Lachancea waltii

RNA-seq analysis

Eric Toro Delgado, Iria Pose Lagoa,
Sara Vega Abellaneda

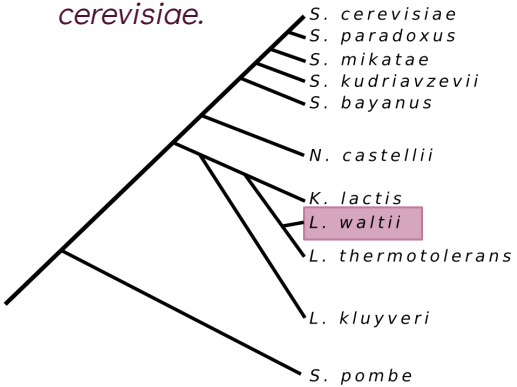
CONTENTS

1. Introduction
2. Analysis strategy
3. Pre-processing
4. Transcripts analysis
5. Novel transcripts characterization
6. Conclusions

1. INTRODUCTION

Species: *Lachancea waltii*

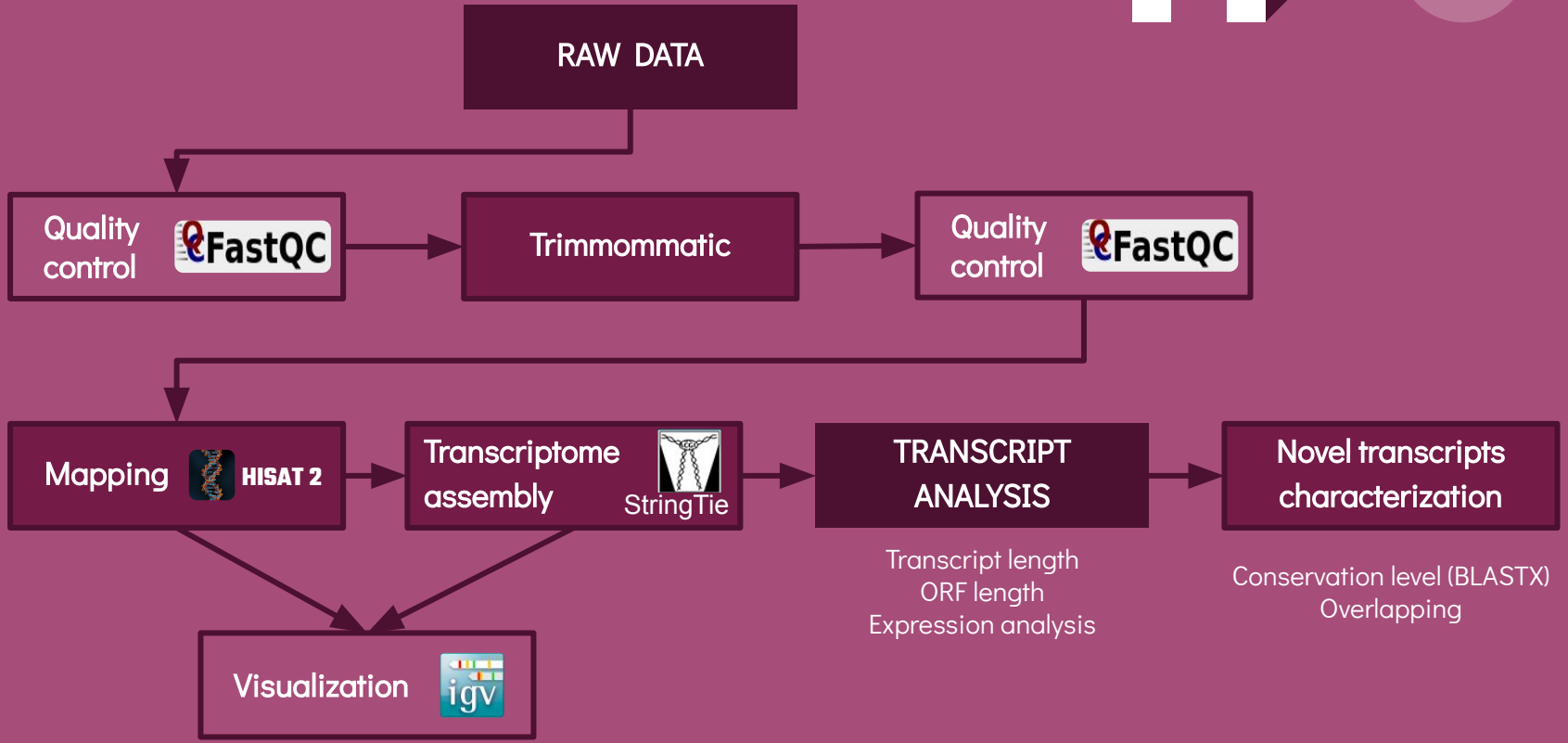
- Protoploid budding yeast species.
- Existing annotation based on predicted ORFs and homology with *S. cerevisiae.*



OBJECTIVES

- Understand the difference between novel and annotated transcripts.
- Understand the patterns in the features related to the origin of novel transcripts, following the framework of *de novo* gene formation proposed by *Carvunis et al. (2012).*

2. ANALYSIS STRATEGY



3. PRE-PROCESSING

RAW DATA

Quality control FastQC

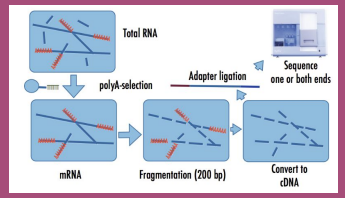
Trimmommatic

Quality control FastQC

Mapping HISAT 2

Transcriptome assembly StringTie

Visualization igv



RAW DATA

ILLUMINA

polyadenylated RNA

RNA-Seq

paired-end reads

forward

24202847 reads

≠

reverse

24728856 reads

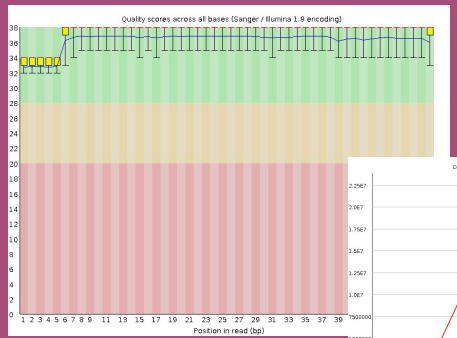
Quality control

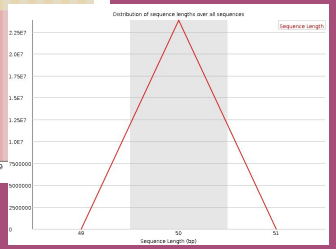
Adapter content ✓

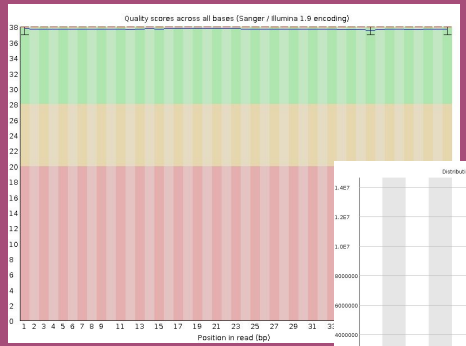
BEFORE TRIMMING

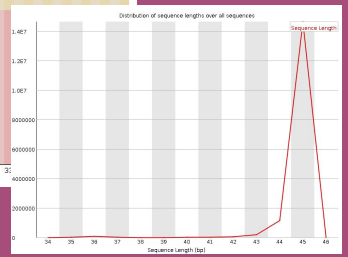
AFTER TRIMMING

READ 1

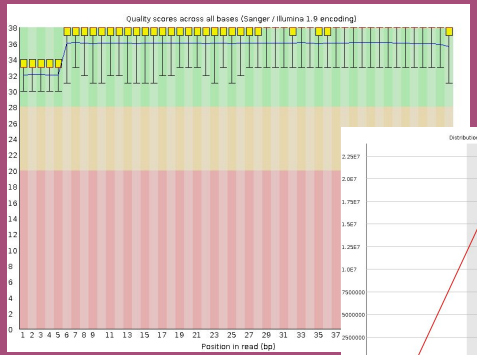


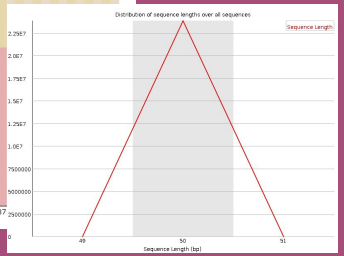


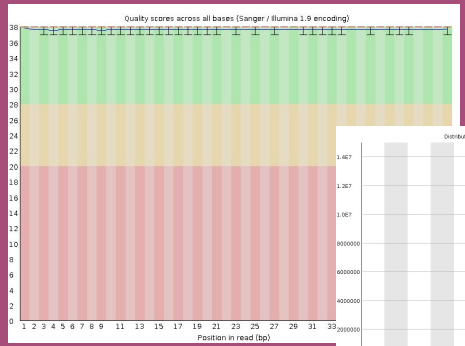


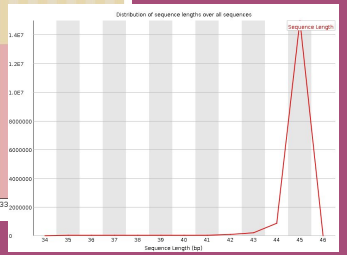


READ 2









forward

16430915 reads

16430915 reads

reverse

Mapping

FASTA files (cleaned reads)

indexes

HISAT 2

96,44% overall alignment rate

reads assigned to a specific location in the genome

SAM file

33707563 reads mapped

SAMTOOLS

BAM file sorted + index

IGV VISUALIZATION

Transcriptome Assembly

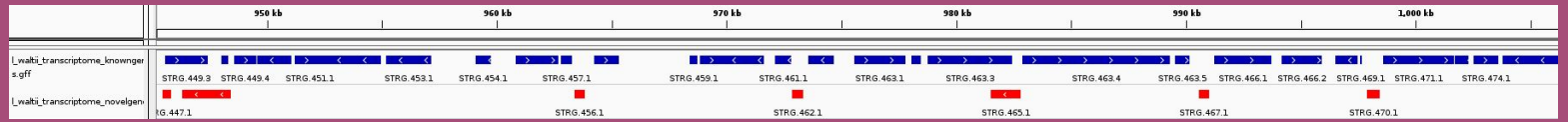
Which transcripts structures are represented by the aligned reads ??

BAM file sorted + index

STRINGTIE

Known transcripts — 5442 transcripts

Novel transcripts — 1557 transcripts



4. TRANSCRIPTS ANALYSIS

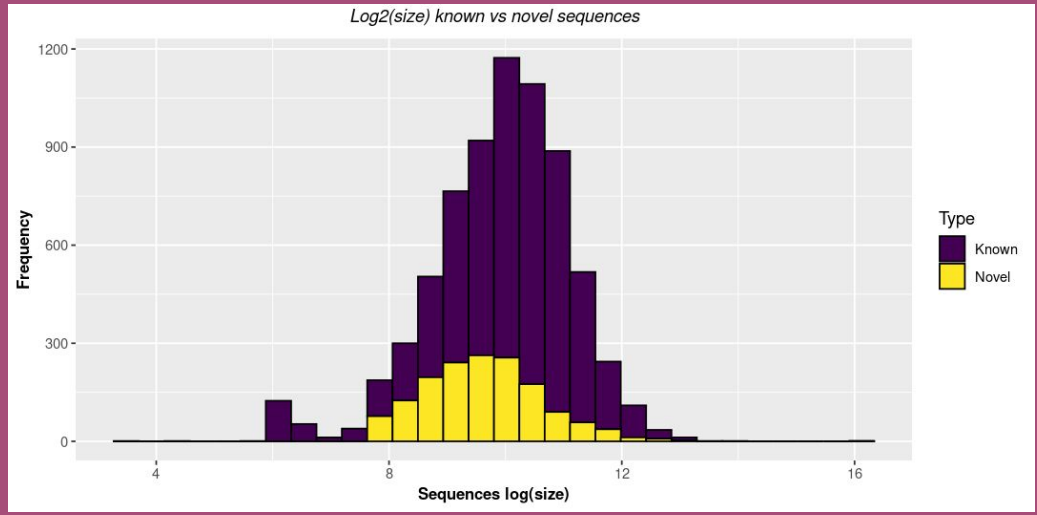
Sequences Length

PERL → sequences lengths → R

	KNOWN	NOVEL
Length Range	12 - 14769	201 - 76379
Mean	1362	1064
Median	1125	767

Mann-Whitney-Wilcoxon test

W=5294271 p-value<2.2e-16

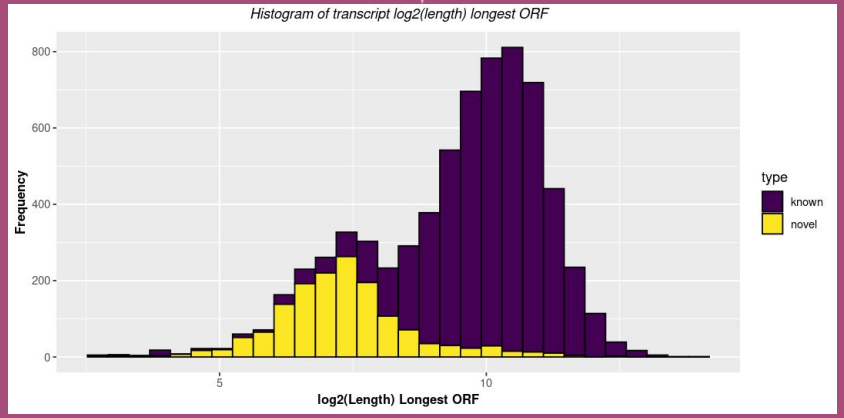


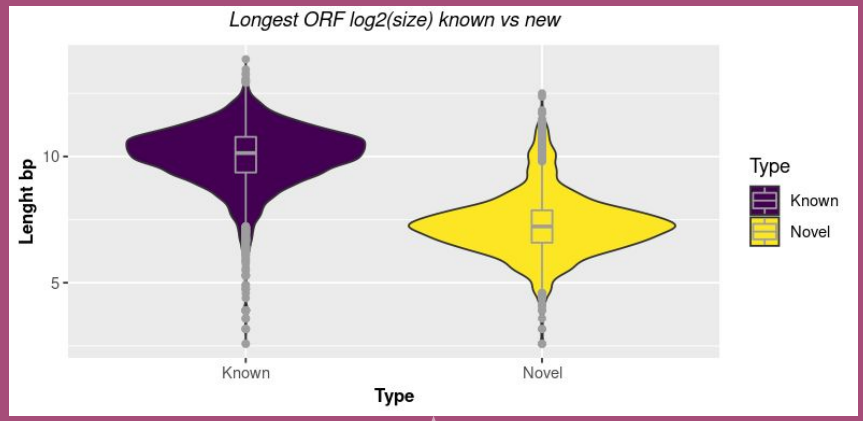
4. TRANSCRIPTS ANALYSIS

ORFs

PERL → ORFs lengths → R

#known > #novel



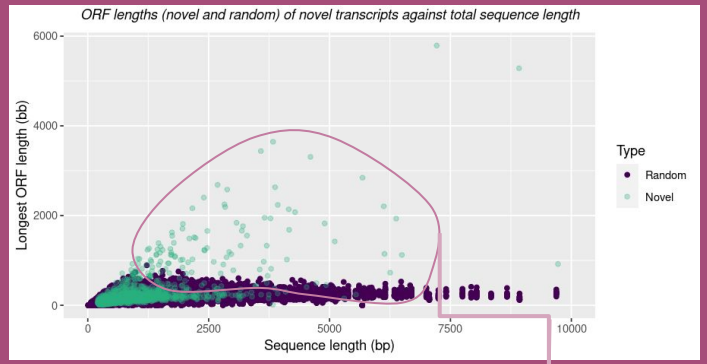


mean: 1318 > 251.8 bp

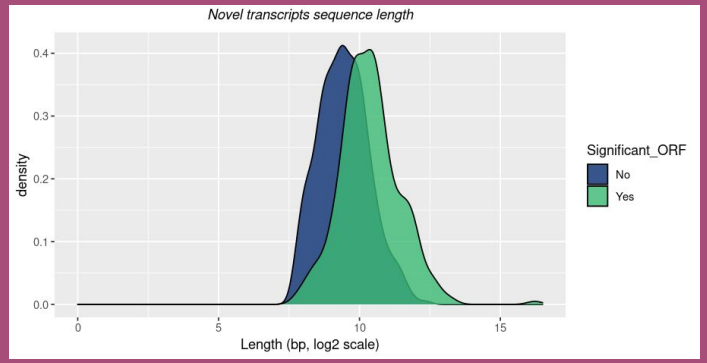
Mann-Whitney-Wilcoxon test

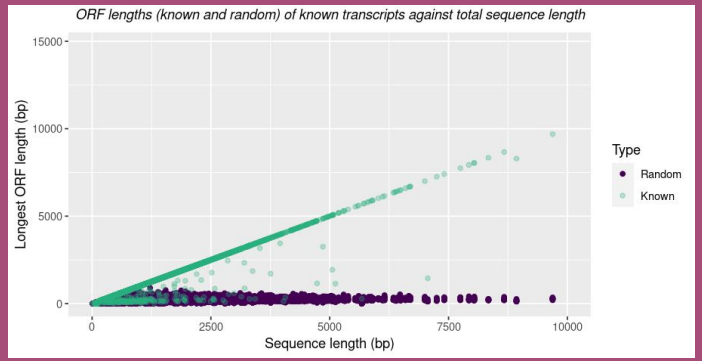
W=7414911 p-value<2.2e-16

PERL → sequences lengths + random sequence length → R



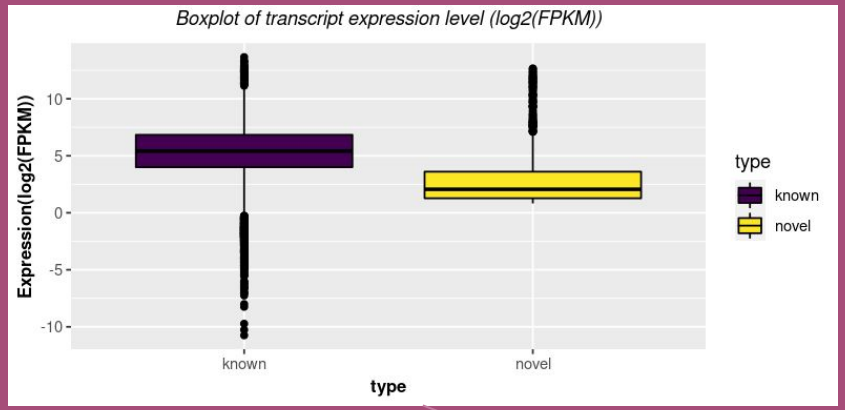
- For random sequences, the length of the ORFs depend of the length of sequences
- 306 out of 1541 (19.86%) novel ORFs of length longer than random ORFs are putative coding sequences
- Sequences annotations come from computational predictions based on ORFs, so that is the reason of the “perfect” correlated line on known transcripts.

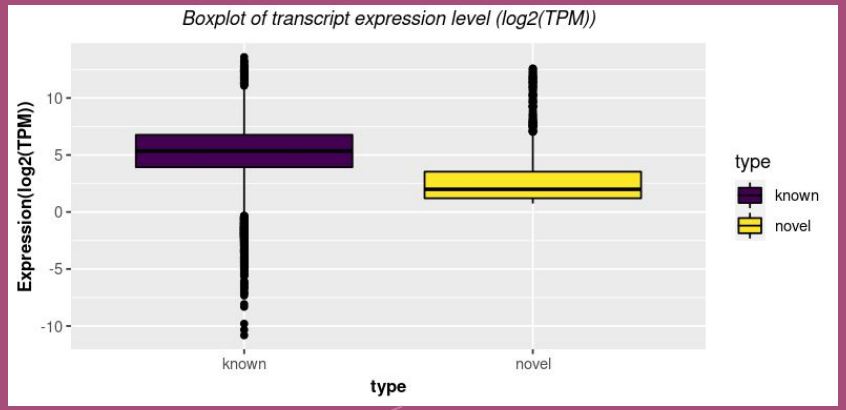




4. TRANSCRIPTS ANALYSIS

Expression





mean known > mean novel

Mann-Whitney-Wilcoxon test

W = 6372692 p-value<2.2e-16

Mann-Whitney-Wilcoxon test

W = 6372692 p-value<2.2e-16

5. Novel transcripts characterization

Homology and conservation analysis

TOOLS

BLASTX

E-value cut-off: 0.00001

SHELL

R

S. cerevisiae

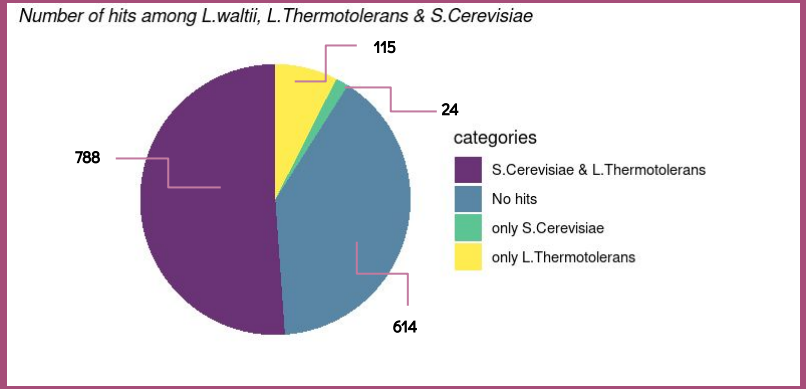
L. thermotolerans

AGE CATEGORY

OLD: hits with *S. cerevisiae*

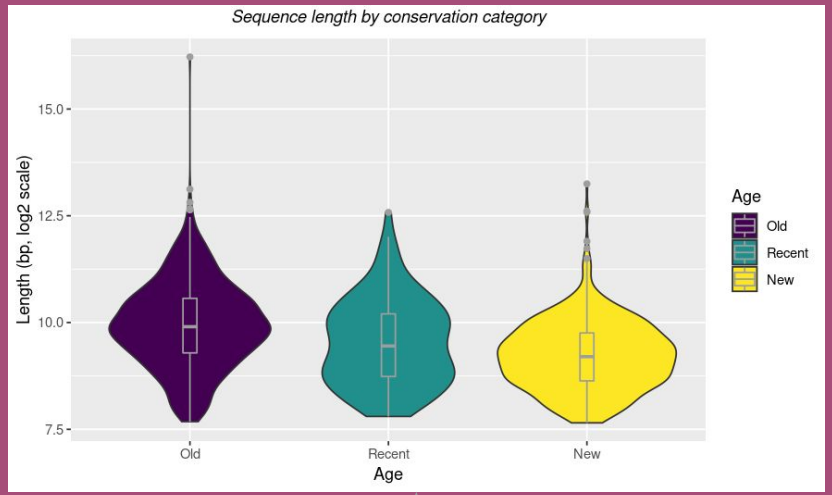
RECENT: hits only with *L. thermotolerans*

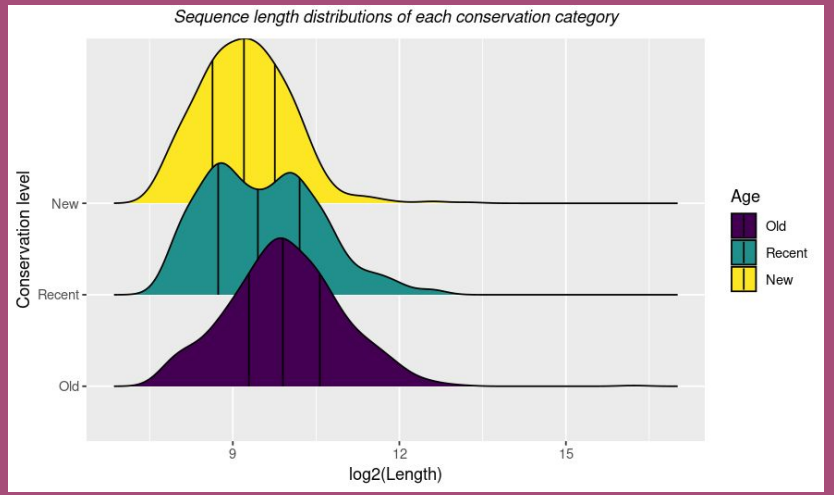
NEW: no hits



specie	hits	no hits	TOTAL
S.Cerevisiae	812	729	1541
L. Thermotolerans	903	638	1541

Sequences Length





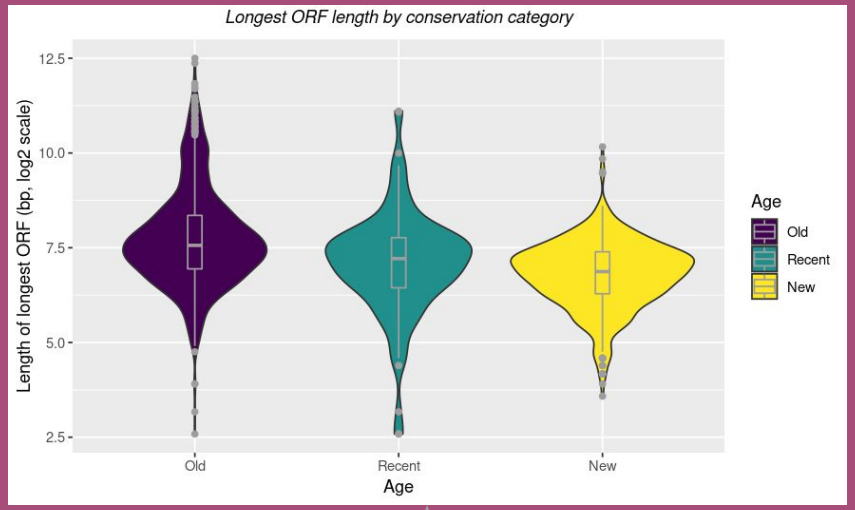
Kruskal-Wallis test

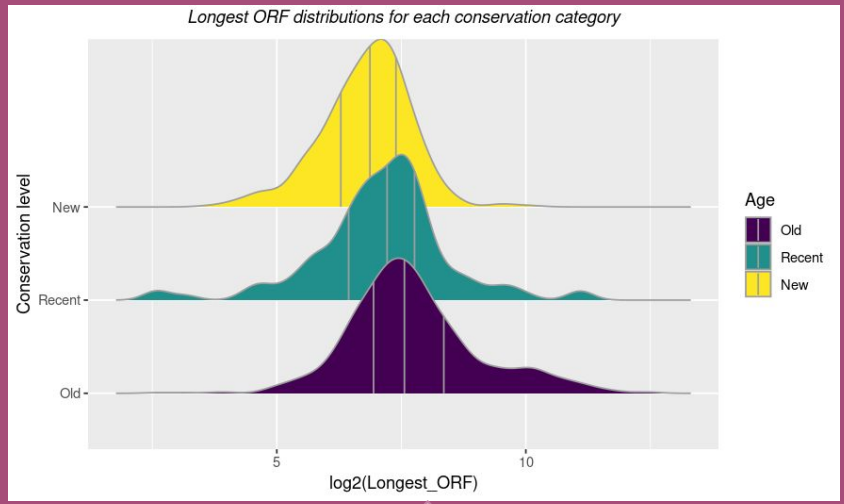
X^2=191.53 p-value<2.2e-16

Mean ± SD

Old:	1340.1 ± 2816.18
Recent	971.5 ± 865.08
New:	717.4 ± 641.33

ORFs





Kruskal-Wallis test

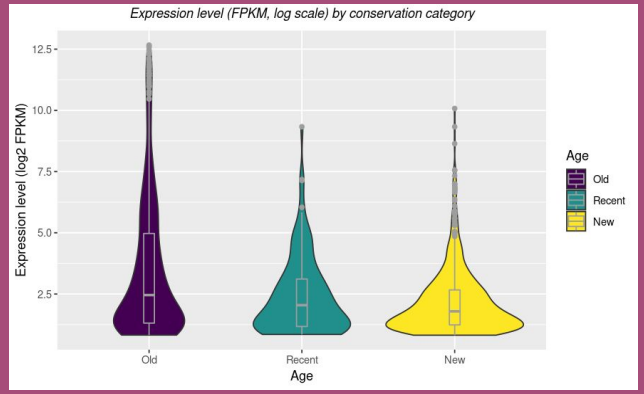
X^2=208.37 p-value<2.2e-16

Mean ± SD

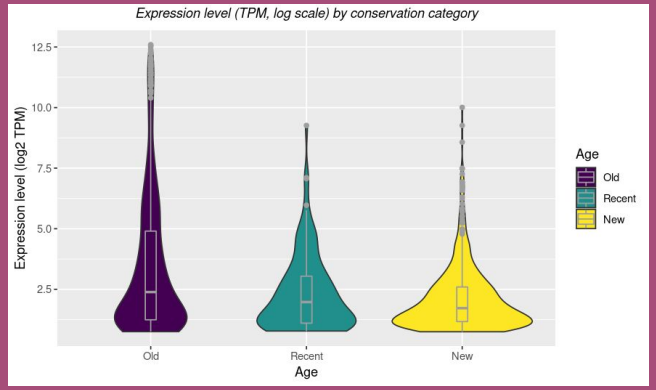
Old:	271.6 ± 522.07
Recent:	248.7 ± 309.17
New:	237.1 ± 95.19

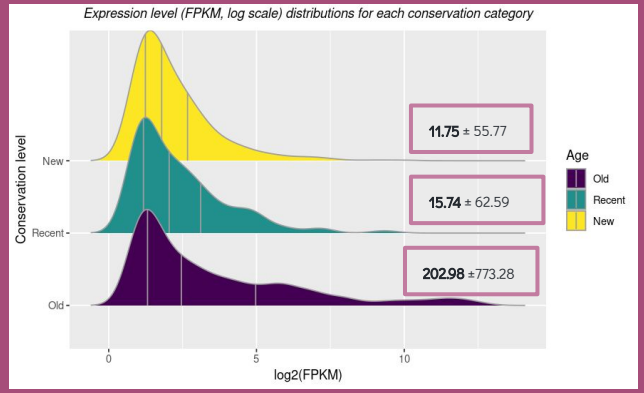
Expression

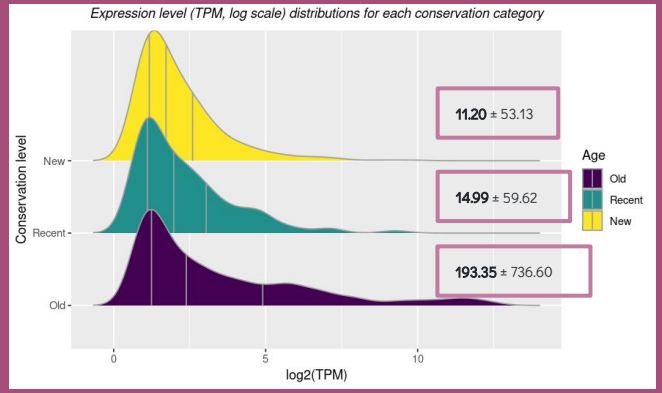
Kruskal-Wallis test
X^2=59.41 p-value=1.255e-13



Kruskal-Wallis test
X^2=59.41 p-value=1.255e-13

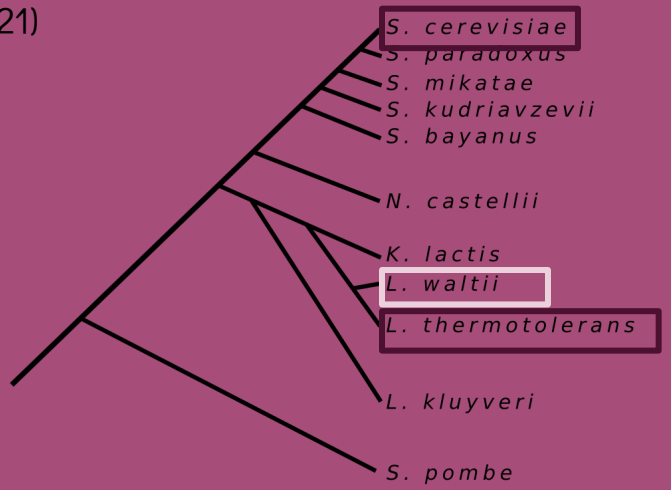






Limitations

- Compared with only 2 species
- Compared with protein DB only → some transcripts may have homologues that have lost function
- Did not check for paralogs nor for synteny (Blevins et al. 2021)
- No Ribo-Seq data → cannot confirm translational activity
- No expression threshold of correct transcript assembly
- Quality of reference genomes and annotations

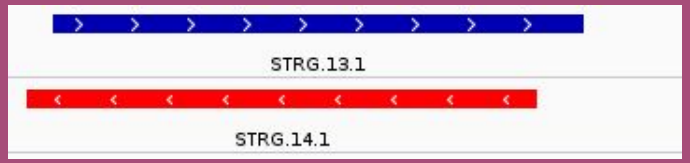


5. Novel transcripts characterization

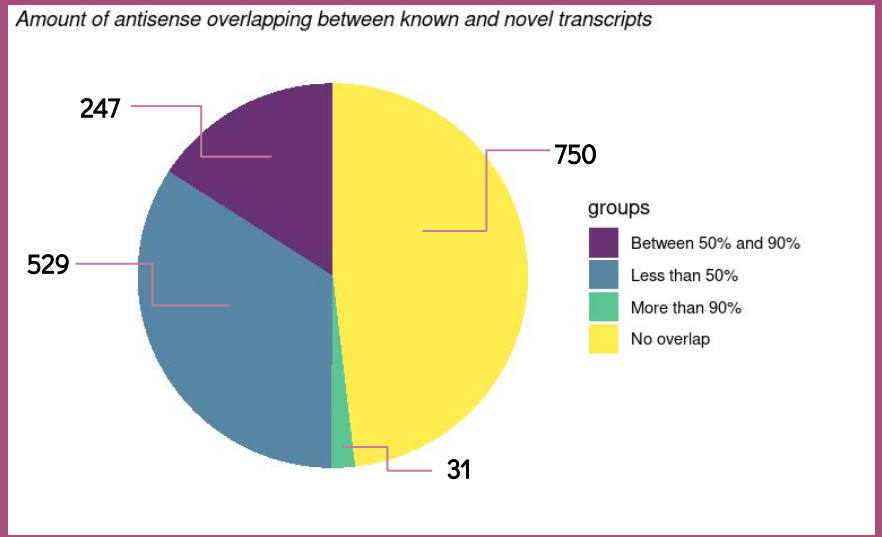
Opposite strand overlap with known genes

TOOLS → SHELL → BEDTOOLS → any overlap / ≥ 50% of reciprocal overlap / ≥ 90% of reciprocal overlap → antisense → R

known
novel



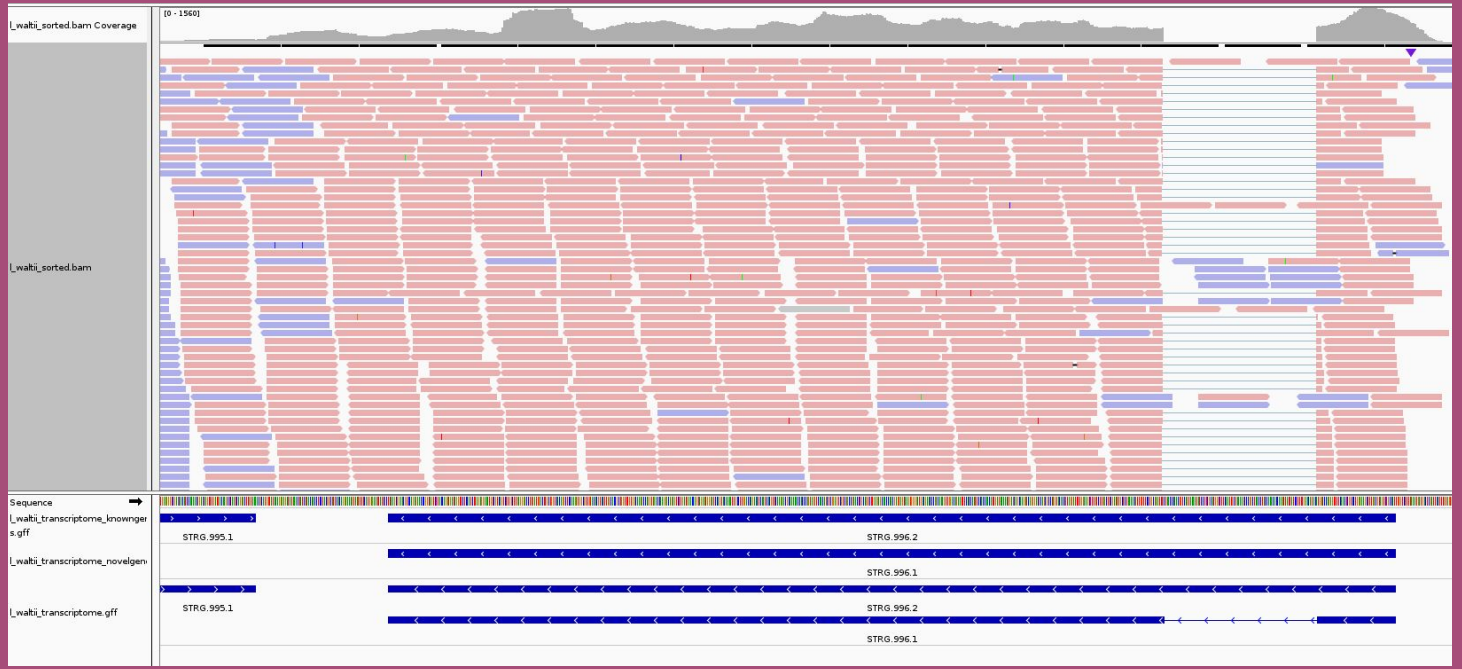
known
novel



5. Novel transcripts characterization

Same strand overlap

Sequences with more or equal to 90% of reciprocal overlap were nearly identical, why ??



6. Conclusions

➔ Identification of 1556 novel transcripts (1541 without rep): 1557/6999 = 22.25% of the total transcriptome catalog for *L. waltii*.

novel vs known

- Novel transcripts are smaller, have lower expression levels and have shorter ORFs than annotated.
- Novel transcripts may not have been previously annotated due to their short ORFs.
- For novel sequences of the same length, ORFs that are longer than ORFs obtained randomly, could be putative coding or truly functional.

novel transcripts features

- Despite limitations of our study, 614 non annotated putative *de novo* transcripts.
- Novel transcripts follow a pattern of increasing total length, ORF length and expression with older dates of origin.
- Many novel transcripts overlapping with already annotated ones in an antisense configuration, but with small overlap.

References

1. Ares M., J., Grate, L., & Pauling, M. H. (1999). A handful of intron-containing genes produces the lion's share of yeast mRNA [2]. In RNA (Vol. 5, Issue 9, pp. 1138–1139). Cambridge University Press. https://doi.org/10.1017/S1355838299991379

2. Blevins, W. R., Carey, L. B., & Albà, M. M. (2019). Transcriptomics data of 11 species of yeast identically grown in rich media and oxidative stress conditions. BMC Research Notes, 12(1), 1–4. https://doi.org/10.1186/s13104-019-4286-0

3. Blevins, W. R., Ruiz-Orera, J., Messeguer, X., Blasco-Moreno, B., Villanueva-Cañas, J. L., Espinar, L., Díez, J., Carey, L. B., & Albà, M. M. (2021). Uncovering de novo gene birth in yeast using deep transcriptomics. Nature Communications, 12(1), 1–13. https://doi.org/10.1038/s41467-021-20911-3

4. Cai, J., Zhao, R., Jiang, H., & Wang, W. (2008). De Novo Origination of a New Protein-Coding Gene in Saccharomyces cerevisiae. Genetics, 179(1), 487–496. https://doi.org/10.1534/GENETICS.107.084491

5. Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., Charloteaux, B., Hidalgo, C. A., Barbette, J., Santhanam, B., Brar, G. A., Weissman, J. S., Regev, A., Thierry-Mieg, N., Cusick, M. E., & Vidal, M. (2012). Proto-genes and de novo gene birth. Nature, 487(7407), 370–374. https://doi.org/10.1038/nature11184

6. Di Rienzi, S. C., Lindstrom, K. C., Lancaster, R., Rolczynski, L., Raghuraman, M. K., & Brewer, B. J. (2011). Genetic, genomic, and molecular tools for studying the protoploid yeast, L. waltii. Yeast, 28(2), 137–151. https://doi.org/10.1002/YEA.1826

7. Kellis, M., Birren, B. W., & Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature 2007 428:6983, 428(6983), 617–624. https://doi.org/10.1038/nature02424

8. Kempken, F. (2013). Alternative splicing in ascomycetes. In Applied Microbiology and Biotechnology (Vol. 97, Issue 10, pp. 4235–4241). Springer. https://doi.org/10.1007/s00253-013-4841-x
Knight, R. D., Freeland, S. J., & Landweber, L. F. (2001). A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biology, 2(4), 1–13. https://doi.org/10.1186/GB-2001-2-4-RESEARCH0010/TABLES/5

9. Ma, C., & Kingsford, C. (2019). Detecting, Categorizing, and Correcting Coverage Anomalies of RNA-Seq Quantification. Cell Systems, 9(6), 589-599.e7. https://doi.org/10.1016/J.CELS.2019.10.005

10. Prat, Y., Fromer, M., Linial, N., & Linial, M. (2009). Codon usage is associated with the evolutionary age of genes in metazoan genomes. BMC Evolutionary Biology, 9(1), 285. https://doi.org/10.1186/1471-2148-9-285

11. Vakirlis, N., Sarilar, V., Drillon, G., Fleiss, A., Agier, N., Meyniel, J. P., Blanpain, L., Carbone, A., Devillers, H., Dubois, K., Gillet-Markowska, A., Graziani, S., Huu-Vang, N., Poirel, M., Reisser, C., Schott, J., Schacherer, J., Lafontaine, I., Llorente, B., … Fischer, G. (2016). Reconstruction of ancestral chromosome architecture and gene repertoire reveals principles of genome evolution in a model yeast genus. Genome Research, 26(7), 918–932. https://doi.org/10.1101/GR.204420.116

Lachancea waltii

RNA-seq analysis

Thank you **very much for your attention**