LncPlankton: A comprehensive collection of plankton long noncoding RNAs

Ahmed Debit[1]*, Pierre Vincens[1], Chris Bowler[1], Helena Cruz de Carvalho[1,2]*

[1] Institut de Biologie de l'ENS (IBENS), École normale supérieure, CNRS, INSERM, Université PSL, Paris, France

[2] Université Paris Est-Créteil (UPEC), Faculté des Sciences et Technologie, Créteil, France

* Correspondence: debit@bio.ens.psl.eu | cruz@bio.ens.psl.eu

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR

Background

For a long time, long noncoding RNAs (lncRNAs) have been viewed as transcriptional noise. However, they're now being recognized as key regulatory molecules across the Eukarya domain, including plants, animals, and fungi. Despite this, we still don't have a clear understanding of how these molecules occur in the marine environment.

To address this knowledge gap, we created LncPlankton, the first comprehensive database of marine planktonic lncRNAs. By combining predictions from ten coding potential tools using majority voting, we identified over 2 million lncRNAs spread across 414 marine planktonic species from more than nine different phyla. We also developed a user-friendly, open-access interface to explore the database (https://www.lncplankton.bio.ens.psl.eu/).

Our goal is for LncPlankton to become a valuable resource for lncRNAs, supporting both small-scale and large-scale research in various marine planktonic species, and enabling comparative analysis that extends far beyond the marine environment.

LncPlankton content

A.

Species	414
Total transcripts	11,623,179
Mean length	1,018
lncRNA-like predicted	2,210,359
sncRNA-like predicted	1,118,450
Coding-like predicted	8,294,370
High-confidence lncRNAs	239,116

B.

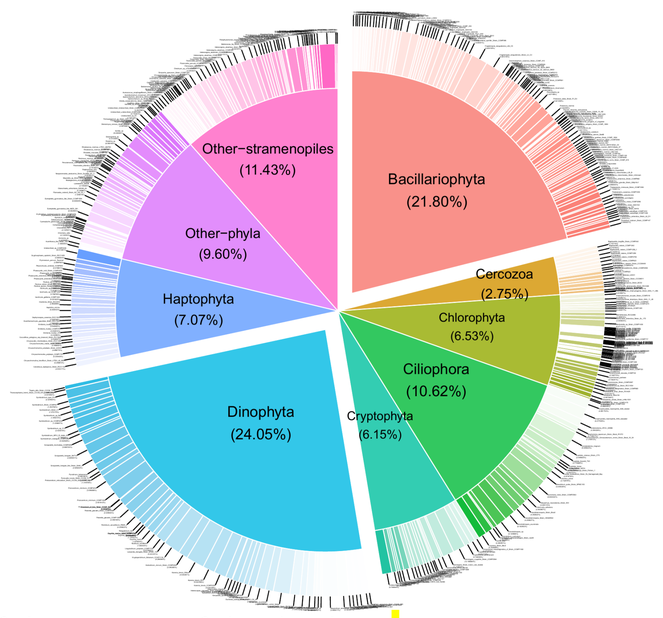

Figure 3. Summary of the content of the database

Materials & Methods

- **Data sources of LncPlankton**

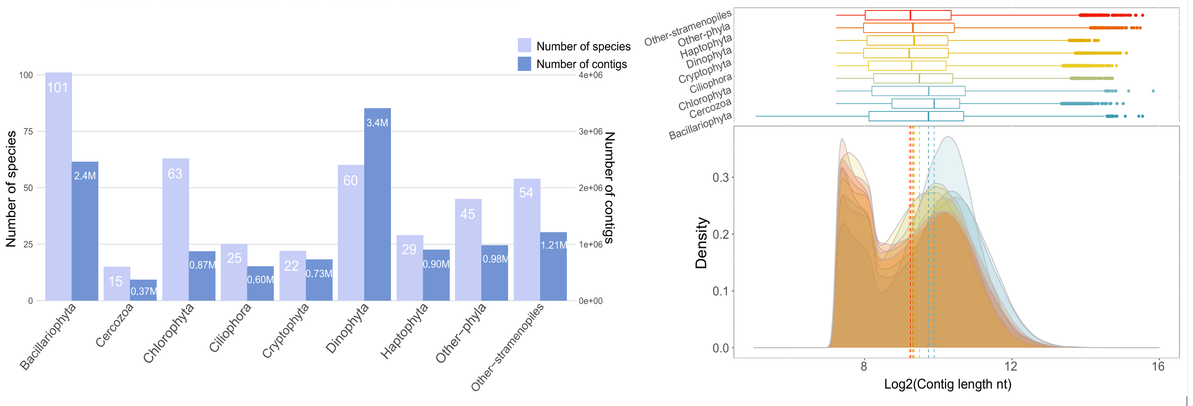

Figure 1. Data used in LncPlankton

- Transcriptomic data of 406 species MMETSP project (Keeling et al., 2014) + 6 diatom species not covered by the MMETSP project.
- Transcriptomes of *P. tricornutum* (Cruz de Carvalho et al., 2016) and *Thalassiosira pseudonana (Goldman J.A. et al., 2019)* used an in-house assembly pipeline.

- **Data analysis pipeline**

Our method shows a high accuracy and low variability

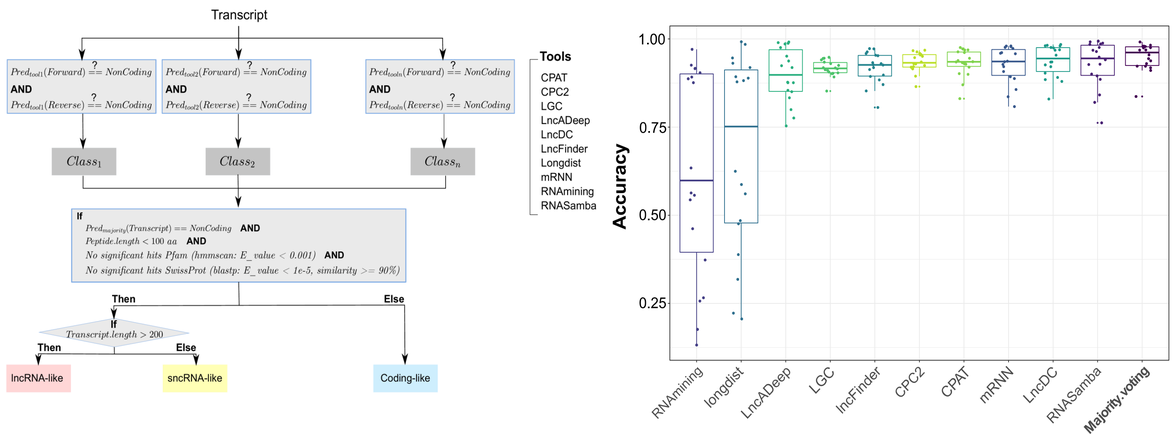

Figure 2. The pipeline used for the prediction of lncRNAs combines 10 coding potential tools in a majority voting setting

Benchmarking data: Different sets of coding and non-coding datasets related to 18 species of different chordate clades. The datasets were independent of the training sets used in the construction of the pre-trained models related to each coding potential tool.

LncPlankton UI

The UI provides modules for browsing, searching, downloading lncRNA data per species and/or per phylum, interactive graphs, and an online BLAST service. A shiny application was also integrated, allowing the user to customize and visualize the classification.

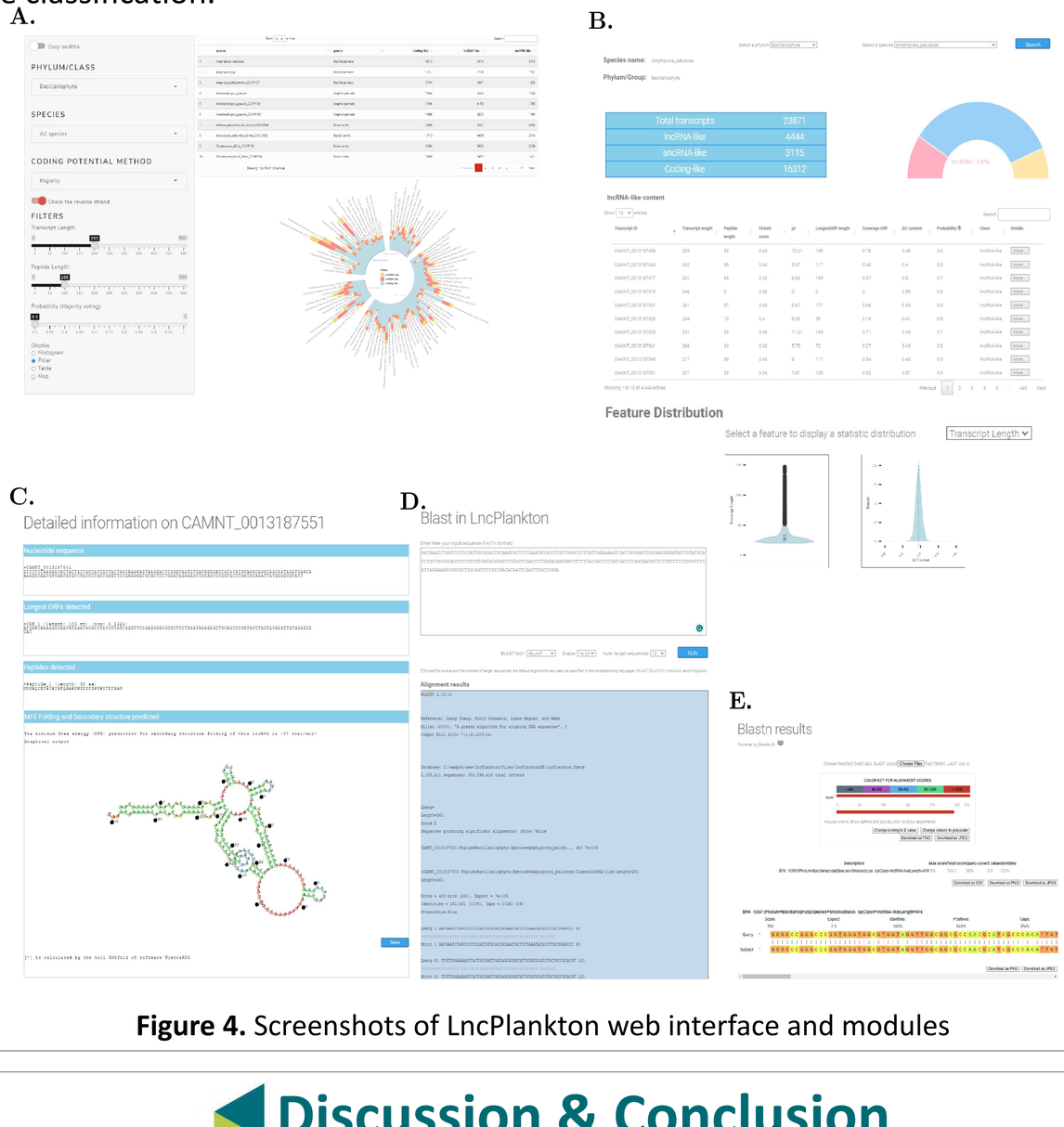

Figure 4. Screenshots of LncPlankton web interface and modules

Discussion & Conclusion

- We implemented a meta-learner using a **majority voting-based** ensemble learning technique for the prediction of lncRNAs in marine planktonic species (https://gitlab.com/a.debit/votinglnc)
- The meta-learner promoted the **heterogeneity** (multiple and diverse ML algorithms), and the **diversity** (integrated of different features describing the transcripts).
- Our method is **robust** and shows a low variability of the prediction across the different testing datasets.
- The majority voting tool offers us a **reliable** choice for lncRNAs identification, beyond what a single tool can offer at the moment.
- LncPlankton is the **largest** data repository on lncRNAs in marine species.
- We anticipate LncPlankton will contribute significantly to future efforts aimed at deciphering the biology of marine lncRNAs**.**

References

- Keeling et al., (2014); The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing; *PLoS Biology, 12.*
- Cruz de Carvalho et al., (2016); Noncoding and coding transcriptome responses of a marine diatom to phosphate fluctuations; *New Phytologist, 210, 497–510.*
- Goldman J.A. et al., (2019); Fe limitation decreases transcriptional regulation over the diel cycle in the model diatom Thalassiosira pseudonana; *PLOS One, 14.*