318.t1 details nucl: 20.5, cyto_nucl: 13, extr: 5, cyto: 4.5, E.R.: 1, golg: 1

g16368.t1 details nucl: 20.5, cyto_nucl: 13, extr: 5, cyto: 4.5, E.R.: 1, golg: 1

Results for diamond:

g4106.t1 details E.R.: 14.5, E.R._golg: 9.5, extr: 7, golg: 3.5, lyso: 3, pero: 2, plas: 1, mito: 1

g12510.t1 details plas: 29, cyto: 3

No nucleus localization was found using diamond

4b. TargetP Server (https://services.healthtech.dtu.dk/service.php?TargetP-2.0)

Results for 34 proteins can be found in TargetP-2.0 (dtu.dk) or here

All predicted signal peptides were widthrawn from analysis.

5. BLAST search

The BLAST search results are represented here. Blast against UniProtKB/Swiss-Prot database in NCBI genbank gave mostly proteins with hypothetical functions.

6. Pfam prediction

We used HMMER (web-version, https://www.ebi.ac.uk/Tools/hmmer/) to search our protein sequences against a collection of profile-HMMs for different protein domains and motifs.
Select “Search” and then select the suitable tool (in our case it is hmmscan), select the Pfam
database, and submit your proteins.

The Pfam prediction results you can find follow the link. For orthologous sequences not founded in databases with Blast the Pfam prediction also did not find any annotations and functions (highlighted by pink in the table above).

7. Integrate your various pieces of evidence

The summary table for selected proteins can be found here.

The yellow lines corresponds to proteins defined by TargetP prediction as nuclear proteins. As you can see just several proteins are trully nuclear. The functions of most proteins are annotated differently in relation to stress tolerance, and only five have been identified as potentially involved in stress tolerance:

Protein	Best blast hit to peptide	e-value	Description	Probable localization(s) in WoLF PSORT	Localization in TargetP
g10513.t1	9	0.003	No significant similarity found	g10513.t1 details nucl: 20, cyto_nucl: 14.5, cyto: 7, extr: 3, E.R.: 1, golg: 1	OTHER
g10514.t1	38	0.6	No significant similarity found	g10514.t1 details nucl: 19, cyto_nucl: 15, cyto: 9, extr: 3, mito: 1	OTHER
g14472.t1	7	0.002	Damage suppressor protein	g14472.t1 details nucl: 28, plas: 2, cyto: 1, cysk: 1	OTHER
g16318.t1	9	4.9	No significant similarity found	g16318.t1 details nucl: 20.5, cyto_nucl: 13, extr: 5, cyto: 4.5, E.R.: 1, golg: 1	OTHER

g16368.t1	9	5.2	No significant similarity found	g16368.t1 details nucl: 20.5, cyto_nucl: 13, extr: 5, cyto: 4.5, E.R.: 1, golg: 1	OTHER