PGB GROUP PROJECT
2021

Bioinformatic Analysis of K. lactis genome

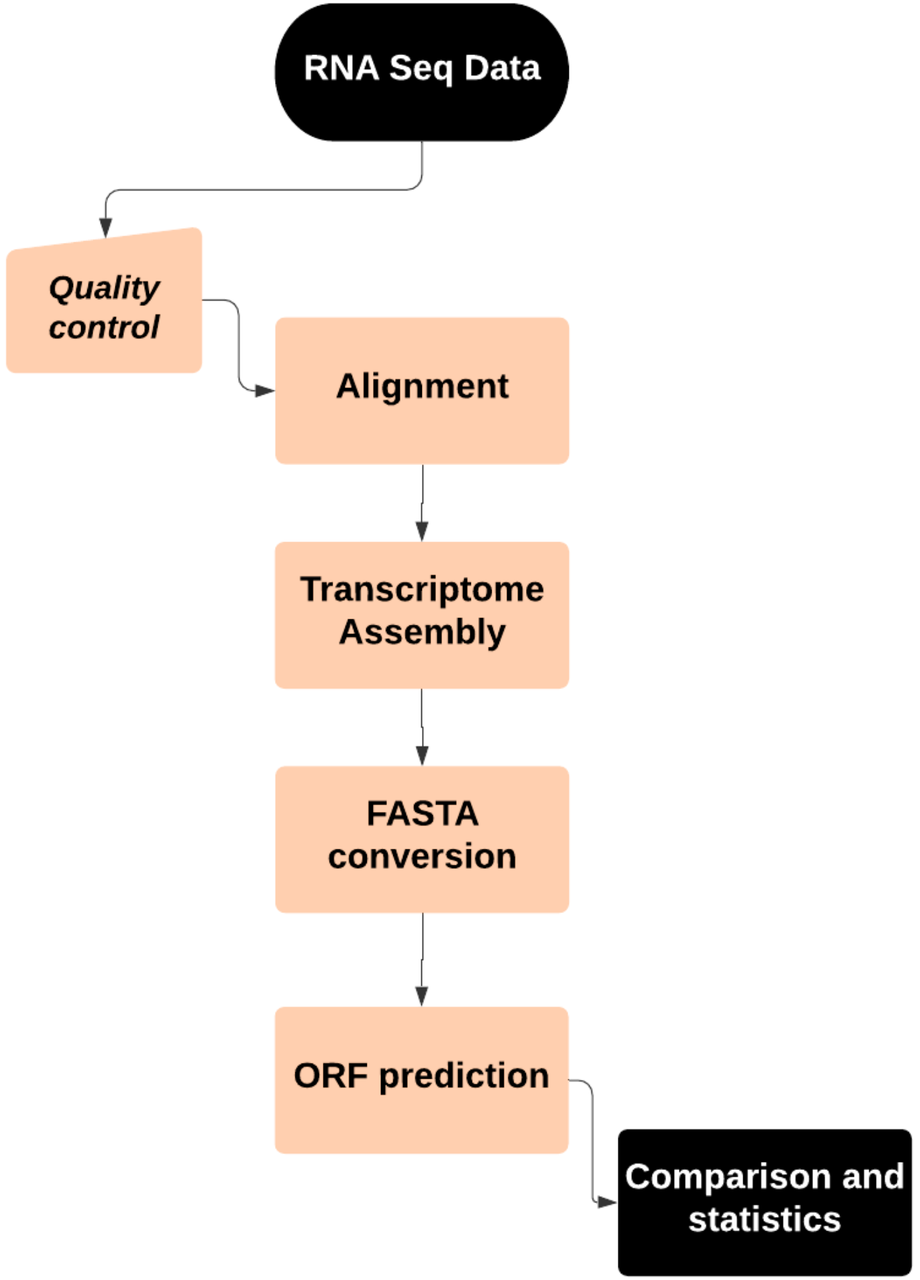
Gerard Romero, Chiara Marchisio, Pau Torren

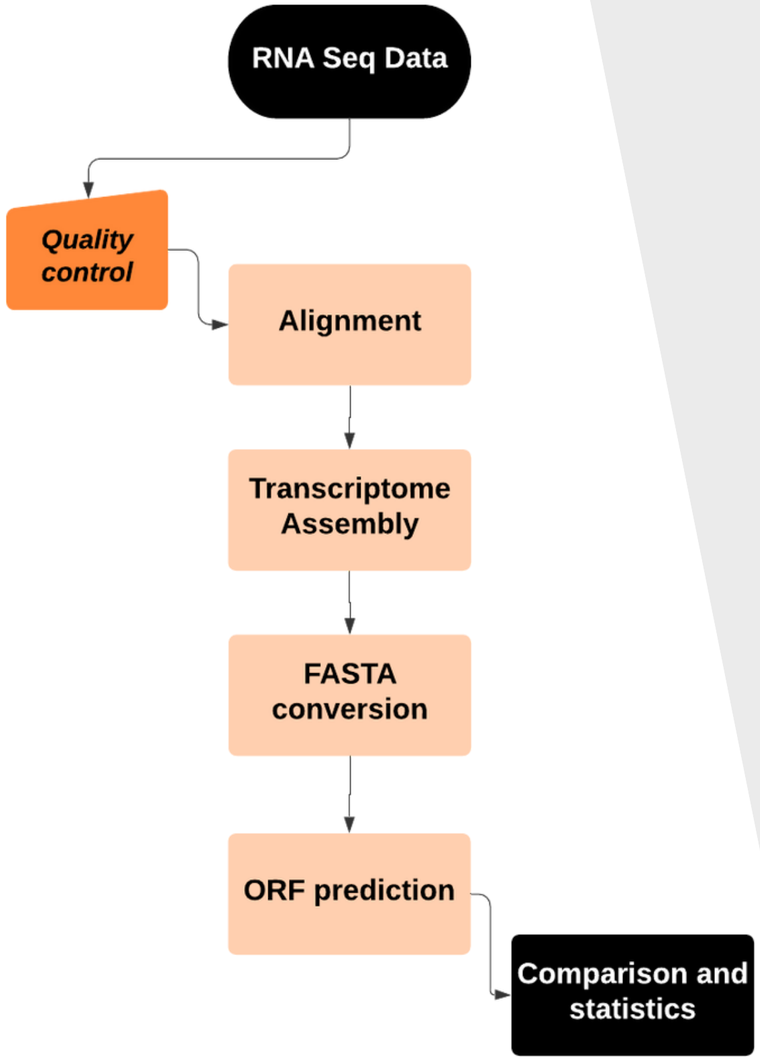
Background

- *Kluyveromyces lactis* is a yeast commonly used in genetic studies and to produce lactic acid
- It is known as a petite-negative yeast which cannot live without its mitochondrial DNA
- Grows by glycolysis and major fermentable sources

(NCBI 2021)

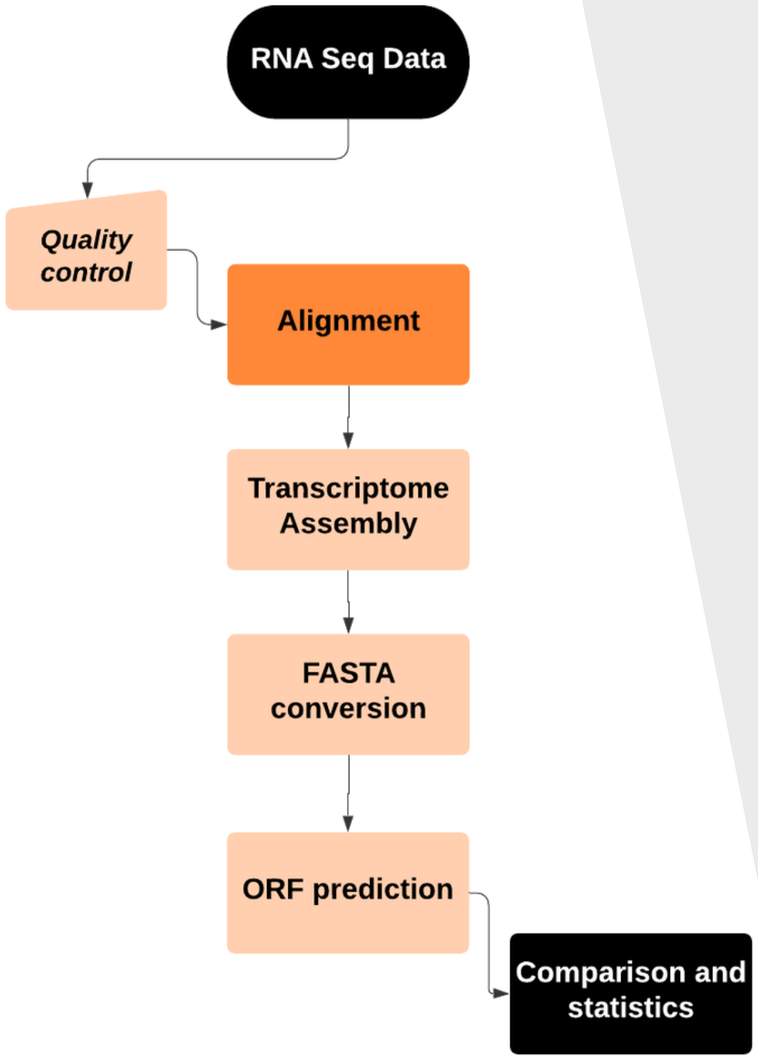
Analysis Pipeline





Quality control

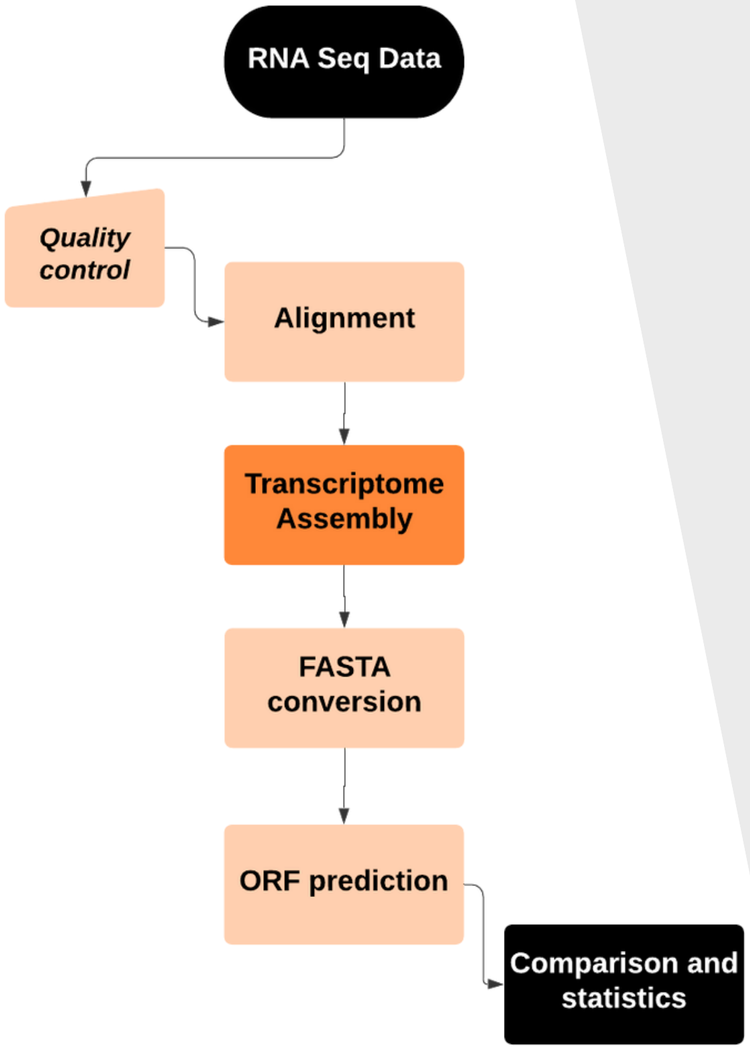
- FastQC was used to filter inaccurate reads
- The number of reads before and after the trimming was determined:
  - Before:
    - 26678609
    - 27321822
  - After:
    - 18841647



Alignment

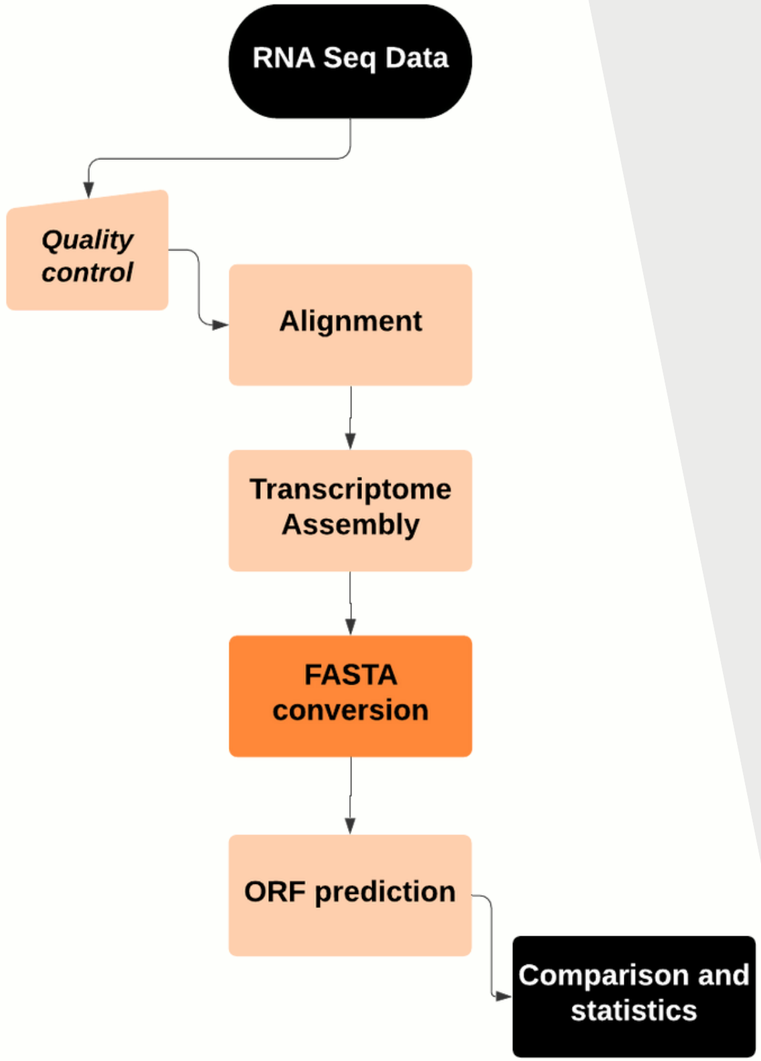
- HISAT2 was used to create a genome index and align the reads to it
- 2 reads were aligned, sense and antisense, here is a summary of results:

Paired Reads	18841647	100%
Aligned concordantly 0 times	314054	1.67%
Aligned concordantly exactly 1 time	18331162	97.29%
Aligned concordantly >1 times	196431	1.04%



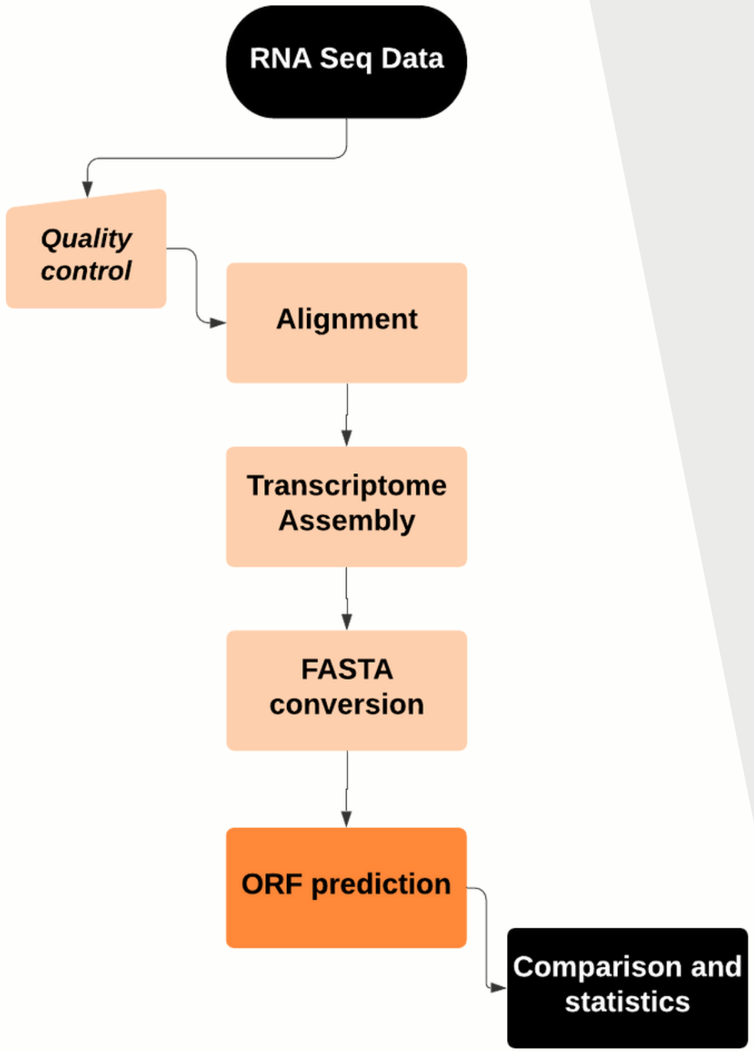
Transcriptome assembly

- Reads were filtered according to whether they had a reference ID or not
- Summary of results:
  - Known transcripts: 5217
  - Novel transcripts: 205



FASTA conversion

- The files obtained were converted to FASTA format to ease further analysis



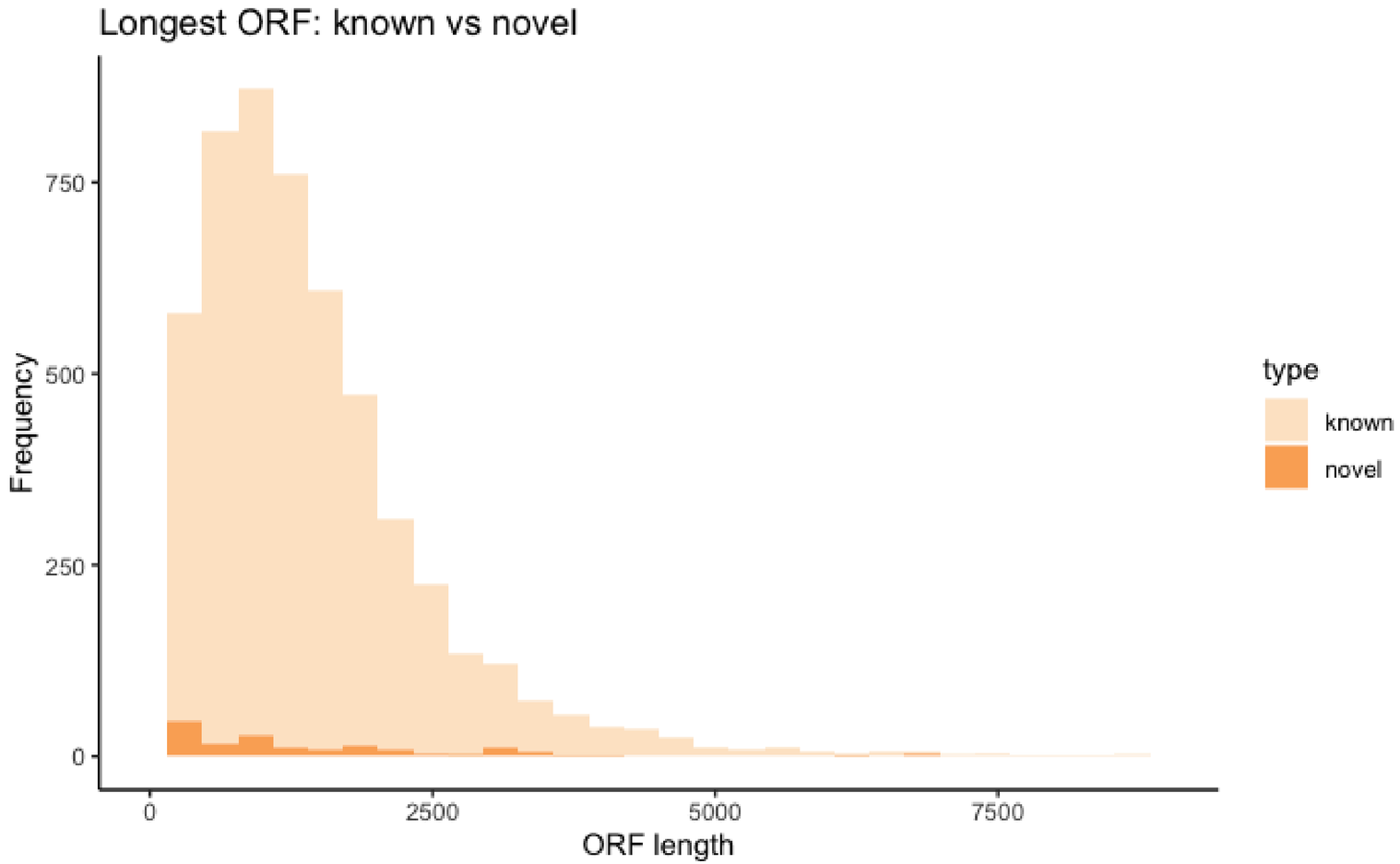
ORF Prediction

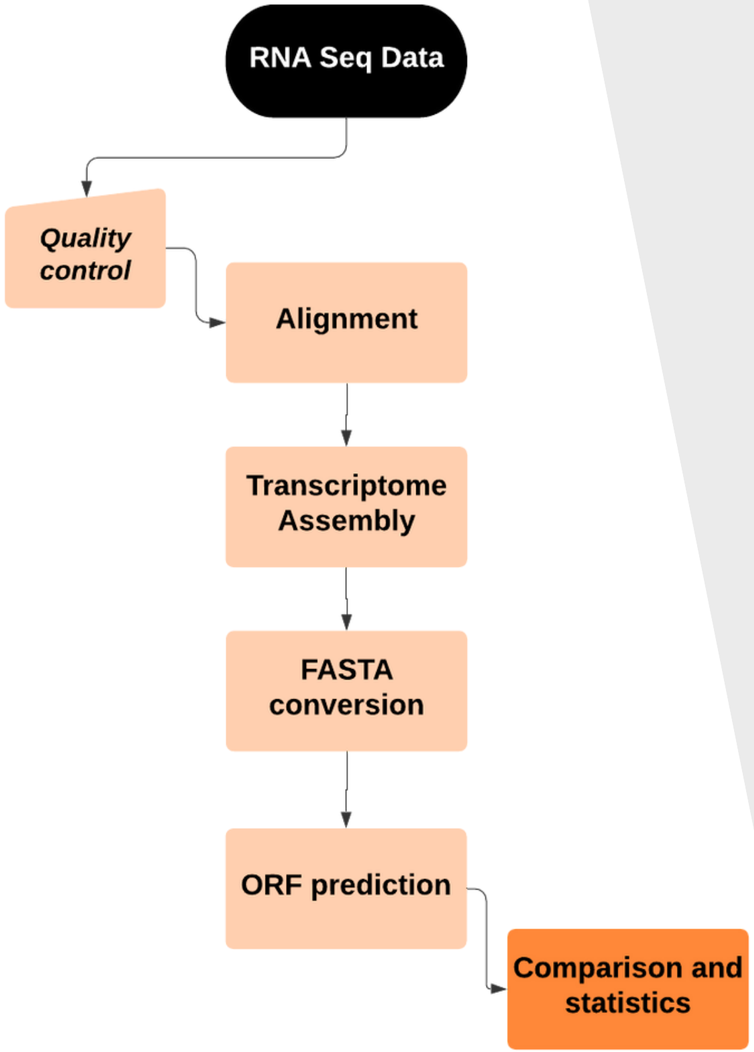
- A given Perl script was used to predict the longest ORF present in the transcriptome
- Random ORFs were also generated and used to compare the distribution with the predicted ORFs
- Wilcoxon test:

Known	p-value <0.001
Novel	p-value <0.001

Longest ORF

- Statistical analysis of known vs novel ORFs was performed.
- Wilcoxon test:
  - p-value <0.001





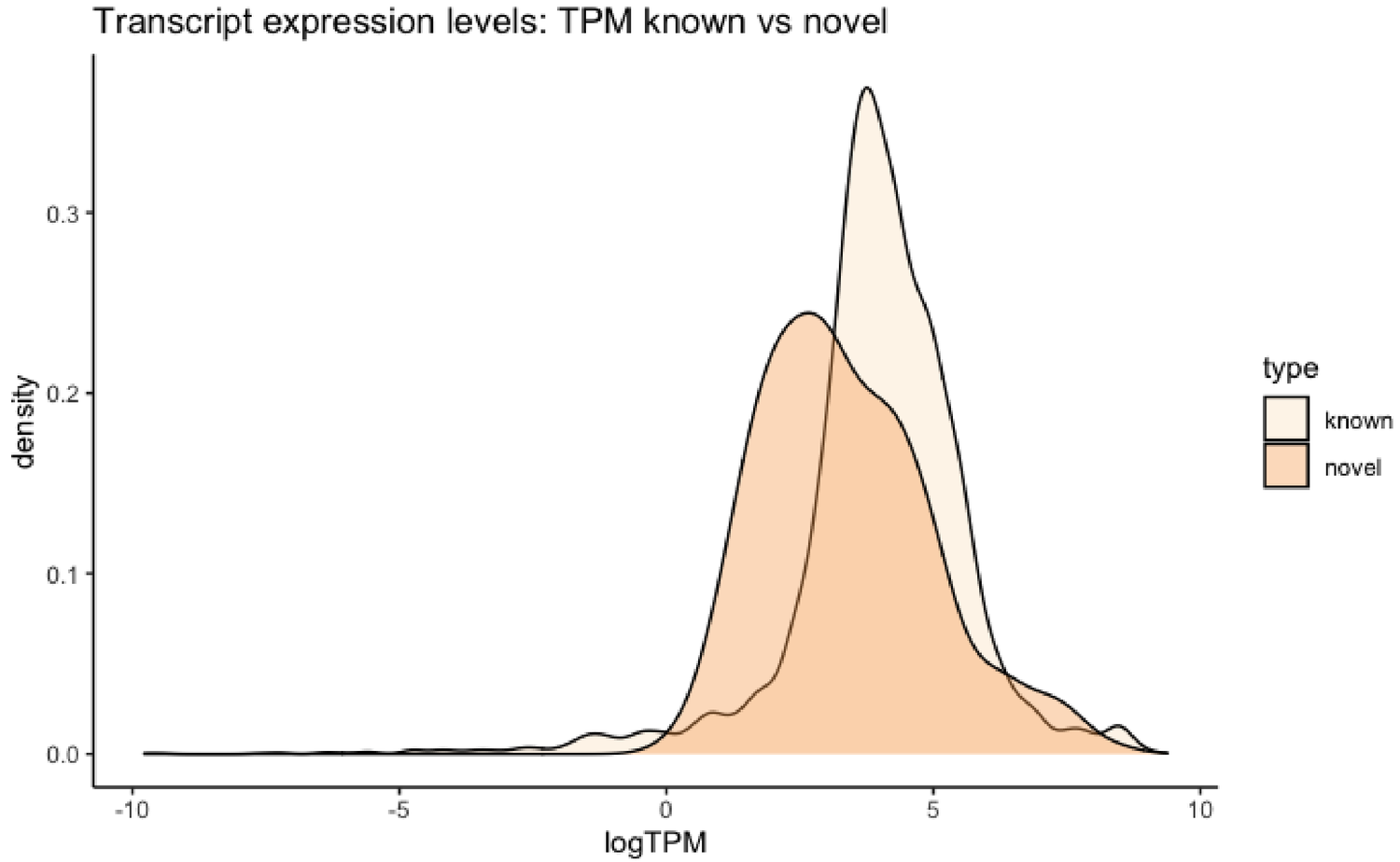
Comparison and Statistics

- Some statystical analysis was performed between novel and known transcripts
- The level of expression of each transcript was determined by their TPMs

TPM

- Statistical analysis of known vs novel TPMs was performed.

- Wilcoxon test:
  - p-value <0.001



Further analysis

01 Transcripts per chromosome

02 Identification of novel transcripts

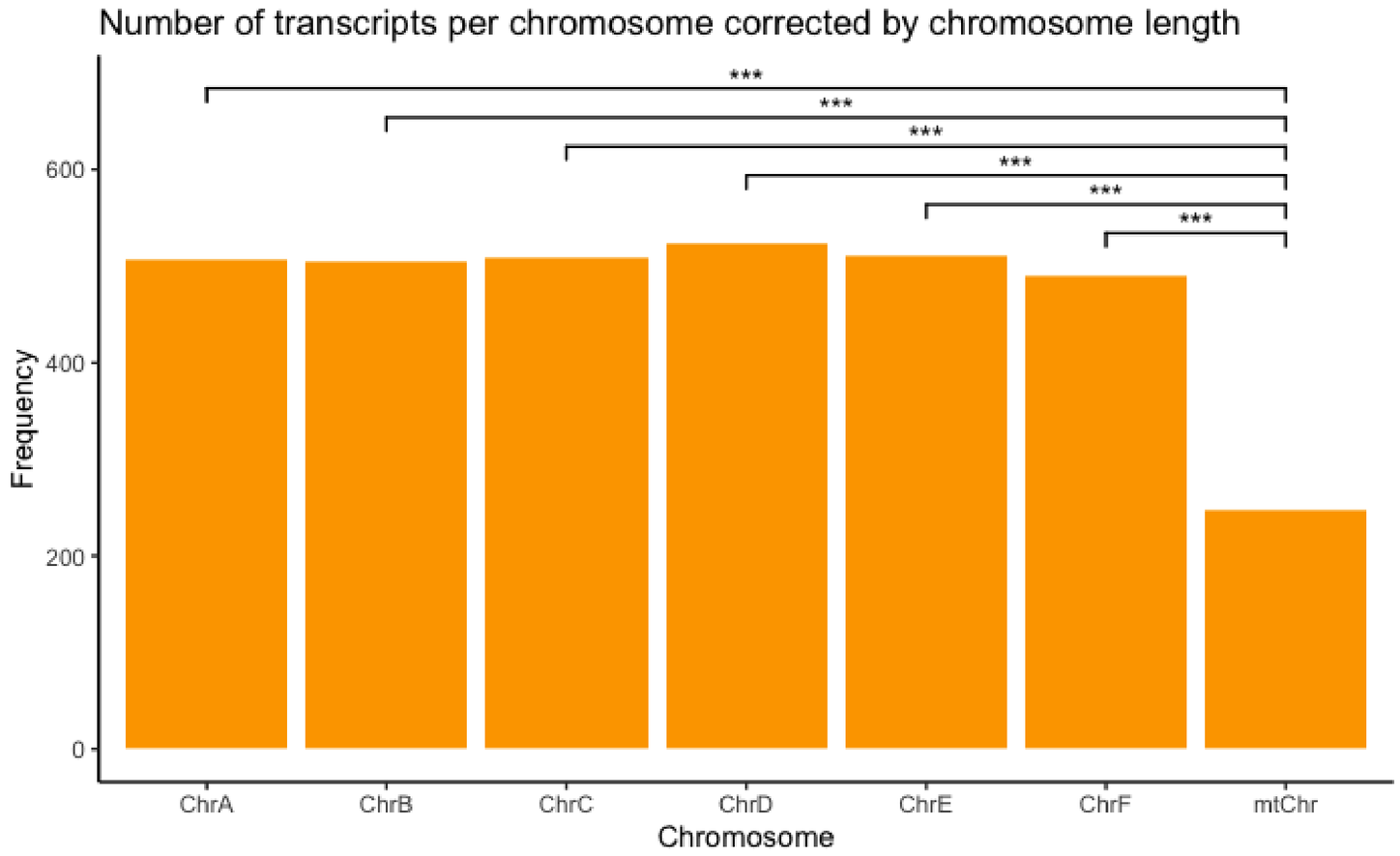
- 2.1 Pre-filtering
- 2.2 BLASTx
- 2.3 Post- BLASTx filtering
- 2.4 Putative protein identification

Transcripts per chromosome

- Statistical analysis of the transcriptome distribution across chromosomes

- Kruskal-Wallis test:
  - p-value <0.001

- Wilcoxon test with Bonferroni correction:
  - 0.05/22 = 0.00227



Identification of novel transcripts

1. Pre-filtering

- Transcript ORF length pre-filtering >500 bp
- TPM filtering >25

2. BLASTx

- The identified transcripts were run against *S. cerevisiae* on BLASTx to identify corresponding proteins.
- Out of 3268 transcripts, 1800 were considered significant

3. Post BLASTx filtering

- From the obtained hits, those with the following parameters were kept:
  - %id and a coverage >70%
  - e-value <0.001
- IGV was then used to discard already annotated transcripts

4. Putative protein identification

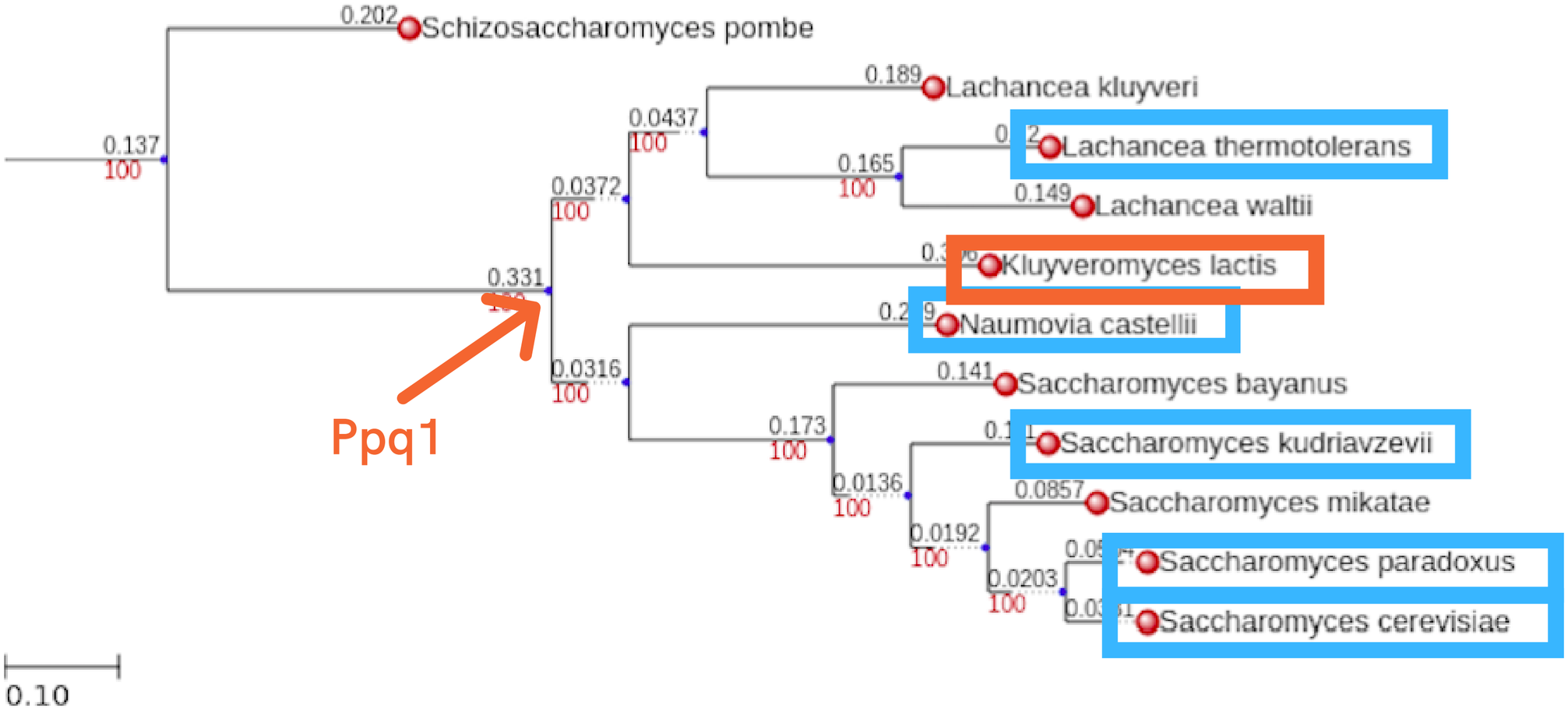
- After the filtering process, two proteins belonging to *S. cerevisiae* were selected:

	S. cerevisiae	*K. lactis*
Name	PHO84	Pho84
Accession ID	NP_013583.1	QEU59767.1

	S. cerevisiae	*K. lactis*
Name	PPQ1	Ppq1
Accession ID	NP_015146.1	QEU61791.1

4. Putative protein identification

- Putative origin of Ppq1:



Discussion

- The GFF file used was from 2016, so the putative proteins we obtained were in fact already annotated (2019) by other research groups (Varela et al. 2019)
- Some transcripts had overlapping issues on IGV

Conclusions

- In summary:
  - significant differences were found between known and novel transcripts in ORF length and level of expression
  - homogenous expression among chromosomes was found with the exception of mitochondrial genome
  - putative proteins were identified and discussed, along with their possible origin
- Possible limitation: the annotation of *K. lactis* genome may have some issues and needs to be checked

References

Geriroso/KLUYVEROMYCES_LACTIS: This repository contains the code used in the project of the kluyveromyces lactis. GitHub. Retrieved November 29, 2021, from https://github.com/Geriroso/Kluyveromyces_lactis.

Kluyveromyces lactis (ID 193) - Genome - NCBI. (2021). Retrieved 25 November 2021, from https://www.ncbi.nlm.nih.gov/genome/?term=kluyveromyces+lactis%5Borgn%5D

Varela, J. A., Puricelli, M., Ortiz-Merino, R. A., Giacomobono, R., Braun-Galleani, S., Wolfe, K. H., & Morrissey, J. P. (2019). Origin of Lactose Fermentation in Kluyveromyces lactis by Interspecies Transfer of a Neo-functionalized Gene Cluster during Domestication. Current biology : CB, 29(24), 4284–4290.e2. https://doi.org/10.1016/j.cub.2019.10.044

National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2021 Nov 29]. Available from: https://www.ncbi.nlm.nih.gov/

Thank you!

Any questions?