DNABERT-based explainable lncRNA identification in plant genome assemblies

Monica F. Danilevicz[1§], Mitchell Gill[1§], Cassandria G. Tay Fernandez[1], Jakob Petereit[1], Shriprabha R. Upadhyaya[1], Jacqueline Batley[1], Mohammed Bennamoun[2], David Edwards[1] and Philipp E. Bayer[1 *]

[1] School of Biological Sciences, University of Western Australia, Australia

[2] School of Physics, Mathematics and Computing, University of Western Australia, Australia

[§] These authors contributed equally to this work.

*To whom correspondence should be addressed: philipp.bayer@uwa.edu.au

Highlights

- We demonstrate for the first time the identification of lncRNAs from genomic sequences, instead of transcriptome sequences, allowing the identification of lowly expressed lncRNAs.
- A deep learning model (natural language processing) was employed for the prediction of lncRNAs in two monocot and five dicot plant species.
- We used explainable machine learning to extract the genomic motifs associated with lncRNA identification, highlighting potentially conserved structures present in more than one plant species.

Abstract

Long non-coding ribonucleic acids (lncRNAs) have been shown to play an important role in plant gene regulation, being involved in both epigenetic and transcript regulation. LncRNAs are transcripts longer than 200 nucleotides that are not translated into functional proteins but can be translated into small peptides. Machine learning and deep learning models have predominantly used transcriptome data with manually defined features to detect lncRNAs, however, they often underrepresent the abundance of lncRNAs and can be biased in their detection. Here we present a study using Natural Language Processing (NLP) models to identify plant lncRNAs from genomic sequences rather than transcriptomic data. The NLP models were trained to predict lncRNAs for seven model and crop species (*Zea mays*, *Arabidopsis thaliana*, *Brassica napus*,

Brassica oleracea, *Brassica rapa*, *Glycine max* and *Oryza sativa*) using publicly available genomic references. We demonstrated that lncRNAs can be accurately predicted from genomic sequences, and that genome assembly quality affects the accuracy of lncRNA identification. Furthermore, we demonstrated that the NLP models are applicable for cross-species prediction as they could predict lncRNAs from a species not used to train the model, with an average of 61% accuracy. Finally, we show that the models can be interpreted using explainable artificial intelligence to identify motifs important for lncRNA prediction and that these motifs were frequently present flanking the lncRNA sequence.

Keywords: lncRNAs; natural language processing; deep learning; genomic motif; cross-species prediction

Introduction

Regulatory non-coding ribonucleic acids (ncRNAs) have the capacity to modulate gene expression, with many ncRNAs reported to regulate flowering time and development (Lee et al., 2010; Waheed and Zeng, 2020), abiotic stress tolerance (Thiebaut et al., 2012; Liu et al., 2017) and interaction with beneficial or pathogenic microorganisms (Thiebaut et al., 2014; Salvador-Guirao et al., 2018; Hu et al., 2020). Long non-coding ribonucleic acids (lncRNA), a subset of ncRNAs, are defined as RNA transcripts longer than 200 nucleotides that are not translated into functional proteins, but can sometimes generate small peptides (Ben Amor et al., 2009; Lin et al., 2020). LncRNA expression is often specific to the plant developmental stage and tissue or cell type (Li et al., 2014b; Zhang et al., 2014b; Ward et al., 2015; Gloss and Dinger, 2016), with some transcripts presenting a remarkably short life (Wang et al., 2011a). Once thought to be an evolutionary oddity, lncRNAs have demonstrated roles in essential biological processes. They may act through regulating the expression of neighbouring genes or across multiple loci in the genome, and are also involved in epigenetic regulation (Heo and Sung, 2011; Guil and Esteller, 2012; Urquiaga et al., 2020). For example in rice, the overexpression of lncRNA LAIR can increase grain yield by upregulating the expression of an LRK (leucine-rich repeat receptor kinase) gene cluster, producing larger primary panicles and more panicles per plant (Wang et al., 2018). The lncRNA Ef-cd, transcribed from the antisense strand of the flowering activator OsSOC1 locus, positively regulates the expression of this gene, shortening time to maturity in late-maturing rice by 7-20 days, without causing a decrease in yield, which is usually associated with early maturing varieties (Fang et al., 2019). Other lncRNAs are associated with heat responsive genes in *Brassica rapa* (Wang et al., 2019), osmotic and salt stress in *Medicago truncatula* (Wang et al., 2015), drought stress response in *Zea mays* and *Oryza sativa* (Zhang et al., 2014a; Chen et al., 2021), stress response in *Glycine max* (Golicz et al., 2018b), sexual reproduction in *O. sativa* (Golicz et al., 2018a), pathogen response in *Triticum aestivum* and *Arabidopsis thaliana* (Xin et al., 2011; Zhu et al., 2014; Zhang et al., 2016), and vernalization-mediated epigenetic silencing in *A. thaliana* (Heo and Sung, 2011). Hence the identification of novel lncRNAs may support the breeding of improved crop varieties.

Multiple approaches are employed for lncRNA detection and characterization. The majority of lncRNA prediction bioinformatics pipelines use a combination of in-house scripts to filter potential lncRNAs from transcriptome datasets based on length, open read frame (ORF) size and comparison to known protein coding sequences, followed by a tool to estimate coding potential, such as CPC2 (Kang et al., 2017), LncMachine (Cagirici et al., 2021), CPAT (Wang et al., 2013), PlncPRO (Singh et al., 2017) and PLEK (Li et al., 2014a) (reviewed in (Li et al., 2020)). For the coding potential models, a common approach is to manually define features to be extracted from transcripts and to use these features as input for machine and deep learning models, such as random forest, to categorise transcripts into mRNA or lncRNA (Fan and Zhang, 2015; Pian et al., 2016). However, handcrafted features used for machine learning can introduce bias and reduce classification performance, depending on the relevance of the chosen features (Eraslan et al., 2019; Shrestha and Mahmood, 2019). Another drawback is that current models rely on transcriptome data or computationally predicted lncRNAs available on CANTATAdb (Szcześniak et al., 2019b), PLncDB (Jin et al., 2020), Greenc (Paytuví Gallart et al., 2016; Di Marsico et al., 2021) or Alnc (Singh et al., 2021) to train their models and output a prediction. Since lncRNAs have a low expression rate that can be highly specific to the plant developmental stage or tissue, the identification of lncRNAs from transcriptome datasets may underestimate lncRNA abundance or fail to detect lncRNA presence (Golicz et al., 2018b). A model that could take advantage of publicly available genomic references rather than transcriptomic data to identify lncRNAs would have the potential to identify all lncRNAs, particularly if trained with learned features, avoiding bias introduced by handcrafted features extraction.

Natural Language Processing (NLP) models, such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) and Universal Language Model Fine-tuning (ULMFiT) (Howard and Ruder, 2018), are deep learning models developed to deal with large sequences of words or text data. Transformers employ an attention-based mechanism to train their encoder-decoder architecture (Vaswani et al., 2017), offering an opportunity to initially pretrain a model on genomic data without labelling the dataset beforehand (unsupervised training), avoiding the introduction of annotation bias. DNABERT, GeneBERT and Genomic-ULMFiT are the genomic versions of the NLP frameworks (https://github.com/kheyer/Genomic-ULMFiT) (Ji et al., 2021; Mo et al., 2021). These NLP models have been fine-tuned for the identification of N4-methylcytosine DNA methylation sites in the *Caenorhabditis elegans* genome (Wahab et al., 2021), identification of gene promoters, splice sites and transcription factor binding sites from human genome data (Ji et al., 2021), and disease risk estimation based on genomic sequences and RNA-splicing (Mo et al., 2021). One of the advantages of using pre-trained genomic model is that it can be fine-tuned to a specific classification task using a smaller dataset with high quality annotations. NLP offers an opportunity to achieve direct genome prediction of lncRNAs from genomic data. An advantage of generating models from genomic data is that models trained with high accuracy on high quality data will likely have cross-species applicability, with previous evidence in transcriptomics prediction indicating machine learning models trained on well studied crops can predict across species (Meng et al., 2021). The NLP

models could then be interpreted to determine important motifs in lncRNA prediction within and across plant species to determine evolutionary conserved regulatory motifs. Visualisation and interpretation techniques have been used by DNABERT (Ji et al., 2021) to pre-train models for genomic input and to visualise important motifs for promoter identification through attention scores (Pierse, 2021) for universal applicability to any NLP model that uses transformers. Packages such as these make the leap from predicting promoter regions to lncRNA transcripts feasible with pre-trained BERT language models for interpretation and visualisation.

This study serves as a foundation to perform a genome-wide search and discovery for lncRNAs using a NLP deep learning model. Here, we trained NLP models to predict lncRNA from genomic data of seven model and crop species (*Zea mays, Arabidopsis thaliana, Brassica napus, Brassica oleracea, Brassica rapa, Glycine max and Oryza sativa*) with 70% average accuracy. The cross-applicability of these NLP models was tested by comparing each species' model's performance with lncRNA data from other species. In addition, we further investigate the model features to visualise the most important motifs for lncRNA prediction. Significant motif conservation across species was also investigated, with the majority of the significant motifs located at the edges of the lncRNA sequence of the species analysed.

Results

Prediction performance of the BERT model for seven crop species

Multiple NLP model instances were fine-tuned separately for each species using genomic sequences, the sequences were split into different k-mer sizes from 3-mers to 6-mers. The models were evaluated using accuracy, AUROC, F1, MCC, precision and recall. The accuracy reported for the highest performing model in each species ranged from 57.9% - 83.4% (73.5% median). The value for MCC varied substantially, ranging between 0.16-0.67 (0.4 median), with 1 being considered a perfect prediction. AUROC varied between 61% and 90%, while F1 ranged between 58% and 83%, the precision score ranged between 57% and 84% and a recall score between 57% and 84% (Table 1). Across all classifiers, we observed 30% performance difference, with the *Z. mays* model consistently showing the highest prediction performance, whereas the *B. rapa* model reported the lowest prediction performance. Assessing the impact of k-mer size on model performance, most species achieved the highest performance using the fine-tuned 3-mer model, except for *B. rapa* and *Z. mays* which had the best accuracy using the 4-mer and 6-mer models respectively.

Table 1. The performance metrics of different k-mer length NLP models are fine-tuned for each species. The best metric model for each species is presented in bold text.

Species	K-mer	Accuracy	AUC	F1	MCC	Precision	Recall
Arabidopsis thaliana	**3**	**65.39**	**0.72**	**0.65**	**0.31**	**0.66**	**0.65**
	4	64.82	0.72	0.65	0.30	0.65	0.65
	5	63.3	0.69	0.63	0.27	0.64	0.63
	6	63.88	0.71	0.64	0.28	0.64	0.64
Glycine max	**3**	**72.77**	**0.79**	**0.73**	**0.46**	**0.73**	**0.73**
	4	71.62	0.80	0.72	0.44	0.72	0.72
	5	71.62	0.80	0.72	0.44	0.72	0.72
	6	71.39	0.80	0.71	0.43	0.72	0.71
Brassica napus	**3**	**74.60**	**0.81**	**0.74**	**0.49**	**0.75**	**0.74**
	4	71.90	0.81	0.71	0.44	0.74	0.71
	5	71.90	0.81	0.71	0.44	0.73	0.71
	6	72.40	0.81	0.72	0.46	0.73	0.72
Brassica oleracea	**3**	**74.15**	**0.81**	**0.74**	**0.49**	**0.75**	**0.74**
	4	72.90	0.80	0.73	0.46	0.73	0.73
	5	70.80	0.79	0.71	0.42	0.71	0.71
	6	73.10	0.80	0.72	0.47	0.74	0.73
Brassica rapa	3	57.7	**0.63**	0.57	**0.16**	**0.58**	**0.58**
	4	**57.86**	0.62	**0.58**	0.16	0.58	0.58
	5	56.86	0.60	0.58	0.14	0.57	0.57

	6	56.49	0.60	0.57	0.13	0.57	0.57
Oryza sativa	**3**	**61.65**	**0.66**	**0.62**	**0.23**	**0.62**	**0.62**
	4	61.11	0.65	0.61	0.22	0.61	0.61
	5	58.75	0.64	0.59	0.18	0.59	0.59
	6	58.86	0.63	0.59	0.18	0.59	0.59
Zea mays	3	81.2	0.90	0.81	0.64	0.83	0.81
	4	83.1	0.91	0.83	0.67	0.84	0.83
	5	81.94	0.90	0.82	0.65	0.83	0.82
	6	**83.42**	**0.90**	**0.84**	**0.67**	**0.84**	**0.84**

Confusion matrix analysis is commonly used in machine learning classification tasks to assess which classes are most often misclassified (or “confused”) by the model, it helps identify whether the misclassified samples share similarities and if either class is more often wrongly predicted. Here, the models accurately identified lncRNAs 59.4 to 88.6% of the time, and detected random stretches of DNA between 55.6 to 78.1% of the time (Figure 1) showing the model was equally accurate to predict either class. The *Z. mays* model had the largest dataset with 14030 genomic sequences and achieved the highest classification performance to identify both lncRNAs and no lncRNAs. The two models trained with *G. max* and *O. sativa* had smaller datasets with 3414 and 3716 sequences respectively but performed on average 18.7% better than the *B. rapa* model with a medium size set (9725 genomic sequences).

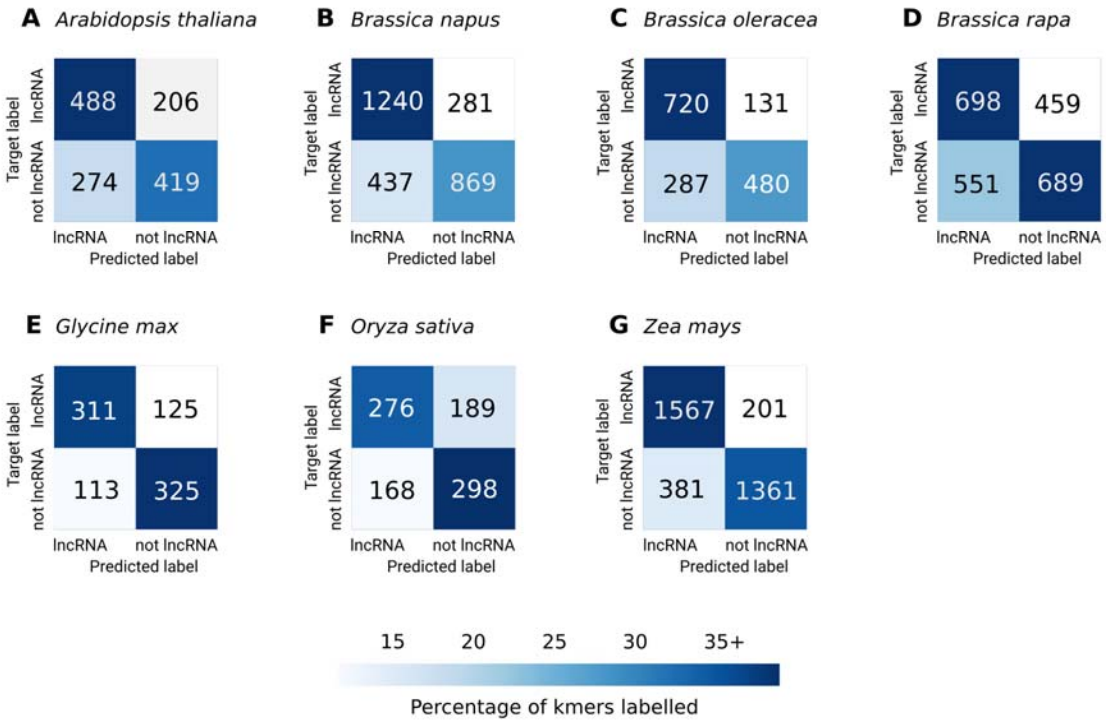

Figure 1: Confusion matrices of the predicted labels from each species' highest performing models. The matrix is coloured based on the percentage of k-mers labelled in each class from the total test set of the species indicated in the title.

Cross-species prediction of lncRNAs using the single species fine-tuned NLP model

NLP models with the highest accuracy for each crop were used to predict the lncRNAs in another species. For example, the 3-mers model trained with *A. thaliana* genomic k-mers was employed to predict lncRNAs in the remaining six species without further fine-tuning. Most of the models performed similarly or slightly worse at predicting lncRNAs for a new species in comparison to identifying lncRNAs in their original species, except for the model trained on *Z. mays* data that performed significantly worse at classifying other species' lncRNAs (Table 2, Figure 2). In all cases, the highest lncRNA classification performance was achieved by models trained in that species instead of cross-species prediction models.

In addition, we assessed how a model trained on a monocot (*Z. mays* and *O. sativa*) or dicot (*A. thaliana*, *B. rapa*, *B. oleracea*, *B. napus* and *G. max*) species would perform when predicting lncRNAs from the other group. Here, the models performed better when predicting lncRNAs from the same group, with average accuracy of 66.2% for monocot-monocot prediction, and 62.7% for dicot-dicot prediction. Models performing lncRNA prediction in a different group (monocot-dicot and dicot-monocot) presented an average accuracy of 60.3%.

Table 2: Performance metrics of cross-species lncRNA prediction.

Trained Species	Test Species	K-mer	Accuracy	AUC	F1	MCC	Precision	Recall
A. thaliana	*B. napus*	3	0.67	0.73	0.67	0.34	0.67	0.67
	B. oleracea	3	0.67	0.74	0.67	0.33	0.67	0.67
	B. rapa	3	0.56	0.55	0.55	0.11	0.55	0.55
	G. max	3	0.65	0.72	0.65	0.30	0.65	0.65
	O. sativa	3	0.53	0.56	0.53	0.06	0.53	0.53
	Z. mays	3	0.58	0.64	0.57	0.18	0.60	0.59
B. napus	*A. thaliana*	3	0.65	0.71	0.65	0.30	0.65	0.65
	B. oleracea	3	0.72	0.81	0.72	0.45	0.73	0.72
	B. rapa	3	0.54	0.56	0.54	0.08	0.54	0.54
	G. max	3	0.69	0.76	0.69	0.38	0.69	0.69
	O. sativa	3	0.58	0.64	0.58	0.17	0.59	0.58
	Z. mays	3	0.66	0.72	0.65	0.32	0.66	0.66
B. oleracea	*A. thaliana*	3	0.64	0.71	0.63	0.28	0.65	0.64
	B. napus	3	0.71	0.79	0.7	0.41	0.71	0.70
	B. rapa	3	0.53	0.55	0.53	0.07	0.53	0.53
	G. max	3	0.66	0.74	0.65	0.32	0.66	0.66
	O. sativa	3	0.59	0.62	0.58	0.18	0.59	0.59
	Z. mays	3	0.69	0.76	0.69	0.38	0.69	0.69
B. rapa	*A. thaliana*	4	0.55	0.58	0.55	0.10	0.55	0.55
	B. napus	4	0.61	0.66	0.61	0.21	0.61	0.61

	B. oleracea	4	0.61	0.65	0.61	0.22	0.61	0.61
	G. max	4	0.59	0.61	0.59	0.18	0.59	0.59
	O. sativa	4	0.57	0.59	0.57	0.14	0.57	0.57
	Z. mays	4	0.58	0.62	0.58	0.16	0.58	0.58
G. max	*A. thaliana*	3	0.60	0.63	0.60	0.19	0.60	0.60
	B. napus	3	0.68	0.74	0.68	0.36	0.68	0.68
	B. oleracea	3	0.66	0.73	0.66	0.33	0.66	0.66
	B. rapa	3	0.51	0.51	0.50	0.01	0.50	0.50
	O. sativa	3	0.58	0.62	0.57	0.16	0.58	0.58
	Z. mays	3	0.67	0.74	0.67	0.35	0.67	0.67
O. sativa	*A. thaliana*	3	0.56	0.60	0.55	0.13	0.57	0.56
	B. napus	3	0.60	0.66	0.60	0.23	0.62	0.61
	B. oleracea	3	0.60	0.66	0.59	0.22	0.62	0.61
	B. rapa	3	0.53	0.53	0.50	0.05	0.53	0.52
	G. max	3	0.60	0.64	0.58	0.20	0.61	0.60
	Z. mays	3	0.60	0.64	0.58	0.19	0.60	0.59
Z. mays	*A. thaliana*	6	0.60	0.64	0.59	0.20	0.61	0.60
	B. napus	6	0.68	0.74	0.67	0.36	0.69	0.67
	B. oleracea	6	0.68	0.72	0.67	0.35	0.68	0.67
	B. rapa	6	0.50	0.52	0.50	0.01	0.50	0.50
	G. max	6	0.69	0.77	0.69	0.39	0.70	0.69
	O. sativa	6	0.61	0.64	0.59	0.25	0.64	0.61

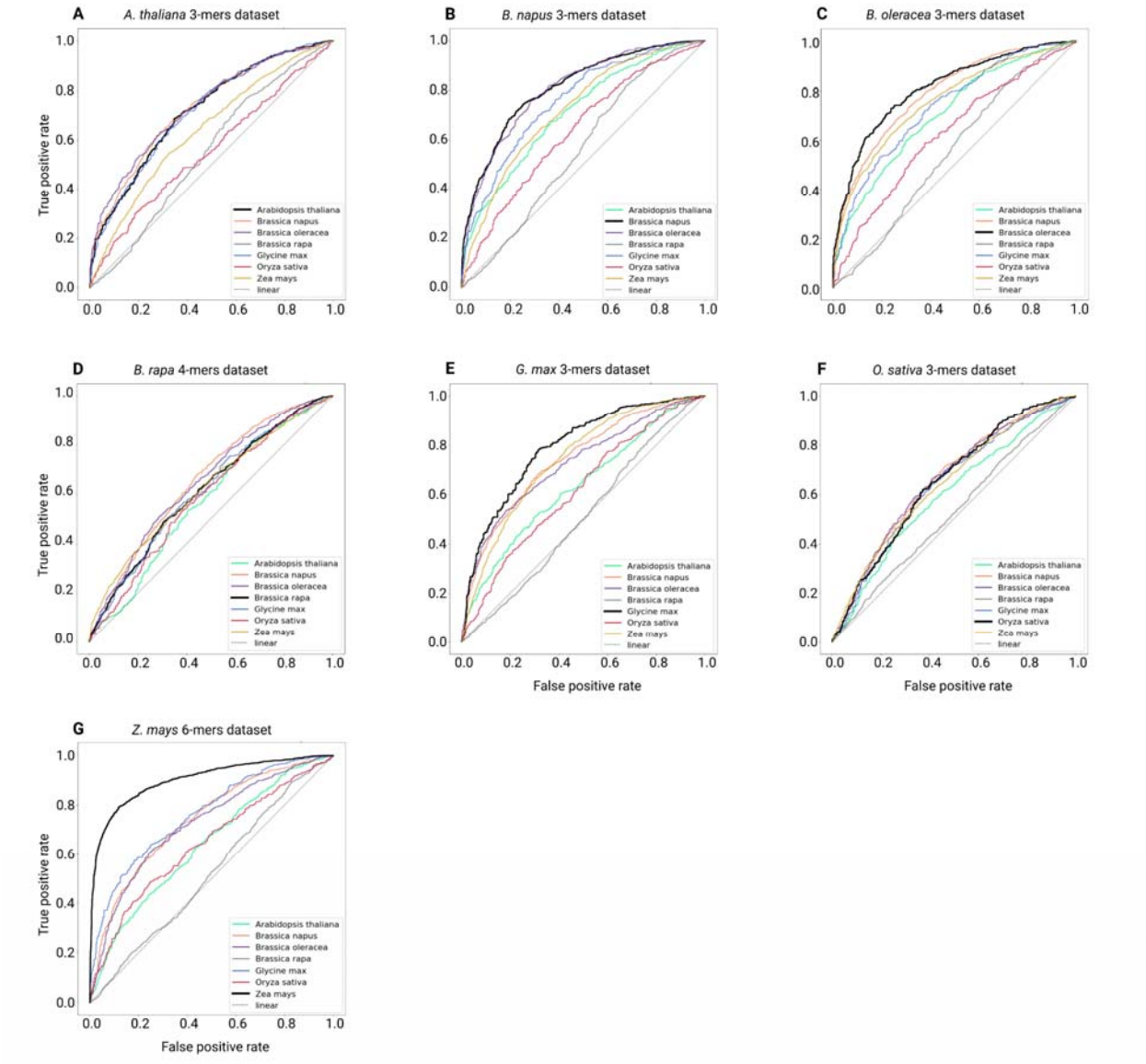

Figure 2: AUROC of each fine-tuned model when performing cross-species prediction of lncRNAs. The black line indicates model performance when predicting lncRNAs in the species it was trained, whereas other species are represented by the colours indicated in the plot legend.

Motif identification and frequency of motifs appearing per genomic location

The characterization of lncRNAs is a challenging task as lncRNAs share many similarities with other RNA types. Thus, to uncover potentially conserved patterns in lncRNA sequences, we extracted and ranked the most relevant genomic features employed by the best performing models for the classification of lncRNA. The best performing model for each crop (ranked by accuracy) was interpreted using the transformers interpret package (Pierse, 2021) using the

lncRNA sequences from the holdout dataset. The 50 most important genomic motifs for lncRNA identification were subjected to a chi-squared test against non-lncRNAs motifs to determine the significant motifs (p-value < 0.05). The resulting 158 unique motifs were compared across species according to sequence similarity (Figure 3, Supplementary Table 1-2). Apart from *O. sativa*, that shared no significant motifs with any of the other species, all species shared two lncRNA motifs (AAAAAAA and AAAAAAAA) belonging to the same motif cluster. The *Brassica* species shared five motifs, whereas *B. napus* and *B. oleracea* shared twelve motifs between them. *Z. mays* and *G. max* shared eleven motifs, the second highest number of motifs shared between two species. Visualising the location of significant motifs in lncRNA divided into 25 bp bins of upstream & downstream flanking regions and bins representing 5% lncRNA sequence, showed that significant motifs important to lncRNA appeared more frequently in the upstream and downstream flanking regions than in the lncRNA sequences themselves (Figure 4, Supplementary Figure 1).

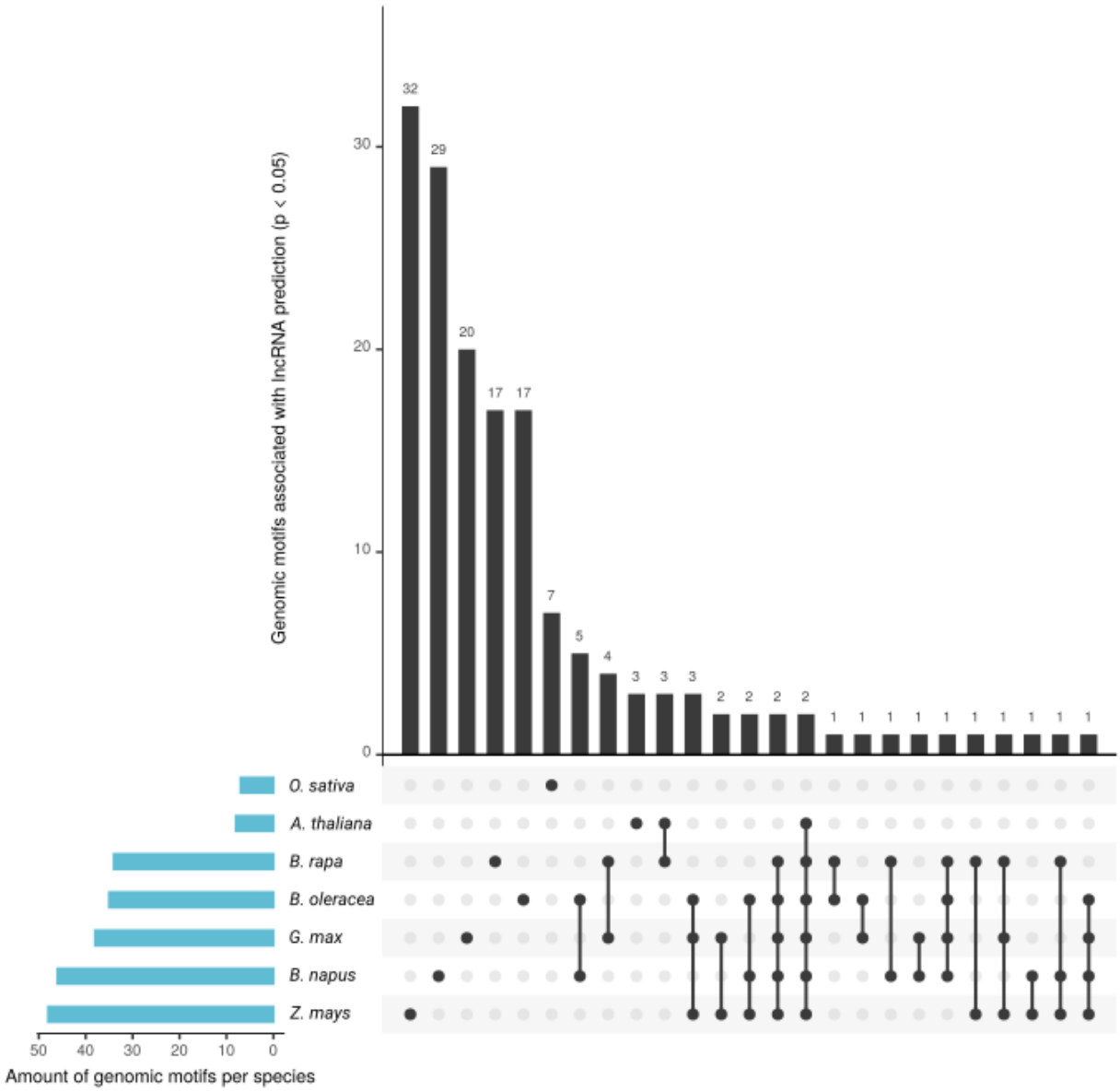

Figure 3: Shared genomic motifs across different plant species. The main plot indicates the number of significant genomic motifs (p-value < 0.05) that are shared or unique to the species. The number of significant genomic motifs identified per species is indicated in the lateral plot in blue.

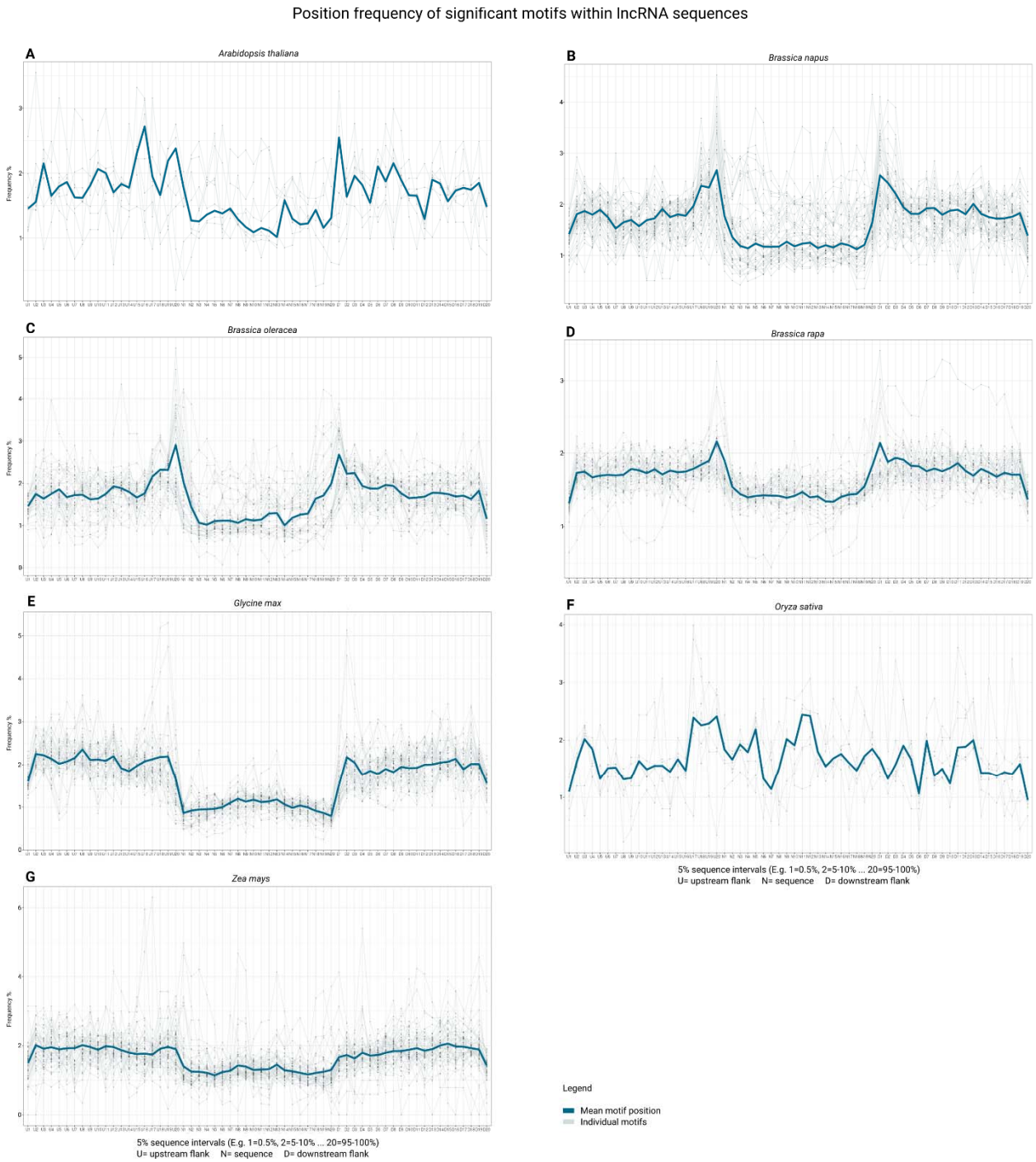

Figure 4: Significant (p-value < 0.05) genomic motif sequence location for the identification of lncRNAs in each species based on the highest performance single species classification model.

Discussion

In this study, we fine-tuned multiple NLP instances to successfully identify lncRNAs within the genome sequence of seven plant species without the need for transcriptome data, and used these models for cross-species prediction of lncRNAs. The code used to develop this project is available on Github (https://github.com/mitchgill16/lncRNA_Prediction_Interpretation) along with the trained model weights, allowing other researchers to implement the model with their own datasets. Identifying lncRNA transcripts is challenging as lncRNAs tend to have lower expression than other transcripts, and lncRNA expression is often specific to the development stage, tissue or cell type (Li et al., 2014b; Zhang et al., 2014b; Ward et al., 2015; Gloss and Dinger, 2016) causing lncRNAs to go undetected when using RNA-seq based models (Li et al., 2014b). As an approach to address this limitation, we identified lncRNAs directly from the genomic sequence using known lncRNAs and genome assemblies. The ability to identify lncRNAs from genomic sequences can further our understanding of their sequence conservation across plant species (Golicz et al., 2018b), providing a foundation to investigate lncRNA evolution and identify novel transcripts (Fico et al., 2019). The NLP models fine-tuned per species presented a median accuracy of 73.5%, with the model trained with *Z. mays* performing 25.6% better than the model for *B. rapa*. Several factors may have contributed to the higher performance of the *Z. mays* model, including the high quality of genome assembly. The *Z. mays* AGP RefGen 4v genome assembly was constructed using advanced single molecule technology, improving contig length and supporting the accurate assembly of intergenic spaces and centromeres (Jiao et al., 2017), providing higher quality sequences for model training. Centromeric regions are reported to encode actively transcribed lncRNAs (Golicz et al., 2018b) that have been associated with centromere maintenance and cellular division (Rošić and Erhardt, 2016). In our study, *Z. mays* also had the largest available dataset of predicted lncRNAs after filtering (Szcześniak et al., 2019a), which supports the accurate training of machine learning models as they may better represent transcript diversity (Barbedo, 2018). The lncRNAs from *B. rapa* and *O. sativa* were consistently misclassified across the different models, which may indicate that different features were more important for these species. The difficulty to accurately identify lncRNAs from these species may also be associated with genome assembly and annotation quality. *B. rapa* was the first species sequenced among the *Brassica* genus in 2011 (Wang et al., 2011b), and the *O. sativa* genome was annotated before RNA sequencing was widely available, and so these genome assemblies and annotations maybe missing lncRNA regions (Kawahara et al., 2013; Golicz et al., 2018a). It was previously reported in the *O. sativa* genome annotation that many lncRNA loci have been incorrectly classed as putative protein coding genes (Paytuví Gallart et al., 2016). Further research will measure the impact of genome assembly on model performance to assess the importance of assembly methods on lncRNA identification. The algorithm architecture applied here is susceptible to dataset errors, as it learns directly from the data presented, thus re-training the model as more validated lncRNA sequences are released will lead to higher model accuracy. Since the model presented here is trained using known lncRNAs identified through bioinformatics pipelines, it is also difficult to assess its

ability to identify novel lncRNAs, though the predicted set may provide a foundation for subsequent validation.

Although lncRNAs do not show the same degree of interspecies sequence conservation as protein-coding genes (Johnsson et al., 2014), a study using ten different monocot and dicot species showed that the majority of lncRNAs have some sequence conservation at the intra- and sub-species level, and the number of lncRNAs identified correlated with the number of genes in each species (Deng et al., 2018). There is also evidence of structural and positional conservation across animal lncRNAs showing similar expression patterns suggesting transcript functional conservation (Hezroni et al., 2015; Tavares et al., 2019). In our study, the models trained on monocot or dicot species had on average a 3-6% higher accuracy than when predicting lncRNAs from other species in the same group, suggesting differences in the features employed for the prediction in these groups. The trained NLP models identified multiple genomic motifs that are significantly associated with lncRNA prediction. For the majority of species, the genomic motifs for lncRNA identification were more frequently located at the boundaries of the lncRNA sequence (Figure 3), and presented a low GC content in comparison to the lncRNA sequences, with several being composed of poly-A or poly-T stretches. Characterising the role of these conserved motifs in lncRNA expression is beyond the scope of this study and the datasets, however these motifs may contribute to the identification of lncRNA expression regulation in future research. In addition, as more conserved motifs are identified, these may act as novel genomic markers allowing a more precise identification of lncRNAs from other plant species.

Conclusion

We present the first study using a NLP model to successfully identify lncRNAs in plant species from genomic sequences instead of transcriptomic data. This method allows for genome-wide detection of lncRNAs regardless of their expression and can contribute to our understanding of lncRNA conservation and function across plant species. The models developed here are able to identify the majority of lncRNAs from the species used for training, and across different species, with the *Z. mays* model presenting the highest accuracy. Further research may include a wider diversity of genomic sequences to increase model robustness and achieve better classification performance in the genome-wide identification of lncRNAs.

Methods

Data Acquisition

GTF files containing locations of lncRNA for *Arabidopsis thaliana, Brassica napus, Brassica oleracea, Brassica rapa, Glycine max, Oryza sativa* and *Zea mays* were downloaded from the cantataDB2.0 (Szcześniak et al., 2019) at the URL: http://cantata.amu.edu.pl/download.php. Reference genome fastas were downloaded from Ensembl Plants (release 37) for each of the above crops with the version matching the used reference on the cantataDB2.0 download page. These were as follows; Arabidopsis_thaliana.TAIR10.dna.toplevel.fa, Brassica_napus.AST_PRJEB5043_v1.dna.toplevel.fa, Brassica_oleracea.v2.1.dna.toplevel.fa, Brassica_rapa.IVFCAASv1.dna.toplevel.fa, Glycine_max.V1.0.dna.toplevel.fa, Oryza_sativa.IRGSP-1.0.dna.toplevel.fa and Zea_mays.AGPv4.dna.toplevel.fa.

BlastX Filtering

The lncRNAs fasta sequences were extracted using bedtools getfasta query (Quinlan and Hall, 2010) using the reference genome fasta for the respective species. A BLAST database for each species was built using makeblastdb query and the genome fasta file (Camacho et al., 2009) Next, the lncRNA fasta sequences for each species was compared to the respective genome database through blastn (outfmt 6, evalue 1e-10) to filter out sequences with high similarity to known mRNAs. The extract_blast.py python script was used to extract lncRNAs with over 90% of its length matched to a mRNA sequence. The blastx_pipeling.sh script can be used to complete these steps efficiently.

Adding Flanking sequence & Generating a Control Set

The lncRNA genomic context was incorporated in the sequence using the add_flanks.py script, which adds the 500bp flanking regions upstream and downstream of the gtf file. The script ensured the flanking regions did not exceed chromosome or scaffold boundaries using the index file generated with samtools faidx (Li et al., 2009).

The control dataset, labelled “non-lncRNA”, was composed from random stretches of genomic. After filtering the lncRNA plus flanking regions from the genome file using bedtools shuffle (-excl lncRNA gtf file), the control sequences were extracted from the reference genome running bedtools getfasta with random intervals of equivalent length to the flanked lncRNA fasta sequences.

Fine-tuning of DNABERT Models

The 3-mer, 4-mer, 5-mer and 6-mer pretrained BERT models for DNA sequences (DNABERT) were cloned from the DNABERT (Ji et al., 2021) repository (https://github.com/jerryji1993/DNABERT). The models were trained on GPU servers using a

container environment using a tensorflow image (version ‘tensorflow:20.03-tf2-py3’) through singularity (Kurtzer et al., 2017). Jupyter notebooks are available at https://github.com/mitchgill16/lncRNA_Prediction_Interpretation along with custom python scripts.

Packages were installed as described in the DNABERT repository with the exception of pytorch (Paszke et al., 2019) version 1.4 and Biopython version 1.78 (Cock et al., 2009). Functions from Genomic-ULMFiT (https://github.com/kheyer/Genomic-ULMFiT) were used to parse fasta files and split datasets into training and testing partitions. The input sequences were converted to k-mers using the ‘seq2k-mer’ function from DNABERT.

The DNABERT models were fine-tuned using run_fine-tune.py with the following settings; model_type=dnalongcat, tokenizer_name=dna<k-mer of model>, task_name=dnaprom,--do_train, --do_eval, max_seq_length=2048, per_gpu_eval_batch_size=4, per_gpu_train_batch_size=4, learning_rate=2e-5, num_train_epochs=4.0, --evaluate_during_training, logging_steps=100, save_steps=4000, warmup_percent=0.1, hidden_dropout_prob=0.1, --overwrite_output, weight_decay=0.01, n_process=8. The resulting fine-tuned model was evaluated on the separate testing set.

For cross species prediction, the best performing k-mer model for each crop was used to predict the separate test set of every other plant with the same evaluation metrics.

Evaluation metrics

Model performance was assessed using accuracy, area under receiver operating characteristics curve (AUROC), F-score (F1), Mathew’s correlation coefficient (MCC), Precision and Recall as the evaluation metrics calculated using the scikit-learn library (Fabian et al., 2011). AUROC is estimated by plotting true positive rate vs false positive rate. The other metrics are defined below, where TP, FP, TN and TP mean true positive, false positive, true negative and true positive:

Accuracy $$\frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

F1 $$\frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (2)$$

MCC $$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)\,(TP + FN)\,(TN + FP)\,(TN + FN)}} \quad (3)$$

Precision $$\frac{TP}{TP + FP} \quad (4)$$

Recall $$\frac{TP}{TP + FN} \quad (5)$$

Interpretation of Models

The transformers interpret repository (https://github.com/cdpierse/transformers-interpret) was used to analyse the model classification results. The scripts atributions.py and explainer.py from this repository were adapted to increase the maximum sequence length of the input from 512 to 2048.

Each lncRNA labelled sequence from the test dataset was iterated over with the transformers interpret sequence classification explainer object. This object returned attribution scores for each k-mer in the sequence. If more than 7 successive k-mers had positive attribution scores for lncRNA prediction, these k-mers were merged and their attribution scores summed until a k-mer with a negative attribution score was present. This merged k-mer would be annotated as a motif important for lncRNA prediction. After iterating over every lncRNA sequence for a given plant species, the frequency of a motif occurring as a positive prediction factor was multiplied by the sum of a motif's attribution score to give a ranking of the motif's importance to lncRNA prediction for each DNABERT model. The top 50 motifs were extracted for visualisation.

Visualisation of Significant Motif Location

The motifs contained in the lncRNA (plus flanking regions) and non-lncRNA sequences were extracted and counted using grep and awk commands. Then, a custom R script (chi_script.R) was used to run a chi-square test to assess whether the motif appeared in lncRNA sequences significantly more than non-lncRNA sequences. Motifs with a p-value less than 0.05 were kept whereas every other motif was excluded.

The parse_fasta function from Genomic_ULMFiT was used to count the locations of significant motifs in lncRNA and non-lncRNA sequences. Each lnRNA sequence was divided into 60 bins. Bins 1-20 were 5% intervals of the upstream flanking region, bins 21-40 were 5% intervals of the lncRNA sequence, and bins 41-60 were 5% intervals of the downstream flanking region. The counts bins 21-40 that refer to the lncRNA sequence, were normalised to counts of bins 1-20 and bins 41-60, as the lncRNA sequence was of variable length whereas the flanks were a consistent 500bp. This ensured motif counts on longer lncRNAs sequences were not biased by the sequence length. The bins and their counts were analysed with the tidyverse suite (Wickham et al., 2019) and plotted with ggplot2 (Wickham, 2011) in R.

Acknowledgments

We would like to thank the Pawsey Supercomputing Centre for computation resources. The Australian Government supported this work through the Australian Research Council (Projects DP210100296, DP200100762, and DE210100398) and the Grains Research and Development Corporation (Projects 9177539 and 9177591). Monica F. Danilevicz, Mitchell Gill and Cassandria G. Tay Fernandez were supported by the Research Training Program scholarship. The Forrest Research Foundation further supported Monica F. Danilevicz.

Conflicts of interest

The authors declare no conflict of interest.

Bibliography

Barbedo JGA (2018) Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. Computers and Electronics in Agriculture **153**: 46–53

Ben Amor B, Wirth S, Merchan F, Laporte P, d'Aubenton-Carafa Y, Hirsch J, Maizel A, Mallory A, Lucas A, Deragon JM, et al (2009) Novel long non-protein coding RNAs involved in Arabidopsis differentiation and stress responses. Genome Res **19**: 57–69

Cagirici HB, Galvez S, Sen TZ, Budak H (2021) LncMachine: a machine learning algorithm for long noncoding RNA annotation in plants. Funct Integr Genomics **21**: 195–204

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics **10**: 421

Chen J, Zhong Y, Qi X (2021) LncRNA TCONS_00021861 is functionally associated with drought tolerance in rice (Oryza sativa L.) via competing endogenous RNA regulation. BMC Plant Biol **21**: 410

Deng P, Liu S, Nie X, Weining S, Wu L (2018) Conservation analysis of long non-coding RNAs in plants. Sci China Life Sci **61**: 190–198

Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Di Marsico M, Paytuvi Gallart A, Sanseverino W, Aiese Cigliano R (2021) GreeNC 2.0: a comprehensive database of plant long non-coding RNAs. Nucleic Acids Res. doi: 10.1093/nar/gkab1014

Eraslan G, Avsec Ž, Gagneur J, Theis FJ (2019) Deep learning: new computational modelling techniques for genomics. Nat Rev Genet **20**: 389–403

Fabian P, Gael V, Alexandre G, Vincent M, Bertrand T (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12, 2825-2830.

Fang J, Zhang F, Wang H, Wang W, Zhao F, Li Z, Sun C, Chen F, Xu F, Chang S, et al (2019) Ef-cd locus shortens rice maturity duration without yield penalty. Proc Natl Acad Sci USA **116**: 18717–18722

Fan X-N, Zhang S-W (2015) lncRNA-MFDL: identification of human long non-coding RNAs by fusing multiple features and using deep learning. Mol Biosyst **11**: 892–897

Fico A, Fiorenzano A, Pascale E, Patriarca EJ, Minchiotti G (2019) Long non-coding RNA in stem cell pluripotency and lineage commitment: functions and evolutionary conservation. Cell Mol Life Sci **76**: 1459–1471

Gloss BS, Dinger ME (2016) The specificity of long noncoding RNA expression. Biochim Biophys Acta

1859: 16–22

Golicz AA, Bhalla PL, Singh MB (2018a) MCRiceRepGP: a framework for the identification of genes associated with sexual reproduction in rice. Plant J **96**: 188–202

Golicz AA, Singh MB, Bhalla PL (2018b) The long intergenic noncoding RNA (lincrna) landscape of the soybean genome. Plant Physiol **176**: 2133–2147

Guil S, Esteller M (2012) Cis-acting noncoding RNAs: friends and foes. Nat Struct Mol Biol **19**: 1068–1075

Heo JB, Sung S (2011) Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. Science **331**: 76–79

Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I (2015) Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. Cell Rep **11**: 1110–1122

Howard J, Ruder S (2018) Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146

Hu G, Hao M, Wang L, Liu J, Zhang Z, Tang Y, Peng Q, Yang Z, Wu J (2020) The Cotton miR477-CBP60A Module Participates in Plant Defense Against Verticillium dahlia. Mol Plant Microbe Interact **33**: 624–636

Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X, Chin C-S, et al (2017) Improved maize reference genome with single-molecule technologies. Nature **546**: 524–527

Jin J, Lu P, Xu Y, Li Z, Yu S, Liu J, Wang H, Chua N-H, Cao P (2020) PLncDB V2.0: a comprehensive encyclopedia of plant long noncoding RNAs. Nucleic Acids Res. **49** D1489-D1495. doi: 10.1093/nar/gkaa910

Ji Y, Zhou Z, Liu H, Davuluri RV (2021) DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics.**37.15** 2112-2120 doi: 10.1093/bioinformatics/btab083

Johnsson P, Lipovich L, Grandér D, Morris KV (2014) Evolutionary conservation of long non-coding RNAs; sequence, structure, function. Biochim Biophys Acta **1840**: 1063–1071

Kang Y-J, Yang D-C, Kong L, Hou M, Meng Y-Q, Wei L, Gao G (2017) CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res **45**: W12–W16

Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu J, Zhou S, et al (2013) Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. Rice (N Y) **6**: 4

Kurtzer GM, Sochat V, Bauer MW (2017) Singularity: Scientific containers for mobility of compute. PLoS ONE **12**: e0177459

Lee H, Yoo SJ, Lee JH, Kim W, Yoo SK, Fitzgerald H, Carrington JC, Ahn JH (2010) Genetic framework for flowering-time regulation by ambient temperature-responsive miRNAs in Arabidopsis. Nucleic Acids Res **38**: 3081–3093

Lin X, Lin W, Ku Y-S, Wong F-L, Li M-W, Lam H-M, Ngai S-M, Chan T-F (2020) Analysis of Soybean Long Non-Coding RNAs Reveals a Subset of Small Peptide-Coding Transcripts. Plant Physiol **182**: 1359–1374

Liu Q, Yang T, Yu T, Zhang S, Mao X, Zhao J, Wang X, Dong J, Liu B (2017) Integrating Small RNA Sequencing with QTL Mapping for Identification of miRNAs and Their Target Genes Associated with Heat Tolerance at the Flowering Stage in Rice. Front Plant Sci **8**: 43

Li A, Zhang J, Zhou Z (2014a) PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. BMC Bioinformatics **15**: 311

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**: 2078–2079

Li J, Zhang X, Liu C (2020) The computational approaches of lncRNA identification based on coding potential: Status quo and challenges. Comput Struct Biotechnol J **18**: 3666–3677

Li L, Eichten SR, Shimizu R, Petsch K, Yeh C-T, Wu W, Chettoor AM, Givan SA, Cole RA, Fowler JE, et al (2014b) Genome-wide discovery and characterization of maize long non-coding RNAs. Genome Biol **15**: R40

Meng X, Liang Z, Dai X, Zhang Y, Mahboub S, Ngu DW, Roston RL, Schnable JC (2021) Predicting transcriptional responses to cold stress across plant species. Proc Natl Acad Sci USA. doi: 10.1073/pnas.2026330118

Mo S, Fu X, Hong C, Chen Y, Zheng Y, Tang X, Shen Z, Xing EP, Lan Y (2021) Multi-modal Self-supervised Pre-training for Regulatory Genome Across Cell Types. arXiv preprint arXiv:2110.05231

Paytuví Gallart A, Hermoso Pulido A, Anzar Martínez de Lagrán I, Sanseverino W, Aiese Cigliano R (2016) GREENC: a Wiki-based database of plant lncRNAs. Nucleic Acids Res **44**: D1161-6

Pian C, Zhang G, Chen Z, Chen Y, Zhang J, Yang T, Zhang L (2016) LncRNApred: Classification of Long Non-Coding RNAs and Protein-Coding Transcripts by the Ensemble Algorithm with a New Hybrid Feature. PLoS ONE **11**: e0154567

Pierse C (2021) Transformers interpret. Model explainability that works seamlessly with transformers, https://github.com/cdpierse/transformers-interpret. Accessed on 21 Feb 2021

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**: 841–842

Rošić S, Erhardt S (2016) No longer a nuisance: long non-coding RNAs join CENP-A in epigenetic centromere regulation. Cell Mol Life Sci **73**: 1387–1398

Salvador-Guirao R, Baldrich P, Weigel D, Rubio-Somoza I, San Segundo B (2018) The MicroRNA miR773 Is Involved in the Arabidopsis Immune Response to Fungal Pathogens. Mol Plant Microbe Interact **31**: 249–259

Shrestha A, Mahmood A (2019) Review of deep learning algorithms and architectures. IEEE Access **7**: 53040–53065

Singh A, Vivek AT, Kumar S (2021) AlnC: An extensive database of long non-coding RNAs in angiosperms. PLoS ONE **16**: e0247215

Singh U, Khemka N, Rajkumar MS, Garg R, Jain M (2017) 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

Szcześniak MW, Bryzghalov O, Ciomborowska-Basheer J, Maka\lowska I (2019a) CANTATAdb 2.0: expanding the collection of plant long noncoding RNAs. Plant Long Non-Coding RNAs. Humana Press, New York, NY, pp 415–429

Szcześniak MW, Bryzghalov O, Ciomborowska-Basheer J, Makałowska I (2019b) Cantatadb 2.0: expanding the collection of plant long noncoding rnas. Methods Mol Biol **1933**: 415–429

Tavares RCA, Pyle AM, Somarowthu S (2019) Phylogenetic Analysis with Improved Parameters Reveals Conservation in lncRNA Structures. J Mol Biol **431**: 1592–1603

Thiebaut F, Rojas CA, Almeida KL, Grativol C, Domiciano GC, Lamb CRC, Engler J de A, Hemerly AS, Ferreira PCG (2012) Regulation of miR319 during cold stress in sugarcane. Plant Cell Environ **35**: 502–512

Thiebaut F, Rojas CA, Grativol C, Motta MR, Vieira T, Regulski M, Martienssen RA, Farinelli L, Hemerly AS, Ferreira PCG (2014) Genome-wide identification of microRNA and siRNA responsive to endophytic beneficial diazotrophic bacteria in maize. BMC Genomics **15**: 766

Urquiaga MC de O, Thiebaut F, Hemerly AS, Ferreira PCG (2020) From trash to luxury: the potential role of plant lncrna in DNA methylation during abiotic stress. Front Plant Sci **11**: 603246

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017). Attention is all you need. In Advances in neural information processing systems 5998-6008

Wahab A, Tayara H, Xuan Z, Chong KT (2021) DNA sequences performs as natural language processing by exploiting deep learning algorithm for the identification of N4-methylcytosine. Sci Rep **11**: 212

Waheed S, Zeng L (2020) The Critical Role of miRNAs in Regulation of Flowering Time and Flower Development. Genes (Basel). **11(3)** 319 doi: 10.3390/genes11030319

Wang A, Hu J, Gao C, Chen G, Wang B, Lin C, Song L, Ding Y, Zhou G (2019) Genome-wide analysis of long non-coding RNAs unveils the regulatory roles in the heat tolerance of Chinese cabbage (Brassica rapa ssp.chinensis). Sci Rep **9**: 5002

Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie BR, Protacio A, Flynn RA, Gupta RA, et al (2011a) A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature **472**: 120–124

Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W (2013) CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res **41**: e74

Wang T-Z, Liu M, Zhao M-G, Chen R, Zhang W-H (2015) Identification and characterization of long non-coding RNAs involved in osmotic and salt stress in Medicago truncatula using genome-wide high-throughput sequencing. BMC Plant Biol **15**: 131

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun J-H, Bancroft I, Cheng F, et al (2011b) The genome of the mesopolyploid crop species Brassica rapa. Nat Genet **43**: 1035–1039

Wang Y, Luo X, Sun F, Hu J, Zha X, Su W, Yang J (2018) Overexpressing lncRNA LAIR increases

grain yield and regulates neighbouring gene cluster expression in rice. Nat Commun **9**: 3516

Ward M, McEwan C, Mills JD, Janitz M (2015) Conservation and tissue-specific transcription patterns of long noncoding RNAs. Journal of Human Transcriptome **1**: 2–9

Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, et al (2019) Welcome to the tidyverse. JOSS **4**: 1686

Wickham H (2011) ggplot2. WIREs Comp Stat **3**: 180–185

Xin M, Wang Y, Yao Y, Song N, Hu Z, Qin D, Xie C, Peng H, Ni Z, Sun Q (2011) Identification and characterization of wheat long non-protein coding RNAs responsive to powdery mildew infection and heat stress by using microarray analysis and SBS sequencing. BMC Plant Biol **11**: 61

Zhang H, Hu W, Hao J, Lv S, Wang C, Tong W, Wang Y, Wang Y, Liu X, Ji W (2016) Genome-wide identification and functional prediction of novel and fungi-responsive lincRNAs in Triticum aestivum. BMC Genomics **17**: 238

Zhang W, Han Z, Guo Q, Liu Y, Zheng Y, Wu F, Jin W (2014a) Identification of maize long non-coding RNAs responsive to drought stress. PLoS ONE **9**: e98958

Zhang Y-C, Liao J-Y, Li Z-Y, Yu Y, Zhang J-P, Li Q-F, Qu L-H, Shu W-S, Chen Y-Q (2014b) Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice. Genome Biol **15**: 512

Zhu Q-H, Stephen S, Taylor J, Helliwell CA, Wang M-B (2014) Long noncoding RNAs responsive to Fusarium oxysporum infection in Arabidopsis thaliana. New Phytol **201**: 574–584