Supplementary Materials - Impact of sequencing technologies on long non-coding RNA computational identification

Alisson G. Chiquitto[1], **Lucas Otávio L. Silva**[1], **Liliane Santana Oliveira**[1], **Douglas S. Domingues**[1,2], **and Alexandre R. Paschoal**[1,*]

[1]Department of Computer Science, Bioinformatics and Pattern Recognition Group, Federal University of Technology - Paraná - UTFPR, Cornélio Procópio-PR, Brazil
[2]Group of Genomics and Transcriptomes in Plants, Institute of Biosciences of Rio Claro, São Paulo State University, Rio Claro-SP, Brazil
[*]corresponding author: paschoal@utfpr.edu.br

Contents

1 Tools and datasets

Table S1. Overview of lncRNA tools used in this report for humans and plants

Tools	Version	Species	Programming language
CNCI[1]	2	*Homo sapiens*	Python
COME[2]	11/2019	*Homo sapiens*	R, Perl
CPAT[3]	3.0.0	*Homo sapiens*	Python, R
CPC2[4]	1.0.1	*Homo sapiens* / Plants	Python
CREMA[5]	06/2020	Plants	Python
LncADeep[6]	1.0	*Homo sapiens*	Python, R
LncMachine[7]	0.1	Plants	Python
lncRNAnet[8]	08/2018	*Homo sapiens*	Python
lncScore[9]	1.0.2	*Homo sapiens*	Python, Perl
PLEK[10]	1.2	*Homo sapiens*	Python
PLncPRO[11]	1.2.2	Plants	Python
RNAmining[12]	1.0.4	*Homo sapiens*	Python
RNAplonc[13]	1.1	Plants	Perl, Python
RNAsamba[14]	0.2.5	*Homo sapiens*	Python

Table S2. Datasets used for the evaluation of the lncRNAs tools

Species	Sample Number	Description/Sample	Bio Sample	Sequences	Database	Data Type	Seq tech
A. trichopoda	Atr Sample 1	Atr leaves[15]	SAMN06320627	91,576	ISOdb	Transcripts	long
A. trichopoda	Atr Sample 2	Atr female flowers[15]	SAMN06320628	55,723	ISOdb	Transcripts	long
A. trichopoda	Atr Sample 3	CANTATAdb 2.0[16]	-	5,511	CANTATAdb	lncRNAs	short
A. thaliana	Ath Sample 1	Ath 4-wk inflorescence 1-2kb[15]	SAMN04456599	17,114	ISOdb	Transcripts	long
			SAMN04456600	27,050			
A. thaliana	Ath Sample 2	Ath 4-wk inflorescence 2-3kb[15]	SAMN04456597	23,579	ISOdb	Transcripts	long
			SAMN04456601	31,069			
A. thaliana	Ath Sample 3	Ath 4-wk inflorescence 3-6kb[15]	SAMN04456598	20,464	ISOdb	Transcripts	long
			SAMN04456602	22,962			
A. thaliana	Ath Sample 4	CANTATAdb 2.0[16]	-	4,373	CANTATAdb	lncRNAs	short
A. thaliana	Ath Sample 5	Ath aboveground parts[17]	SRR10611193	8,258,511	NANOPORE (SRA/NCBI)	Transcripts	long
			SRR10611194	7,849,574			
			SRR10611195	7,025,983			
H. sapiens	Hsa Sample 1	Release 21[18]	-	26,414	GENCODE	lncRNAs	short
H. sapiens	Hsa Sample 2	Release 38[18]	-	48,751	GENCODE	lncRNAs	long
H. sapiens	Hsa Sample 3	Hsa reference sample[19]	NA12878	10,302,647	NANOPORE	Transcripts	long
H. sapiens	Hsa Sample 4	Hsa blood, african male[15]	SAMN00001695	33,651	ISOdb	Transcripts	long
H. sapiens	Hsa Sample 5	Hsa blood, african female[15]	SAMN00001694	64,638	ISOdb	Transcripts	long
			SAMN00001696	79,649			
T. aestivum	Tae Sample 1	Tae leaves 1-2kb[15]	SAMN04456603	12,836	ISOdb	Transcripts	long
T. aestivum	Tae Sample 2	Tae leaves 2-3kb[15]	SAMN04456604	18,830	ISOdb	Transcripts	long
T. aestivum	Tae Sample 3	Tae leaves 3-6kb[15]	SAMN04456605	2,753	ISOdb	Transcripts	long
T. aestivum	Tae Sample 4	NONCODEv6[20]	-	12,427	NONCODE	lncRNAs	short

2 Results for plant datasets

Table S3. Sensitivity on plant lncRNA datasets based on short sequencing technologies

Dataset/Bio Sample	Total	RNAplonc	PLncPRO-mono	PLncPRO-dico	Crema	LncMachine	CPC2
Amborella trichopoda	5,511	97.59%	97.71%	95.84%	62.87%	90.73%	**98.66%**
Arabidopsis thaliana	4,373	98.87%	-	96.34%	38.78%	**99.66%**	98.61%
Triticum aestivum	12,427	96.71%	99.07%	-	41.55%	92.16%	**99.40%**
Mean		97.72%	98.39%	96.09%	47.73%	94.18%	98.89%

Table S4. Relative Frequency of transcripts classified as lncRNA on ISOdb Plants Datasets from long read technologies[15]

Dataset/Bio Sample	Total	RNAplonc	PLncPRO-mono	PLncPRO-dico	Crema	LncMachine	CPC2
Amborella trichopoda							
Atr Sample 1	91,576	68.58%	**76.32%**	61.04%	26.51%	51.08%	65.14%
Atr Sample 2	55,723	77.18%	**84.80%**	70.95%	34.65%	56.75%	74.31%
Arabidopsis thaliana							
Ath Sample 1†	44,164	**38.55%**	-	28.56%	7.49%	33.59%	33.23%
Ath Sample 2†	54,648	**43,67%**	-	35,44%	8,36%	30,64%	36,72%
Ath Sample 3†	43,426	**52.01%**	-	42.94%	11.94%	35.36%	44.84%
Triticum aestivum							
Tae Sample 1	12,836	34.76%	25.23%	-	10.27%	**36.64%**	34.54%
Tae Sample 2	18,830	23.21%	**28.62%**	-	12.36%	25.68%	26.76%
Tae Sample 3	2,753	35.78%	**50.42%**	-	15.03%	25.50%	44.57%

(† mean)

Table S5. Relative Frequency of lncRNAs on plant NANOPORE Datasets

Dataset/Bio Sample	Total	RNAplonc	Crema	LncMachine	CPC2
Ath Sample 5†	23,134,068	**95.02%**	47.71%	84.76%	94.91%

(† average)

3 Results for *Homo sapiens* datasets

Table S6. Performances of Each Tool Applied on GENCODE v21 and v38 Datasets

Tools	v21 Short reads		v38 Long reads		Average	Gain/Lost
	Total	Sensitivity	Total	Sensitivity		
CNCI	26,413	97.40%	48,751	97.54%	97.47%	+0.14%
COME	26,414	95.67%	48,751	95.88%	95.77%	+0.22%
CPAT	26,414	86.74%	48,751	87.67%	87.20%	+1.07%
CPC2	26,414	94.19%	48,751	94.11%	94.15%	-0.08%
LncADeep	26,413	96.58%	48,751	97.68%	96.96%	**+1.14%**
lncRNAnet	26,412	96.48%	48,749	97.34%	96.91%	+0.89%
lncScore	26,124	93.40%	48,463	93.68%	93.54%	+0.30%
PLEK	26,414	98.28%	48,470	93.70%	95.99%	-4.66%
RNAmining	26,414	**98.88%**	48,751	**99.56%**	**99.22%**	+0.69%
RNASamba	26,413	93.20%	48,751	93.58%	93.39%	+0.41%
Mean		95.08%		95.07%	95.08%	-0.01%

Table S7. Relative Frequency of lncRNAs on *Homo sapiens* on ISOdb Datasets

Dataset/Bio Sample	Total	CNCI	LncADeep	lncRNAnet	PLEK	RNAmining
Hsa Sample 4	33,651	63.80%	58.03%	58.66%	49.36%	**97.58%**
Hsa Sample 5†	144,287	61.50%	55.20%	54.99%	41.48%	**97.82%**

(† average)

Table S8. Relative Frequency of lncRNAs on *Homo sapiens* NANOPORE Dataset

Dataset/Bio Sample	Total	CNCI	LncADeep	lncRNAnet	PLEK	RNAmining
Hsa Sample 3	10,302,647	91.06%	77.45%	77.42%	59.15%	**96.82%**

Table S9. Misclassification transcripts results in both GENCODE version dataset from *Homo sapiens*

	v21 short-reads		v38 long-reads	
Dataset size	26,414	-	48,751	-
CNCI	687	2.60%	1,198	2.46%
LncADeep	903	3.42%	1,129	2.32%
lncRNAnet	931	3.52%	1,298	2.66%
PLEK	449	1.70%	3,052	6.26%
RNAmining	295	1.12%	214	0.44%
Total Misclassifcation Entries	2,190	8.29%	5,573	11.43%

The first line of Table S9 shows the number of transcripts in the datasets GENCODE v21 and v38. The following lines show the amount (and percentage) of transcripts misclassified by these tools. The last line presents the sum of the number of sequences without redundancy misclassified among all tools (see Venn Diagram in Figure 7 of the paper). We present the tools in alphabetical order.

4 Commands to run the tools

This section shows de commands to run the tools used in this study. According to the tools authors, we executed all commands in a Linux environment.

CNCI

```
cd CNCI_DIR
python CNCI/CNCI.py -f input.fa -o output_dir -m ve -p 8
```

COME

```
cd COME_DIR/bin
Bin_dir=`pwd|awk '{print $1}'`
bash $Bin_dir/COME_main.sh input.gtf output_dir $Bin_dir human human.model
```

CPAT

```
cd CPAT_DIR
cpat.py -x Human_Hexamer.tsv --antisense -d Human_logitModel.RData \
    --top-orf=5 -g input.fa -o output_dir
```

CPC2

```
cd CPC2_DIR
python ./bin/CPC2.py -i input.fa -o output_dir
```

CREMA

```
cd CREMA_DIR
python predict.py -f input.fa -c output.txt -d diamond_output.txt
```

LncADeep

```
cd LNCADEEP_DIR
python LncADeep.py -MODE lncRNA -f input.fa \
    -o output_dir -th 4
```

LncMachine

```
cd LNCMACHINE_DIR
python lncMachine.py -c input.fa \
    --model prediction_models/features.csv.DecisionTree_model.sav \
    - o output.csv
```

lncRNAnet

```
cd LNCMACHINE_DIR
python lncMachine.py -c input.fa \
    --model prediction_models/features.csv.DecisionTree_model.sav \
    - o output.csv
```

lncScore

```
cd LNCSCORE_DIR
lncScore.py -f input.fa -g input.gtf -o output.csv \
    -p 1 -x dat/Human_Hexamer.tsv -t dat/Human_training.dat
```

PLEK

```
cd PLEK_DIR
lncScore.py -f input.fa -g input.gtf -o output.csv \
    -p 1 -x dat/Human_Hexamer.tsv -t dat/Human_training.dat
```

PLncPRO

```
cd PLNCPRO_DIR
python prediction.py -i input.fa -p pred_res -o output_dir \
    -m models/monocot.model -d ../blastdb/swissprot/swissprot -t 10
python prediction.py -i input.fa -p pred_res -o output_dir \
    -m models/dicot.model -d ../blastdb/swissprot/swissprot -t 10
```

RNAmining

```
cd RNAMINING_DIR
python rnamining.py -f input.fa -organism_name Homo_sapiens \
    -prediction_type coding_prediction -output_folder output_dir
```

RNAplonc

```
cd RNAPLONC_DIR
perl 200nt.pl input.fa
txCdsPredict input_.fasta output.cds
perl feature_extraction.pl input_.fasta output.cds > output.arff
java -cp weka.jar weka.classifiers.trees.REPTree \
    -l RNAplonc/RNAplonc.model -T output.arff -p 0 > result.txt
python FilterResults.py -c output.cds -r result.txt \
    -o result2.txt -p 0.0 -t 1
```

RNAsamba

```
cd RNASAMBA_DIR
classify -v 1 output.tsv input.fa data/full_length_weights.hdf5
```

References

1. Sun, L. *et al.* Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.* **41**, e166, DOI: 10.1093/nar/gkt646 (2013).
2. Hu, L., Xu, Z., Hu, B. & Lu, Z. J. COME: a robust coding potential calculation tool for lncRNA identification and characterization based on multiple features. *Nucleic Acids Res.* **45**, e2, DOI: 10.1093/nar/gkw798 (2017).
3. Wang, L. *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41**, e74, DOI: 10.1093/nar/gkt006 (2013).
4. Kang, Y.-J. *et al.* CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* **45**, W12–W16, DOI: 10.1093/nar/gkx428 (2017).
5. Simopoulos, C. M. A., Weretilnyk, E. A. & Golding, G. B. Prediction of plant lncRNA by ensemble machine learning classifiers. *BMC genomics* **19**, 316, DOI: 10.1186/s12864-018-4665-2 (2018).
6. Yang, C., Zhou, M., Xie, H. & Zhu, H. LncADeep performance on full-length transcripts. *Nat. Mach. Intell.* **3**, 197–198, DOI: 10.1038/s42256-019-0108-2 (2021).
7. Cagirici, H. B., Galvez, S., Sen, T. Z. & Budak, H. LncMachine: a machine learning algorithm for long noncoding RNA annotation in plants. *Funct. & Integr. Genomics* **21**, 195–204, DOI: 10.1007/s10142-021-00769-w (2021).
8. Baek, J., Lee, B., Kwon, S. & Yoon, S. LncRNAnet: long non-coding RNA identification using deep learning. *Bioinformatics* **34**, 3889–3897, DOI: 10.1093/bioinformatics/bty418 (2018).
9. Zhao, J., Song, X. & Wang, K. lncScore: alignment-free identification of long noncoding RNA from assembled novel transcripts. *Sci. Reports* **6**, 34838, DOI: 10.1038/srep34838 (2016).
10. Li, A., Zhang, J. & Zhou, Z. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinforma.* **15**, 311, DOI: 10.1186/1471-2105-15-311 (2014).
11. Singh, U., Khemka, N., Rajkumar, M. S., Garg, R. & Jain, M. PLncPRO for prediction of long non-coding RNAs (lncRNAs) in plants and its application for discovery of abiotic stress-responsive lncRNAs in rice and chickpea. *Nucleic Acids Res.* **45**, e183, DOI: 10.1093/nar/gkx866 (2017).
12. Ramos, T. A. R. *et al.* RNAmining: A machine learning stand-alone and web server tool for RNA coding potential prediction. *bioRxiv* (2021). Type: article.
13. Negri, T. d. C. *et al.* Pattern recognition analysis on long noncoding RNAs: a tool for prediction in plants. *Briefings Bioinforma.* **20**, 682–689, DOI: 10.1093/bib/bby034 (2019).
14. Camargo, A. P., Sourkov, V., Pereira, G. & Carazzolle, M. RNAsamba: neural network-based assessment of the protein-coding potential of RNA sequences. *NAR Genomics Bioinforma.* **2**, DOI: 10.1093/nargab/lqz024 (2020). Lqz024, https://academic.oup.com/nargab/article-pdf/2/1/lqz024/34054199/lqz024.pdf.
15. Xie, S.-Q. *et al.* ISOdb: A Comprehensive Database of Full-Length Isoforms Generated by Iso-Seq. *Int. J. Genomics* **2018**, 1–6, DOI: 10.1155/2018/9207637 (2018).
16. Szcześniak, M. W., Bryzghalov, O., Ciomborowska-Basheer, J. & Makałowska, I. CANTATAdb 2.0: Expanding the Collection of Plant Long Noncoding RNAs. *Methods Mol. Biol. (Clifton, N.J.)* **1933**, 415–429, DOI: 10.1007/978-1-4939-9045-0_26 (2019).
17. Cui, J. *et al.* Analysis and comprehensive comparison of PacBio and nanopore-based RNA sequencing of the Arabidopsis transcriptome. *Plant Methods* **16**, 85, DOI: 10.1186/s13007-020-00629-x (2020).
18. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923, DOI: 10.1093/nar/gkaa1087 (2020). https://academic.oup.com/nar/article-pdf/49/D1/D916/35363795/gkaa1087.pdf.
19. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345, DOI: 10.1038/nbt.4060 (2018).
20. Zhao, L. *et al.* NONCODEV6: an updated database dedicated to long non-coding RNA annotation in both animals and plants. *Nucleic Acids Res.* **49**, D165–D171, DOI: 10.1093/nar/gkaa1046 (2020). https://academic.oup.com/nar/article-pdf/49/D1/D165/35364620/gkaa1046_supplemental_file.pdf.