Project 4. Pudgy wudgies’s genome

1. Obtaining data. Genome sequence.

Downloading assembeled genome from NCBI genomes database

```
wget 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
```

2. Functional annotation.

```
sudo apt install augustus augustus-data augustus-doc
```

```
wget http://augustus.gobics.de/binaries/scripts/getAnnoFasta.pl
```

```
chmod +x getAnnoFasta.pl
```

```
./getAnnoFasta.pl augustus.whole.gff
```

Number of obtained proteins:

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

3. Physical localisation

Downloaded **list** of peptides obtained with use of mass spectrometry:

https://disk.yandex.ru/d/xJqQMGX77Xueqg

To find whole sequences of obtained proteins we blasted those proteins on our genome

1. Creating database with ‘makeblastdb’

   ```
   makeblastdb -in augustus.whole.aa -dbtype prot -out tardigrade_db
   ```

2. Blastp with outfmt6 output:

   Documentation: https://www.metagenomics.wiki/tools/blast/blastn-output-format-6

   Added additional column with coverage

   ```
   blastp -db tardigrade_db -query peptides.fa -outfmt "6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bits
   ```

3. Result file with formating. Proteins selected for further analysis are highlighted with blue color:

   https://s3-us-west-2.amazonaws.com/secure.notion-static.com/d3305e95-4794-4438-9fc3-eb2561de05d7/outmft6_result_selection.xlsx

4. Parsing for whole protein sequences with samtools

```
samtools faidx augustus.whole.aa
```

all unique names were placed in file unique_proteins.txt with space as delimiter (otherwise it won’t work)

```
xargs samtools faidx augustus.whole.aa < unique_proteins.txt > unique_proteins_seq.faa
```

4. Localization prediction

1. Used WoLF PSORT to predict subcellular localisation

Wolf output:

> 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
>
> g1285.t1 details extr: 25, plas: 5, mito: 1, lyso: 1
>
> 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
>
> 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
>
> g702.t1 details extr: 29, plas: 2, lyso: 1 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

We selected poteins that have major presence in nucleus:

> g10513.t1 g10514.t1 g11806.t1 g11960.t1 g14472.t1 g15484.t1 g16318.t1 g16368.t1 g5927.t1 g7861.t1 g8100.t1 g8312.t1

2. Used TargetP 1.1 to also predict subcellular localization

5. BLAST search (online)

Results of alignment

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/9436abc9-53d0-4b42-9442-f6f7493b3c59/TF282WS1016-Alignment_(2).txt

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/1dc86fa1-77e6-4afc-810f-76f8bd7d3342/TF282WS1016-Alignment.txt

Selected hits presented here:

Protein	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident
g10513.t1							
g10514.t1							
g11806.t1							
g11960.t1	E3 ubiquitin-protein ligase BRE1B	Rattus norvegicus	334	334	0,96	0	26.96%
g14472.t1	Damage suppressor protein	Ramazzottius varieornatus	814	814	100%	0.0	100.00%
g15484.t1	Vacuolar protein sorting-associated protein 51 homolog	Danio rerio	592	592	78%	0.0	45.03%
g16318.t1	Eukaryotic translation initiation factor 3 subunit A	Xenopus laevis	57.4	590	40%	4,00E-08	36.11%
g16368.t1	Eukaryotic translation initiation factor 3 subunit A	Xenopus laevis	57.4	590	40%	4,00E-08	36.11%
g5927.t1	Glucosamine 6-phosphate N-acetyltransferase	Caenorhabditis elegans	87.4	87.4	0,14	0	38.64%
g7861.t1	Inositol monophosphatase 2	Rattus norvegicus		293	0,99	0	37.21%

g8100.t1	Inositol monophosphatase 3	Danio rerio	173	173	0,22	0	36.04%
g8312.t1	Vacuolar protein sorting-associated protein 41	Mus musculus	554	554	0,84	0.0	40.84%

6. Pfam prediction

https://www.ebi.ac.uk/Tools/hmmer/

Protein	Id	Accession	Clan	Description	Cross-references	Start	End
g10513.t1	None						
g10514.t1	None						
g11806.t1	None						
g11960.t1	zf-C3HC4	PF00097.28	CL0229	Zinc finger, C3HC4 type (RING finger)	dgPsbyl	927	965
g14472.t1	None						
g15484.t1	Vps51	PF08700.14	CL0295	Vps51/Vps67	dgPsbyl	10	96
g16318.t1	None						
g16368.t1	None						
g5927.t1	None						
g7861.t1	SNF2-rel_dom	PF00176.26	CL0023	SNF2-related domain	dgPsbyl	269	566
g7861.t1	HARP	PF07443.16	n/a	HepA-related protein (HARP)	dgPsbyl	173	228
g8100.t1	Inositol_P	PF00459.28	CL0171	Inositol monophosphatase family	dgPsbyl	449	788
g7861.t1	MKLP1_Arf_bdg	PF16540.8	n/a	Arf6-interacting domain of mitotic kinesin-like protein 1	dgPsbyl	1183	1287
g8312.t1	Clathrin	PF00637.23	CL0020	Region in Clathrin and VPS	dgPsbyl	652	792

7. Integrated table

Protein	Best BLAST hit	E value	Organism	Pham domains	Probable localization (WoLF PSORT)	TargetP
g10513.t1	None				Nuclear	OTHER
g10514.t1	None				Nuclear	OTHER
g11806.t1	None				Nuclear	OTHER
g11960.t1	E3 ubiquitin-protein ligase BRE1B	0	Rattus norvegicus	zf-C3HC4	Nuclear	OTHER
g14472.t1	Damage suppressor protein				Nuclear	OTHER
g15484.t1	Vacuolar protein sorting-associated protein 51 homolog	0.0	Danio rerio	Vps51	Nuclear	OTHER
g16318.t1	Eukaryotic translation initiation factor 3 subunit A	4,00E-08	Xenopus laevis		Nuclear	OTHER

g16368.t1	Eukaryotic translation initiation factor 3 subunit A	4,00E-08	Xenopus laevis		Nuclear	OTHER
g5927.t1	Glucosamine 6-phosphate N-acetyltransferase	0	Caenorhabditis elegans		Nuclear	OTHER
g7861.t1	Inositol monophosphatase 2	0	Rattus norvegicus	SNF2-rel_dom, HARP	Nuclear	OTHER
g8100.t1	Nuclear	Other		Inositol_P, MKLP1_Arf_bdg	Nuclear	OTHER
g8312.t1	Inositol monophosphatase 3	0	Danio rerio	Clathrin	Nuclear	OTHER