RNA SEQUENCING DATA ANALYSIS OF *Naumovozyma castellii*

PRINCIPLES OF GENOME BIOINFORMATICS (UPF)

Xavier Soler

Sergio Suárez

Marina Vallejo

01/12/2021

RAW DATA

Transcriptomics data of 11 species of yeast identically grown in rich media and oxidative stress conditions

William R. Blevins, Lucas B. Carey & M. Mar Albà

FILES
illumina_truseq_adapters.fa
n_castellii.fa
n_castellii.gff
n_castellii _read1.fastq.gz
n_castellii _read2.fastq.gz

https://bmcresnotes.biomedcentral.com/articles/10.1186/s13104-019-4286-0/tables/1

ANALYSIS

BASIC:

- Read filtering
- RNA-Seq alignment
- Transcript assembly
- FASTA and ORF obtention

EXTENDED:

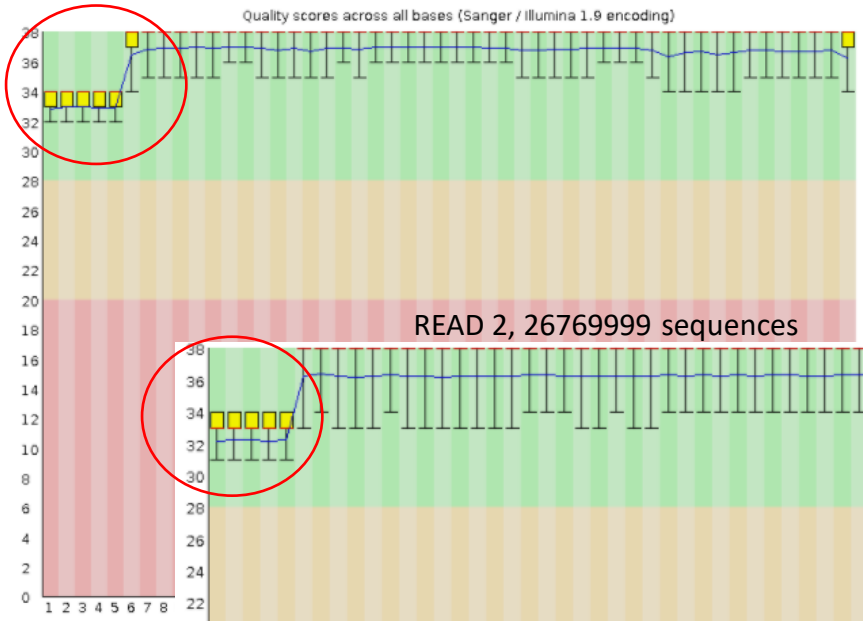
- Codon usage
- GC content
- BLASTp
- Sequence length extended
- Coding Score (CIPHER)
- TPM extended

READ FILTERING

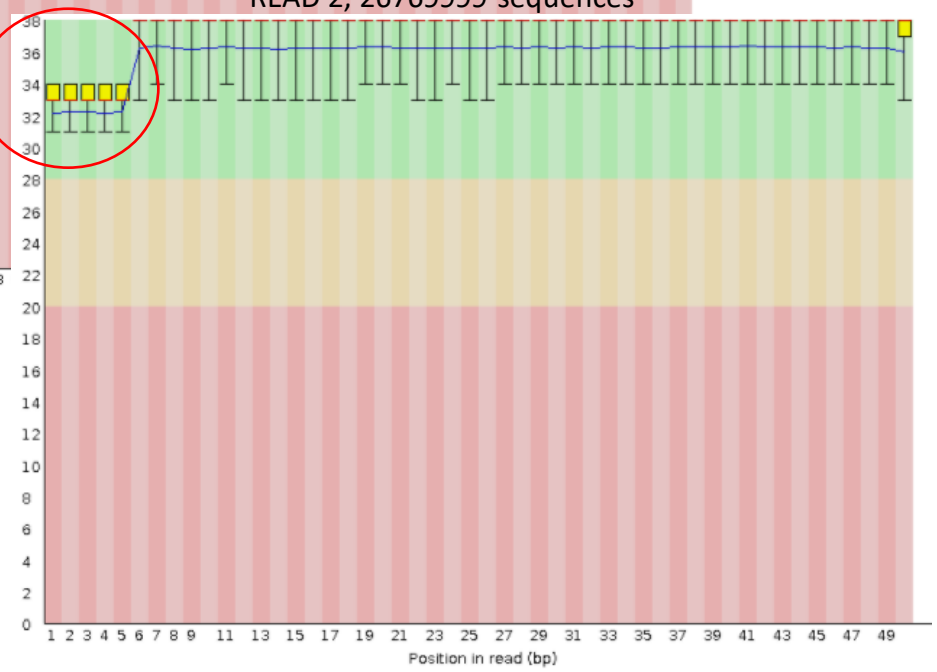
BEFORE TRIMMOMATIC

Per base sequence quality

READ 1, 26769999 sequences



READ 2, 26769999 sequences

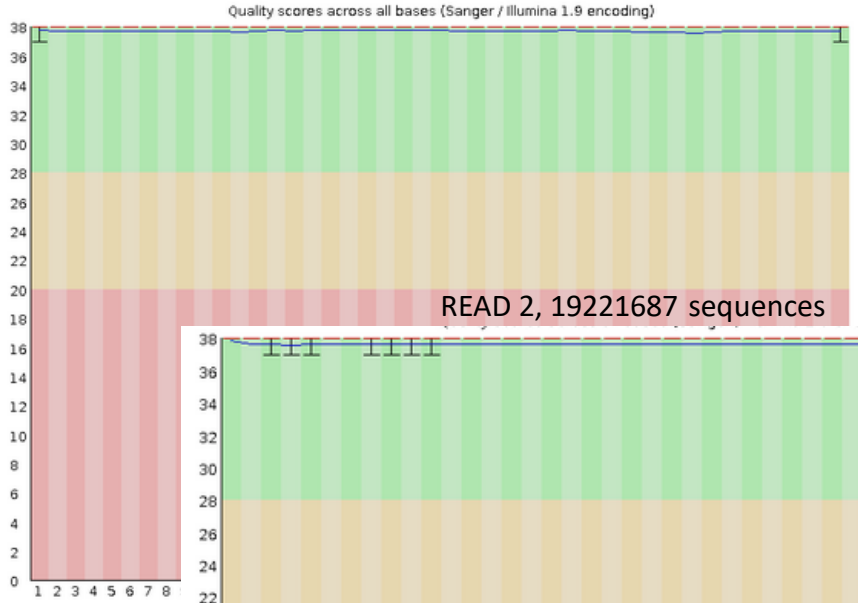


AFTER TRIMMOMATIC

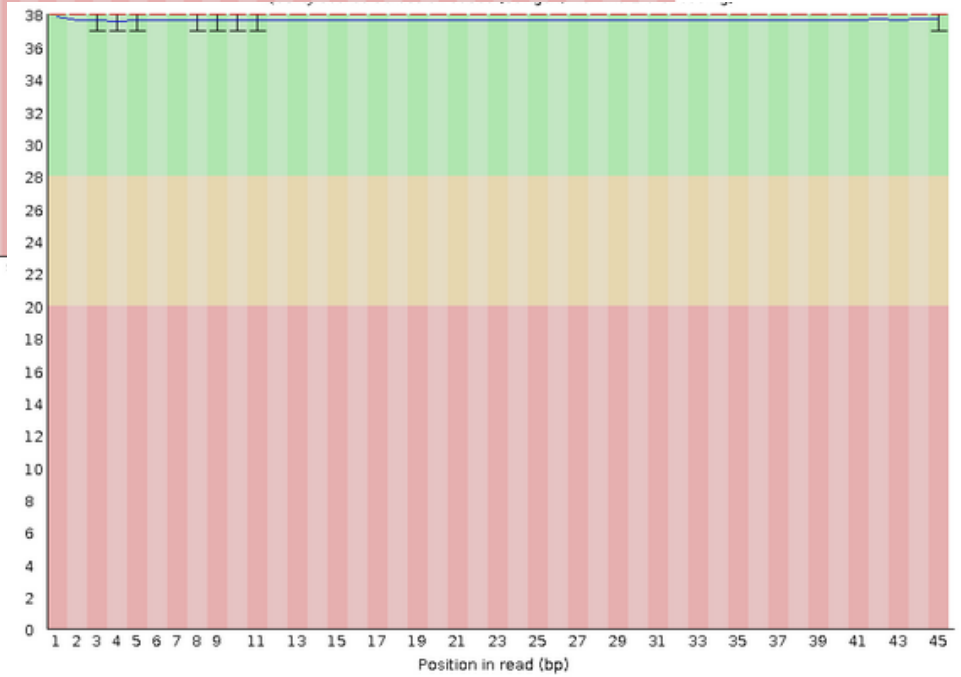
Drop 7,548,312 sequences

Per base sequence quality

READ 1, 19221687 sequences



READ 2, 19221687 sequences



RNA-SEQ ALIGNMENT

hisat2, samtools, IGV

TRANSCRIPT ASSEMBLY

stringtie + awk

TRANSCRIPT
815 Novel (12%)
5969 Known (88%)
6511 Total

Known = annotated

FASTA AND ORF OBTENTION

FASTA

getfasta (bedtools)

Input: global FASTA+ known/novel gff

Output: known/novel FASTA

```
>NC_016491.1:350-1271(-)
ATGGCGTCCGCTTTTCTGGGACAAACGAAGAAAAACAGAATATATAACTG
>NC_016491.1:3019-3643(-)
ATGTCAGATACAAATCTTCAAAAAAGCCAGGAGATTGATAAAAGCGCAAA
>NC_016491.1:6066-6759(+)
ATGACAAACCAGCGTGCTATCCTTTACACCCACGCTGAATTTACAAGACC
```

Number of sequences:

```
ubunti@ubunti:~/Desktop$ grep '>' known.fasta | wc -l
5696
ubunti@ubunti:~/Desktop$ grep '>' novel.fasta | wc -l
815
```

ORF

Perl script

Input: known/novel FASTA

Output: longest ORF FASTA +

randomized ORF FASTA +

lengths

MAX ORF LENGTH (nt)

Known: 14805

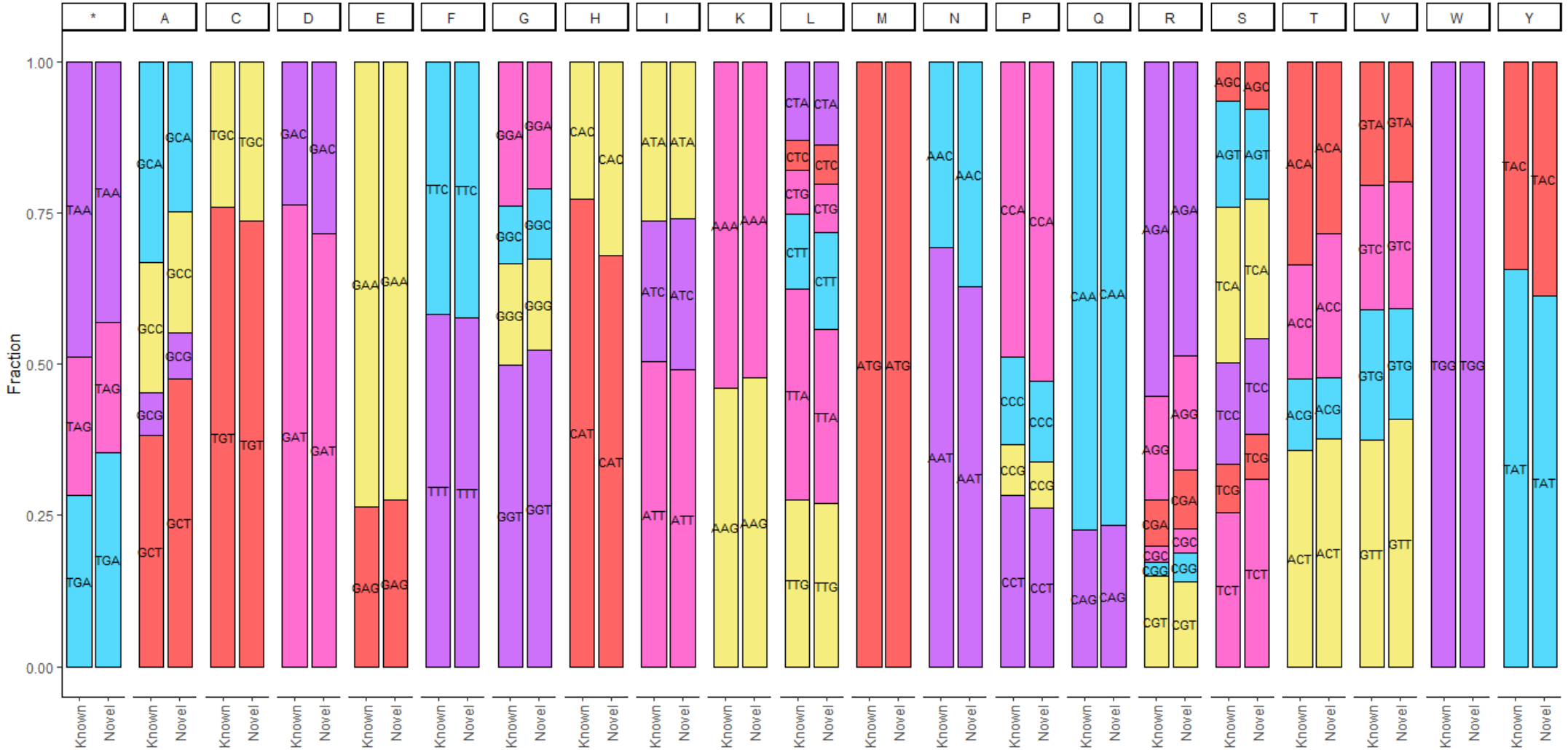
Known random: 564

Novel: 5367

Novel random: 666

CODON USAGE

cusp (EMBOSS)



GC CONTENT

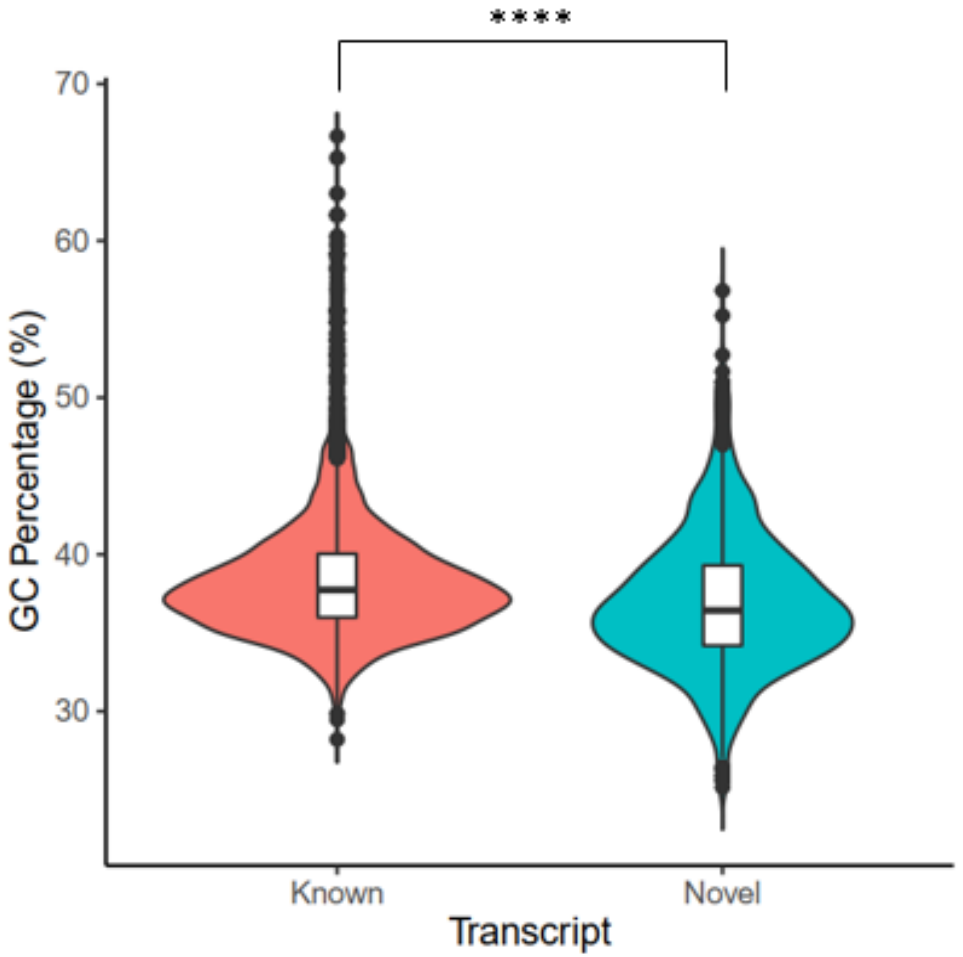
infoseq (EMBOSS)

Descriptive statistics:

	TRANSCRIPT	
	KNOWN	NOVEL
Min.	28.21	25.15
1st Qu.	35.99	34.19
Median	37.74	36.43
Mean	38.59	37.01
3rd Qu.	40.04	39.30
Max.	66.67	56.82

Kolmogorov-Smirnov test:

	TRANSCRIPT
P value	<0.0001
P value summary	****
Significantly different (P < 0.05)?	Yes
Kolmogorov-Smirnov D	0.2043

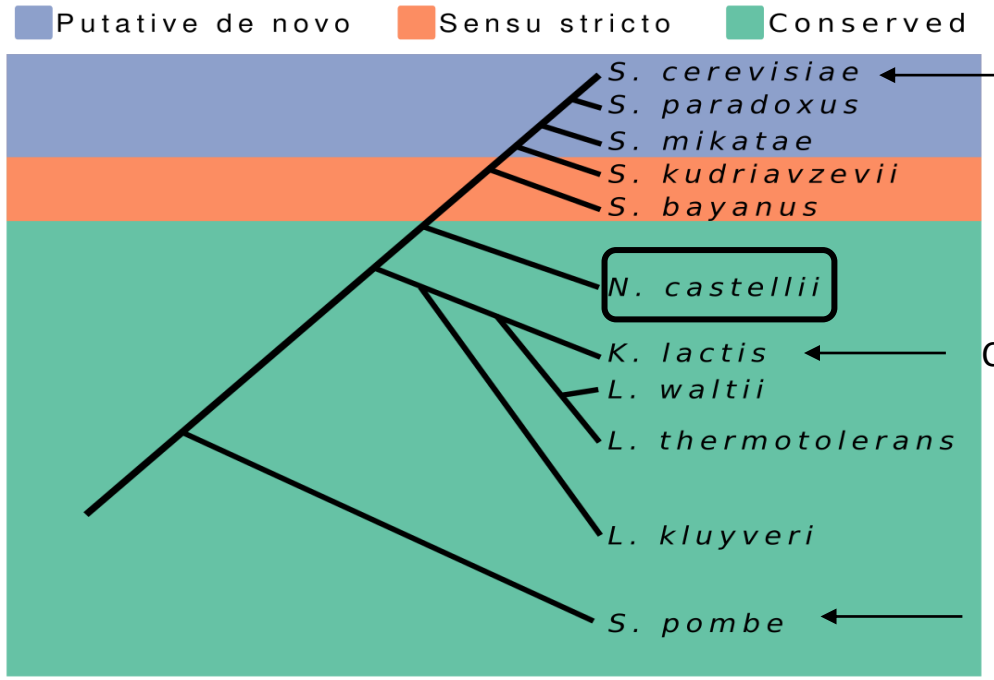


GC content lower in Novel transcript

Higher GC content related with mRNA stabilization

New gens evolution positive to increase GC content

BLASTp



Will Blevins

BLASTp - Compare our transcript with proteins of other organisms

N. castellii

- 5867 gens
- 95.3% coding

S. Cerevisae

- 6464 gens
- 93.1% coding

K. lactis

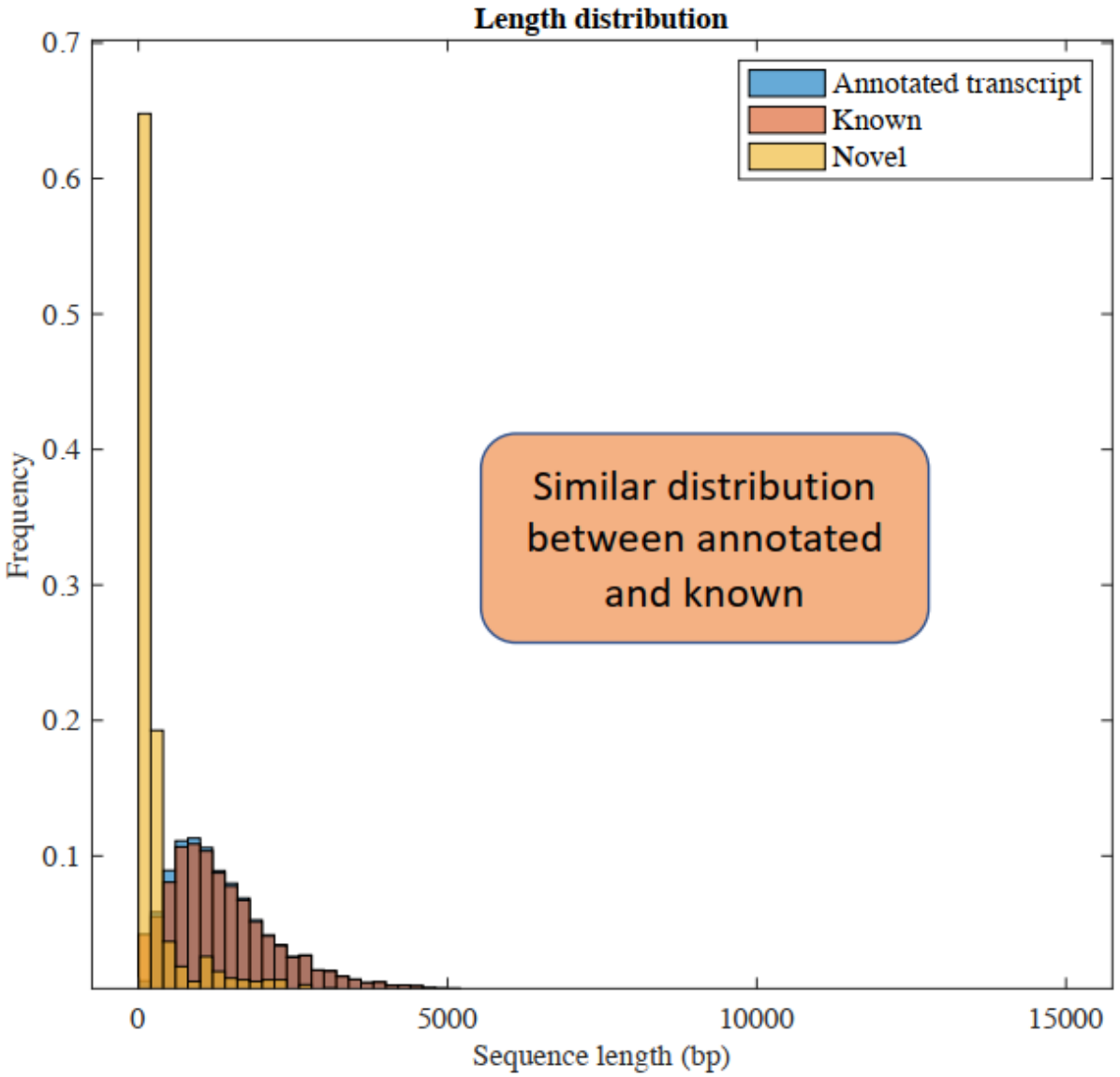
- 5335 gens
- 95.3% coding

S. pombe

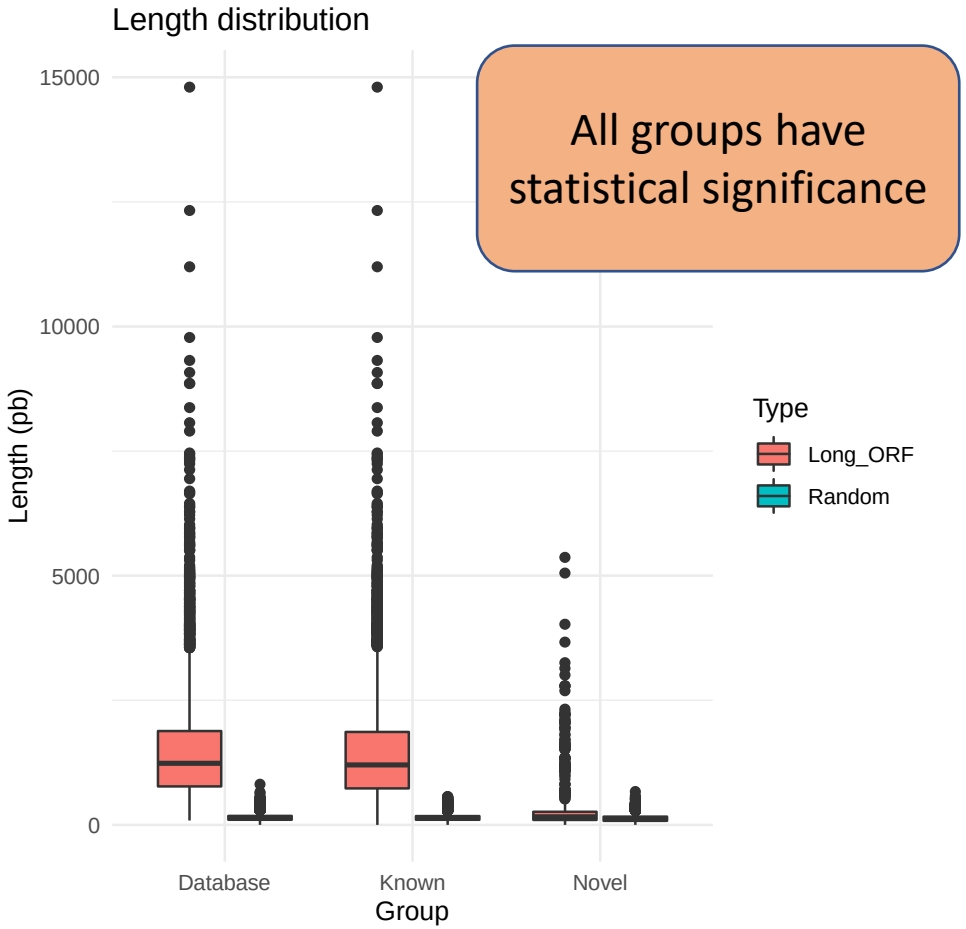
- 6974 gens
- 73.6% coding

	Transcript	Matches	e-value > 0.1
S. cerevisiae	Known	573	3 – 0.5%
	Novel	84	0 – 0%
	Database	541	5 – 0.9%
K. lactis	Known	644	2 – 0.3%
	Novel	117	1 – 0.9%
	Database	591	2 – 0.3%
S. pombe	Known	628	4 – 0.6%
	Novel	134	0 – 0%
	Database	607	4 – 0.7%

SEQUENCE LENGTH



Similar distribution between annotated and known



All groups have statistical significance

Novel most part sort
Probably no-coding

CODING SCORE (CIPHER)

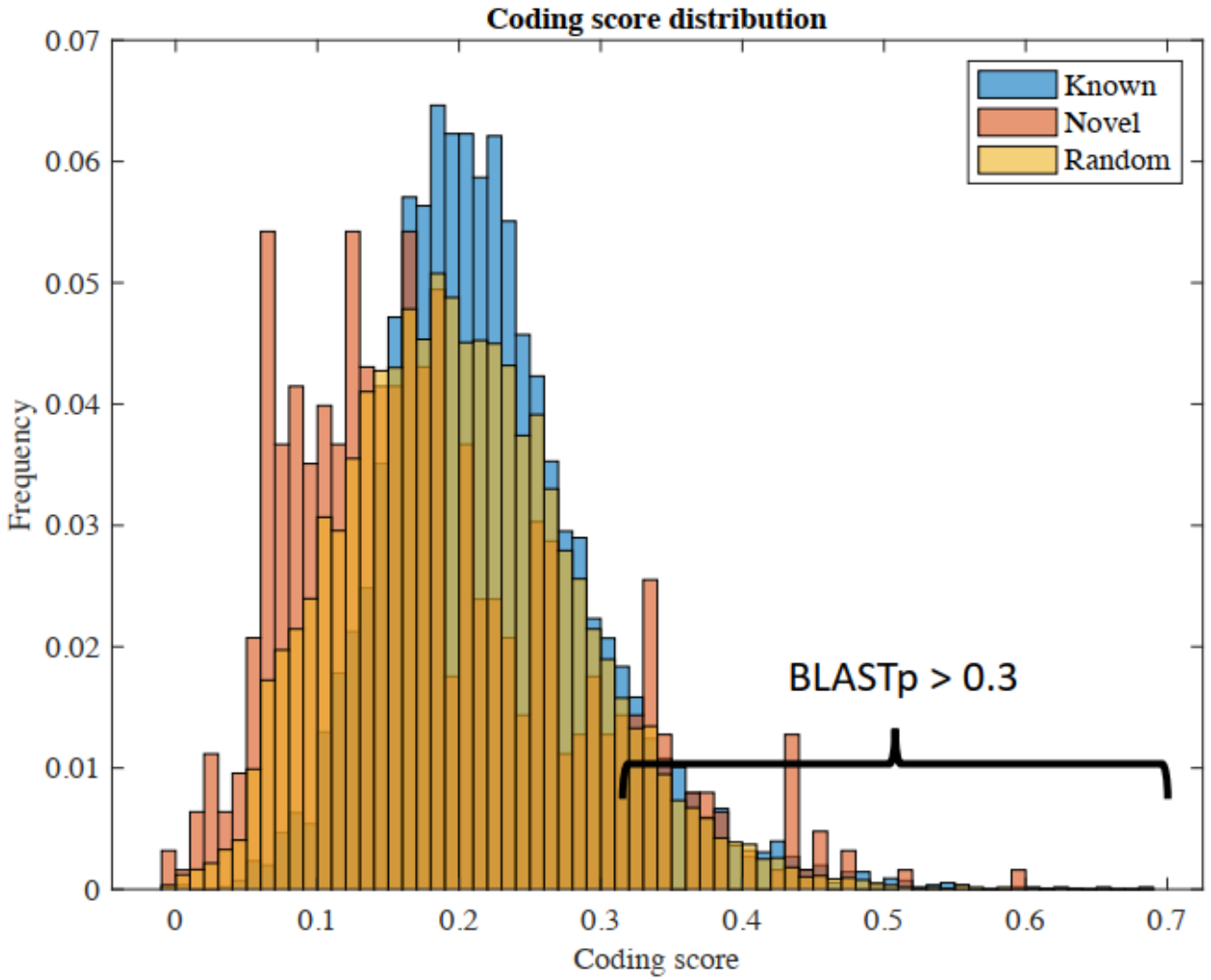
$$CS_{hexamer(i)} = \log\left[freq_{coding}\left(hexamer\left(i\right)\right) \Big/ freq_{non-coding}\left(hexamer\left(i\right)\right)\right]$$

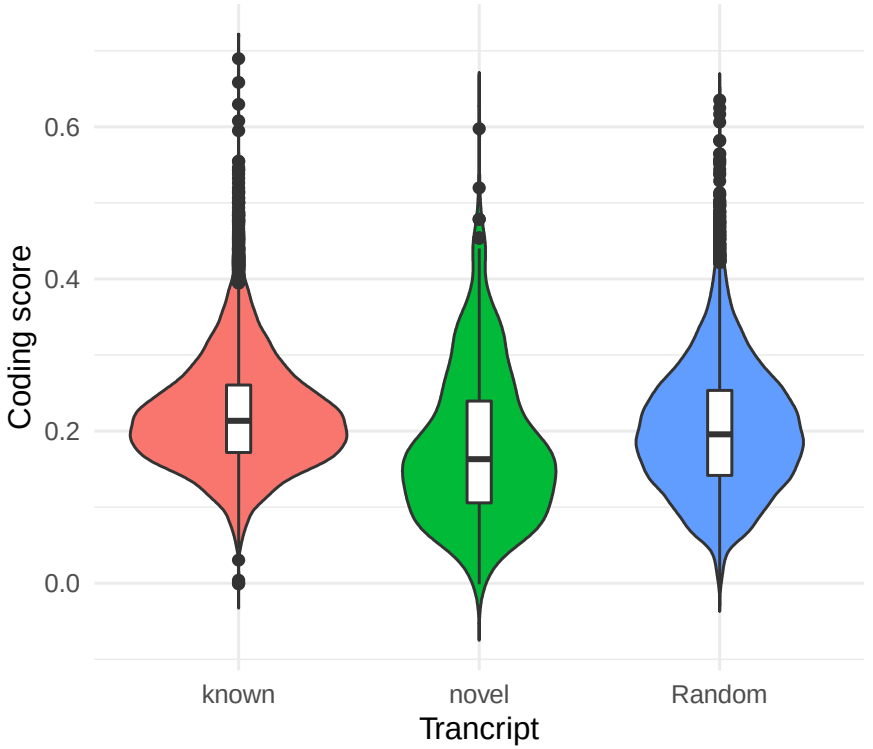
$$CS_{ORF} = \frac{\sum_{i=1}^{i=n} CS_{hexamer(i)}}{n}$$

https://github.com/jorruior/CIPHER

CODING SCORE (CIPHER)

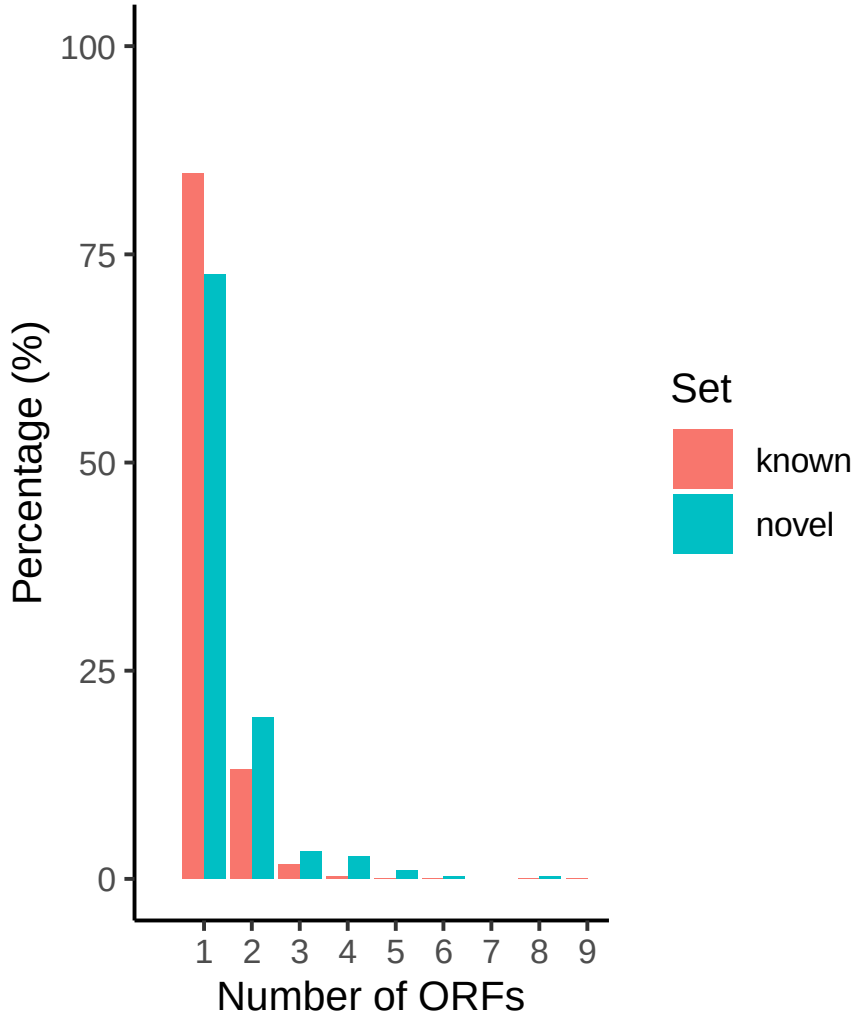
Novel transcripts with high coding capacity are specie-specific



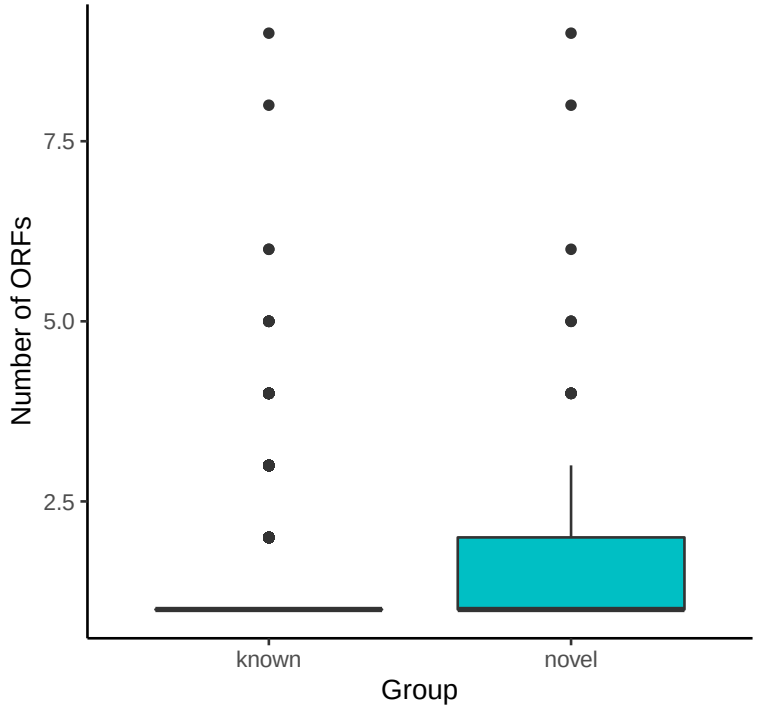


ORFS PER TRANSCRIPT

Pervasive translation of the transcriptome reflected in the novel transcripts

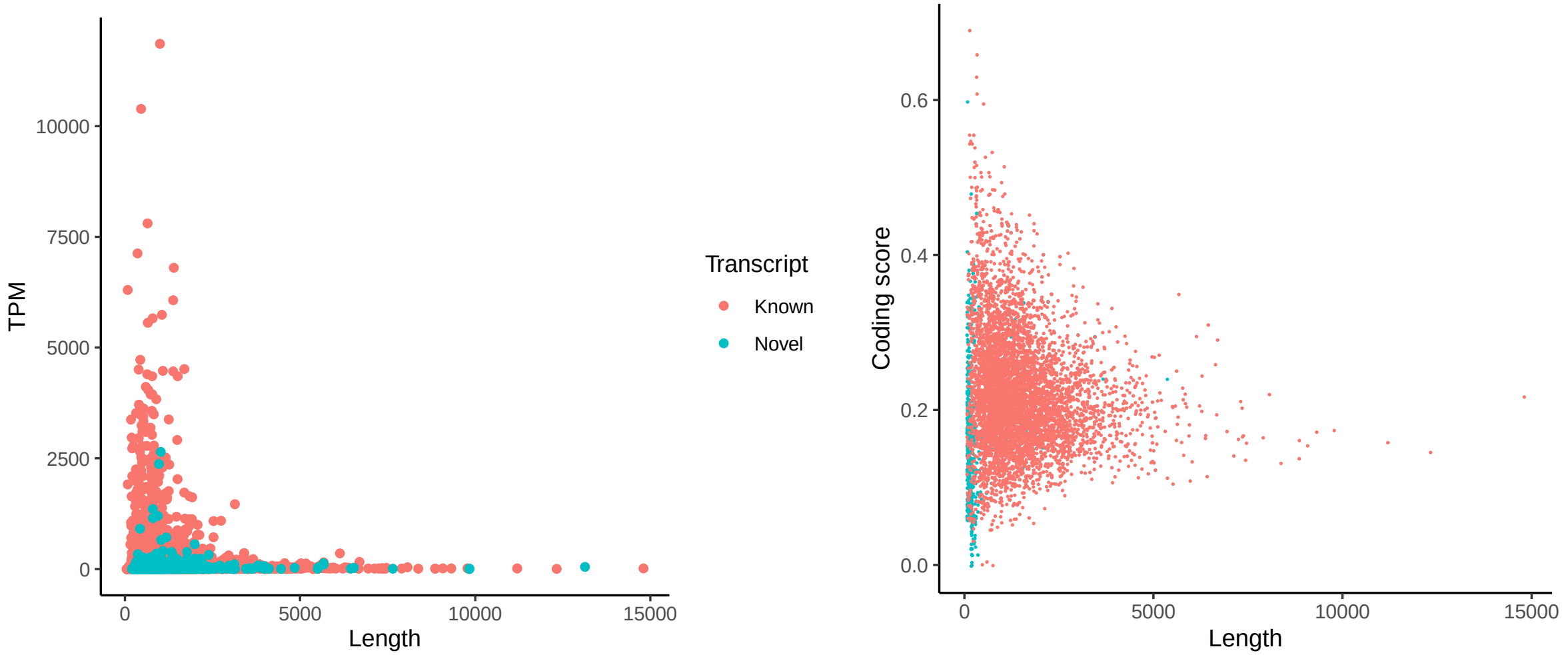


95% CI for difference:
(0.1561422 - 0.3842968)



TRANSCRIPT PER MILION (TPM)

We cannot be sure of the coding capacity of the short transcripts



CONCLUSIONS

Novel genes seem to be developed in novel transcript

- Lower GC content
- Low homology in BLAST
- Higher ORF number

Uncertainty about the coding capacity of the short transcripts

- Low homology in BLAST
- High variability in TPM and CS

MACHINES AND TOOLS

MACHINE 1	MACHINE 2	MACHINE 3
macOS 11.6 - Darwin 20.6.0	MacOS Monterey – 12.0.1	20.04.2-Ubuntu
Intel Core i5 CPU 2.3 GHz	Intel Core i5 CPU 2.3 GHz	Intel Core i7 CPU 2.7 GHz
8 GB RAM	8 GB RAM	12 GB RAM
2 Cores	2 Cores	2 Cores

TOOLS BASIC ANALYSIS	TOOLS EXTENDED ANALYSIS
fastQC 0.11.5	**EMBOSS 6.6.0**
SRA 2.4.1	**blastp 2.12.0+**
stringtie 1.3.4d	**CIPHER 1.0.0**
Trimmomatic 0.36	**Python 2.7**
PERL v5.30.0	**R 4.0.2**
hisat2-2.2.0	**MATLAB_R2021a**
bedtools-2.28.0	