

LncPlankton: a comprehensive database of candidate lncRNAs from marine microbial eukaryotes

Ahmed Debit ^{1,*}, Pierre Vincens ¹, Chris Bowler ¹, Helena Cruz de Carvalho ^{1,2,*}

¹Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, Paris 75005, France

²Faculté des Sciences et Technologie, Université Paris Est-Créteil (UPEC), 61, avenue du Général De Gaulle, Créteil 94000, France

*To whom correspondence should be addressed. Email: debit@bio.ens.psl.eu

Correspondence may also be addressed to Helena Cruz de Carvalho. Email: cruz@bio.ens.psl.eu

Abstract

Historically neglected or considered to be mere transcriptional noise, long non-coding RNAs (lncRNAs) are now emerging as central, regulatory molecules in a multitude of eukaryotic species, from animals to plants to fungi. Yet, our knowledge about the occurrence of these molecules in the marine environment is still elusive. To help fill this knowledge gap, we have developed LncPlankton, a comprehensive database of candidate marine lncRNAs. By integrating the predictions derived from 10 distinctive coding potential prediction tools in a majority voting setting, we have identified over 2M potential lncRNAs distributed across 414 marine plankton species from over nine different phyla. A user-friendly, open-access web interface of the database has been implemented to facilitate exploration (<https://www.lncplankton.bio.ens.psl.eu/>). We believe LncPlankton will serve as a rich resource for studies of lncRNAs, which will contribute to small- and large-scale analyses in a wide range of marine plankton species and allow comparative studies between them and well beyond the marine environment.

Introduction

Often referred to as the ‘dark matter’ of genomes, the non-protein coding DNA portion of eukaryotic genomes has, in the last decade, been shown to be much larger than what was previously thought. Furthermore, with the advent of deep sequencing, it has become clear that genomes are pervasively transcribed and that the non-coding fraction generates large numbers of transcripts that cannot be accounted for as simple ‘junk’ or transcriptional noise [1]. Among this non-coding fraction, long non-coding RNAs (lncRNAs) represent the most abundant and prevalent class [2, 3]. lncRNAs are arbitrarily defined as transcripts of >200 nucleotides (nt) in length that lack large open reading frames [2]. Like messenger RNAs (mRNAs), lncRNAs are generally polyadenylated, capped, and processed [2, 3]. However, besides their lack of, or low, protein coding potential, lncRNAs exhibit certain characteristics that distinguish them from mRNAs. These include a lower GC content, fewer exons, shorter sequence length, lower sequence conservation, and lower expression levels compared to mRNAs [4]. They have also been shown to have more restrictive expression patterns, being expressed in specific cell types or in response to specific stress cues [1, 2]. In terms of function, lncRNAs have been shown to play key regulatory roles in a multitude of biological processes and diseases, which involve the control of epigenetic modifications as well as other mechanisms of gene expression and protein regulation [3]. Their intrinsic capacity to interact with proteins, DNA, and other RNAs seems to enable a diversity of regulatory mechanisms, involving changes in the chromatin landscape

via interactions with histone modifying complexes, modulation of the expression of neighbouring or distant genes, by direct or indirect interaction with transcription factors and/or RNA polymerase II or by influencing RNA maturation and stability [1, 3]. Furthermore, some lncRNAs have been shown to have dual functions, functioning as RNA transcripts and also having the potential to encode small regulatory peptides [5, 6].

With the booming of high-throughput sequencing techniques and the exponential rise of transcriptomics data in public repositories, several databases dedicated to lncRNAs have been developed. Some are exclusively devoted to human lncRNAs, such as GeneCaRNA [7] and LNCipedia [8], while other databases have grouped together a significant number of lncRNAs coming from photosynthetic organisms such as CANTATAdb [9] and GRENC v.2 [10]. On the other hand, the NONCODE knowledge database v6.0 compiled lncRNAs from 16 animal and 23 plant species [11], and the RNAcentral catalogue is an extensive database of lncRNA sequences from an even broader range of organisms [12]. Despite such extensive work on the identification of lncRNAs in animals and plants, the ocean realm remains largely unexplored. Although invisible to the naked eye, the highly abundant marine microbial eukaryotes are at the base of all marine food webs and are ultimately global drivers of biogeochemical cycles, helping also to regulate the Earth's climate. As unicellular eukaryotes, they are likely to hold clues to the evolutionary processes that gave rise to multicellular organisms, and their genomes represent valuable sources of

Received: May 13, 2025. Revised: September 26, 2025. Accepted: October 6, 2025

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License

(<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the

original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact

journals.permissions@oup.com.

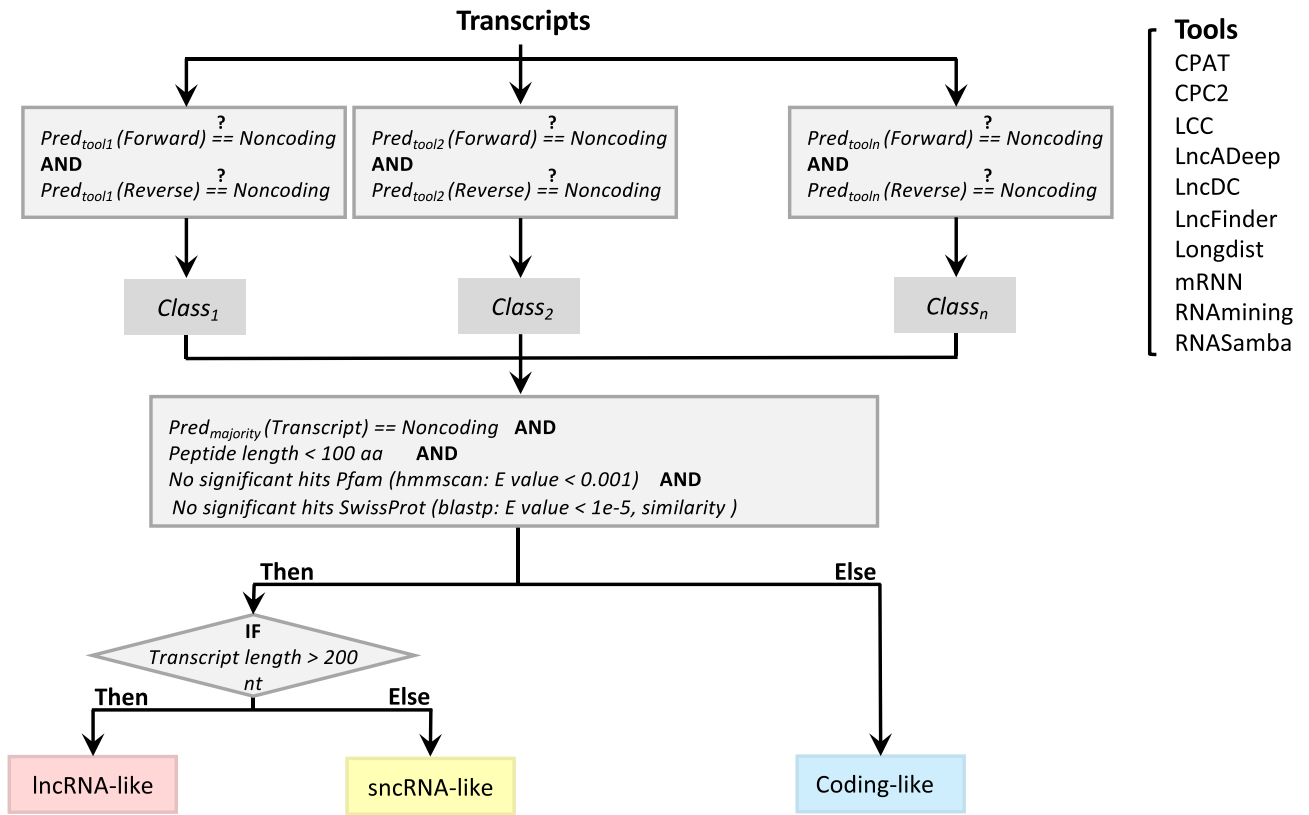


Figure 1. Overview of the majority voting meta-learner pipeline (VotingLNC) for the prediction of lncRNAs in LncPlankton. The 10 coding potential prediction tools included in the procedure are displayed on the right panel.

information for understanding the evolution and functions of lncRNAs. To fill this knowledge gap, our study has used the publicly available highly diverse transcriptomic datasets from the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) [13] to perform a thorough large-scale screen to identify and characterize lncRNAs in marine plankton.

Determining the coding potential of a transcript is a crucial step in the identification of putative lncRNAs, yet it represents a complex task due to overlapping characteristics and functions that exist between coding and non-coding RNAs [14]. To overcome this challenge, a plethora of computational methods have been developed, many of which use features derived from the nucleotide sequence of the transcripts such as the Fickett score, GC content, and k -mer composition. Some of the tools also use secondary structure properties to try to improve the discrimination between putative non-coding RNAs and coding RNAs. Despite the satisfactory performance of these tools reported in many studies, they generate a significant number of false positives and false negatives [9], and their reliability is questionable, adding uncertainty to the results. Taking this into account, we propose a novel pipeline for candidate lncRNAs detection that combines multiple tools in a majority voting setting, inspired by the ensemble method technique used in machine learning. The goal was to increase the reliability and promote the diversity of the ensemble model, since a joint prediction is likely to behave better than any single model [15, 16] and is likely to achieve higher cross-species prediction performance [17].

Materials and methods

Data sources

Previously published transcriptomic data of 414 marine micro-planktonic species, all generated by Illumina RNA-Sequencing, 406 of which derived from the MMETSP [13] plus the transcriptomes of eight other diatoms, including the two reference diatom species, *Phaeodactylum tricoratum* [18] and *Thalassiosira pseudonana* [19], assembled using an in-house assembly pipeline (Supplementary Fig. S1), were used.

Prediction of candidate lncRNAs using transcriptomic data

The assembled transcriptomes were screened for the prediction and identification of lncRNA-like transcripts in each species (Fig. 1), using a majority voting-based procedure inspired by the ensemble machine learning methods. In order to diversify the ensemble model, 10 protein coding potential prediction tools were selected based on the following criteria: (i) the tools must use a different artificial intelligence (AI) approach; (ii) the tools must provide pretrained models with different fine-tuned classification features; and (iii) the tools should achieve good performance, especially for cross-species prediction, within a reasonable usability score as defined in [20]. In addition, the tools should accept sequences presented in a FASTA format. Hence, among the 10 tools, 3 used a SVM-based model (CPC2 [21], LncFinder [22], and Longdist [23]), 2 were based on XGBOOST (LncDC [24] and

RNAmining [25]), 2 on maximum likelihood estimation and logistic regression (LGC [26] and CPAT [27], respectively), and 3 other were based on deep learning approaches with different intrinsic architectures: convolutional neural network in RNASamba [28], recurrent neural network (RNN) in mRNN [29], and deep belief network in LncADeep [30]. The key characteristics of each selected tool are summarized in the [Supplementary Table S1](#). The selected tools were applied with equal weight, using their default parameters and their default pretrained classification models. The Python package ezLncPred (V.1.0) [31] was used to run the following tools: CPC2, CPAT (*-p Human*), LGC, and Longdist. The remaining tools were run using an in-house script with the following parameters: LncADeep (*-MODE lncRNA*), and the coding potential probability of a transcript was calculated as the mean of probabilities generated by the 21 intrinsic models of the algorithm), lncFinder (*svm.model='human'*), mRNN (mRNN_ensemble, which used the weighted average prediction of the five best single mRNN models), RNAmining (*-organism_name Homo_sapiens*), and RNASamba (full and partial weighted models were included). The FASTA files containing the sequence contigs were used as an input to each tool that predicted both the forward and the reverse strand (with the exception of two diatom species, *P. tricornutum* and *T. pseudonana*, for which the transcriptome was generated from strand-specific RNA-seq data, and thus only the sequence of forward strand was predicted), and were labelled as ‘Coding-like’ or ‘Non-coding-like’. The majority label output from all the tools was submitted to several stringent filters (Pfam and SwissProt hits, putative ORF size, transcript length), and the output was considered as the final coding potential class.

Expression quantification of candidate lncRNAs

The expression level of each lncRNA-like transcript predicted was quantified using Salmon (version 1.10.2) using RNA-seq data downloaded from (<https://ftp.sra.ebi.ac.uk/vol1/fastq>). The expression values were given in transcript per million.

Database implementation

The three-tier client/server architecture model containing data, logic, and presentation layers has been implemented for LncPlankton ([Supplementary Fig. S2](#)). The data layer represents the data storage part that is handled by a relational database ([Supplementary Fig. S2](#)) setup with the popular MySQL (version 5.7.36) open-source relational database management system. The data layer is expanded with NoSQL file storage. The logic layer represents the core of the architecture and is responsible for the communication between the user queries from the presentation layer, fetching the data from the data layer, processing the data, and formatting the response to the presentation layer. The JSON (JavaScript Object Notation)-based data structure is mainly the most used format. In addition, the logic layer is integrated with the following components:

- A BLAST program implemented via the interface rBLAST (<https://github.com/mhahsler/rBLAST>) for online similarity search.
- An RNAfold program implemented via the package LncFinder [22] and RNAPlot implemented via the package RRNA [32] for the calculation and the visualization of secondary structure.

- ORFfinder implemented via LncFinder [22] for the exploration of lncRNA containing sORFs, seqinr package (<https://github.com/cran/seqinr>) for the translation to peptide sequences.

These functionalities were implemented via a web API program provided by the R package Plumber (<https://github.com/rstudio/plumber>). A shiny server function was also developed and was integrated into the logic layer. This function processes the request of the shiny prediction app from the presentation layer and uses the static part of the data layer in addition to the SQL part. The presentation layer contains several modules based on AJAX (Asynchronous JavaScript and XML), jQuery (JavaScript Query system version 3.5.1), and the PHP server-side scripting language (version 7.1.26), as well as the CSS (Cascading Style Sheets) code to describe how HTML elements are to be displayed on user-side web interface. JQuery and AJAX provide methods to perform asynchronous call requests to the logic tier using GET and POST methods, parsing the JSON response, and dynamically rendering the browser display.

Implementation of a web application for lncRNAs prediction

The shiny application was built in R (V.4.3.1) using the shiny framework. The app currently depends on the following R packages: shiny, shinyWidgets, reshape2, wesanderson, dplyr, ggplot2, ggthemes, tidyverse, ggrepel, shinybusy, shinyjs, DT, plotly, leaflet, and RMySQL.

Results and discussion

Determining the coding potential of a transcript is a crucial step in the identification of potential lncRNAs, yet it represents a complex task due to overlapping characteristics and functions that exist between coding and non-coding RNAs [14]. To overcome this challenge and avoid the generation of false positives and false negatives, we decided to use a combination of multiple tools in a majority voting setting to increase the robustness and confidence of the screening.

Establishment of a majority voting meta-learner tool to identify candidate lncRNAs

Ten coding potential prediction classifiers ([Supplementary Table S1](#)) were combined into a single meta-learner. The selected tools used different AI algorithms, provided models pretrained with diverse fine-tuned features, and achieved good performance within a reasonable runtime. Furthermore, two tools considering full and partial sequence length were included, mRNN [29] and LncADeep [30]. In addition, the tools selected presented a good usability score, which is based on ease of use in installation and running of the tool [20]. A transcript was considered as ‘non-coding-like’ within a single tool only if both strands were labelled ‘non-coding-like’, otherwise it was assigned the ‘coding-like’ label. The majority label output from all the tools was considered as the final coding potential class. Alongside the majority class, a non-coding potential score was calculated as the number of tools labelling the transcript as ‘non-coding-like’ divided by 10. This score was used to calculate the reliability (confidence level) of the lncRNA transcripts identified. The threshold was set to a value = 0.6, meaning that if at least 6 out of 10 tools recognized a tran-

script as non-coding RNA, we considered it more likely to be a non-coding RNA rather than a coding RNA. A score of 1 means that all the tools recognize the transcript as a non-coding RNA, making it the highest-ranking category. We believe the majority voting tool offers a reliable choice for candidate lncRNA identification, beyond what a single tool can offer at this time. In order to increase the robustness of the results, the transcripts classified as lncRNAs by the majority voting approach but with significant hits in either the Pfam (V.35.0) or SwissProt (V.2023.01) databases were filtered out. Transcripts which were ‘non-coding-like’ but had a predicted putative ORF peptide length >100 aa were also filtered out. Finally, ‘non-coding-like’ transcripts were classified as candidate lncRNAs or candidate small non-coding (sns)RNAs based on transcript length (using the conventional 200 nt as threshold; Fig. 1).

Benchmarking the majority voting-based procedure with each of its individual tools

The majority voting-based pipeline (VotingLNC) was tested on different sets of coding and non-coding datasets originating from 18 different species (from *Homo sapiens* to *Caenorhabditis elegans*; Supplementary Table S2) and compared to the state-of-the-art coding potential prediction tools. Each tested dataset was independent of the training sets used in the construction of the pretrained classification models related to each method. Among the organisms tested, 10 datasets were perfectly balanced, containing the same number of coding and non-coding RNA transcripts (Supplementary Table S2). Furthermore, in order to test the robustness of the tools towards biased datasets, one imbalanced plus seven highly imbalanced testing sets were also considered (Supplementary Table S2). Each tool was applied to each tested dataset to predict the classes of the transcripts using the default parameters and pretrained models for each tool. Tool outputs were parsed and analysed with custom R scripts. A cross-tabulation of observed and predicted classes was generated, and the performance metrics including the accuracy, sensitivity, and specificity were calculated using the ‘*confusionMatrix*’ method of the caret package (<https://CRAN.R-project.org/package=caret>). The receiver operating characteristic (ROC) curve as well as its AUC (area under the curve) value was also computed using the ROCR package (<http://rocr.bioinf.mpi-sb.mpg.de>). All tested tools considered coding transcripts as positive and non-coding transcripts as negative sets. Our method showed the highest mean accuracy across all the datasets with a mean = 0.95 (Fig. 2A), and low inter-dataset variability (variance: 0.0014). At the species level, our method outperformed the other tools in seven organisms and performed well in cases where the other tools displayed poor performances; e.g. for *H. sapiens*, the majority voting procedure yields an accuracy of 0.95 while Longdist and RNAmMining yield only 0.56 and 0.54, respectively. For the remaining organisms, our method achieves approximately the same accuracy as the other tools (Supplementary Table S3). In addition, our method showed comparable AUC performance (AUC: 0.92–1.00) with the best performing coding potential methods such as CPAT and CPC2 (Fig. 2B and C). Among the 18 datasets tested, our method showed the highest AUC score in 13 datasets (Fig. 2C). The detailed results regarding accuracy, sensitivity, specificity, and the AUC of all the tools can be found in Supplementary Tables S3–S6, respectively.

Mining for candidate lncRNAs across nine plankton phyla

Using our pipeline, we screened transcriptomic libraries originating from 414 species of marine plankton (microbial eukaryotes). These included the three dominant plankton groups of the modern ocean (diatoms, dinoflagellates, and haptophytes), as well as green algae, ciliates, choanoflagellates, and many other groups, totalling nine different phyla. With the exception of a few species such as the diatom *P. tricoratum* [18], as well as the dinoflagellate *Prorocentrum cordatum* (minimum) CCMP1329 [33], to our knowledge, none of the other species used in this study have previously been examined for their lncRNA portfolio. Globally, we screened 11 623 813 contigs across 414 species, the dinoflagellates (Dinophyta) and diatoms (Bacillariophyta) groups representing almost half of the contigs screened, with 3.4 and 2.4 million contigs, respectively (Fig. 3A). The average contig length varied between 593 nt (dinoflagellates) and 949 nt (Cercozoa) with a median = 651 nt across all species (Fig. 3B). Using the majority voting-based method that combined the 10 coding potential tools with strict prediction criteria (Fig. 1), 2 210 359 candidate lncRNA transcripts were identified, which corresponds to 19% of the full input set (Fig. 4 and Supplementary Fig. S3). Of these transcripts, 239 116 were predicted as non-coding by all the tools and have a non-coding potential score = 1. This corresponds to 2% of the total transcriptomic input set that falls into the ‘high-confidence lncRNAs’ category, as their non-coding potential status is indisputable by the 10 tools used (Fig. 4A). More than 45% of the total candidate lncRNA transcripts identified belong to the two most abundant microeukaryote groups screened, the dinoflagellates (531 029 lncRNAs, 24.1%) and the diatoms (482 473 lncRNAs, 21.8%) (Fig. 4B). Interestingly, the ciliates (Ciliophora) were the group that presented the highest density of high confidence candidate lncRNAs, showing the opposite trend from all the other groups (Supplementary Fig. S4). It would be extremely interesting to characterize further these highly consensual lncRNA candidates by performing conservation analysis, which could provide insights into the evolutionary significance of such occurrence in this specific alveolate group of protists. At the species level, the highest number of candidate lncRNAs was identified in the dinoflagellate *Karenia brevis* Wilson (23 686 lncRNAs) followed by the diatom *Fragilariopsis kerguelensis* (21 011 lncRNAs), while the lowest numbers were found in the Rhizarian *Minchimia chitonis* (83 lncRNAs) and the dinoflagellate *Thoracosphaera heimii* (with only 10 lncRNAs) (see Supplementary Table S7). However, these numbers do not directly relate to the proportion of coding to non-coding RNAs of a given species’ transcriptome. For instance, the Cercozoa species for which the pipeline detected the lowest number of lncRNA candidate (*Minchimia chitonis*, 83 lncRNAs), the fraction of candidate lncRNAs in the total transcriptome screened for this species was 25.6% whereas the one for which the highest number of candidate lncRNAs was detected (*Lotharella globosa* Strain CCCM811, 5 522 lncRNAs), the fraction of lncRNAs was 20.4% (see Supplementary Table S7). In total, over 2 million candidate lncRNAs have been predicted and characterized, around 10% of which with high confidence (i.e. detected by all the prediction tools) (Fig. 4A). Regarding the model diatom *P. tricoratum*, the pipeline detected 6765 candidate lncRNAs, 1064 of which with high confidence. Among the previously annotated 1510 lincRNAs [18], 1344 (89%) overlapped with lncRNAs

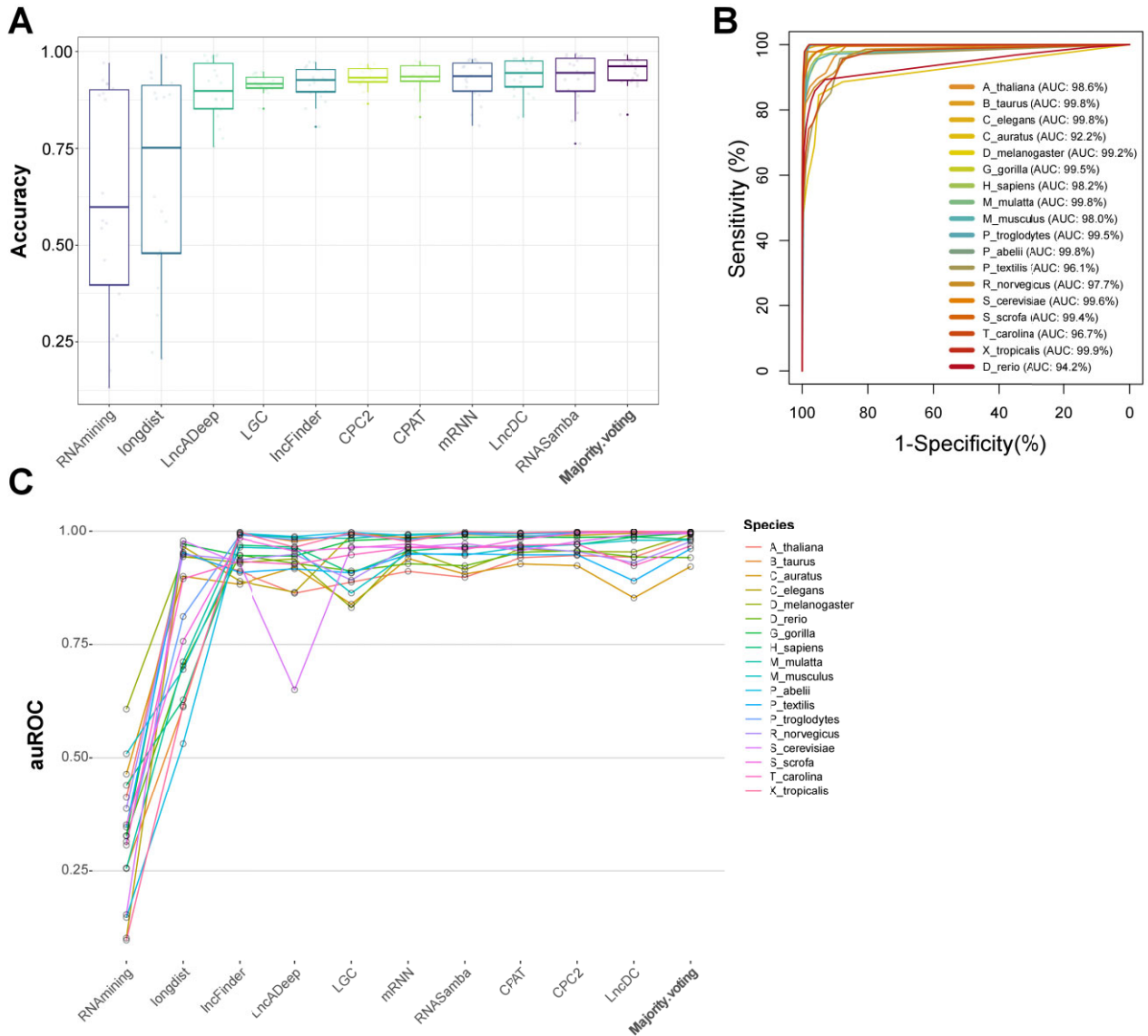


Figure 2. Benchmarking of 10 coding potential prediction methods and the majority voting-based procedure using 18 independent testing sets. **(A)** Distribution of the accuracy across the 11 methods tested; the majority voting method on the last column showed the highest mean accuracy with a small variability. **(B)** ROC curves of the majority voting-based procedure on the 18 datasets tested; the AUC values corresponding to the curves are also reported. **(C)** Distribution of the AUC scores obtained by each method on the 18 datasets.

detected by VotingLNC, however, only 26% were classified as high confidence. A recent report identified a total of 48 039 potential lncRNAs in three dinoflagellate species, one of which (*Prorocentrum minimum/cordatum* CCMP1329) [33] was also screened in LncPlankton. The authors identified 27 568 potential lncRNAs, whereas in LncPlankton, 20 878 lncRNA-like transcripts have been characterized, including 1411 high-confidence lncRNAs. Of the 27 568 candidate lncRNAs identified [33], only 1745 (6.32%) overlap with those present in LncPlankton, of which 168 have high confidence. The small percentage of overlap can be due not only to the difference in the prediction tools used but also to the nature of the transcriptomic library sets that were screened. In particular, the MMETSP transcriptomic datasets from *Prorocentrum minimum/cordatum* CCMP1329 that were used in the present work originate from nutrient stress experiments [13], whereas in [33] they were obtained under heat-stress. lncRNAs are

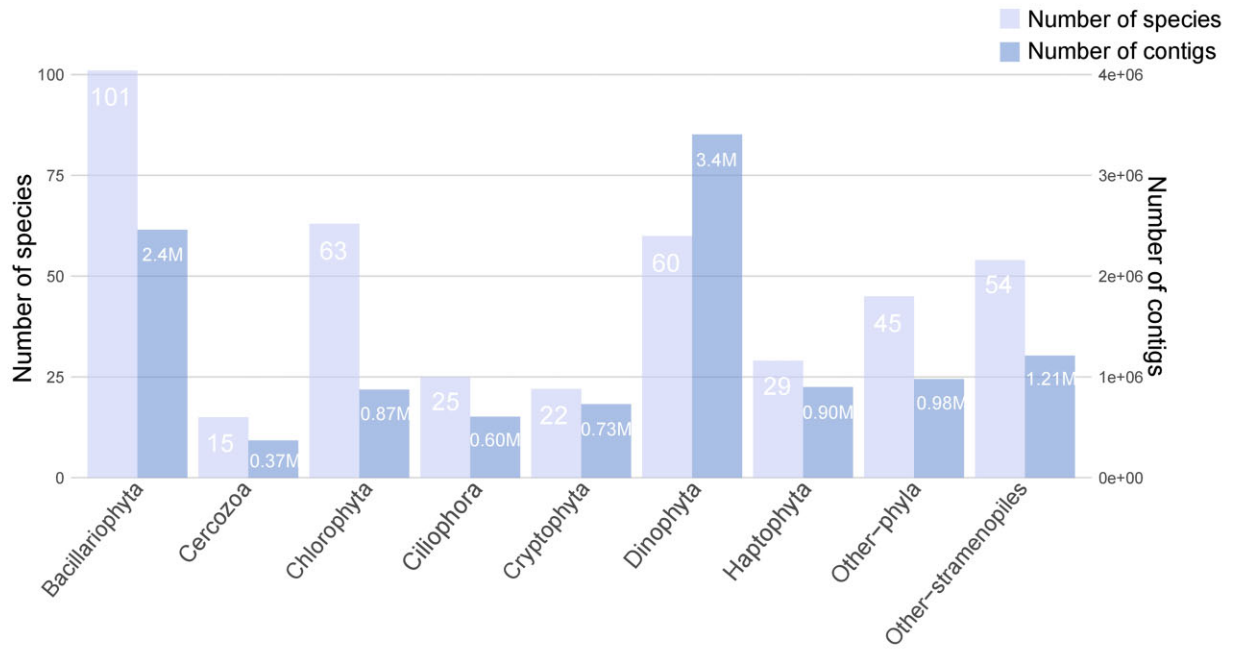
transcripts that are known to be expressed under specific conditions [2], more so than mRNAs [34, 35], which also contributes to explain the little overlap that was found between the two approaches.

All the data generated in LncPlankton are available at <https://www.lncplankton.bio.ens.psl.eu> and the user can freely browse, BLAST, and download the data deposited in FTP bulk or programmatically through dedicated APIs.

Functional modules of LncPlankton website

LncPlankton is available as a user-friendly interface that offers various ways to browse and search lncRNA resources (Fig. 5). The current release of LncPlankton allows querying a given species to explore the content of its transcriptome. The user is required to select a phylum and a species from drop-down lists (Fig. 5A). A page will then appear (Fig. 5B) displaying a summary table with the number of transcripts classified as

A



B

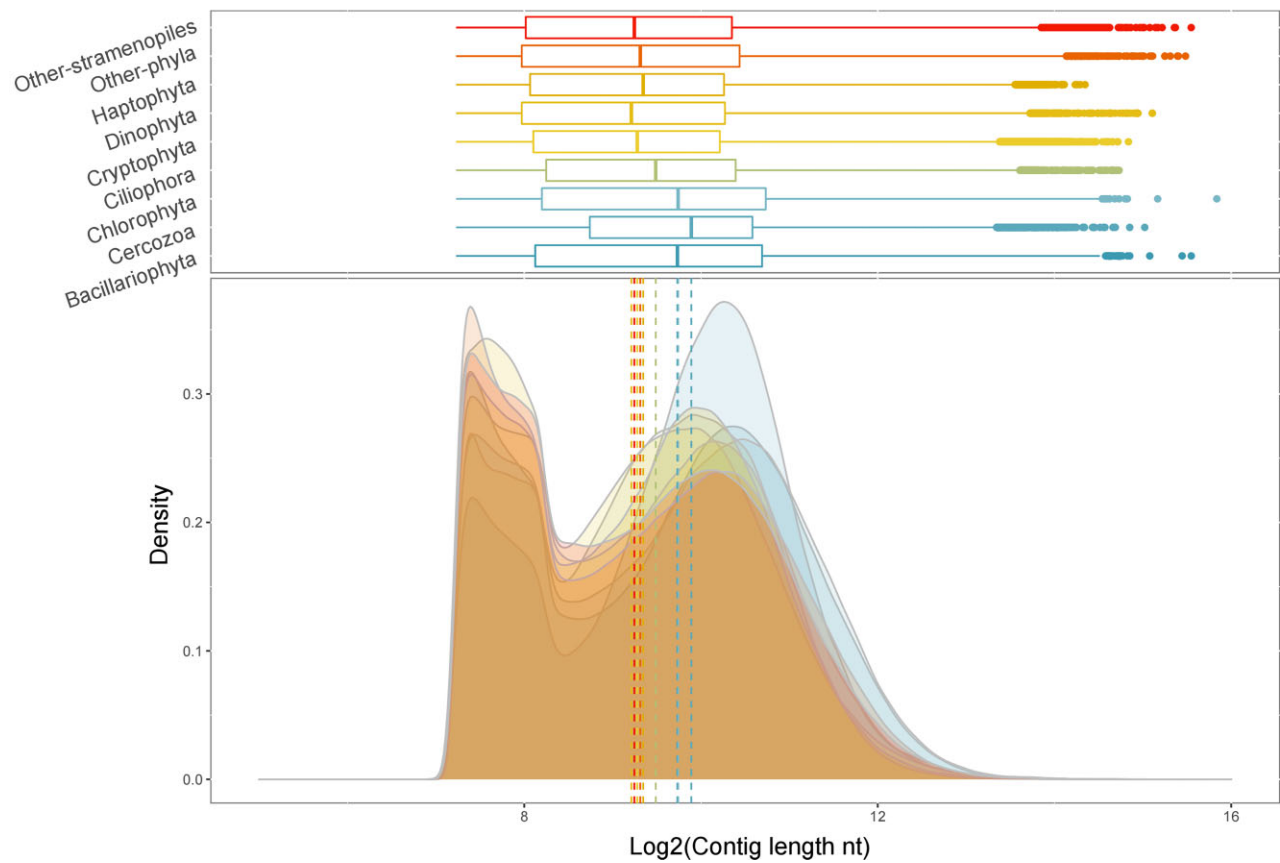


Figure 3. Species and transcript (contigs) used in LncPlankton. **(A)** The distribution of the number of species and the number of assembled contigs across the different phyla; the number of contigs is displayed in millions and reported on the right scale y axis of the graph and **(B)** the density distribution of the scaled \log_2 contig length across phyla; the dotted vertical lines denote the median length.

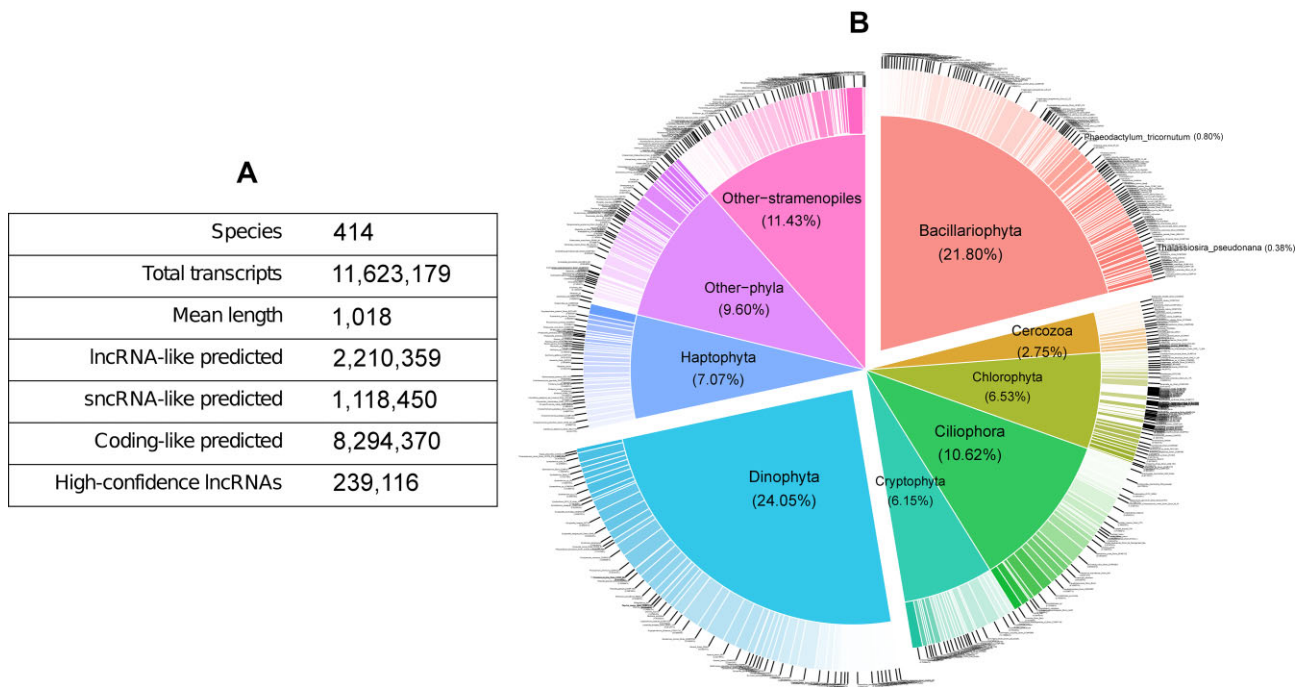


Figure 4. Dataset content of LncPlankton. **(A)** Summary statistics of the content of the database, and **(B)** percentage of transcripts predicted as lncRNA-like by the majority voting pipeline distributed across phyla.

lncRNA-like, sncRNA-like, and coding-like by the majority voting procedure described above (which has been set as the default query). A pie chart reporting the percentage of lncRNAs found is also displayed. The user can also visualize the distribution of lncRNA features such as transcript length, peptide length, GC content, and Fickett score. Additionally, all the information about the lncRNAs found for the selected species are reported in a table with one row per lncRNA. The presented data include lncRNA ID, length, peptide length (if any), Fickett score, iso-electric point, longest ORF length (if any), coverage ORF (if any), GC content, class determined by the pipeline above, and the probability of coding potential generated by the majority voting procedure. To get access to more information on the selected lncRNA, the user can click upon the ‘More...’ button in the last column of the table and an lncRNA detail webpage is displayed with detailed information about the lncRNA selected (Fig 5C). In addition to the basic details of lncRNA sequence, the page displays the expression levels, ORFs, and conceptual translation products of the sequence, alongside with the length of ORFs detected, their coverage information, and the length of the translated peptides. Furthermore, the page shows the one-dimensional dot-bracket notation and the two-dimensional rendering secondary structure of the lncRNA. The minimum free energy (MFE) of the structure folding was also calculated by the RNAfold tool (V.2.5.1) and displayed. A ‘save’ button was added to allow the user to download and save the structure.

As noted above, the lncRNA content of the current release of LncPlankton was generated using the majority voting procedure. To allow the user to modulate the outcome of the screening and customize their predictions (i.e. using a preferred coding potential prediction tool, and/or to modify the cut-off value of the filters), a shiny application was developed and embedded to the UI. The application contains two panels (Fig. 5A): (i) the input panel in which the user can select

the phylum, the species, and the coding potential tool from dedicated drop-down lists. Other inputs can be set by the user such as the transcript length, the peptide length, and whether or not to include the reverse strand in the prediction. The input panel also offers the possibility to select the way the result will be displayed, namely as a histogram, polar plot, table, or map (with the coordinates of the sampling location of the selected species), and (ii) the output panel, which renders the results and displays them according to the user’s choice. All the generated figures and tables can optionally be downloaded and saved. The output panel was implemented using the plotly package (V4.10.1).

We also introduced the possibility for the user to perform a sequence-based search of data stored in LncPlankton using BLAST (Fig. 5D), either through BLASTN or MEGABLAST. Except for the expectation value (E-value) and the number of target-hit sequences, which can be selected in dedicated drop-down lists, the default arguments were used. The BLAST search outputs a raw report that includes pairwise alignment, BLAST hits based upon alignment scores, and other measures of statistical significance. To interactively display the BLAST result, a viewer module was implemented using the BlasterJS library [36] and was integrated to LncPlankton (Fig. 5E).

Both the full and the high-confidence collections of lncRNA-like transcripts can be found on the download page of LncPlankton. The lncRNA collections related to each species can also be separately downloaded in a FASTA format. In addition, the majority voting R package used for the prediction can be retrieved from this page. A Representational State Transfer Application Programming Interface (REST-API) has been implemented using the plumber package and made available to allow programmers to interface directly with LncPlankton. The API returns documents in the JSON format and can be used in any programming language. In the current version of LncPlankton, three APIs

Acknowledgements

This work has used the computational resources of the Bioclust cluster (<https://bioclustg01.bioclust.biologie.ens.fr/>). Hosting and logistic resources have been provided by the computational platform of the ENS Biology department.

Author contributions: Ahmed Debit (Conceptualization [equal], Data curation [equal], Validation [lead], Visualization [lead], Writing – original draft [lead], Writing – review & editing [equal]), Pierre Vincens (Methodology [supporting], Validation [supporting], Visualization [supporting], Writing – review & editing [supporting]), Chris Bowler (Data curation [supporting], Funding acquisition [supporting], Resources [supporting], Writing – review & editing [supporting]), and Helena Cruz de Carvalho (Conceptualization [equal], Data curation [equal], Funding acquisition [lead], Supervision [lead], Writing – review & editing [lead])

Supplementary data

Supplementary data is available at NAR Genomics & Bioinformatics online.

Conflict of interest

None declared.

Funding

French Agence Nationale de la Recherche (ANR DiaLincs 19-CE43-0011-01 to H.C.C.), European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Diatomic; grant agreement no. 835067 to C.B.), French Government 'Investissements d'Avenir' program (MEMO LIFE, ANR-10-LABX-54) and PSL Research University (ANR-11-IDEX-0001-02 to C.B.).

Data availability

The package of the majority voting procedure was implemented in R and can be downloaded on the LncPlankton website at <https://www.lncplankton.bio.ens.psl.eu/files/dwd/tools/votingLNC.tar.gz>. The source code can be found at <https://gitlab.com/a.debit/votingLNC> and has been deposited in figshare under the DOI: <https://doi.org/10.6084/m9.figshare.24799632.v2>. The database is freely available without restrictions for use by academics and non-commercial researchers. The web server is publicly available at <https://www.lncplankton.bio.ens.psl.eu/>. Inquiries concerning the database may be directed to debit@bio.ens.psl.eu or cruz@biologie.ens.fr.

References

- Mattick JS, Amaral PP, Carninci P *et al.* Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat Rev Mol Cell Biol* 2023;24:430–47. <https://doi.org/10.1038/s41580-022-00566-8>
- Derrien T, Johnson R, Bussotti G *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* 2012;22:1775–89. <https://doi.org/10.1101/gr.132159.111>
- Mattick JS, Rinn JL. Discovery and annotation of long noncoding RNAs. *Nat Struct Mol Biol* 2015;22:5–7. <https://doi.org/10.1038/nsmb.2942>
- Cabili MN, Trapnell C, Goff L *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* 2011;25:1915–27. <https://doi.org/10.1101/gad.17446611>
- Dinger ME, Gascoigne DK, Mattick JS. The evolution of RNAs with multiple functions. *Biochimie* 2011;93:2013–8. <https://doi.org/10.1016/j.biochi.2011.07.018>
- Wu P, Mo Y, Peng M *et al.* Emerging role of tumor-related functional peptides encoded by lncRNA and circRNA. *Mol Cancer* 2020;19:22. <https://doi.org/10.1186/s12943-020-1147-3>
- Barshir R, Fishilevich S, Iny-Stein T *et al.* GeneCaRNA: a Comprehensive Gene-centric Database of Human Non-coding RNAs in the GeneCards Suite. *J Mol Biol* 2021;433:166913. <https://doi.org/10.1016/j.jmb.2021.166913>
- Volders P-J, Verheggen K, Menschaert G *et al.* An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res* 2015;43:D174–80. <https://doi.org/10.1093/nar/gku1060>
- Szczeniak MW, Rosikiewicz W, Makalowska I. CANTATAdb: a collection of plant long non-coding RNAs. *Plant Cell Physiol* 2016;57:e8. <https://doi.org/10.1093/pcp/pcv201>
- Gallart AP, Pulido AH, Lagrán IAMd *et al.* GREENC: a Wiki-based database of plant lncRNAs. *Nucleic Acids Res* 2016;44:D1161–6.
- Zhao L, Wang J, Li Y *et al.* NONCODEV6: an updated database dedicated to long non-coding RNA annotation in both animals and plants. *Nucleic Acids Res* 2020;49:D165–71. <https://doi.org/10.1093/nar/gkaa1046>
- RNAcentral Consortium T, Petrov AI, Kay SJE *et al.* RNAcentral: a comprehensive database of non-coding RNA sequences. *Nucleic Acids Res* 2017;45:D128–34. <https://doi.org/10.1093/nar/gkw1008>
- Keeling PJ, Burki F, Wilcox HM *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol* 2014;12:e1001889. <https://doi.org/10.1371/journal.pbio.1001889>
- Singh D, Roy J. A large-scale benchmark study of tools for the classification of protein-coding and non-coding RNAs. *Nucleic Acids Res* 2022;50:12094–111. <https://doi.org/10.1093/nar/gkac1092>
- Duan Y, Zhang W, Cheng Y *et al.* A systematic evaluation of bioinformatics tools for identification of long noncoding RNAs. *RNA* 2021;27:80–98. <https://doi.org/10.1261/rna.074724.120>
- Simopoulos CMA, Weretilnyk EA, Golding GB. Prediction of plant lncRNA by ensemble machine learning classifiers. *BMC Genomics* 2018;19:316. <https://doi.org/10.1186/s12864-018-4665-2>
- Hu J, Andrews B. Distinguishing long non-coding RNAs from mRNAs using a two-layer structured classifiers. In: *IEEE 7th International Conference on Computational Advances in Bio and Medical Sciences (ICCBMS)*. Orlando, FL, USA: IEEE, 2017, 1–5. <https://doi.org/10.1109/iccabs.2017.8114304>
- Cruz de Carvalho MH, Sun H, Bowler C *et al.* Noncoding and coding transcriptome responses of a marine diatom to phosphate fluctuations. *New Phytol* 2016;210:497–510. <https://doi.org/10.1111/nph.13787>
- Goldman JAL, Schatz MJ, Berthiaume CT *et al.* Fe limitation decreases transcriptional regulation over the diel cycle in the model diatom *Thalassiosira pseudonana*. *PLoS One* 2019;14:e0222325. <https://doi.org/10.1371/journal.pone.0222325>
- Ammunét T, Wang N, Khan S *et al.* Deep learning tools are top performers in long non-coding RNA prediction. *Brief Funct Genom* 2022;21:230–41.

21. Kang Y-J, Yang D-C, Kong L *et al.* CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res* 2017;45:W12. <https://doi.org/10.1093/nar/gkx428>
22. Han S, Liang Y, Ma Q *et al.* LncFinder: an integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property. *Brief Bioinform* 2018;20:2009–27. <https://doi.org/10.1093/bib/bby065>
23. Schneider HW, Raiol T, Brigido MM *et al.* A support vector machine based method to distinguish long non-coding RNAs from protein coding transcripts. *BMC Genomics* 2017;18:804. <https://doi.org/10.1186/s12864-017-4178-4>
24. Li M, Liang C. LncDC: a machine learning-based tool for long non-coding RNA detection from RNA-Seq data. *Sci Rep* 2022;12:19083. <https://doi.org/10.1038/s41598-022-22082-7>
25. Ramos TAR, Galindo NRO, Arias-Carrasco R *et al.* RNAmining: a machine learning stand-alone and web server tool for RNA coding potential prediction. *F1000Res* 2021;10:323. <https://doi.org/10.12688/f1000research.52350.2>
26. Wang G, Yin H, Li B *et al.* Characterization and identification of long non-coding RNAs based on feature relationship. *Bioinformatics* 2019;35:2949–56. <https://doi.org/10.1093/bioinformatics/btz008>
27. Wang L, Park HJ, Dasari S *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res* 2013;41:e74. <https://doi.org/10.1093/nar/gkt006>
28. Camargo AP, Sourkov V, Pereira GAG *et al.* RNAsamba: neural network-based assessment of the protein-coding potential of RNA sequences. *NAR Genom Bioinform* 2020;2:lqz024. <https://doi.org/10.1093/nargab/lqz024>
29. Hill ST, Kuintzle R, Teegarden A *et al.* A deep recurrent neural network discovers complex biological rules to decipher RNA protein-coding potential. *Nucleic Acids Res* 2018;46:8105–13. <https://doi.org/10.1093/nar/gky567>
30. Yang C, Yang L, Zhou M *et al.* LncADeep: an *ab initio* lncRNA identification and functional annotation tool based on deep learning. *Bioinformatics* 2018;34:3825–34. <https://doi.org/10.1093/bioinformatics/bty428>
31. Xu X, Liu S, Yang Z *et al.* A systematic review of computational methods for predicting long noncoding RNAs. *Brief Funct Genomics* 2021;20:162–73. <https://doi.org/10.1093/bfpg/elab016>
32. Bida JP, Maher LJ. Improved prediction of RNA tertiary structure with insights into native state dynamics. *RNA* 2012;18:385–93. <https://doi.org/10.1261/rna.027201.111>
33. Chen Y, Dougan KE, Nguyen Q *et al.* Genome-wide transcriptome analysis reveals the diversity and function of long non-coding RNAs in dinoflagellates. *NAR Genom Bioinform* 2024;6:lqae016. <https://doi.org/10.1093/nargab/lqae016>
34. Debit A, Charton F, Pierre-Elies P *et al.* Differential expression patterns of long noncoding RNAs in a pleiomorphic diatom and relation to hyposalinity. *Sci Rep* 2023;13:2440. <https://doi.org/10.1038/s41598-023-29489-w>
35. Cruz de Carvalho MH, Bowler C. Global identification of a marine diatom long noncoding natural antisense transcripts (NATs) and their response to phosphate fluctuations. *Sci Rep* 2020;10:14110. <https://doi.org/10.1038/s41598-020-71002-0>
36. Blanco-Míguez A, Fdez-Riverola F, Sánchez B *et al.* BlasterJS: a novel interactive JavaScript visualisation component for BLAST alignment results. *PLoS One* 2018;13:e0205286. <https://doi.org/10.1371/journal.pone.0205286>