Identification of putative novel ncRNAs in *S. kudriavzevii*

by Vicente Ledesma Martín *and* Joan Pallarès Albanell

Table of contents

Universitat Pompeu Fabra Barcelona

MSc Bioinformatics for Health Sciences

Introduction

Comparative Genomics Between *Saccharomyces kudriavzevii* and *S. cerevisiae* Applied to Identify Mechanisms Involved in Adaptation

Laura G. Macías[1,2], Miguel Morard[1,2], Christina Toft[1,2†] and Eladio Barrio[1,2]*

[1] *Departament de Genètica, Universitat de València, Valencia, Spain,* [2] *Departamento de Biotecnología, Instituto de Agroquímica y Tecnología de Alimentos IATA, CSIC, Valencia, Spain*

dominant yeast in most fermentations and it has been widely used as a model eukaryotic organism. Recently, other species of the *Saccharomyces* genus are gaining interest to solve the new challenges that the fermentation industry are facing. One of these species is *S. kudriavzevii*, which exhibits interesting physiological properties compared to *S. cerevisiae*, such as a better adaptation to grow at low temperatures, a higher glycerol synthesis and lower ethanol production. The aim of this study is to understand the molecular basis behind these phenotypic differences of biotechnological interest by using a species-based comparative genomics approach. In this work, we sequenced,

(Macías *et al.*, 2019)

Review

The long non-coding RNA world in yeasts☆

Akira Yamashita [a,b,*], Yuichi Shichino [a], Masayuki Yamamoto [a,b]

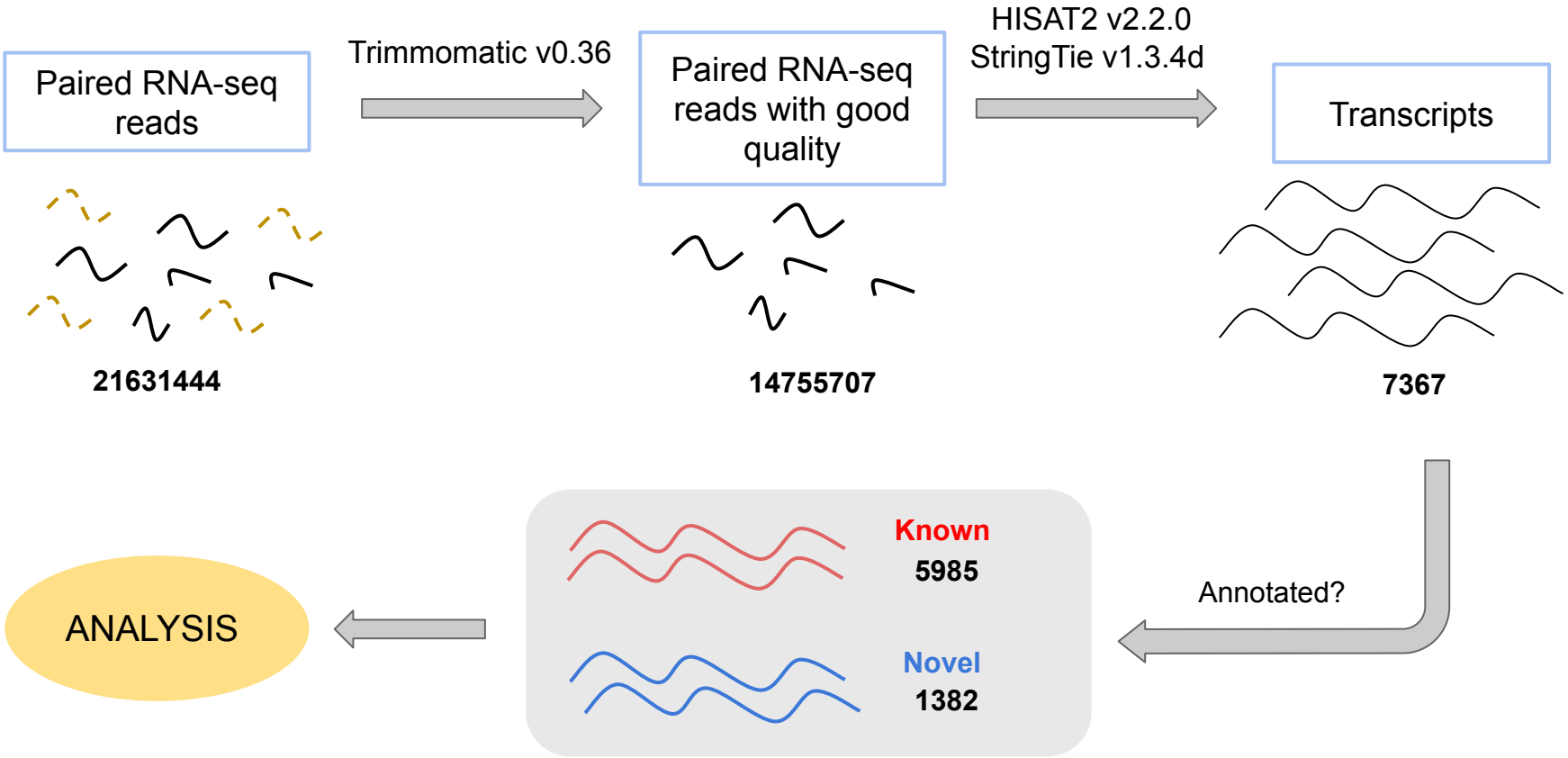
(RNA-Seq) have revealed that most of the eukaryotic genomes are transcribed. RNA polymerase II has been shown to stay outside of coding regions in the yeast and human genomes [1,2]; indeed, at least 75% of the genomes of the budding yeast *Saccharomyces cerevisiae* and fission yeast *Schizosaccharomyces pombe* are transcribed [3–6]. Numerous transcripts are so-called non-coding RNAs, which do not encode a protein. Non-coding RNAs that are more than 200 nucleotides in length are conven-

(Yamashita *et al.*, 2016)

Introduction

Working Hypothesis: Functional ncRNAs are among the novel genes identified in *S. kudriavzevii*.

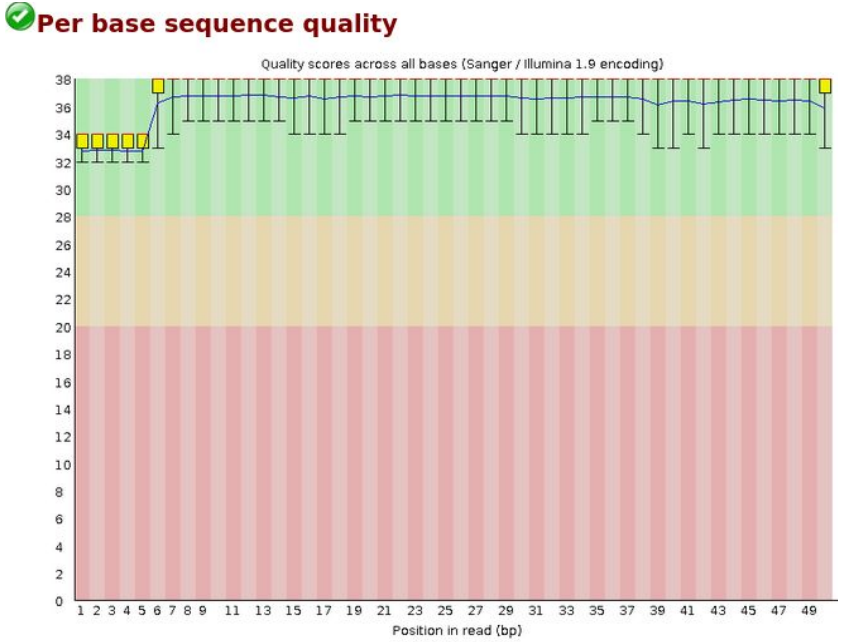
Read Quality and Transcript Assembly



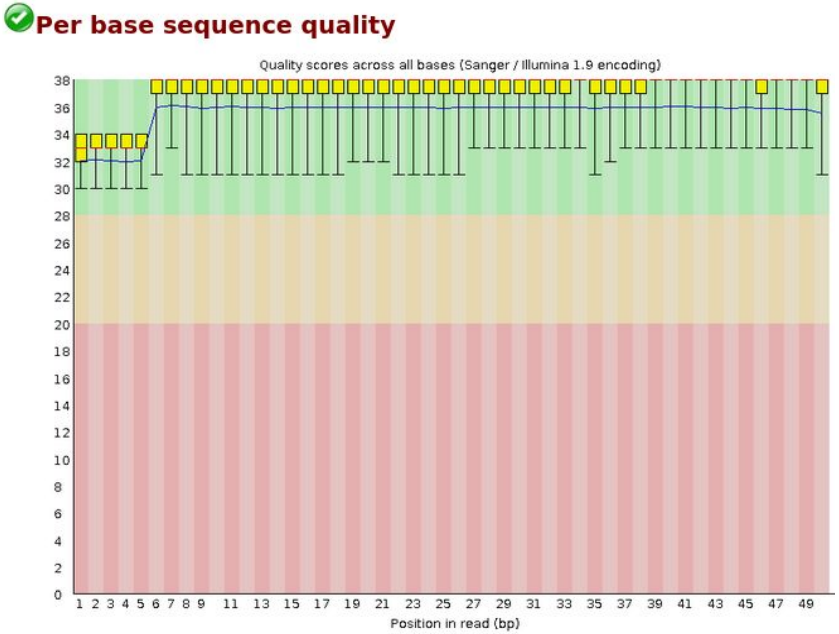
Read Quality

FastQC v0.11.9

Read 1



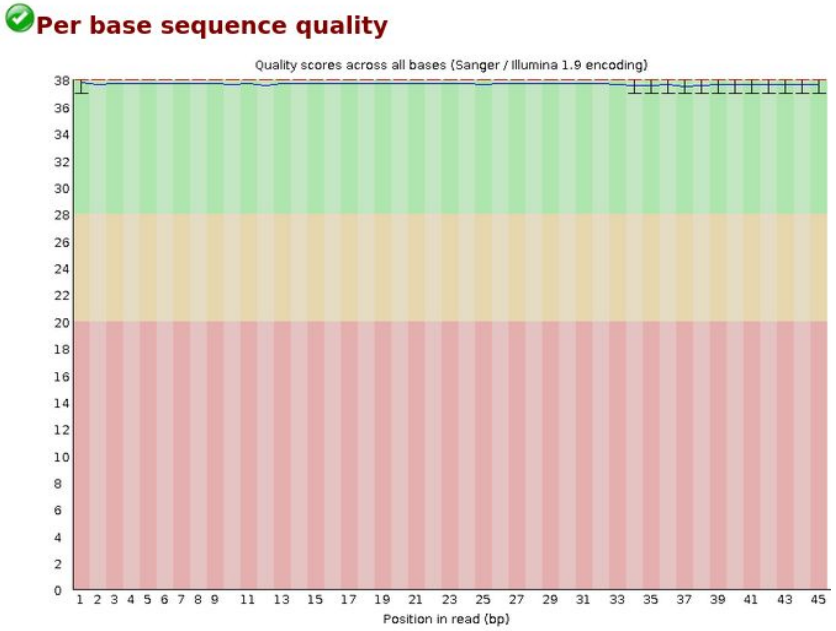
Read 2



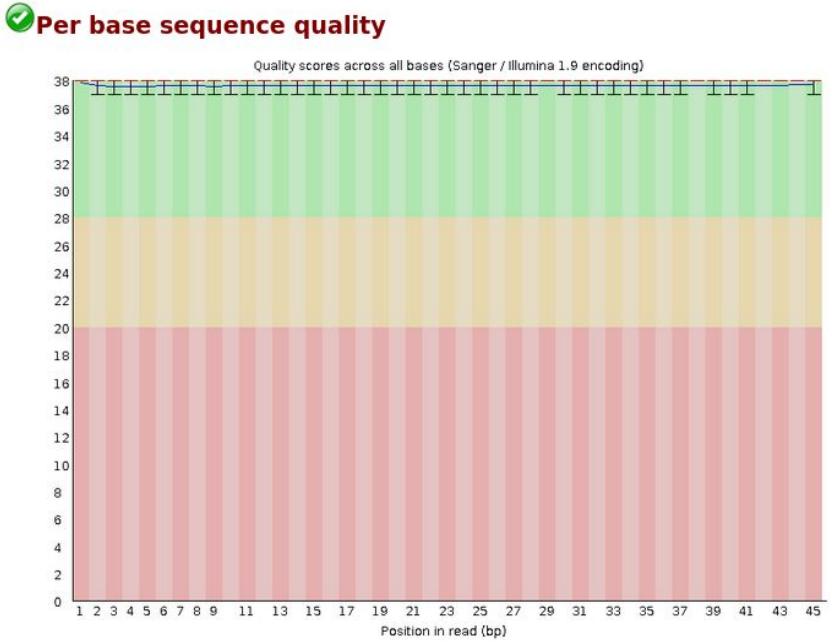
Read Quality

Trimmomatic v0.36 + FastQC v0.11.9

Read 1



Read 2



Transcript Assembly: Visualization

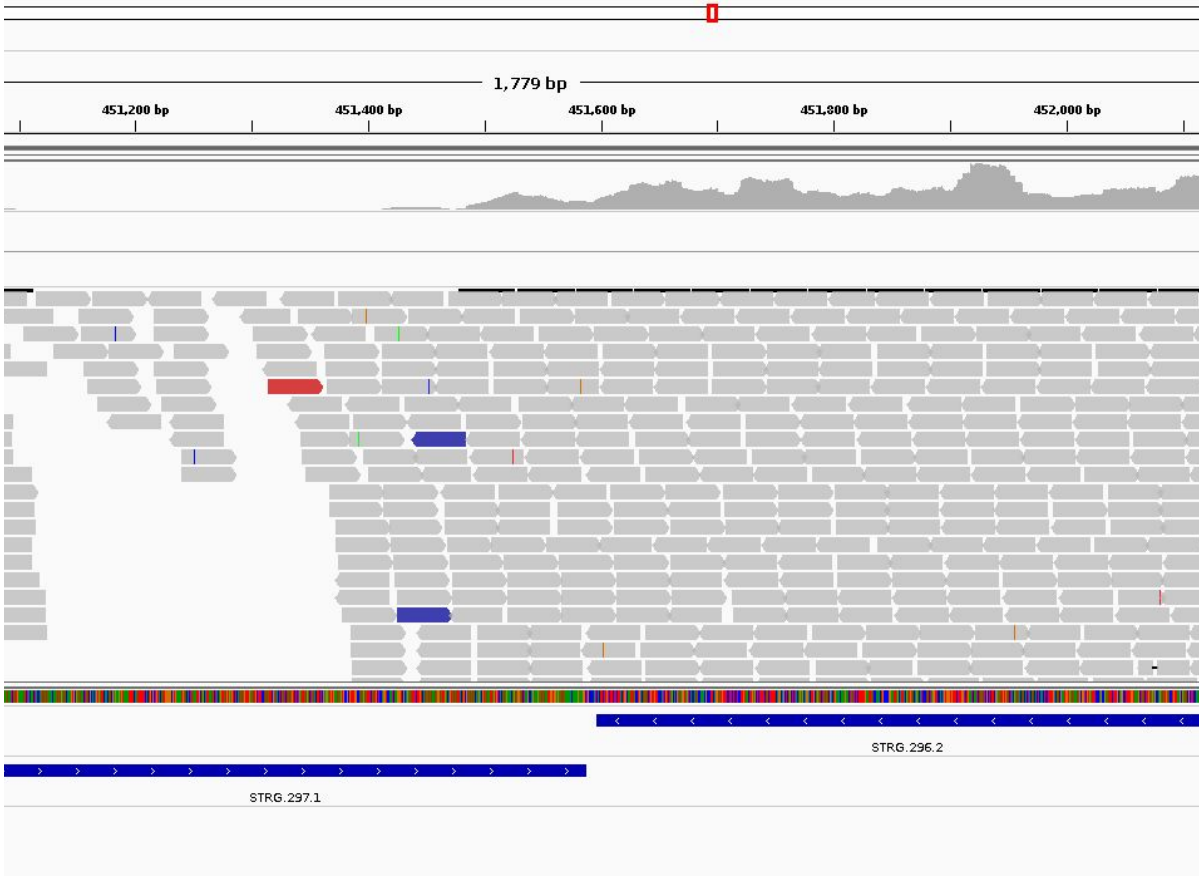
IGV v2.11.4

S. kudriavzevii reference genome

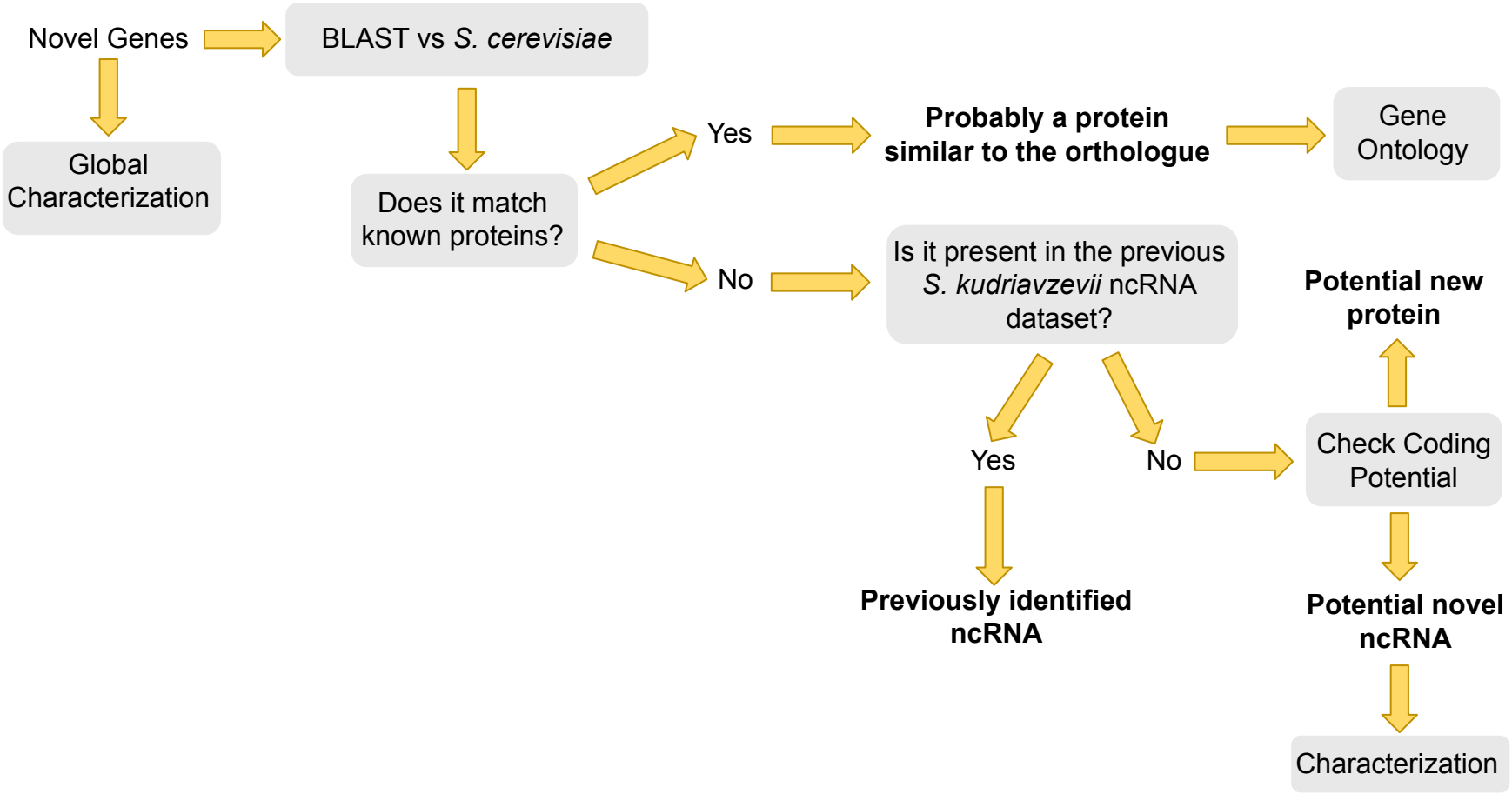
Expression

Known transcripts

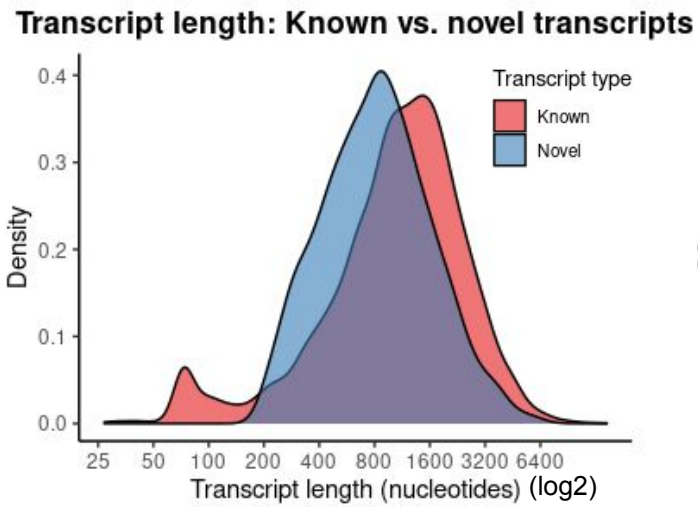
Novel transcripts



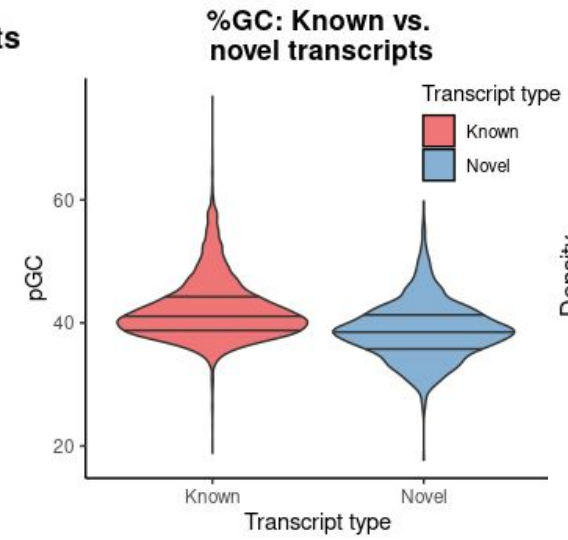
Analysis of Novel Transcripts Workflow



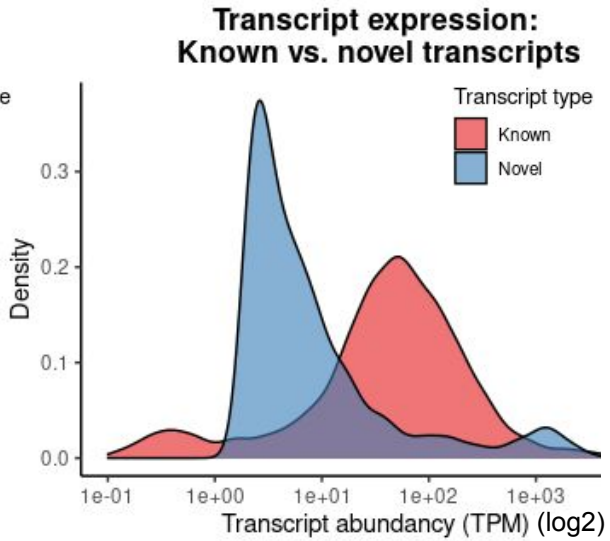
Novel vs. Known Transcripts



Wilcoxon test
W = 5046034, p-value < 2.2e-16

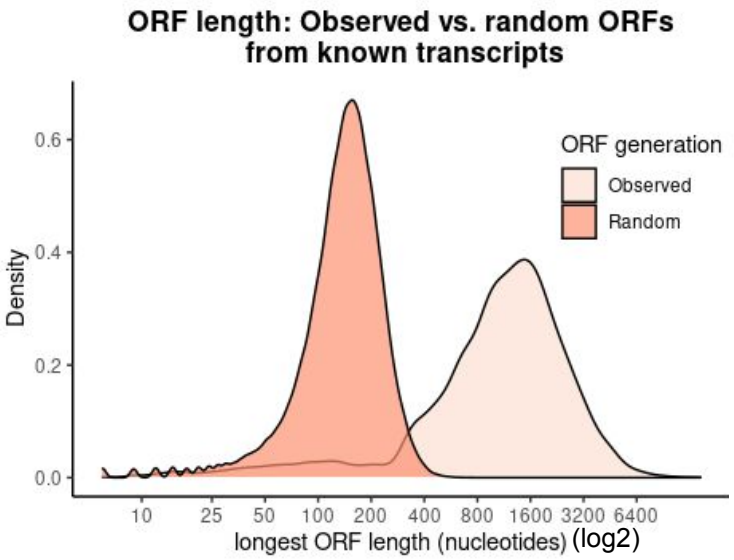


Wilcoxon test
W = 5790808, p-value < 2.2e-16

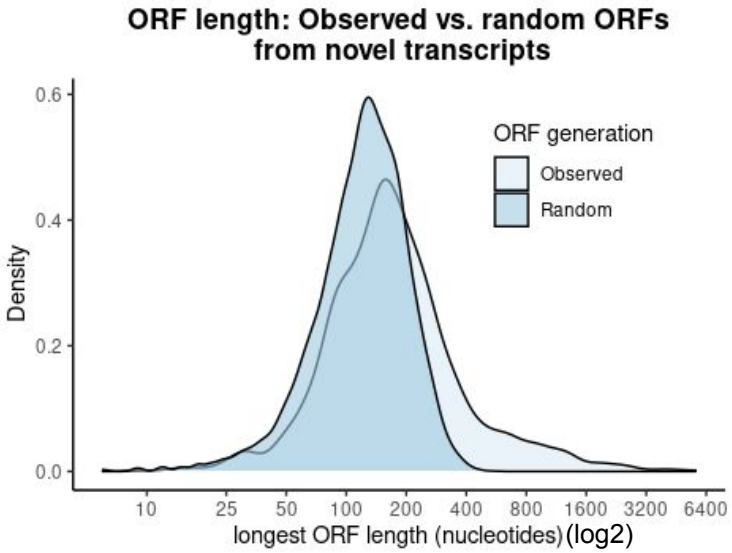


Wilcoxon test
W = 6285324, p-value < 2.2e-16

ORF Generation

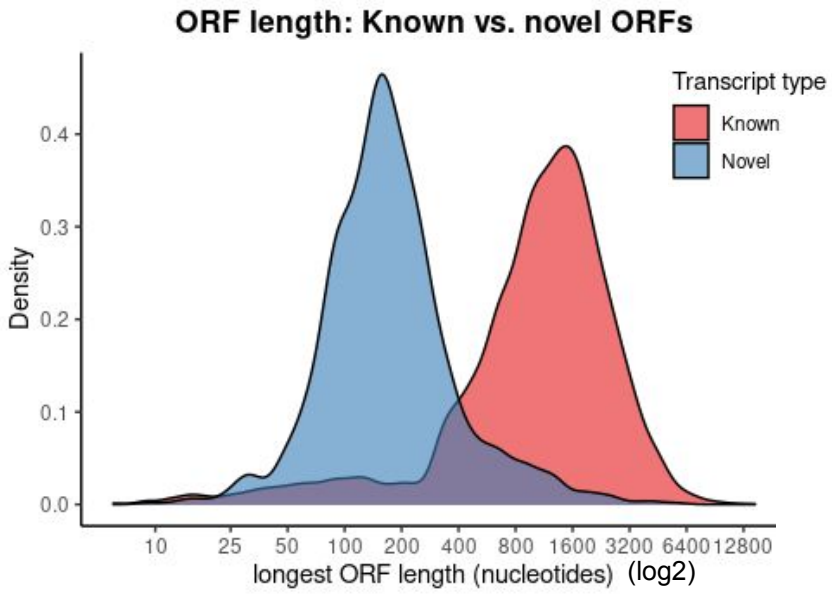


Wilcoxon test
W = 37663806, p-value < 2.2e-16

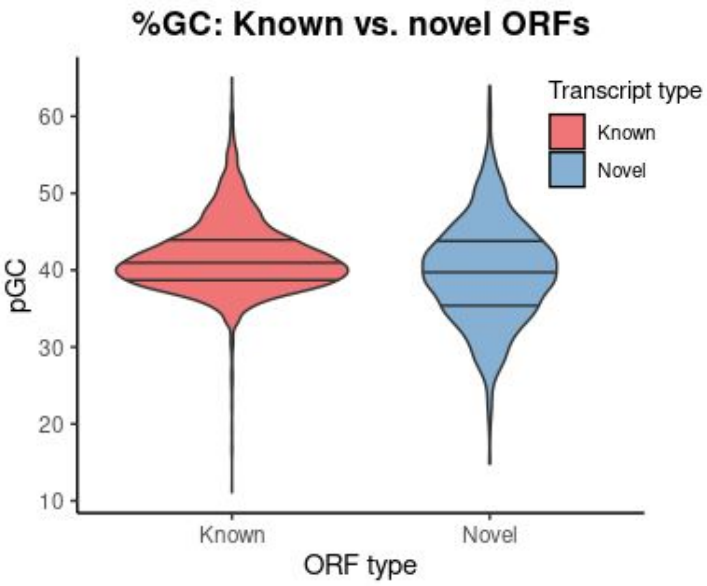


Wilcoxon test
W = 6901369, p-value < 2.2e-16

ORF Characteristics: Novel vs. Known

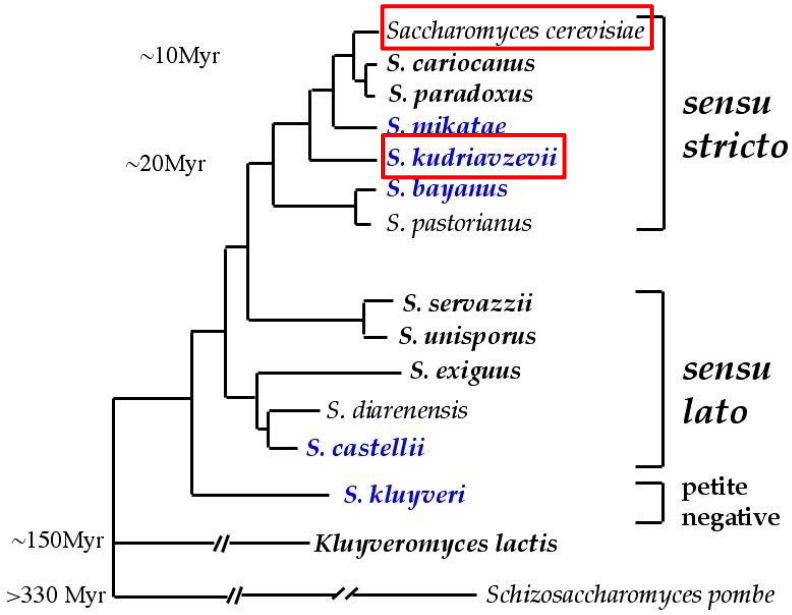


Wilcoxon test
W = 7114308, p-value < 2.2e-16

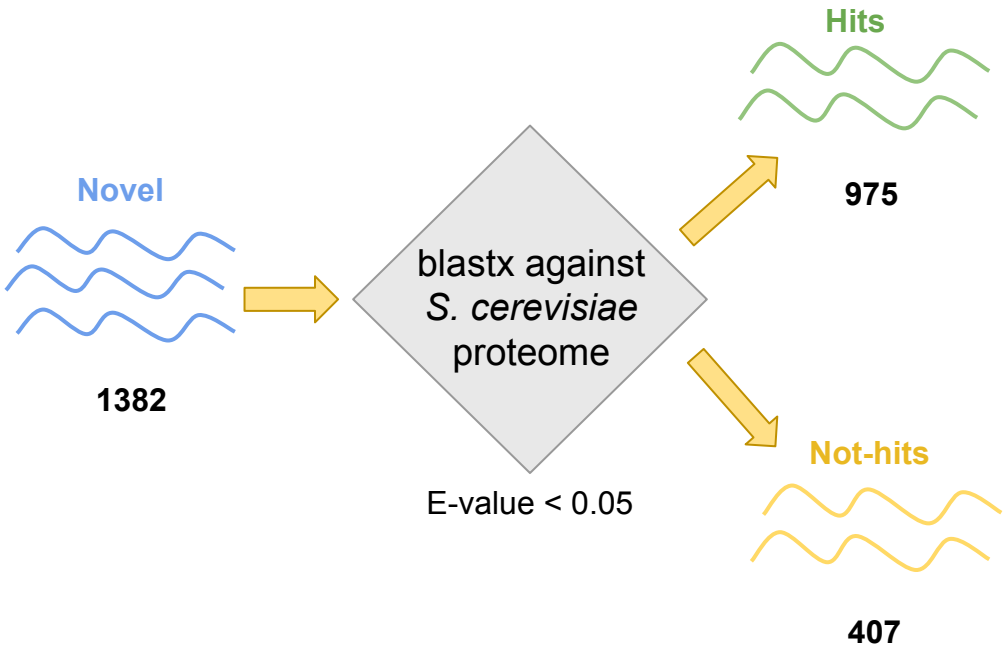


Wilcoxon test
W = 4653452, p-value < 2.2e-16

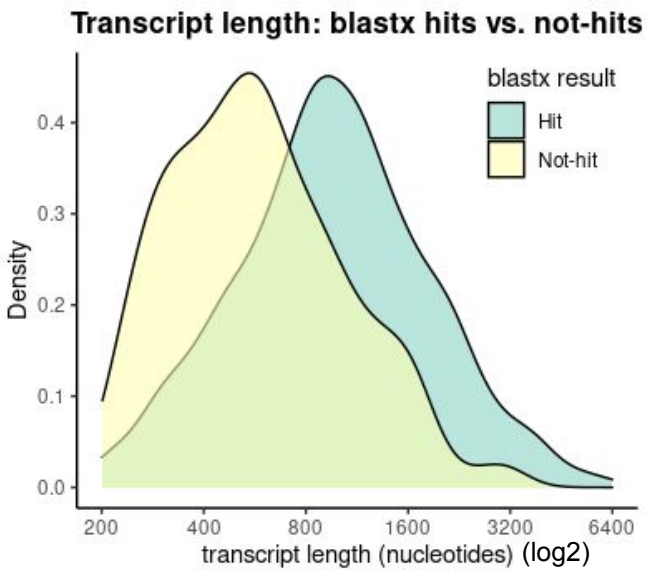
Can we find homologues for the novel genes in *S. cerevisiae*?



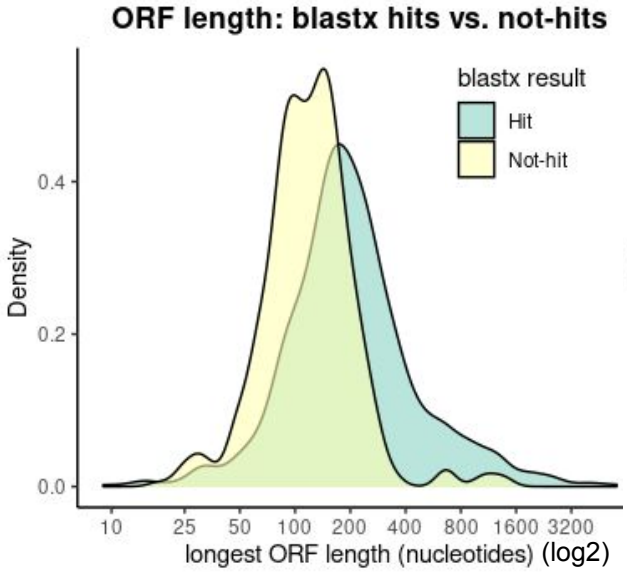
(Cliften *et al.*, 2003)



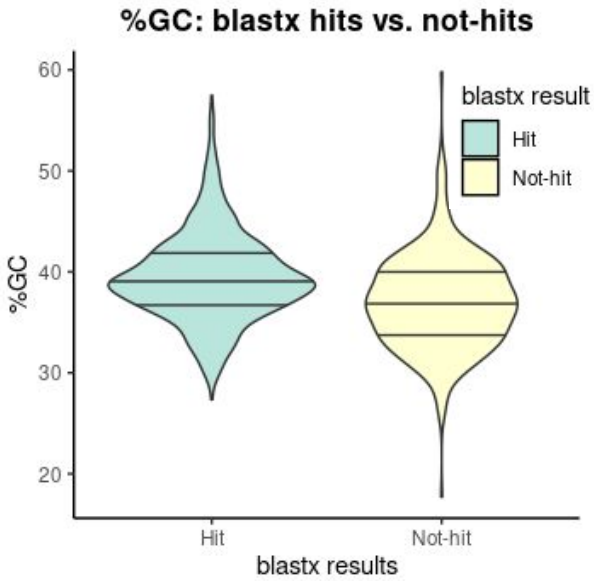
Hits vs not-hits sequence analysis



Wilcoxon test
W = 300190, p-value < 2.2e-16

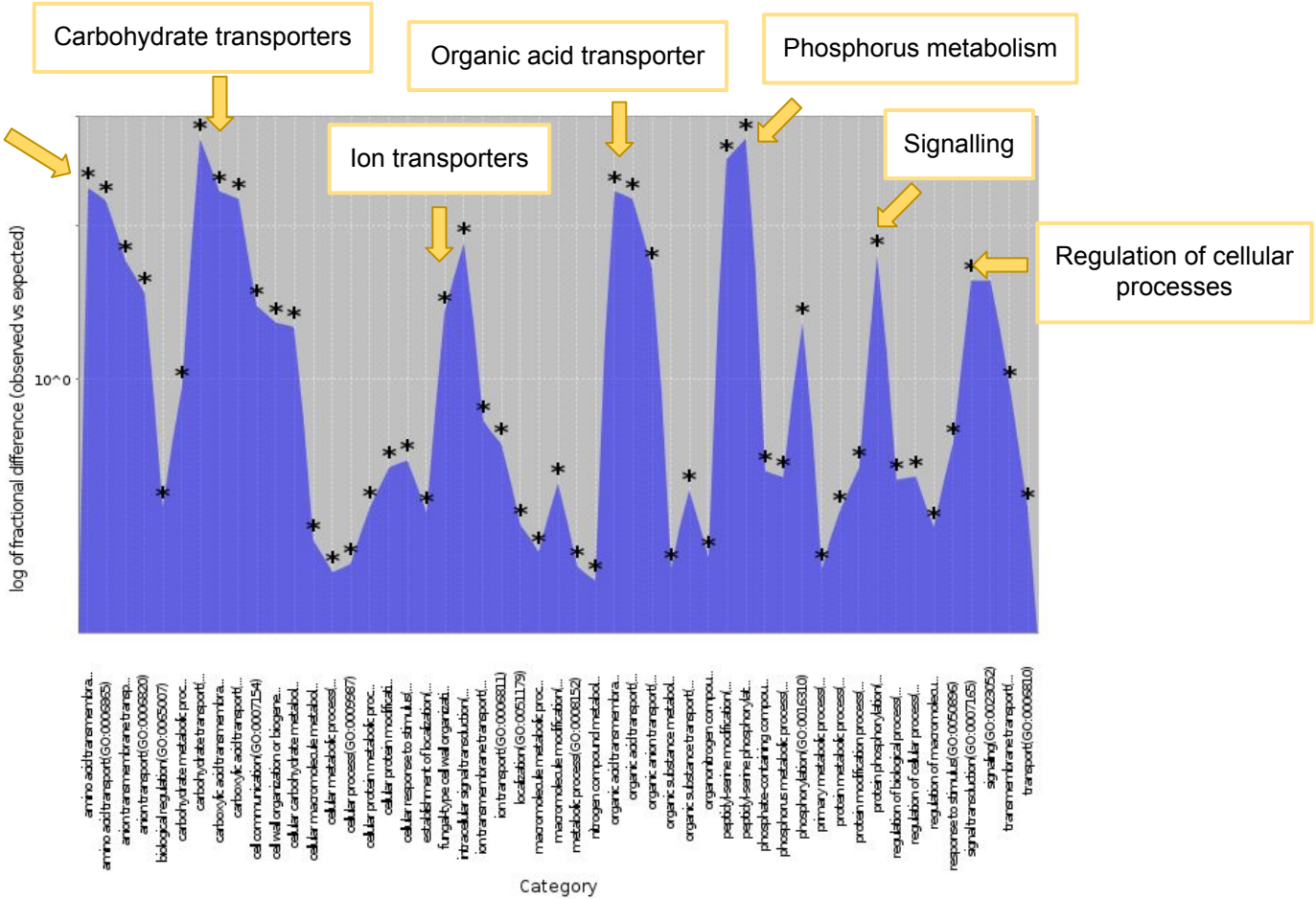


Wilcoxon test
W = 303125, p-value < 2.2e-16



Wilcoxon test
W = 273075, p-value < 2.2e-16

What functions are enriched in hits?



GO enrichment analysis

- Biological processes
- FDR < 0.05

Non-coding prediction

RNAsamba: neural network-based assessment of the protein-coding potential of RNA sequences

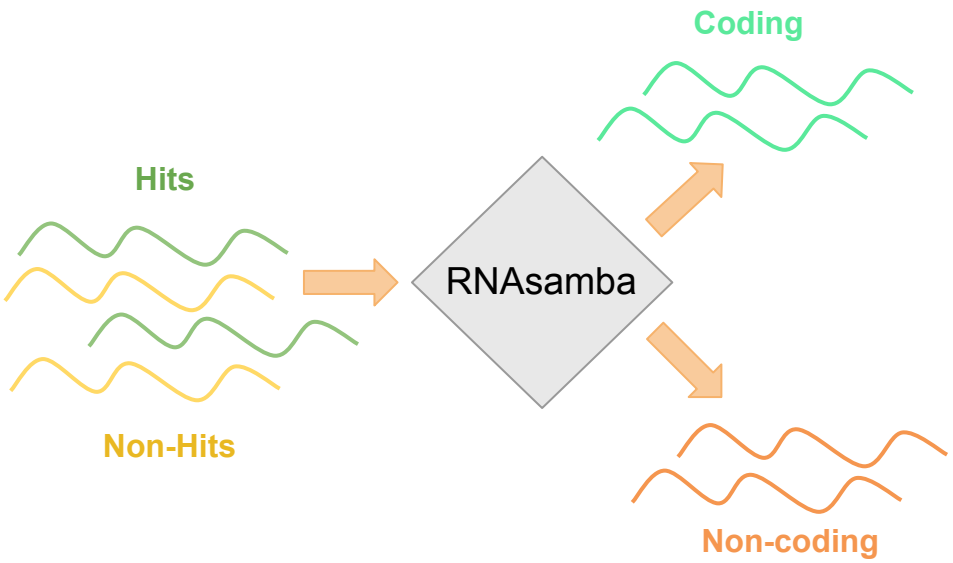
Antonio P. Camargo[1], Vsevolod Sourkov[2], Gonçalo A. G. Pereira[1] and Marcelo F. Carazzolle[1,*]

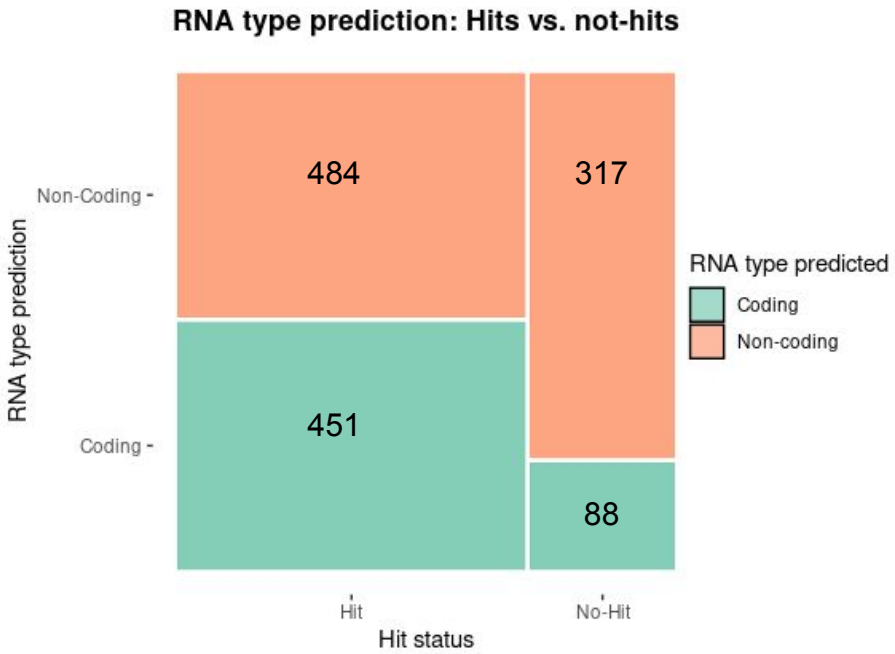
https://rnasamba.lge.ibi.unicamp.br/

The power of multi-layered neural networks to identify deep patterns has made them the *de facto* standard in many machine learning applications, such as image and text analysis, and have been extensively employed in bioinformatics to provide new biological insights (19). Contrasting to conventional machine learning algorithms, deep learning approaches do not necessarily depend on human-designed features and can be used to capture concealed sequence signals that are fundamentally different between mRNAs and lncRNAs.

(Camargo *et al.*, 2020)

Are non-hits more likely to be non-coding?





Chi-squared test
$X^2 = 95.142$, df = 1, p-value < 2.2e-16

Can we find non-hits in a ncRNA dataset?

EnsemblFungi ▾ HMMER | BLAST | BioMart | Tools | Downloads | Help & Docs | Blog

Saccharomyces kudriavzevii IFO 1802 (GCA_000167075) (Saccharomyces_kudriavzevii_strain_IFO1802_v1.0) ▾

Search

Search ... Go

e.g. **YPL216W** or **JH798041:1843-5775** or **synthetase**

Gene annotation

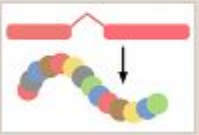
What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

More about this genebuild

Download genes, cDNAs, ncRNA, proteins - FASTA - GFF3

Update your old Ensembl IDs

Example gene



Example transcript

Can we find non-hits in a ncRNA dataset?

BLASTN of Novel Genes without homologues in *S. Cerevisiae* vs ncRNA *S. kudriavzevii* dataset.

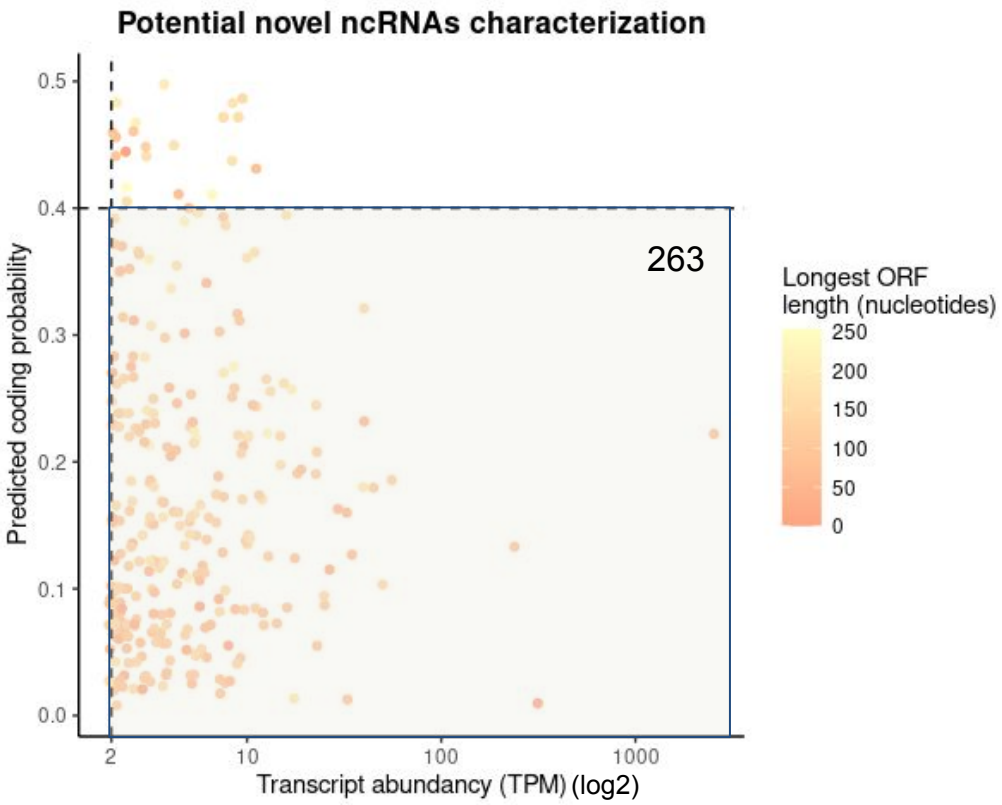
BLASTN

405 unique Non-hits for *S. cerevisiae*

34 Hits on the ncRNA *S. kudriavzevii* dataset

Potential novel ncRNAs characterization

Potential novel ncRNAs characterization



Future Directions

- Ribo-seq Data → *To confirm/discard they code for proteins*
- Loss of Function Studies (e.g. CRISPR) → *To assess functionality*
- RNA Purification Studies (e.g. ChIRP) → *To identify potential partners*

These approaches shall shed light on whether identified transcripts are bona fide functional ncRNAs, spurious transcription or new protein-coding genes.

Concluding Remarks

- We were capable of efficiently processing *S. kudriavzevii* RNA-seq data
- We identified **1382** novel Genes, from which:
    - **975** were identified as protein-coding homologues of *S. cerevisiae*
    - **34** were identified as previously reported ncRNAs
    - **263** were considered new potential novel functional ncRNAs