LncPlankton V1.0: a comprehensive collection of plankton long non-coding RNAs

Ahmed Debit, Pierre Vincens, Chris Bowler, Helena Cruz de Carvalho

Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR 8197 INSERM U1024, Paris, France

Background

Long-time considered as transcriptional noise, long noncoding RNAs (lncRNAs) are emerging as central, regulatory molecules in a multitude of eukaryotic species, from plants to animals to fungi. Yet, our knowledge about the occurrence of these molecules in the marine environment, namely in planktonic protists, is still extremely elusive. To fill that gap of knowledge we developed LncPlankton V1.0, which is the first comprehensive database of marine planktonic lncRNAs. By integrating the predictions derived from ten distinctive coding potential tools in a majority voting setting, we identified 2.188.411 lncRNAs distributed across 414 marine planktonic species from over 9 different phyla. A user-friendly, open-access web interface for the exploration of the database was implemented (https://www.lncplankton.bio.ens.psl.eu/). We believe LncPlankton V1.0 will serve as a rich resource of lncRNAs that will contribute to small- and large-scale studies in a wide range of marine planktonic species and allow comparative analysis well beyond the marine environment.

LncPlankton V1.0 content

A.

Phyla/Groups	≥ 9
Species	414
Total transcripts	11.537.434
Mean length (nt)	1018
lncRNA-like predicted	2.188.411
sncRNA-like predicted	1.110.310
Coding-like predicted	8.238.713
High-confidence lncRNAs	236.155

B.

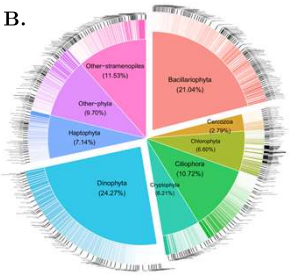
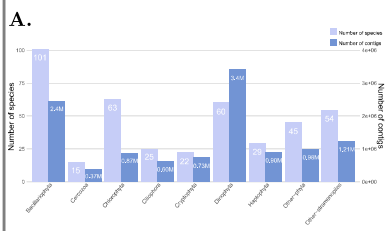


Figure 3. Summary statistics of the content of the database

Materials & Methods

- Data sources of LncPlankton V1.0

A.



B.

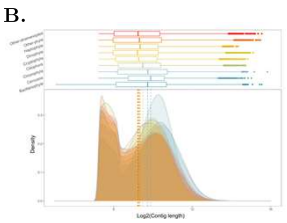


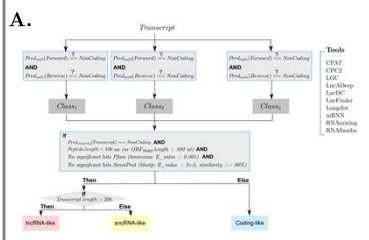
Figure 1. Data used in LncPlankton V1.0.

- Transcriptomic data of 406 species MMETSP project + 6 species not covered by the MMETSP projec
- Two transcriptomes (Phatr and Thaps) assembled used an in-house assembly pipeline.

- Data analysis pipeline

Our method shows a high accuracy and low variability

A.



B.

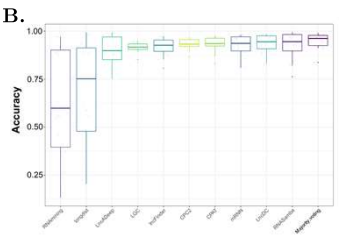


Figure 2. The pipeline used for the prediction of lncRNAs combines 10 coding potential tools in a majority voting setting

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

LncPlankton v1.0 UI

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

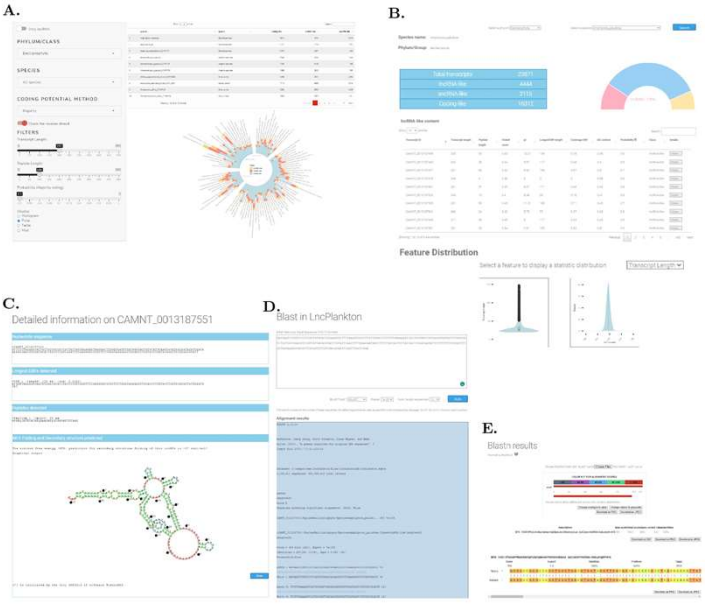


Figure 4. Screenshots of LncPlankton V1.0 web interface and modules

Discussion & Conclusion

- We implemented a meta-learner using a **majority voting-based** ensemble learning technique for the prediction of lncRNAs in marine planktonic species.
- 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** showing a low variability of the prediction across the different testing datasets
- We believe the majority voting tool offers us a **reliable** choice for lncRNAs identification, beyond what a single tool can offer at the moment
- We developed LncPlankton V1.0 which 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.
- **Future works:** genomic coordinates, expression patterns, conservation and evolutionary relationship, regular updates (new species)

debit@bio.ens.psl.eu