Analysis of RNA-sequencing from *Schizosaccharomyces pombe*
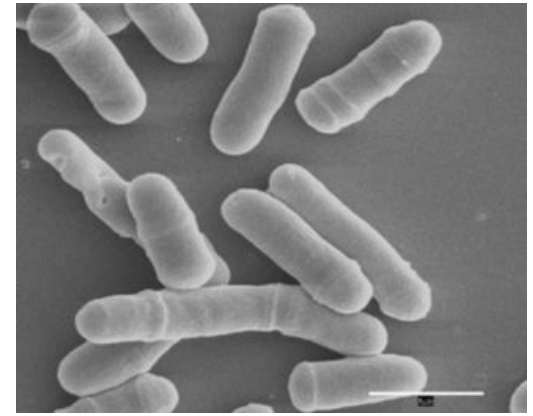
Laura Ciaran, Álvaro Ciudad,
Tatiana Gusinskaia

Table of contents

- *Schizosaccharomyces pombe*

- Procurement of the novel transcripts

- Comparison of known and novel transcripts

- Analysis and characterization of novel transcripts
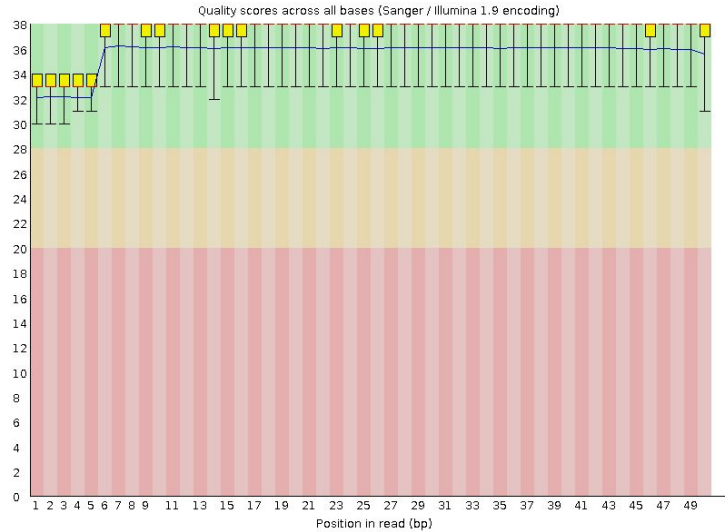
- Conclusion

Schizosaccharomyces pombe

- Unicellular ascomycota

- Rod shape cells with sizes varying between 3–5 × 5–24 μm

- Reproduction by fission instead of budding → Fission yeast

- Life cycle predominantly haploid only makes diploids when starved for nitrogen

- Genome size: approximately **14.1 million base pairs**, with estimated **4,970 protein-coding genes** and at least **450 non-coding RNAs**, distributed in three chromosomes: **I (5,7 Mb), II (4,6 Mb) y III (3,5 Mb)**.

- Used to study the cellular responses to DNA damage and the process of DNA replication.

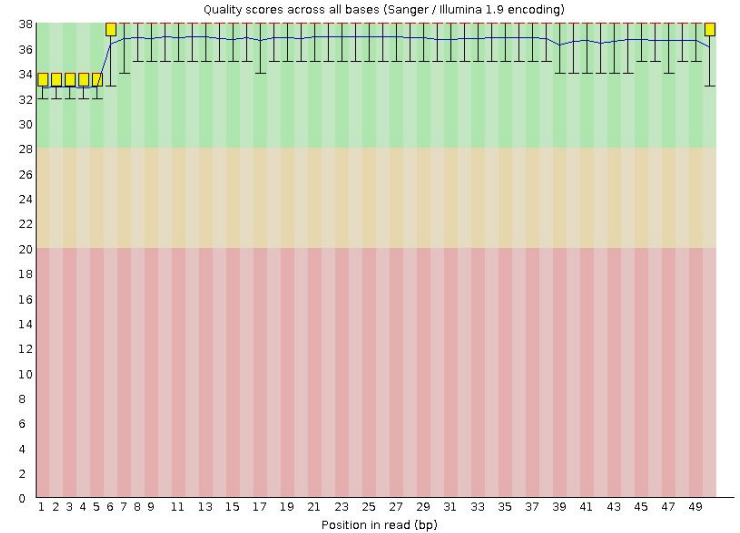# Analysis of the quality of the sequencing reads

Read 1   Read 2

22.724.832 reads   23.242.560 reads

FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at:
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2015), "FastQC," https://qubeshub.org/resources/fastqc.

Analysis of the quality of the sequencing reads

Bolger, A. M., Lohse, M., & Usadel, B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, (2014), btu170.
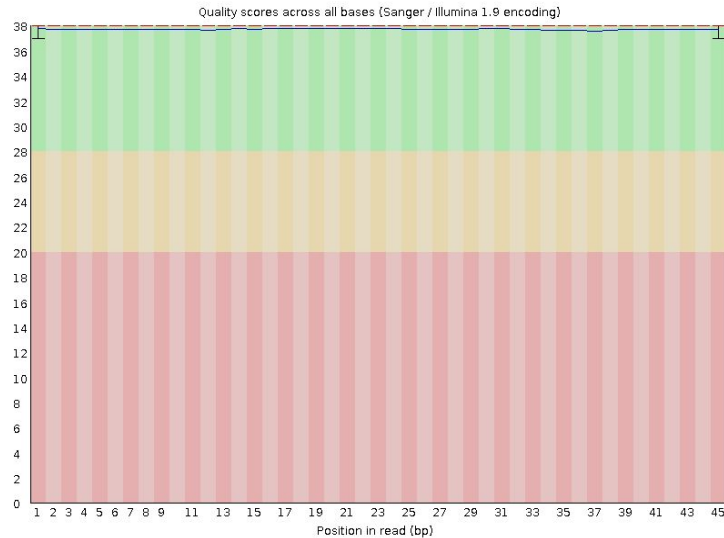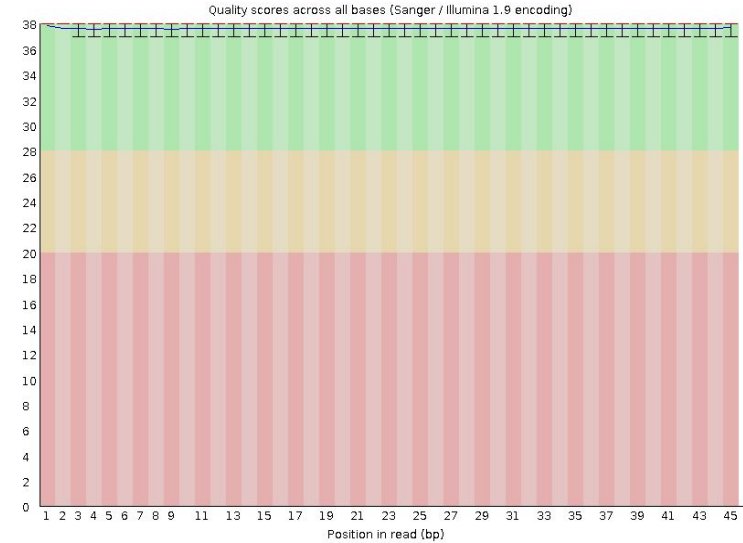
Aligning RNA-Seq reads to the genome

Assembling of transcripts using mapped reads

Obtaining sequences in fasta format using the genomic coordinates from a gff file

Known transcripts: 17045

Novel transcripts: 1617

- Unique novel transcripts: 1457

Kim, D., Langmead, B. & Salzberg, S. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods.* 12, 357–360 (2015).
Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295 (2015).
Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, Heng Li. Twelve years of SAMtools and BCFtools. *GigaScience.* 10(2) (2021), giab008.
Aaron R. Quinlan, Ira M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics.* 26(6): 841–842 (2010).
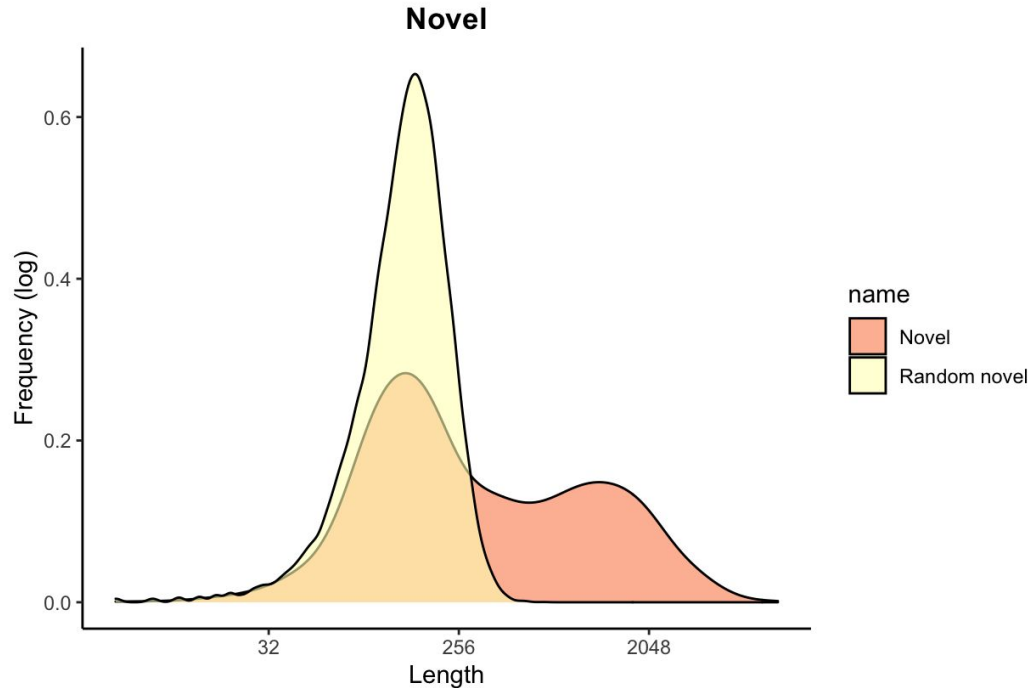
Longest ORFs length comparison- Known transcripts

Known median = 129
Known random median = 101
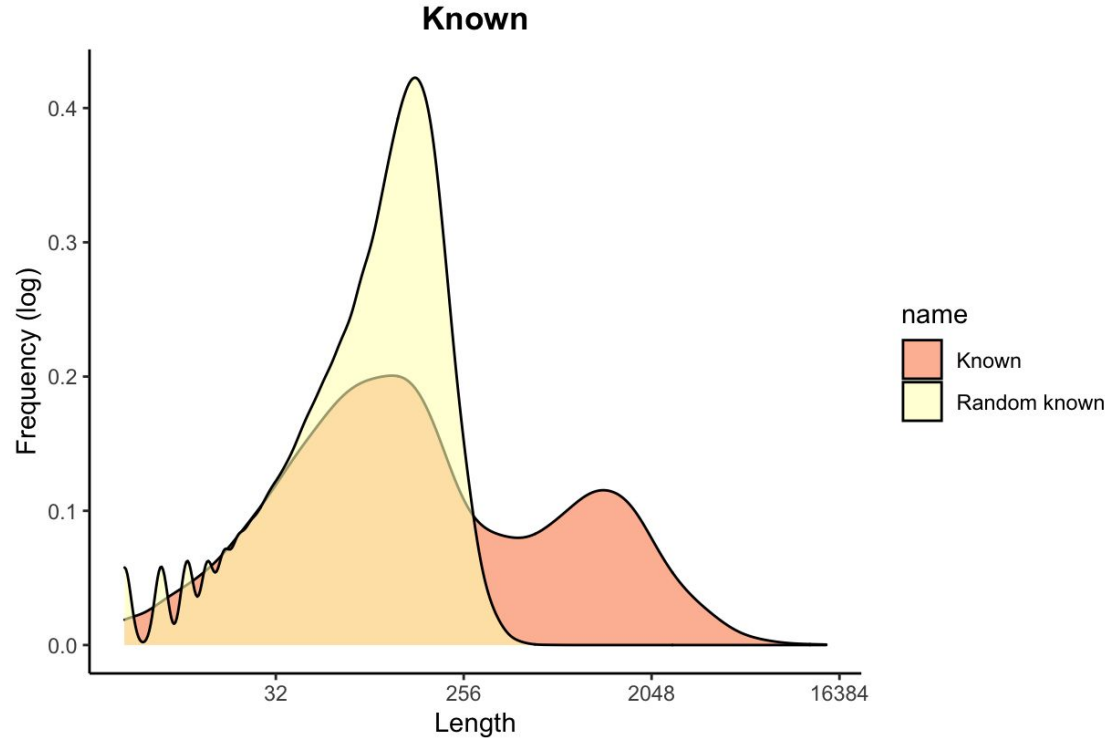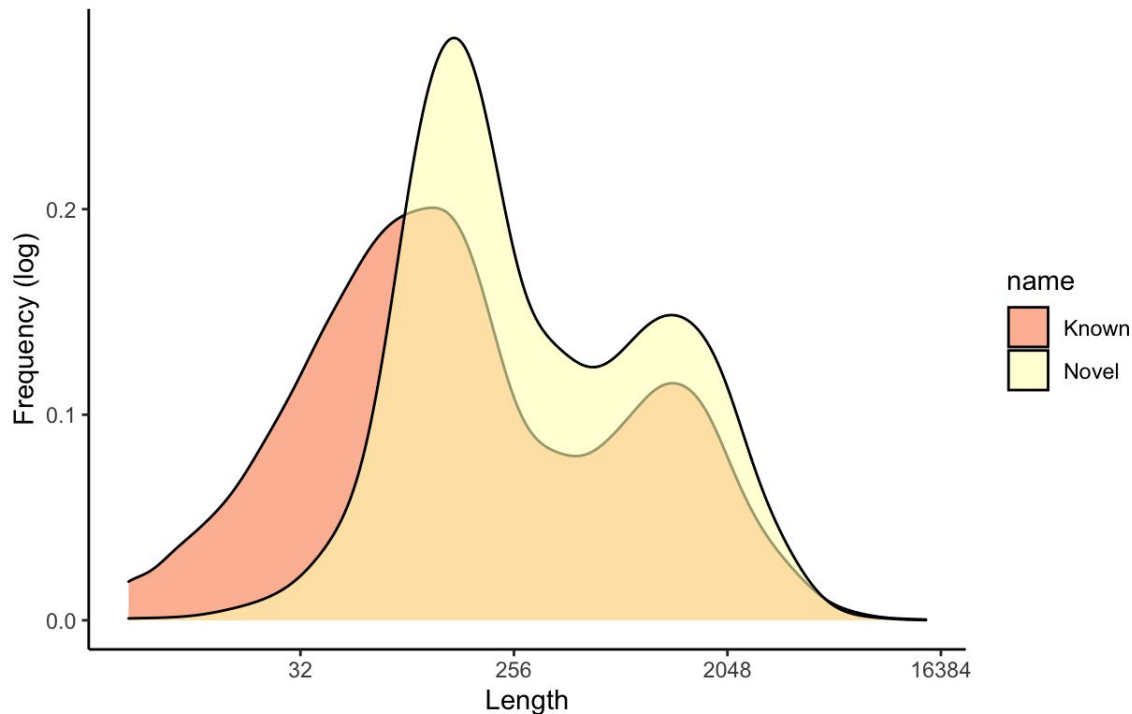
Longest ORFs length comparison- Novel transcripts
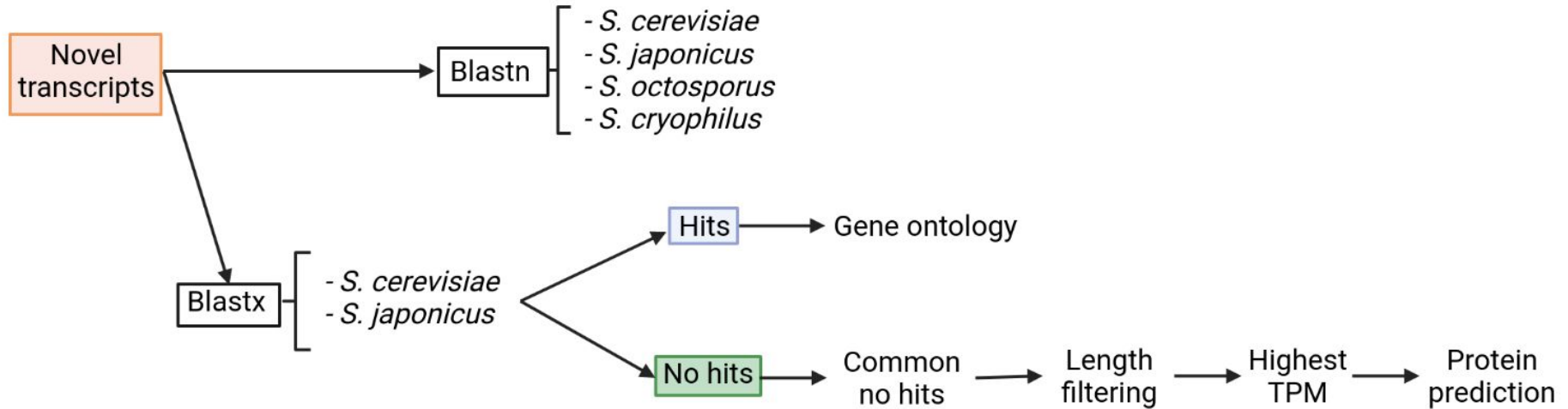
Novel median = 234
Novel random median = 147

Longest ORFs length comparison- Known vs Novel

Known median = 129
Novel median = 234

Longest ORF comparison
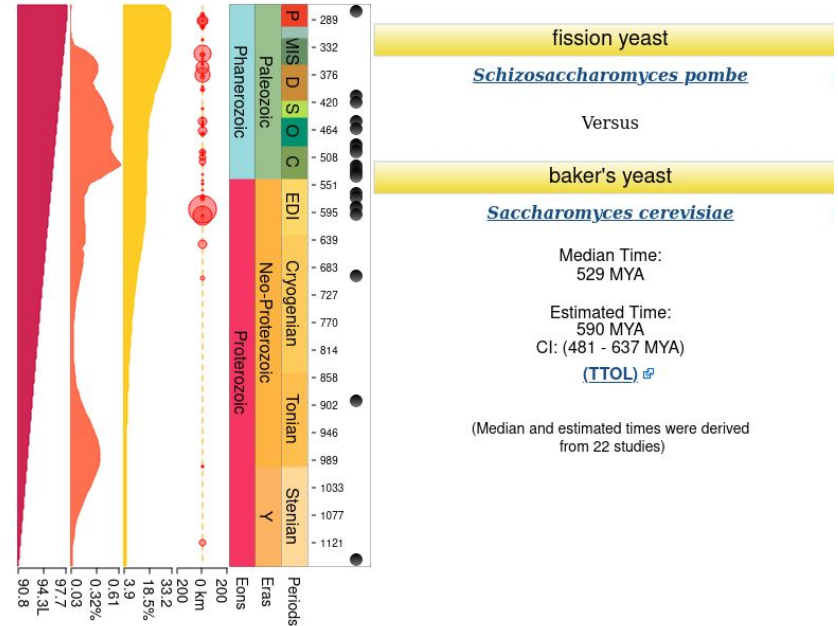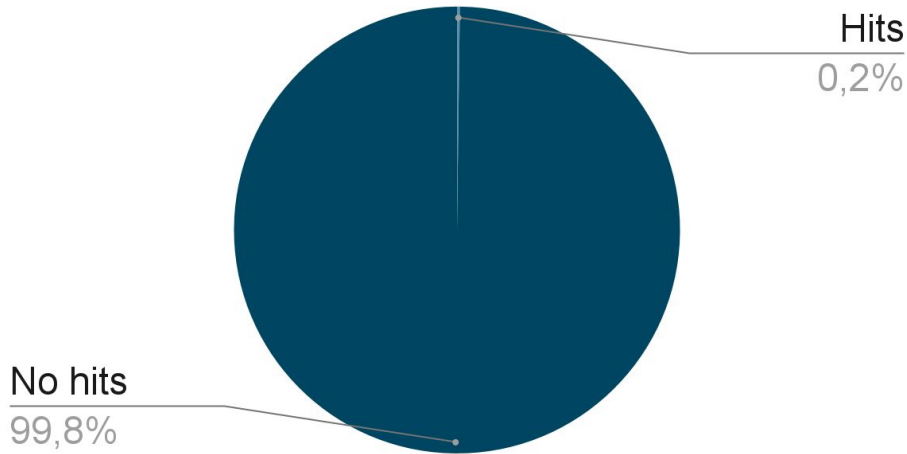
Analysis of novel transcripts

Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, David J. Lipman, Basic local alignment search tool.
Journal of Molecular Biology, 215(3):403-410 (1990).

Analysis of novel transcripts – Blastn

S. cerevisiae

Hits
0,2%

No hits
99,8%

fission yeast

Schizosaccharomyces pombe

Versus

baker's yeast

Saccharomyces cerevisiae

Median Time:
529 MYA

Estimated Time:
590 MYA
CI: (481 - 637 MYA)
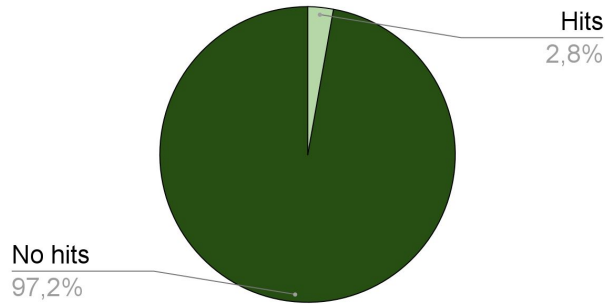
(TTOL)

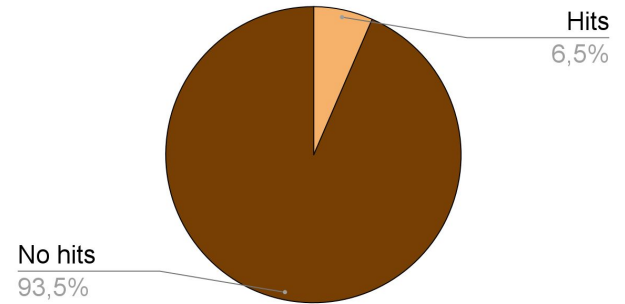(Median and estimated times were derived
from 22 studies)

S. Kumar, G. Stecher, M. Suleski, and S.B. Hedges, TimeTree: a resource for timelines, timetrees, and divergence times. M*olecular Biology and Evolution* 34: 1812-1819 (2017).

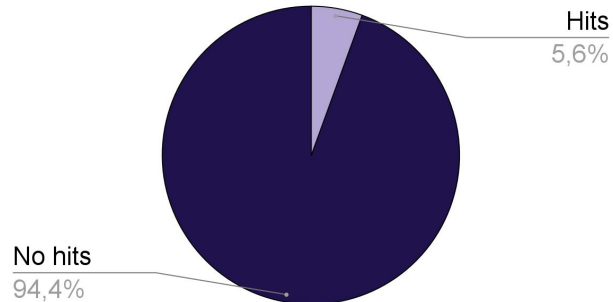# Analysis of novel transcripts - Blastn

S. japonicus

Hits
2,8%
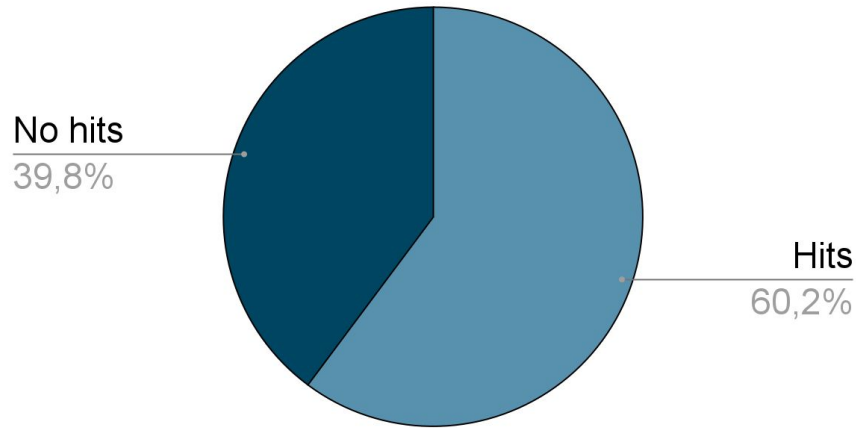
No hits
97,2%
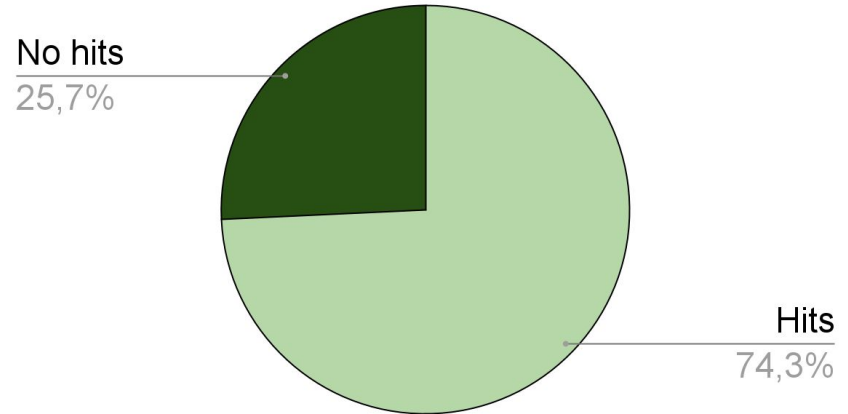
S. octosporus

Hits
6,5%

No hits
93,5%

S. cryophilus

Hits
5,6%

No hits
94,4%

Analysis of novel transcripts – Blastx

S. cerevisiae

No hits
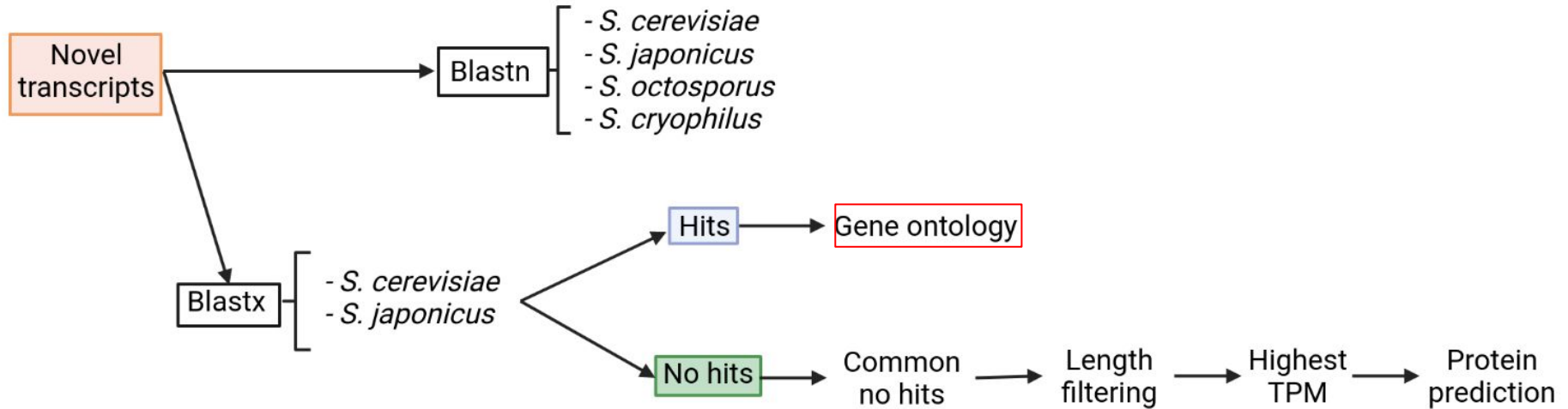39,8%

Hits
60,2%

877 hits

S. japonicus

No hits
25,7%

Hits
74,3%

1082 hits

Analysis of novel transcripts – Hit Ontology

Analysis of novel transcripts – Hit Ontology

S. cerevisiae

Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, Jaak Vilo. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists. *Nucleic Acids Research* (2019).
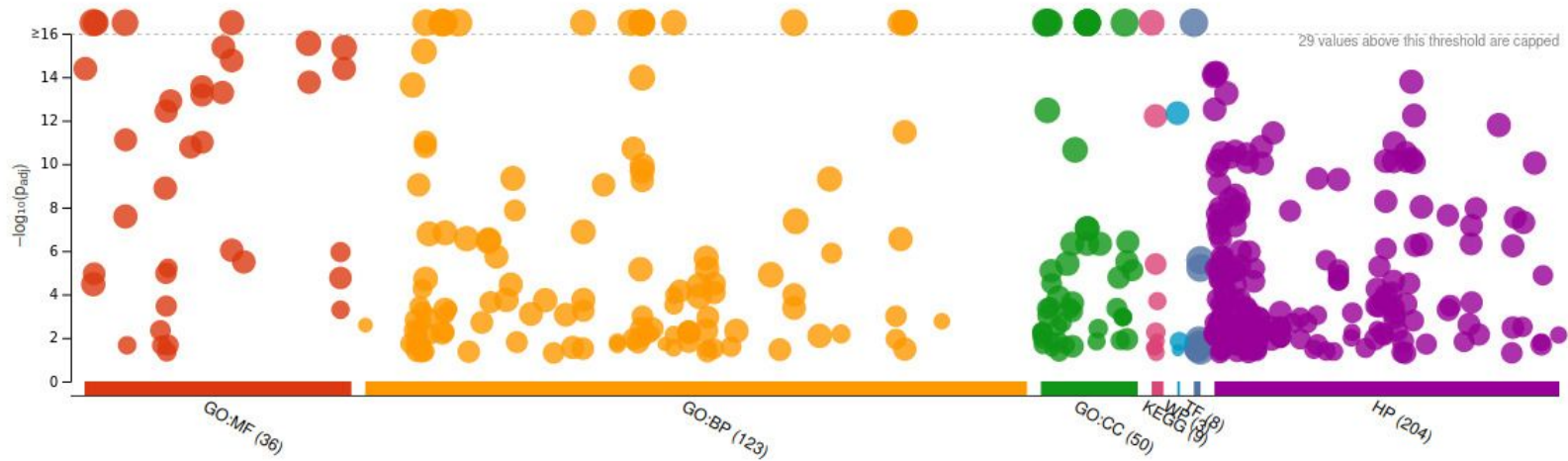
Analysis of novel transcripts – Hit Ontology

S. japonicus

Analysis of novel transcripts – Blastx
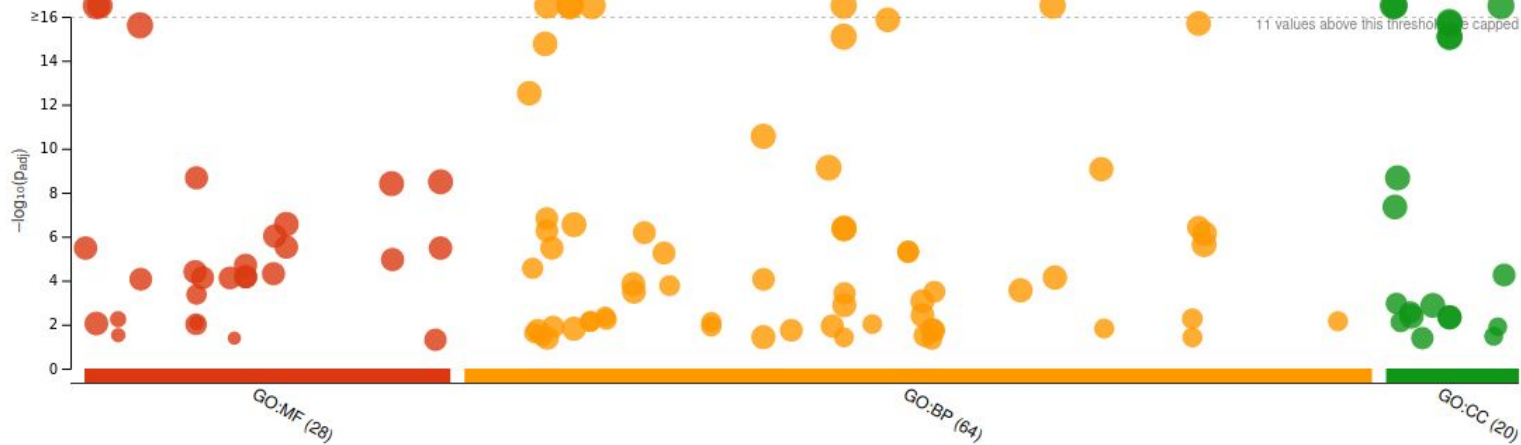
S. cerevisiae

No hits
39,8%

Hits
60,2%

S. japonicus

No hits
25,7%

Hits
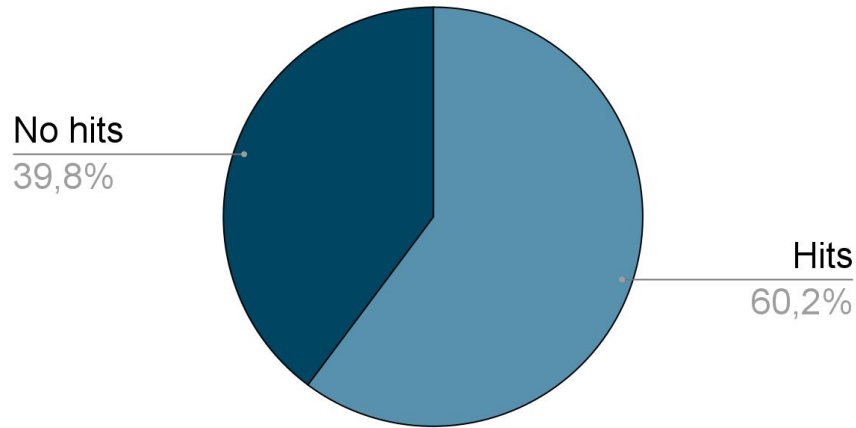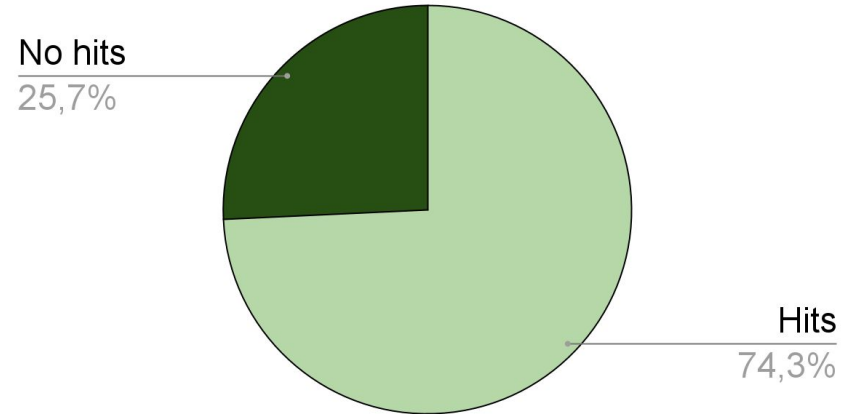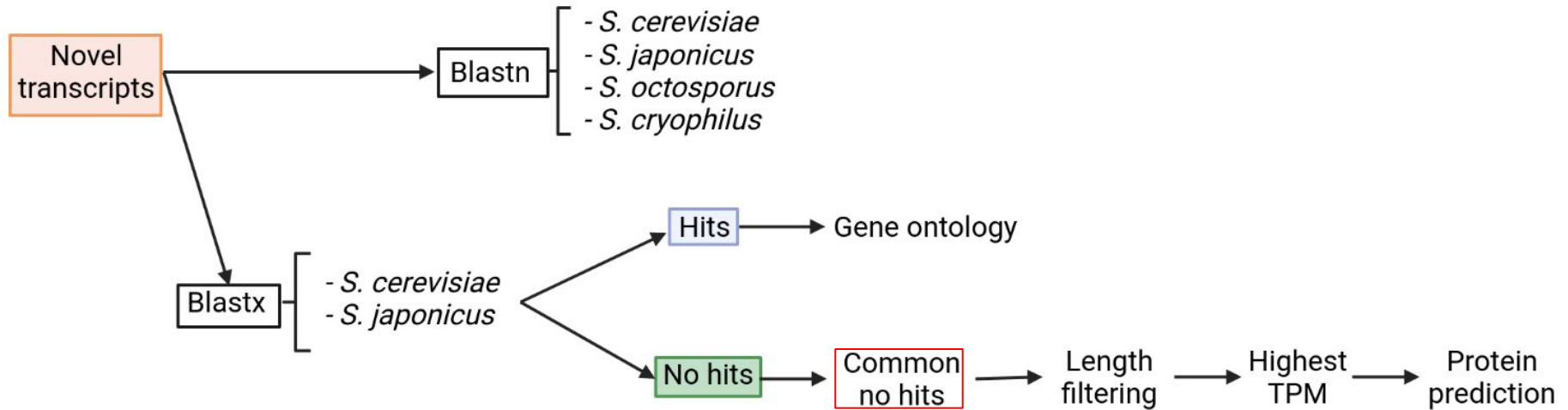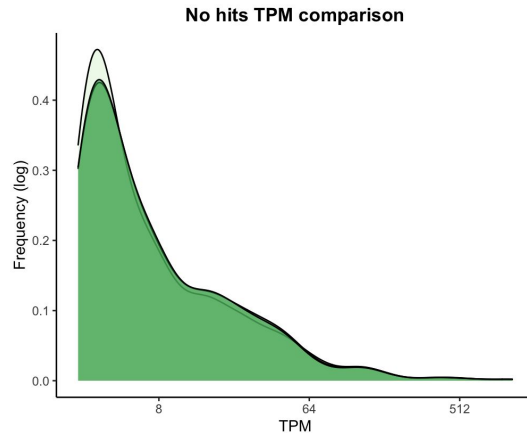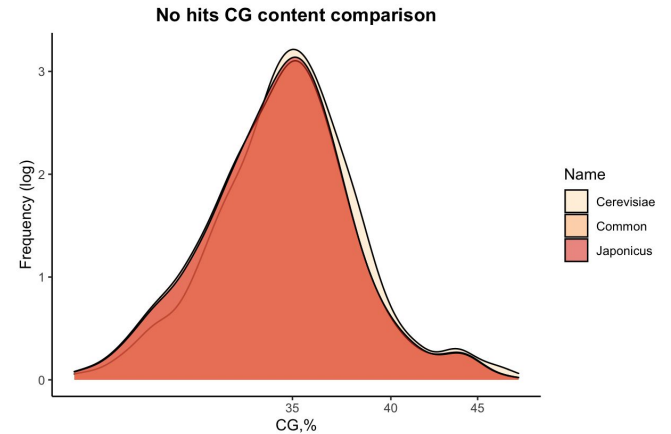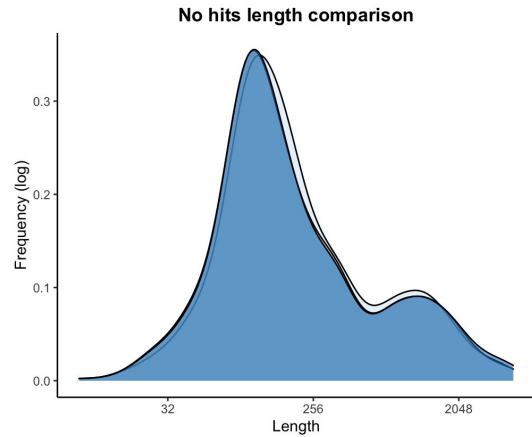74,3%

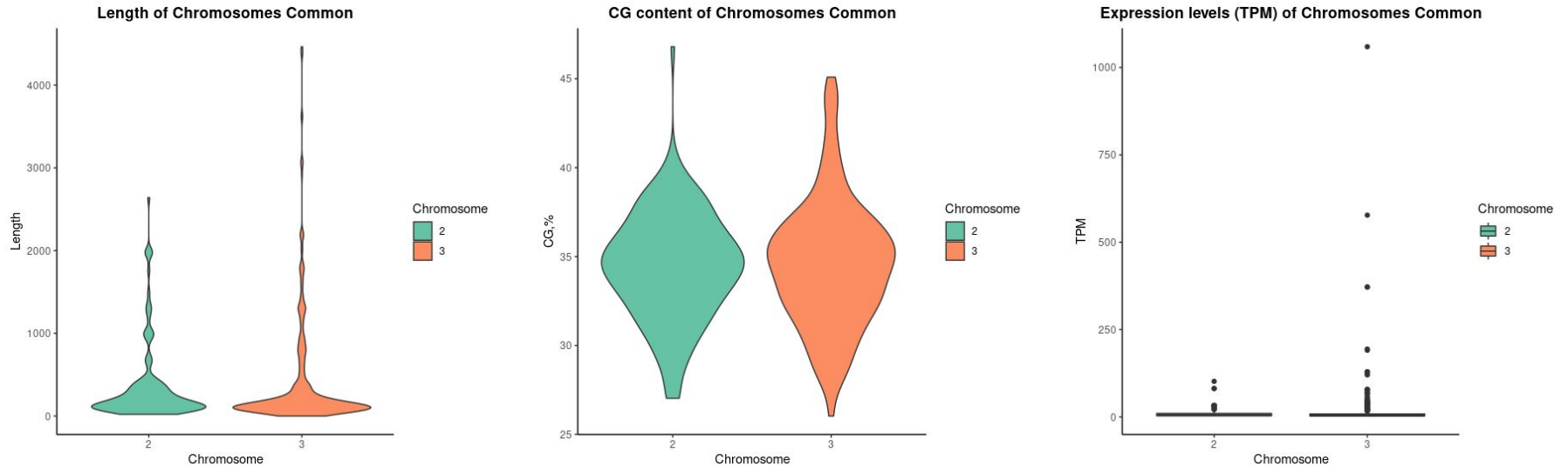516 transcripts didn't align with either proteome of the species

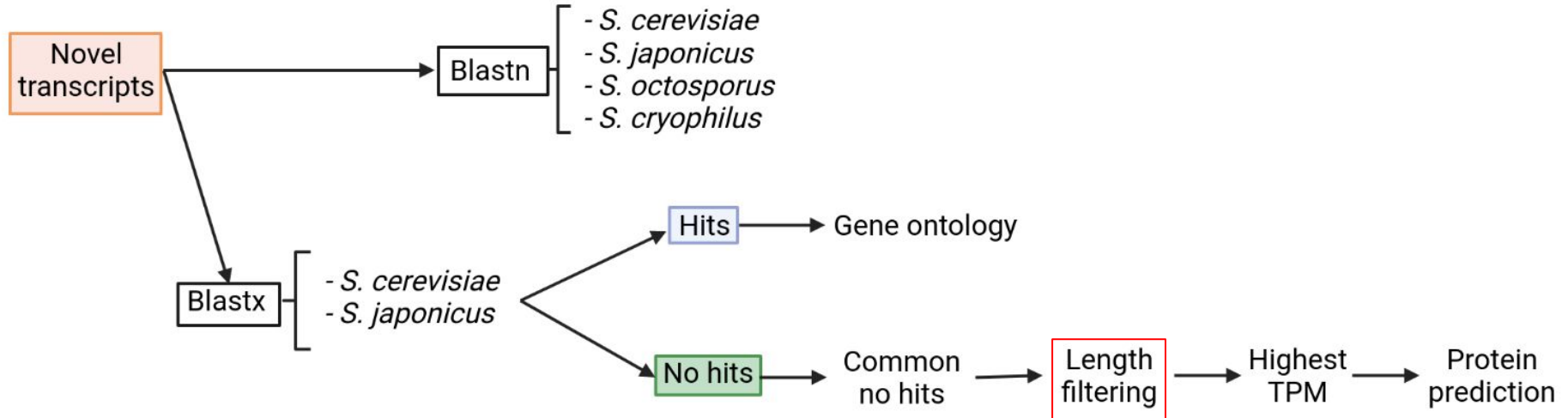# Analysis of novel transcripts – Hit Ontology

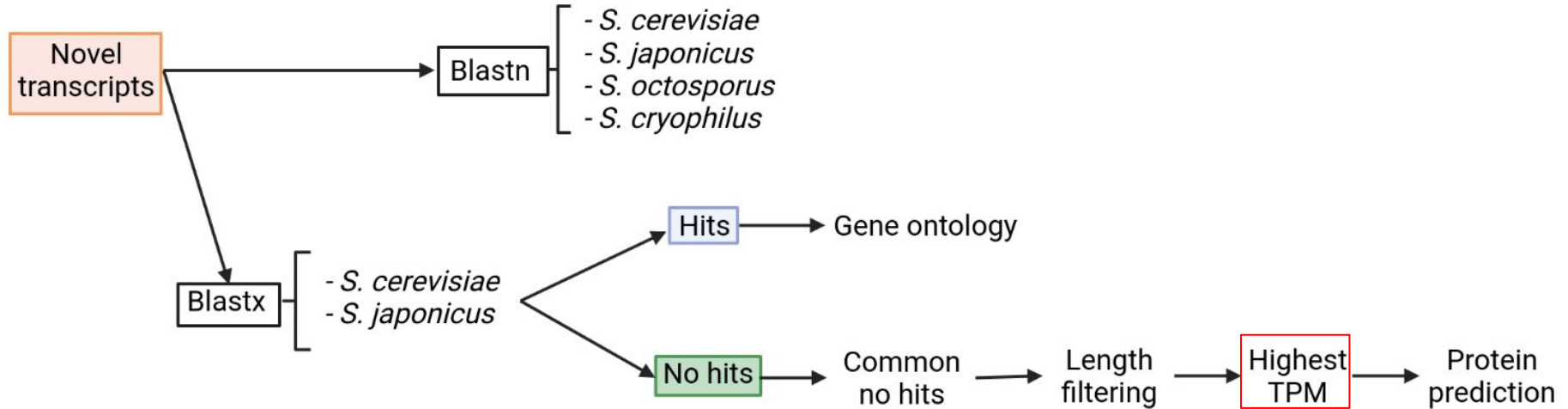# Analysis of novel transcripts – No hits comparison
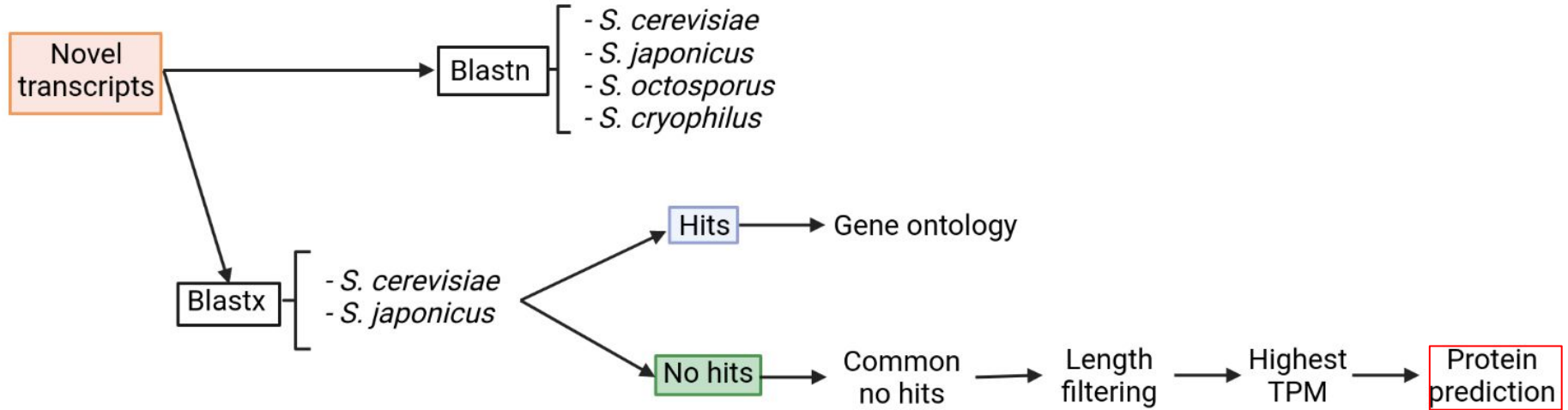
Analysis of novel transcripts - No hits

Analysis of novel transcripts – Length filtering

Analysis of novel transcripts – Highest TPM

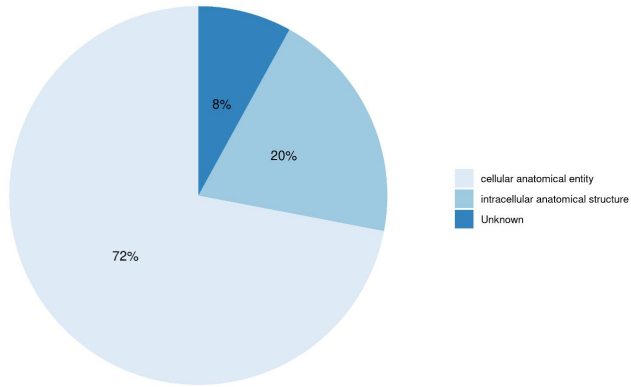# Analysis of novel transcripts – Protein Prediction

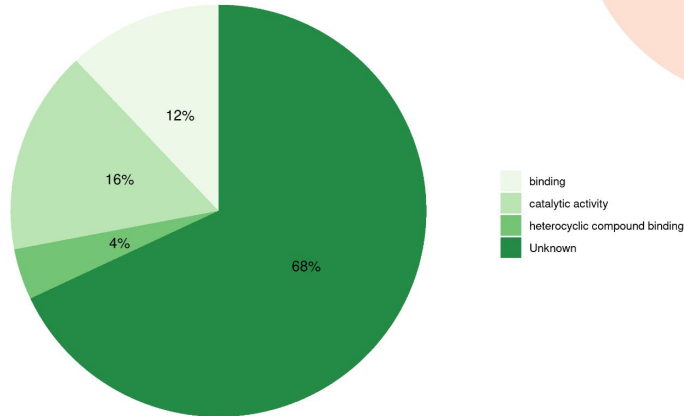# Analysis of novel transcripts – Protein prediction

	Gene	Predicted Protein	BitScore		Gene	Predicted Protein	BitScore
1	NC_003424.3:4523297-4527764(-)	Unknown	N/A	13	NC_003421.2:2412337-2414790(+)	PUR3_YEAST	157
2	NC_003424.3:634751-636959(+)	MU161_SCHPO	470	14	NC_003423.3:3617962-3619663(+)	TAD2_SCHPO	144
3	NC_003424.3:5015958-5018021(-)	ANM1_SCHPO	184	15	NC_003423.3:478916-483814(-)	ALP6_SCHPO	1200
4	NC_003424.3:3617070-3623784(-)	SUB2_SCHPO	855	16	NC_003424.3:2057514-2059140(+)	Unknown	N/A
5	NC_003423.3:3326688-3329592(+)	YBX1_SCHPO	835	17	NC_003424.3:345623-347672(-)	Unknown	N/A
6	NC_003423.3:1714944-1715583(-)	Unknown	N/A	18	NC_003424.3:4523297-4527764(-)	Unknown	N/A
7	NC_003423.3:4220586-4222599(-)	SCS22_SCHPO	48.9	19	NC_003423.3:1156782-1158424(+)	PUR7_SCHPO	605
8	NC_003423.3:1775168-1777010(-)	SCE3_SCHPO	563	20	NC_003424.3:925110-926289(-)	COQ4_SCHPO	253
9	NC_003424.3:1283089-1286338(+)	OST3_SCHPO	138	21	NC_003424.3:3837568-3838456(+)	YAY6_SCHPO	89.4
10	NC_003423.3:1307691-1308797(+)	Unknown	N/A	22	NC_003424.3:308137-309938(-)	GPI18_SCHPO	843
11	NC_003423.3:1455518-1456130(+)	Unknown	N/A	23	NC_003423.3:969007-970357(-)	YN84_SCHPO	276
12	NC_003423.3:2932932-2934291(+)	YPT1_SCHPO	150	24	NC_003421.2:1616620-1619316(-)	YCVC_SCHPO	194

Kulmanov M, Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics*, 36(2):422-429 (2020).

Analysis of novel transcripts - Protein prediction

Celular Component Distribution

- cellular anatomical entity
- intracellular anatomical structure
- Unknown

8%
20%
72%

Molecular Function Distribution

- binding
- catalytic activity
- heterocyclic compound binding
- Unknown

12%
16%
4%
68%

Biological Process Distribution

- cellular process
- metabolic process
- Unknown

44%
4%
52%
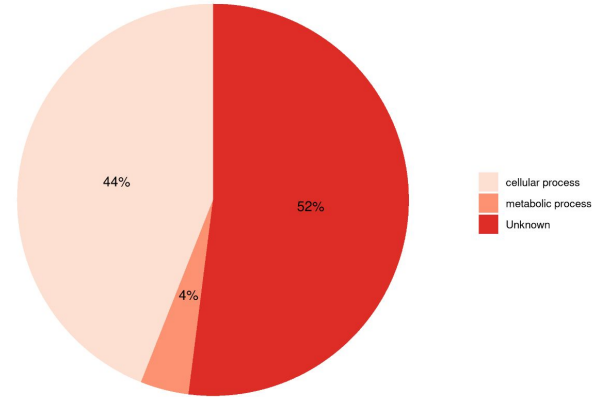
Conclusions

- Known and novel transcripts were compared
    - Novel transcripts were longer.
- Metabolic processes are overrepresented in the hits.
- Protein function was predicted for the best candidates.