LIFT: lncRNA identification and function-prediction tool

Deshpande, S., Shuttleworth, J., Yang, J., Taramonli, S. & England, M.

DOI 10.1504/IJBRA.2021.120535
ISSN 1744-5485
ESSN 1744-5493

Publisher: Inderscience

LIFT: lncRNA identification and function-prediction tool

Sumukh Deshpande

Central Biotechnology Services (CBS),
College of Biomedical and Life Sciences,
Cardiff University,
Sir Geraint Evans Building (Room 1/14),
Heath Park, Cardiff, CF14 4XN, UK
Email: deshpandes1@cardiff.ac.uk

James Shuttleworth

School of Computing, Electronics and Maths,
Coventry University,
Coventry, CV1 2JH, UK
Email: csx239@coventry.ac.uk

Jianhua Yang

Department of Computer Science,
University of Warwick,
6, Lord Bhattacharyya Way,
Coventry, CV4 7EZ, UK
Email: jianhua.email@gmail.com

Sandy Taramonli

Faculty of Engineering, Environment and Computing,
School of Computing, Electronics and Mathematics,
Coventry University,
Coventry, CV1 2JH, UK
Email: ab7680@coventry.ac.uk

Matthew England

Faculty of Engineering, Environment and Computing,
School of Computing, Electronics and Mathematics,
Faculty Research Centre for Fluid and Complex Systems,
Coventry University,
Coventry, CV1 2JH, UK
Email: ab9797@coventry.ac.uk

AQ1: Please indicate who the corresponding author is.

Abstract: Long non-coding RNAs (lncRNAs) are a class of non-coding RNAs which play a significant role in several biological processes. Accurate identification and sub-classification of lncRNAs is crucial for exploring their characteristic functions in the genome as most coding potential computation (CPC) tools fail to accurately identify, classify and predict their biological functions in plant species. In this study, a novel computational framework called LncRNA identification and function prediction tool (LIFT) has been developed, which implements least absolute shrinkage and selection operator (LASSO) optimisation and iterative random forests classification for selection of optimal features, a novel position-based classification (PBC) method for sub-classifying lncRNAs into different classes and Bayesian-based function prediction approach for annotating lncRNA transcripts. Using LASSO, LIFT selected 31 optimal features and achieved 15–30% improvement in the prediction accuracy on plant species when evaluated against state-of-the-art CPC tools. Using PBC, LIFT successfully identified the intergenic and antisense transcripts with greater accuracy in *A. thaliana* and *Z. mays* datasets. The predicted functions were verified with published experimental results. The source code is publicly available together with relevant data on GitHub: https://github.com/deshpan4/LIFT.

AQ2: Please reduce abstract of no more than 150 words.

Keywords: lncRNA; long non-coding RNAs; LASSO; least absolute shrinkage and selection operator; iterative random forests; PBC; position-based classification; BMRF; Bayesian Markov random fields; function prediction.

Reference to this paper should be made as follows: Deshpande, S., Shuttleworth, J., Yang, J., Taramonli, S. and England, M. (xxxx) 'LIFT: lncRNA identification and function-prediction tool', *Int. J. Bioinformatics Research and Applications*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Sumukh Deshpande is currently a Core Bioinformatician at Cardiff University with PhD in Computing. He has worked in several academic and industrial organisations in the field of next-generation sequencing data analysis. His expertise spans across bioinformatics application development, bioinformatics programming, NGS data analysis, web programming, data visualisation and machine learning.

James Shuttleworth received his PhD from Coventry University, Coventry, UK, in 2008. He is currently a Director of Institute of Coding Programme for the School of Computing, Coventry University. During his PhD studies, he was engaged in research on colour texture analysis in automated systems for assisting in dysplasia analysis. His research interests include data visualisation and processing in pervasive systems and wireless sensor networks, mapping of sensed phenomena, 3D visualisation and wireless sensor network middleware.

Jianhua Yang is currently a Senior Teaching Fellow at University of Warwick. He has also worked as a lecturer in the School of Computing, Electronic and Mathematics at Coventry University. Educated in computational intelligence background, his expertise span across soft computing, system biology, and web technologies. He is a member of the British Computer Society and IEEE Computer Society.

Sandy Taramonli received her BSc in Computer Science and an MSc in Networks Security. She holds a PhD in Engineering from the University of Warwick. She is currently a Senior Lecturer in Cyber Security and Digital Forensics at Coventry University. Her research is focused on the use of stochastic methods in low energy encryption and network forensics.

Matthew England is an Associate Professor in Faculty of Engineering, Environment and Computing at Coventry University, UK, with PhD in Mathematics. His research expertise is in symbolic computation and computer algebra systems; their applications both direct (e.g., biology, chemistry) and in other areas of computer science (satisfiability checking, programming over complex numbers); and the application of machine learning to improve them.

This paper is a revised and expanded version of a paper entitled [title] presented at [name, location and date of conference].

AQ3: If a previous version of your paper has originally been presented at a conference please complete the statement to this effect or delete if not applicable.

1 Background

Recent advances in genome sequencing have led to the discovery of thousands of non-coding RNA transcripts. Using RNA sequencing (RNA-seq) and epigenome sequencing, a new class of RNA transcripts i.e., long non-coding RNAs (lncRNAs) is defined as those having transcript length > 200 nucleotides. Although this class of RNA lacks protein-coding ability, they have been found involved in the regulation of biological processes such as enzymatic activity regulation, genomic loci imprinting, transcription, translation, cellular differentiation (Liu et al., 2015). Several lncRNA databases such as GENCODE and NONCODE have been developed for storage of lncRNAs (Harrow et al., 2012; Zhao et al., 2016). These databases provide valuable resources for further identification of novel lncRNAs from genomic sequences. Even though NGS techniques such as RNA-seq are actively used for identification and discovery of novel lncRNAs, identification of lncRNAs and functions of lncRNAs in non-model plant organisms need to be discovered.

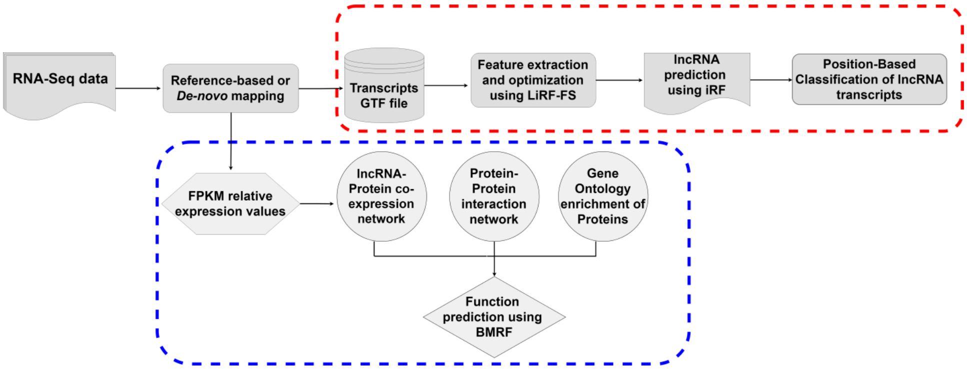
Computational prediction of lncRNAs has been viable for the past few years. These methods generally use machine learning approaches to classify RNAs into different classes. Several tools have been developed including: coding potential calculator (CPC) (Kong et al., 2007), coding-non-coding index (CNCI) (Sun et al., 2013), coding potential assessment tool (CPAT) (Wang et al., 2013), predictor of lncRNAs and messenger RNAs based on improved k-mer scheme (PLEK) (Li et al., 2014) for computational prediction of lncRNAs. The CPC is based on a support vector machine (SVM). Some tools such as CPAT and lncScore (Zhao et al., 2016) classified protein-coding and non-coding transcripts using sequence-based features such as open-reading frame (ORF) size, ORF length, ORF coverage, GC content, Fickett score and Hexamer score whereas others such as CNCI and LncRNA-MFDL (Fan and Zhang, 2015) classified lncRNAs using adjoining nucleotide triplets (ANT) features to identify most-like CDS (MLCDS) regions in each transcript.

Currently developed sequence alignment-based approaches often require significant computational resources due to which the usage of such tools becomes computationally impractical. In contrast, alignment-free methods compute the coding potential scores (CPS) depending on the intrinsic features of the input RNA transcript sequences such as relative oligonucleotide frequencies or *k*-mer. However, CPS tools relying on computation of *k*-mer frequencies require longer computation times and computational resources similar to alignment-based approaches. While, most CPS tools perform well on reference datasets such as GENCODE and NONCODE, they often fail to perform reasonably on the sequences derived from RNA-seq datasets. Identification of lncRNA sub-class (i.e., intergenic, antisense, etc.) provides valuable details about functional mechanisms and regulatory functions. Currently developed machine-learning based methods for sub-classification of lncRNA genes utilises annotated lncRNA transcripts for constructing learning set. Prediction on the lncRNA sequences derived from the plant species often fluctuates due to limited availability of the confirmed lncRNA transcripts from model and non-model plant datasets. Additionally, recent advances in lncRNA function prediction primarily focus on mammalian datasets where genome annotation and co-expression data are easily available (Jiang et al., 2015; Xiao et al., 2015; Perron et al., 2017). Therefore, less attention has been paid on lncRNA identification and functional prediction on model and non-model plant transcriptome datasets.

In recent years, several experiments have shown regulatory roles of lncRNAs in various fundamental and biological processes including cell differentiation, proliferation, apoptosis, epigenetic regulation, transcription, translation, genomic splicing and more (Guttman et al., 2009; Khalil et al., 2009; Mercer et al., 2009). Based on the assumption that similar lncRNA functions are associated with similar diseases, several computational methods have been reported for computing functional similarity or determining lncRNA-disease association (Chen, 2015; Chen et al., 2016a, 2016b). Chen (2015) proposed a lncRNA functional similarity calculation tool based on information of miRNA (LFSCM) which integrates disease semantic similarity, known lncRNA-miRNA interactions and miRNA-disease association. For prediction of disease association, a hypergeometric distribution test was implemented which calculates a P-value indicating significance of commonly shared miRNAs between lncRNA and a disease. Using a similar method, Chen et al. later implemented Improved Random Walk with Restart (IRWR) for predicting lncRNA-disease association (Chen et al., 2016a). Another similar computational tool called FMLNCSIM proposed by Chen et al. (2016b) was developed for determining function similarities of lncRNAs by combining fuzzy measure with the concepts of information content. These were used for calculating similarities among the diseases. The aforementioned methods need prior information of known experimentally verified lncRNA-miRNA interactions and are primarily developed for humans thereby limiting there usage on plant species. Based on co-expression of lncRNAs, Guo et al. (2013) proposed bi-coloured network-based global function predictor for function prediction of lncRNAs (lnc-GFP). Based on re-annotated microarray data (Affymetrix Mouse Genome 430 2.0 Array), coding-noncoding co-expression network was constructed.

In this work, we have developed the LncRNA identification and function prediction tool (LIFT) for lncRNA identification, genomic sub-classification and functional prediction of lncRNAs (Figure 1) in plant RNA-seq datasets. For lncRNA identification, LIFT implements 73 sequence and codon-bias based features. The framework implements an optimisation module called LASSO iterative random forest-feature selection (LiRFFS) (Tibshirani, 1996; Basu et al., 2018) which selects an optimal feature set from training and validation set features. The selected feature set can be applied on the test dataset for sequence prediction using an iRF classifier. For sub-classification of lncRNAs, LIFT implements a position-based classification (PBC) algorithm which classifies the sequences on direction, type and position categories by mapping the exonic (E) and intronic (I) lncRNA sequence coordinates to mRNA E and I coordinates. Inspired by the work undertaken by Kourmpetis et al. (2010) for function prediction of protein-coding genes, the function prediction module of LIFT utilises the lncRNA-mRNA co-expression data from transcriptomic datasets. Based on the co-expression and protein-protein interaction data, molecular and regulatory functions of lncRNAs can be inferred based on Bayesian Markov random fields (BMRF) approach (Kourmpetis et al., 2010). We benchmarked the accuracy of LIFT with existing tools and demonstrated its applicability on lncRNA sequences with diverse lengths. We also assessed the sub-classification performance of PBC and function prediction results with published lncRNA annotations. The transcript sequences classified and annotated by the computational methods will provide an extensive catalogue of molecular regulatory mechanisms.

Figure 1 Workflow of LIFT framework for identification and functional prediction of lncRNAs . The first component (coloured red) identifies lncRNAs by feature matrix construction and classification by iRF from RNA-seq GTF file produced from Cufflinks (Trapnell et al., 2012). The prediction lncRNA sequences can be sub-classified using PBC algorithm. The second component of LIFT (coloured blue) predicts functions of the lncRNAs using co-expression regulatory network and BMRF method (Kourmpetis et al., 2010) (see online version for colours)


RNA-Seq data
Reference-based or *De-novo* mapping
Transcripts GTF file
Feature extraction and optimization using LiRF-FS
lncRNA prediction using iRF
Position-Based Classification of lncRNA transcripts
FPKM relative expression values
lncRNA-Protein co-expression network
Protein-Protein interaction network
Gene Ontology enrichment of Proteins
Function prediction using BMRF

2 Methods

2.1 Reference sequence datasets

Since a reliable dataset is important for model training and prediction, a random selection of protein-coding and lncRNA transcripts from plant species were obtained from Refseq database (O'Leary et al., 2016). Transcript sequences for Arabidopsis thaliana (ATH), Brassica rapa (BRA), Brassica napus (BNA), Brassica oleracea (BOL), Zea mays (ZM), Oryza sativa (OS), Solanum tuberosum (ST) and Solanum lycopersicum (SL) were downloaded from RefSeq database. lncRNA sequences were filtered by applying a threshold cutoff of 200bp on non-coding RNA (ncRNA) FASTA files.

2.2 RNA-seq datasets

Two RNA-seq datasets were used for identification, genomic annotation, and functional prediction of lncRNAs. The first dataset consists of 10 samples derived from the apical shoot meristem time-series dataset from the *A. thaliana* genome obtained from the NCBI SRA database (Project ID: PRJNA268115) (Klepikova et al., 2015). 7–16 days old plants were harvested to obtain a synchronised and representative sample at different developmental stages, denoted by S7 to S16 respectively. The dataset consists of 10 samples from Day-7 to 16 with two replicates from 9-14 days. 9 sample pairs were constructed by comparing samples from Day 8-16 against Day-7. The second dataset consisted of 11 time-series samples (from 0 to 20 days with interval of two days between sample obtained) from whole seed of *Z. mays* inbred line B73 which was obtained from the SRA database (Project ID: SRP037559) (Chen et al., 2014). The whole dataset consists of 53 samples from embryo, endosperm and whole seed stages. However, for this study, only 11 samples were selected. 10 sample pairs were constructed by comparing samples against Day-0.

2.3 RNA-seq data analysis

The first 15 base pairs of the sequence reads are trimmed using Cutadapt to remove adapter and low-quality sequences with Q-score greater than or equal to 30. For *A. thaliana* reads, trimmed reads are aligned to Arabidopsis genome (TAIR10) using Tophat2 mapper (Kim et al., 2013) with custom parameter values (minimum intron length = 40, maximum intron length = 5000, segment length = 20, segment mismatches = 2, max multi-hits = 1, minimum normalised depth = 0, minimum anchor length = 10) for optimal read alignment of ATH sequence reads. For *Z. mays* sequence reads, the sequence reads were aligned to Maize genome (AGPv4) with max intron length set to 60,000 and min intron length set to 5. Remaining parameters were kept to default. Once the reads are aligned, assembly of exonic and splice junction reads was performed for individual sample pairs using Cufflinks for generating gene transfer format (GTF) file (Figure 1).

2.4 Feature extraction for lncRNA classification

For extraction of features from the RNA-seq derived genomic sequences, the transcript sequences were first extracted from the binary alignment map (BAM) file produced by Tophat2 mapper (Kim et al., 2013). Based on reference alignment of sample reads, a consensus FASTA sequence for each transcript coordinate was constructed by a two-step process:

- SNP and INDEL calling of BAM file using SAMtools mpileup (Li and Durbin, 2009) that generated a variant call format (VCF) file
- sequence extraction from the genome and consensus sequence generation using variants from VCF by the SAMtools *faidx* tool (Li and Durbin, 2009).

The mpileup function collects the information from the BAM file and computes the likelihood. This is stored in a Binary VCF (BCF) format. The Bcftools consensus function creates a consensus FASTA transcript sequence based on reference genome by applying the VCF variants. The sequence obtained can be used for extraction of features for lncRNA classification.

Features extracted from FASTA sequences can be categorised into either ORF-based features or codon bias features. These features constitute a feature set $F = \{f_1, f_2, \ldots., f_n\}$, where f_n denotes the nth feature. The features were selected based on the published results of sequence measures and codon bias measures (Fickett and Tung, 1992; Roth et al., 2012).

2.4.1 ORF and sequence based features

Three ORF-based features were extracted: maximum ORF length (f_1), ORF coverage (f_2) and mean ORF coverage (f_3) and four sequence-based features: transcript length (f_4), GC content (f_5), Fickett score (f_6) and Hexamer score (f_7). f_1 is one of the most fundamental feature used to distinguish lncRNA from mRNA as the majority of protein-coding genes have ORFs greater than 100 amino acids (Frith et al., 2006). f_2 is the ORF coverage defined as length of the longest ORF divided by transcript length. This feature has also shown to produce good classification performance when compared to ORF length (Wang et al., 2013; Zhao et al., 2016). f_3 is the overall ORF coverage defined as average of total ORF lengths divided by transcript length for sequence. f_5 is the GC content, which is also a common measure to differentiate lncRNA from protein-coding transcripts as coding sequences have been reported to have higher GC content in exons over introns (Amit et al., 2012). f_5 is simply calculated as absolute total number of GC motifs in a sequence. f_6 is the Fickett score (Fickett, 1982) obtained by calculating four base pair position values in transcript sequence. f_7 is the Hexamer score which is computed by making a Hexamer table of 4096 (64 × 64 hexamers) *k*-mers using reference set of coding and non-coding sequences. f_7 is calculated by first measuring frequencies of hexamers in the test set sequences. Logarithmic ratio of coding and non-coding sequences were then computed for each hexamer having non-zero frequency in the test set. Positive f_7 indicates higher probability of protein-coding sequence whereas negative score indicates higher probability of non-coding RNA sequence.

2.4.2 Codon bias features

In protein-coding genes, the translational mapping process of codons (or nucleotide triplets) to amino acids involve usage of synonymous codons which codes same amino acids that is non-distinguishable at protein level. However, it has been reported that there exists a non-uniform codon usage in most genes i.e., codon bias (Clarke, 1970; Ikemura, 1982). Many indices have been proposed for measuring codon bias, among which we carefully selected six codon-bias measures which are important in distinguishing lncRNAs from mRNAs. These includes frequency of optimal codons (f_8) (Fickett, 1982; Amit et al., 2012), codon usage bias (f_9) (Karlin and Mrázek, 1996), relative codon bias (f_{10}) (Roymondal et al., 2009), weighted sum of relative entropy (f_{11}) (Suzuki et al., 2004), synonymous codon usage order (f_{12}) (Wan et al., 2004) and relative synonymous codon usage (RSCU) (f_{13}) (Sharp et al., 1986).

f_8 is the frequency of optimal codons (Fop) which is calculated as ratio of total number of optimal codons to the total number of synonymous codons. Fop was also one of the measures proposed by Ikemura (1982). The number of optimal codons is calculated as:

$$O_{opt} = \sum_{c \in C_{opt}} O_c \tag{1}$$

where C_{opt} is defined as subset of optimal codons from all codons C and O_{tot} is the total number of codons in the sequence. Therefore, f_8 is calculated as:

$$f_8 = \frac{O_{opt}}{O_{tot}} \tag{2}$$

f_9 is the Codon Usage Bias (CUB) which assesses codon bias in test set sequence relative to reference set of sequences based on weighted sum of distances of relative codon usage frequencies between the reference set and test set sequences (Karlin and Mrázek, 1996). The reference set is used as standard to which other sequences can be compared. f_9 is defined as:

$$f_9 = \sum_{a \in \mathrm{A}} F_a d\left(f_a, f_a^{ref}\right) \tag{3}$$

where f_a is frequency of amino acid a in the test set sequence whereas f_a and f_a^{ref} are codon frequencies for amino acid a in test and reference sets, respectively and d is the L1 norm or manhattan distance for the codon frequency f_a and f_a^{ref} vectors which is calculated as:

$$d\left(f_a, f_a^{ref}\right) = \sum_{c \in C_a} \mid f_{ac}, f_a^{ref} \mid \tag{4}$$

where f_{ac} is the frequency of codon c encoding amino acid a in test set sequences and f_a^{ref} is the frequency of amino acid a in reference set sequences.

f_{10} is relative codon bias (RCB) (Roymondal et al., 2009) which is a measure that defines contribution of codons as:

$$w_c^{RCB} = \frac{O_c - E\left[O_c\right]}{E\left[O_c\right]} \tag{5}$$

where $E\left[O_c\right]$ is the expected number of codon occurrences in three codon positions. Once w_c^{RCB} is determined f_{10} is calculated by the following method for each sequence:

$$f_{10} = \exp\left(\frac{1}{O_{tot}} \sum_{c \in \mathrm{C}} \log w_c^{RCB}\right) - 1 \tag{6}$$

f_{11} feature used is the weighted sum of relative entropy (Ew) which measures the degree of deviation from equal codon usage (Suzuki et al., 2004). Therefore, f_{11} is defined as sum of relative entropy of each amino acid weighted by its relative frequency in the test sequence which is given by:

$$f_{11} = \sum_{a \in \mathrm{A}} F_a E_a \tag{7}$$

where F_a is the relative frequency of amino acid a in the test sequence and E_a is computed as:

$$E_a = \frac{H_a}{\log_2 k_a} \tag{8}$$

where k_a is number of synonymous codons observed in the test sequence and H_a is the entropy which measures uncertainty of codon usage in the test sequence for amino acid a and is computed as:

$$H_a = -\sum_{c \in C_a} f_{ac} \log_2 f_{ac} \tag{9}$$

f_{12} is the Synonymous Codon Usage Order (SCUO) is also an entropy-based codon bias measure and is similar to *Ew* which differs only by the way entropy is calculated for each amino acid (Wan et al., 2004). Instead of calculating the relative entropy, normalised difference between maximum and observed entropy is computed:

$$E_a = \frac{\log_2 k_a - H_a}{\log_2 k_a} \tag{10}$$

and the f_{12} is computed as:

$$f_{12} = \sum_{a \in \mathrm{A}} F_a E_a \tag{11}$$

f_{13} is the RSCU score which defined the relationship between observed codon frequencies and number of times codon observed when synonymous codon usage is random with no codon bias (Sharp et al., 1986). This is calculated as:

$$RSCU_{ac} = \frac{O_{ac}}{\frac{1}{k_a} \sum_{c \in C_a} O_{ac}} \tag{12}$$

where O_{ac} is the frequency of codon c for amino acid a. $RSCU_{ac}$ is the RSCU score (f_{13}) for each codon c encoding amino acid a and is computed for 61 codons individually by the above equation. Methionine (M), Tryptophan (W) and stop codons were excluded from the analysis as M and W do not have any synonymous codons and stop codons do not contribute any information. Therefore, in total f_{13} provided 61 features for the classification.

2.5 Feature selection using LASSO and iRF classifier

Selection of optimal features is an important optimisation approach for classification. Wrapper-based feature selection (FS) methods such as sequential forward selection (SFS) (Pudil et al., 1994) or SVM-recursive feature elimination (SVM-RFE) (Huang et al., 2014) are computationally inefficient and fail to identify optimal feature subsets. Whereas filter-based FS methods such as mRMR (Peng et al., 2005), Chi-square (Chen and Chen, 2011), Information Gain (Lee and Lee, 2006) assigns relevance score or rank to each feature by considering each feature separately and ignoring dependencies between features which leads to worse classification performance. Regression based approaches such as least-squares estimate method often produces larger variance during model fitting which leads to overfitting and poor generalisation. Least absolute shrinkage and selection operator (LASSO) is a feature selection method which combines least-square loss with $\ell 1$ norm constraint and produces sparse features by shrinking coefficients to zero. Other approaches such as ridge regression (Marquardt, 1970;

Tibshirani, 1996) uses $\ell 2$ norm due to which it produces non-zero coefficients and therefore becomes inefficient for feature selection. Usage of ℓq norm (with $q < 1$ or $q > 1$) approaches for optimisation are generally non-convex and makes the minimisation computationally challenging.

The LIFT framework implements LASSO and iterative Random Forest for FS (*LiRFFS*) for identifying principal set of collective features yielding the highest accuracy which works by iterative selection of features based on varying lambda (λ) values. As λ changes, non-zero beta coefficients are generated which corresponds to the selection of features using $\ell 1$-regularised optimisation of LASSO (Tibshirani, 1996). Using training and test sets obtained from feature normalisation, coefficients (β) for each feature are calculated by the following formula:

$$\beta = (X^T X)^{-1} X^T Y \tag{13}$$

For estimating β^{lasso} coefficients in each iteration, coordinate-descent minimisation is performed (Wu and Lange, 2008) and the coefficients are obtained by the following objective function:

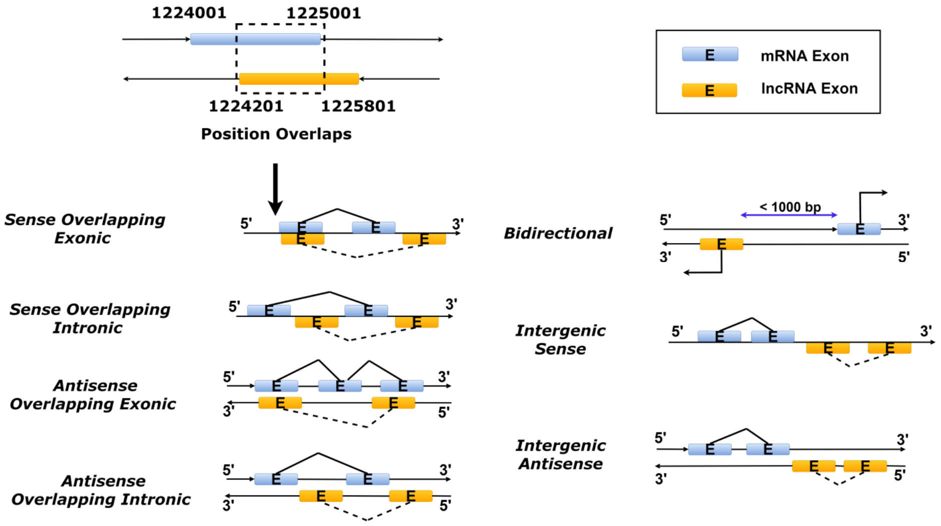
$$\beta^{lasso} = {}^{argmin}_{\beta} \frac{1}{2n} \| X\beta - y|\,|_2^2 + \lambda |\,|\beta|\,|_1 \tag{14}$$

where $\lambda \geq 0$, $\|y - X\beta\|_2^2$ is the loss function (i.e., sum of squares), $\|\beta\|_1$ is the penalty term and λ is the tuning parameter which controls the strength of the penalty. Features extracted from the coding and noncoding sequences are divided into training and validation sets. β coefficients are calculated on each λ value. The selected features for each λ are iteratively applied on the validation set to obtain the accuracy vector. The optimal feature set is obtained by selecting the feature set that produces the prediction accuracy between the tolerance accuracy value and the maximum prediction accuracy value. The optimal feature set can be used for building the model for classification of test set transcript sequences. Detailed implementation of the LiRFFS algorithm has been provided in Supplementary Material (Section S1).

2.6 *Position-based classification (PBC) of lncRNA sequences*

Genomic annotation of lncRNAs is essential for classification based on their position in the genome. For sub-classification of lncRNAs, a PBC algorithm was implemented (Figure 2) for finding the optimal overlaps of lncRNA exonic and intronic sequences. The algorithm extracts the ORFs for each transcript sequence. Using the ORF sequence, the algorithm extracts the exonic (E) and intronic (I) sequences based on exon-intron boundaries (GT-AG). Position-based mapping is performed by overlapping the coordinates of the lncRNA E and I sequences with the genomic coordinates of the protein-coding E and I sequences. Using this strategy, LIFT annotates lncRNAs into various sub-classes which involves direction of overlap (sense or antisense), type of overlap (exonic or intronic) and position of lncRNA sequence (bidirectional or intergenic). The classification was implemented on the *A. thaliana* and *Z. mays* lncRNA transcript sequences.

Figure 2 Position-based classification description Sub-classification of lncRNA sequences is performed based on positional coordinates. Sense and antisense-overlapping is performed based on GT-AG exonic and intronic sequences from the ORFs. Intergenic classification is performed by scanning lncRNA sequences between protein-coding genes. Bidirectional lncRNA sequences are classified by finding lncRNA exonic sequences less than 1000 bp from the protein coding exonic sequences (see online version for colours)



2.7 *Function prediction of lncRNA sequences using BMRF method*

Relative expression values from the Fragments Per Kilobase of transcript per Million mapped reads (FPKM) individual sample pairs were computed by dividing the sample pair read count by the maximum read count value from all other sample pairs to obtain a relative expression value between 0 and 1. A co-expression similarity matrix of lncRNA and protein sequences was constructed. An expression similarity matrix was constructed between FPKM values of each pair of lncRNA and mRNA using Pearson's correlation coefficient (PCC) ≥0.9 and ≤–0.9. This is named as the lncRNA-protein co-expression similarity (LPCS) matrix. lncRNAs and protein-coding sequences with zero FPKM values in ≥70% of samples pairs were excluded from the analysis.

A protein-protein interaction (PPI) network was constructed using protein-protein interaction data obtained from the STRING database (Szklarczyk et al., 2015) for proteins having higher expression profile correlation with predicted lncRNAs. This resulted in protein-protein interaction pairs with interacting proteins represented by nodes and interactions represented by edges. Resulting lncRNA-protein and protein-protein interacting pairs were concatenated for functional association of lncRNAs.

To predict functional association of lncRNAs, a Bayesian Markov Random Fields (BMRF) method was used which has been previously used for predicting protein functions of unannotated proteins (Kourmpetis et al., 2010). BMRF is originally based on the Markov Random Fields (MRF) approach (Deng et al., 2002) where the nodes are coloured and encoded in the binary vector with $X_i = 1$ if *i*th protein performs a particular function, and $X_i = 0$ if *i*th protein protein does not possess any function. We substituted lncRNAs where $X_i = 0$ if *i*th protein not having functions in the network. The conditional probability across all the unannotated lncRNA nodes in the network is computed by a Pseudo-Likelihood Function (PLF). It is a function which possesses properties similar to a likelihood function and therefore helps in determining the conditional probability of the state of the lncRNA. The conditional probability of an unannotated lncRNA i is given by a logistic function $\left(1+\exp\left(-v_i\right)\right)^{-1}$. Each state of the unannotated lncRNA is sampled using this logistic function. Once the PLF is computed, Gibbs-Sampling (GS) is performed by iterating over all the states of the unannotated lncRNA sequences. In each iteration, nodes connected to the lncRNAs are updated conditionally with parameter values corresponding to $(\alpha, \beta^0, \beta^1)$. The Differential Evolution Markov Chain method implemented in the BMRF updates the conditional probabilities. This process is repeated until convergence is reached. For input to BMRF, LPCS matrix, PPI matrix and protein-coding Gene Ontology (GO) annotations are required.

Functions of significantly expressed lncRNAs were obtained by applying false discovery rate (FDR) cutoff of 0.05 and log2 Fold Change (FC) ≥1 or ≤–1 on *A. thaliana* and *Z. mays* sample pairs.

2.8 K-fold cross-validation benchmarking

For evaluating the prediction accuracy of LIFT against CPAT, CPC2, lncScore and PLEK tools, a 10-fold Cross Validation (CV) benchmarking was performed on the lncRNA sequences. Test set sequences annotated in the TAIR10 dataset and expressed sequence tags (EST) derived sequences for *A. thaliana* were obtained from the PLncDB database (Jin et al., 2013) whereas sequences for *Z. mays* were obtained from the Ensembl Genomes 38 AGPv4 annotation file. 10% test set sequences and 90% training set sequences were selected in each fold consisting of balanced lncRNA and protein-coding sequences. The TAIR10-annotated and EST-derived sequences were shuffled to perform fair comparison of prediction accuracy. The datasets were labelled as D1 (TAIR10-based), D2 (EST-derived) and D3 (ZM Ensembl 38 AGPv4) for further analysis.

2.9 Performance evaluation criteria

To assess classification performance of lncRNAs and mRNA transcripts, Accuracy, Sensitivity, Specificity, Precision, F1-Score, NPV and MCC metrics were used which were defined as:

$$\text{Accuracy (ACC)} = \frac{TP+TN}{TP+FP+FN+TN},$$

$$\text{Sensitivity or Recall (SENS)} = \frac{TP}{TP+FN},$$

$$\text{Specificity (SPEC)} = \frac{TN}{FP+TN},$$

Precision (PRES) = $\frac{TP}{TP+FP}$,

F1-Score (F1) = $\frac{2*(Precision*Recall)}{Precision+Recall}$,

NPV = $\frac{TN}{TN+FN}$,

MCC = $\frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)*(FN+TN)*(FP+TN)*(TP+FN)}}$

with TP = true positive, TN = true negative, FP = false positive and FN = false negative.

2.10 data availability

The source code is publicly available together with relevant data on GitHub: https://github.com/deshpan4/LIFT.

3 Results

3.1 Performance of LIFT feature groups on reference datasets

To evaluate performance of LIFT using 73 features on different species, prediction accuracies of LIFT on reference datasets were tested. An area under receiver operating characteristic (AUC) curve gives better insight about the ability of a classifier to separate two classes. From the reference datasets, an average AUC of 99.23% for plants was observed using an iRF classifier. Table 1 shows prediction accuracies of LIFT on plant species. The prediction accuracies ranged from 94.51% to 97.25% whereas specificity values ranged from 94.12% to 97.76%. An average MCC of 0.91 was observed for the 8 plant species. The higher accuracy exhibited with these datasets demonstrated that LIFT can predict the lncRNA sequences in plants with reasonable accuracy without overfitting the training data.

Table 1 Performance of LIFT on identification of lncRNA test set transcript sequences in multiple plant species

Species	*ACC*	*SENS*	*SPEC*	*F1*	*MCC*	*AUC*
A. thaliana	94.34	93.19	95.57	94.45	0.887	98.90
Z. mays	94.62	95.20	94.08	94.45	0.892	98.84
B. napus	96.47	97.29	95.68	96.41	0.929	99.54
B. rapa	95.76	96.58	94.93	95.82	0.915	99.21
B. oleracea	96.27	96.008	96.53	96.28	0.925	99.20
O. sativa	96.91	95.80	98.003	96.83	0.938	99.62
S. lycopersicum	96.37	96.17	96.57	96.32	0.927	99.46
S. tuberosum	96.15	97.25	95.11	96.07	0.923	99.27

To evaluate the predictive power of LIFT, its performance was benchmarked against four other popular coding-potential alignment-free tools: CPAT (Wang et al., 2013), PLEK (Li et al., 2014), lncScore (Zhao et al., 2016) and CPC2 (Kang et al., 2017). Prediction on test set data in individual species shows that in general, LIFT achieves higher accuracy and presents better performance than other tools. Specifically, LIFT performed exceptionally accurately on *Z. mays* and *O. sativa* datasets with prediction accuracies of 94.71% and 96.95%, respectively (Table 2). Prediction accuracies of CPAT and lncScore produced similar accuracies, however, the specificity was comparatively lower.

Table 2 Performance (percentage accuracy) comparison of LIFT against other CPC tools on multiple plant species obtained from refseq dataset

Species	*LIFT*	*PLEK*	*CPAT*	*CPC2*	*lncScore*
A. thaliana	94.51	80.82	97.28	95.99	97.04
Z. mays	94.71	65.8	94.71	91.82	94.36
B. napus	96.73	56.77	96.86	94.64	96.35
B. rapa	95.77	61.29	96.9	94.73	96.34
B. oleracea	96.35	54.98	96.78	92.45	96.43
O. sativa	96.95	24.61	93.7	49.78	19.63
S. lycopersicum	97.25	67.94	97.98	95.85	97.92
S. tuberosum	95.69	62.29	95.36	93.73	95.43

Source: O'Leary et al. (2016)

LIFT exhibited highest accuracy values in 5 plant species when compared with PLEK and CPAT. When compared with CPC2, the LIFT displayed superior performance in all the species except ATH where higher metrics were observed for CPC2. An average prediction accuracy difference of 1–4% between the LIFT and CPC2 was detected in ZM, BNA, BRA, BOL, SL and ST species. OS displayed an accuracy difference of 47.17% between the LIFT and CPC2. Accuracy difference between the LIFT and PLEK showed an average difference of ~30–40% in BNA, BRA, BOL and ZM datasets, ~7–15% in ATH, whereas a significant difference of 72.34% was observed in OS species.

3.2 Selection of optimal features using LiRFFS

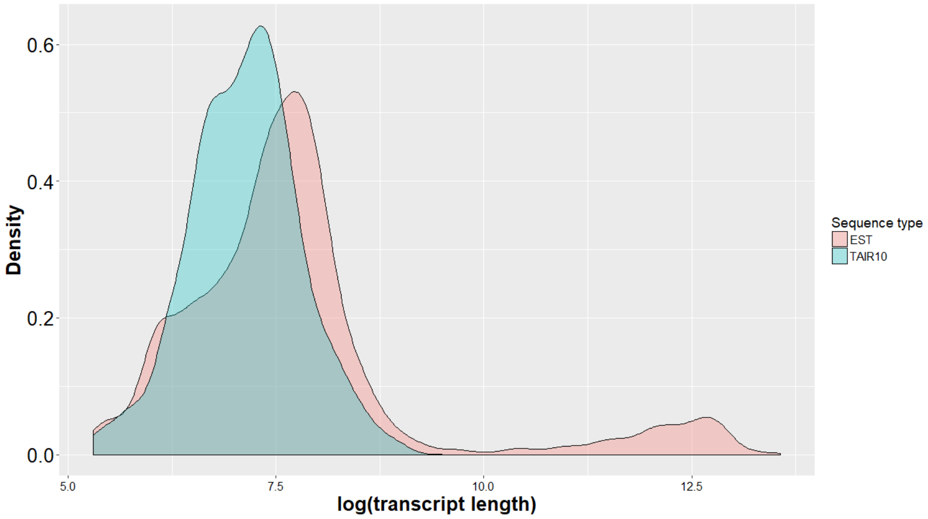
The selection of optimal features was performed on a unified dataset of 6 plant species *A. thaliana*, *Z. mays*, *O. sativa*, *B. napus*, *B. rapa* and *B. oleracea*). The dataset consisted of 22,468 (lncRNA and mRNA) transcript sequences selected as training set and 7,532 sequences selected as validation set. An optimal feature set was selected based on λ values ranging from 0.1 to 1×10^{-5}. Based on $tolerance$ cutoff value of 0.5, two feature sets, namely, the 7 feature set (7F) and the 31 feature set (31F) were selected having minimal and maximal optimal features. 7F is selected based on least number of features producing higher prediction accuracy having accuracy within the *tolerance* threshold value from the maximum prediction accuracy λ value. Whereas 31F is selected based on the maximum number of features having prediction accuracy within the $tolerance$ threshold value from the maximum prediction accuracy λ value. Prediction of test set sequences was performed based on the optimal feature sets obtained from LiRFFS computation.

3.3 K-fold cross validation benchmarking on plant RNA-seq datasets

To assess the performance of LIFT in the *A. thaliana* and *Z. mays* RNA-seq datasets, a 10-fold CV performance benchmarking was performed. The prediction accuracy of LIFT against CPAT, PLEK, CPC2 and lncScore was tested on each fold. A total of 994 and 1,878 transcripts test set transcripts in Arabidopsis and Maize, respectively. To evaluate the classification performance of LIFT, 31F was used for comparing the prediction accuracies of test sets.

Transcript length distribution of TAIR10-annotated and EST-derived lncRNA transcripts demonstrates the degree of sequence length variation in lncRNA transcripts (Figure 3). Sequences derived from the TAIR10 annotation data ranges between 200 bp and 8000 bp whereas sequences derived from EST analysis ranges widely between 200 bp and 7.8×10^5 bp. Additionally, ORF count of EST-lncRNA sequences reveals counts greater than 700 ORFs per frame. Such extremely long lncRNA sequences are generally misclassified as protein-coding transcripts, due to which the overall prediction accuracy decreases.

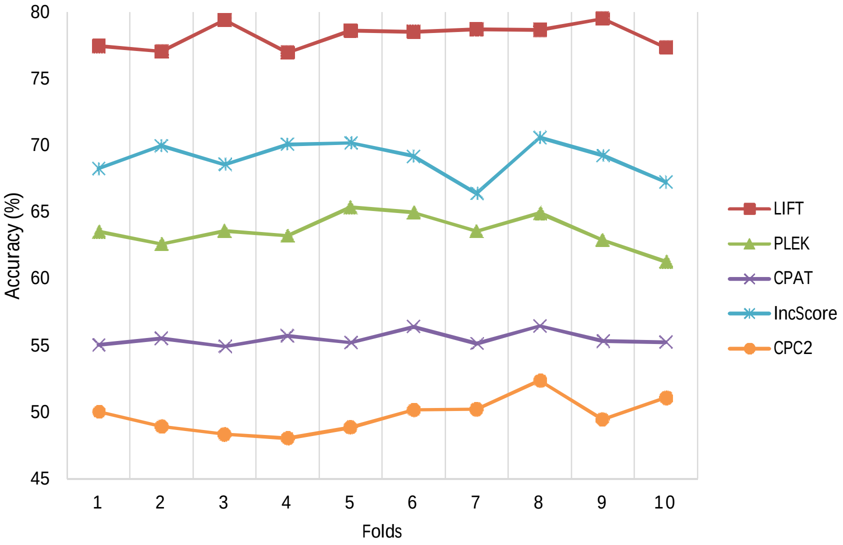
Figure 3 Density distribution of transcript lengths of lncRNA sequences in ATH TAIR10-annotated and EST-predicted results. X-axis is log of transcript lengths and y-axis is density (see online version for colours)



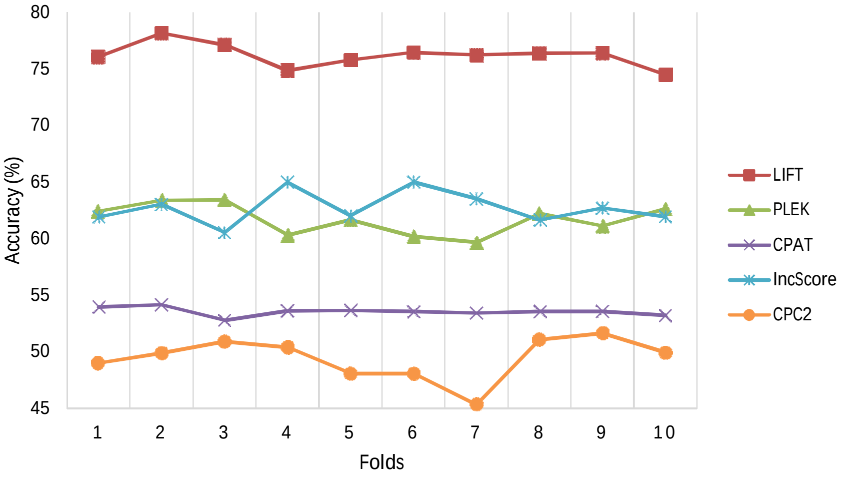
Results from the CV benchmarking on *A. thaliana* and *Z. mays* indicates that LIFT outperformed other tools with greater precision in identifying the lncRNA transcripts (Figure 4). LIFT identified the lncRNA transcripts with an average accuracy of 78.22% and 76.2% for *A. thaliana* D1 and D2, whereas 96.24% for *Z. mays* D3 data along with higher sensitivity, specificity, NPV and MCC values. lncScore produced an accuracy of 68.98% on D1 and 62.75% on D2. CPAT, on the other hand, identified the mRNA and lncRNA test set sequences with an average of 55.51% on D1 and 53.57% on D2 in *A.*

thaliana test set data. PLEK identified lncRNA transcripts with an average of 63.6% on D1 and 61.72% on D2. In *Z. mays* dataset, prediction accuracies of CPAT show 2.79% difference on the first fold. On the second fold, this difference increases to 3.05%. Whereas prediction accuracy difference between LIFT and PLEK is comparatively much higher with an average difference of 11.58% against LIFT from Folds-1 to 7.

Figure 4 Plots illustrating performance of LIF against other existing tools based on k-fold cross validation benchmarking analysis for: (a) *A. Thaliana*; (b) *A. Thaliana*-EST derived lncRNA sequences and (c) *Z. Mays*. X-axis represents folds whereas y-axis represents percentage accuracy (see online version for colours)

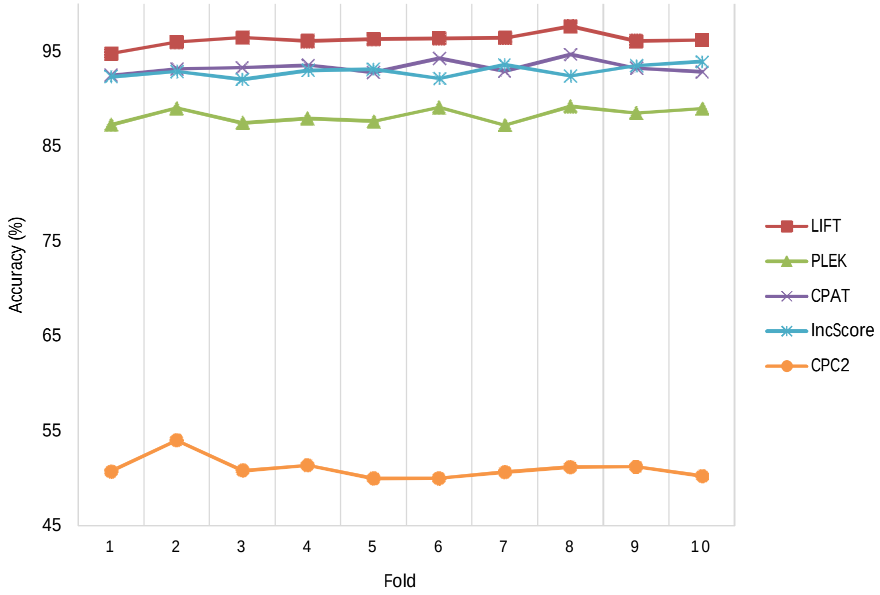


(a)



(b)

Figure 4 Plots illustrating performance of LIF against other existing tools based on k-fold cross validation benchmarking analysis for: (a) *A. Thaliana*; (b) *A. Thaliana*-EST derived lncRNA sequences and (c) *Z. Mays.* X-axis represents folds whereas y-axis represents percentage accuracy (see online version for colours) (continued)



(c)

CPC2, on the other hand, exhibited lowest prediction accuracies in folds-1 to 7 with a mean value of 47.1%. The accuracies increased to 68.63% and 70.64% in folds 9 and 10 with accuracy differences of 51.14%, 37.06% and 22.42% against LIFT. Interestingly, we noted that LIFT had higher accuracy, specificity, F1 and MCC as compared to other tools, thus indicating good quality of classification. CPAT produced higher sensitivity and NPV but performed poorly on accuracy, specificity, F1 and MCC. Contrastingly, lncScore exhibited comparable specificity performance but generated lower sensitivity, NPV and MCC values. PLEK showed lower NPV and MCC but produced higher accuracy, specificity, F1 and MCC than CPAT. On D3, LIFT displayed much better prediction accuracy, specificity, sensitivity and MCC than other tools. CPAT produced higher specificity and NPV but generated lower prediction accuracy than LIFT. PLEK showed similar sensitivity as LIFT, however, produced much lower specificity, NPV, F1 and MCC.

3.4 LIFT PBC module for genomic annotation of lncRNA transcripts

Based on coordinate mapping, 478 *A. thaliana* and 2511 *Z. mays* lncRNAs were classified into sense-overlapping exonic, sense-overlapping intronic, antisense-overlapping exonic, antisense-overlapping intronic, intergenic and bidirectional sub-classes. 478 lncRNAs in *A. thaliana* were classified based on 35,343 protein-coding transcripts into 3 sense-overlap exonic, 3 sense-overlap intronic, 70 antisense-overlap exonic, 69 antisense-overlap intronic, 252 antisense lncRNA, 5 intergenic and 306 bidirectional promoter lncRNA. In *Z. mays*, 2511 lncRNAs were classified based on 39,646 protein-coding transcripts out of which 119 sense-overlap exonic, 118 sense-overlap intronic, 13 antisense-overlap exonic and intronic, 909 antisense lncRNA, 1682 intergenic and 166 bidirectional promoter lncRNA were classified, respectively.

Classification results from the PBC annotation were compared against the experimentally annotated lncRNA transcripts from the TAIR9 and Ensembl Genomes 38 AGPv4 databases. Classification performance across the chromosomes measured an average accuracy of 72.55% for *A. thaliana* Natural Antisense Transcripts (NATs) and 90.86% for *Z. mays* with long intergenic lncRNAs (lincRNAs) sequences. The computational time required to classify the lncRNA transcripts took 2 h 52 min for *A. thaliana* and ~12 h for *Z. mays*. The computation was performed on an Intel(R) Xeon(R) CPU X5650 @ 2.67 GHz Linux station running on a single compute node.

3.5 *Function prediction of lncRNA genes based on lncRNA-mRNA co-expression patterns*

To predict the functions of lncRNA sequences from *A. thaliana* and *Z. mays* data, BMRF was implemented in LIFT for predicting the functions of Differentially Expressed (DE) lncRNA genes. Based on co-expression of lncRNA and mRNA FPKM values, 5,923 correlations for 18 lncRNA sequences in *A. thaliana* and 97,443 correlations for 93 lncRNA sequences were obtained in *Z. mays*. By applying probability ≥ 0.8, 18 lncRNAs were associated with 545 GOTerms in *A. thaliana* whereas 93 lncRNAs displayed a high association probability with 10 GOTerms in *Z. mays* (Table 3). Results from the BMRF analysis were validated against experimental data to detect similar lncRNA-functional association. A summary of experimentally reported lncRNA-function association data in plants (Liu et al., 2015) was used for validation of the results.

Function annotation results were filtered based on a dictionary of keywords extracted from the experimentally-derived lncRNA regulatory functions. From the analysis, 17 lncRNA genes were found to have association with 44 regulatory functions in *A. thaliana*, some of which included histone modification, regulation of transcription from RNA polymerase II promoter, DNA-templated transcription initiation, single-stranded DNA binding and alternative RNA splicing. DE lncRNA sequences in *Z. mays* were mainly annotated with biosynthetic process, cellular amino acid metabolic process, dopamine neurotransmitter receptor activity, intracellular functions, oxidation-reduction process, pyridoxal phosphate binding, ribosomal functions, ribosome biogenesis, structural constituent of ribosome and translation. Therefore, keyword filtering analysis did not generate any match in the *Z. mays* DE geneset.

Experimental studies reveal that *Alternative Splicing Competitor lncRNA* (ASCO-lncRNA) found in *A. thaliana* forms an alternative splicing regulatory module with *nuclear speckle RNA-binding protein* (NSR). The *AtNSR* is primarily expressed in the lateral root meristem development (Bardou et al., 2014). BMRF function prediction on the complete lncRNA geneset in *A. thaliana* (i.e., DE and non-DE geneset) confirms the presence of *ASCO-lncRNA* (AT1G67105) involved in the regulation of alternative mRNA splicing with a probability of 0.99. Certain lncRNAs found in *A. thaliana*, *O. sativa* and *S. lycopersicum* such as *IPS1*, *Cis-NAT*$_{PHO1;2}$, *OsPI1* and *TPS11* have been found to be involved in phosphate homeostasis as translational enhancers (Liu et al., 1997; Wasaki et al., 2003; Franco-Zorrilla et al., 2007; Jabnoune et al., 2013). Annotation results from the BMRF analysis demonstrated 89 lncRNA genes predicted to be involved in cellular phosphate ion homeostasis with an average probability of 0.943. The *Heat Stress transcription Factor* (HSF) lncRNA expressed in *A. thaliana* are principal regulators of heat stress response. The *antisense HSF B2a* (asHSFB2a) lncRNA has been found to be involved in the regulation of vegetative and gametophytic development (Wunderlich et al., 2014). Prediction results indicate that 15 DE lncRNA genes regulate transcriptional activity through positive and negative regulatory mechanisms.

Table 3 List of specific lncRNAs identified in *A. Thaliana* and *Z. Mays* associated with intra-nuclear functions using LIFT

lncRNA ID	*Biological function*
AT1G76892	Functions in poly(U) RNA binding
AT5G34871	Involved in regulation of response to DNA damage stimulus
AT1G78265	Involved in post-replication repair
AT1G76892	Functions in core promoter proximal region sequence-specific DNA binding
Zm00001d000474	Involved in DNA-directed RNA polymerase III complex
Zm00001d000738	Involved in mismatch repair
Zm00001d000972	Involved in nucleotide binding

4 Discussion

In this study, we developed a new tool, called LIFT for accurate identification, classification and function prediction of lncRNA genes. LIFT implements a set of sequence and codon-bias features and provides a feature selection-based approach for identifying the set of optimal features using LASSO and iRF classifier. For sub-classification of lncRNA sequences, a position-based strategy has been applied which classifies the lncRNA transcripts based on their relative genomic coordinates with the protein-coding transcript sequences. Inspired from the work undertaken by Kourmpetis et al. (2010) and Guo et al. (2013), LIFT integrates co-expression data derived from RNA-seq datasets. The co-expression data is generated by computing similarities in the expression profiles of lncRNA and mRNA transcripts. Based on the co-expression of lncRNA and mRNA sequences, a network of closely connected nodes is constructed. Implementation of BMRF in LIFT provides identification of the unannotated states of lncRNAs in the network with associated conditional probability values. The Gene Ontology term and probability-associated lncRNA genes can be filtered from the prediction set based on the dictionary of keywords extracted from the experimentally-verified lncRNA regulatory functions. Benchmarking results from K-fold CV analysis demonstrated that LIFT performed much better than existing CPC tools, including, CPAT, CPC2, lncScore and PLEK. Moreover, LIFT identified exceptionally longer lncRNA sequences with higher accuracy, specificity, F1 and MCC on all test datasets. Performance of existing tools on EST lncRNA data (*A. thaliana*) produced much lower prediction accuracy as compared to LIFT generating a difference of ~15–20%.

The selection of optimal features determines the classification performance of machine learning classifier. Implementation of LiRFFS in LIFT selected 24 codon-bias and 7 sequence features. These results provide insights into the preferential selection of synonymous codons in the classification process. LiRFFS produces a minimal and maximal set of optimal feature sets from the training and validation datasets. Application of the optimal feature sets on *A. thaliana* and *Z. mays* test set data demonstrated comparatively higher performance of 31 features when compared with 7 features using K-fold CV. LncRNA sub-classification using PBC achieved an accuracy of 90.86% in *Z. mays* and 72.55% in *A. thaliana* with published lncRNA annotated datasets. Function prediction results of the DE *A. thaliana* lncRNA sequences confirms the regulatory mechanisms of certain lncRNAs. The keyword-based search strategy implemented in LIFT provides filtering of genes based on experimentally published regulatory functions. Several genes were found to positively and negatively regulate the transcriptional activity as well as modification of histones. This approach provides a comprehensive list of associated molecular functions, which can serve as a useful resource for annotating lncRNA genes in non-model plants.

Compared with other tools, LIFT provides unique set of features for accurate identification of lncRNAs transcripts. For lncRNA classification, LIFT employs a coordinate-based classification approach which annotates the transcript sequences based on overlaps of the exonic and intronic lncRNA sequences. Many research studies undertaken for classification of lncRNAs rely on machine-learning approaches which often fail due to limitation of data. Present novel approach removes the dependency of training set which not only enhances the classification accuracy but also annotates wide range of lncRNA classes. LIFT also integrated a novel function prediction approach for annotating lncRNAs predicted with iRF classifier and network-based probabilistic approach in LIFT which implemented LPCS and PPI data from STRING database. The method employs ‘guild-by-association’ strategy which says that if non-functional lncRNA is significantly co-expressed with a protein associated with some function and the protein is physically connected to another protein with known function, then the lncRNA can possibly be connected to its nearest neighbours and involved in similar function with associated probability. In summary, LIFT outperformed other tools on testing datasets as well as RNA-seq datasets with full-length transcript sequences. Moreover, LIFT showed better performance specifically on accuracy, sensitivity, F1-score, NPV and MCC metrics which shows that LIFT can predict more precisely when applied on other species.

LIFT has several advantages over other tools. First, apart from commonly known distinguishing sequence-based features such as ORF length, GC content and Fickett score, it takes advantage of codon-biased features to increase discriminative power. Second, LIFT implements a powerful semi-supervised optimisation approach for selection of principal features which can be applied to any species. Third, integrative approach of LiRFFS and codon-biased features by LIFT provides insights into preferential selection of species-specific synonymous codons in the classification process. Fourth, it implements novel position-based mapping algorithm for sub-classification which can provide valuable insights into different features of lncRNAs and their underlying functional mechanisms in non-model species. Fifth, it provides functional annotation for the predicted lncRNAs using BMRF which takes advantage of PPI and LPCS networks. Sixth, LIFT can work with less user-provided *a priori* information which does not require any score cutoff values or customised parameterisation of the classifier for lncRNA identification. Compared with existing tools such as CPAT, PLEK, CPC2 and lncScore, LIFT not only performed better on reference datasets but also provided higher prediction accuracy with selection of relevant features on *A. thaliana* and *Z. mays* datasets. Altogether, LIFT is stable, accurate and robust tool for identifying, classifying and functionally annotating lncRNAs in multiple species.

5 Conclusions

In this work, we developed a novel tool, LIFT for accurate identification, classification and function prediction of lncRNA transcripts that is suitable for plant species. LIFT exceeds the prediction performance over other tools on various parameters. The ability to identify and differentiate various lncRNA transcripts was demonstrated with cross-validation tests on Arabidopsis and Maize datasets. Using LiRFFS, optimal features were identified generating higher prediction accuracy of tremely longer lncRNA sequences in *A. thaliana* datasets. Implementation of iRF classifier in LIFT additionally helps in determining prevalent feature interactions. Additionally, the PBC module of LIFT provides accurate classification of lncRNAs and identified several other classes not yet recognised in published datasets. Considering the complexity of various features implemented in the framework, LIFT requires a considerable amount of time for PBC-based classification. However, the amount of time required to classify can be reduced by the implementation of multi-threading which can bring down the computation time from ~12 h to ~2 h. Altogether, LIFT is stable, accurate and robust tool for identifying, classifying and functionally annotating lncRNAs in plant species.

References

Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., Lev-Maor, G., Burstein, D., Schwartz, S., Postolsky, B., Pupko, T. and Ast, G. (2012) 'Differential GC content between exons and introns establishes distinct strategies of splice-site recognition', *Cell Reports*, Vol. 1, No. 5, pp.543–556, doi: 10.1016/j. celrep.2012.03.013.

Bardou, F., Ariel, F., Simpson, C.G., Romero-Barrios, N., Laporte, P., Balzergue, S., Brown, J.W.S. and Crespi, M. (2014) 'Long noncoding RNA modulates alternative splicing regulators in arabidopsis', *Developmental Cell*, Vol. 30, No. 2, pp.166–176, doi: 10.1016/j. devcel.2014.06.017.

Basu, S., Kumbier, K., Brown, J.B. and Yu, B. (2018) 'Iterative random forests to discover predictive and stable high-order interactions', *Proceedings of the National Academy of Sciences of the United States of America*, National Academy of Sciences, Vol. 115, No. 8, pp.1943–1948, doi: 10.1073/pnas.1711236115.

Chen, J., Zeng, B., Zhang, M., Xie, S., Wang, G., Hauck, A. and Lai, J. (2014) 'Dynamic transcriptome landscape of maize embryo and endosperm development', *Plant Physiology*, Vol. 166, No. 1, pp.252–264, doi: 10.1104/pp.114.240689.

Chen, X. (2015) 'Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA', *Scientific Reports*, p.5, doi: 10.1038/srep.13186.

Chen, X., Huang, Y-a., Wang, X-s., You, Z-h. and Chan, K.C.C. (2016a) 'FMLNCSIM: fuzzy measure-based lncRNA functional similarity calculation model', *Oncotarget*, Vol. 7, No. 29, doi: 10.18632/oncotarget.10008.

Chen, X., You, Z-H., Yan, G-Y. and Gong, D-W. (2016b) 'IRWRLDA: improved random walk with restart for lncRNA-disease association prediction', *Oncotarget*, Vol. 7, No. 36, pp.57919–57931, doi: 10.18632/oncotarget.11141.

Chen, Y.T. and Chen, M.C. (2011) 'Using chi-square statistics to measure similarities for text categorization', *Expert Systems with Applications*, Vol. 38, No. 4, pp.3085–3090, doi: 10.1016/j.eswa.2010.08.100.

Clarke, B. (1970) 'Darwinian evolution of proteins', *Science*, Vol. 168, pp.1009–1011, doi: 10.1126/science.168.3934.1009.

Deng, M., Zhang, K., Mehta, S., Chen, T. and Sun, F. (2002) 'Prediction of protein function using protein-protein interaction data', *Proceedings/IEEE Computer Society Bioinformatics Conference. IEEE Computer Society Bioinformatics Conference*, Vol. 1, No. 6, pp.197–206, doi: 10.1089/106652703322756168.

Fan, X.N. and Zhang, S.W. (2015) 'lncRNA-mFDL: identification of human long non-coding RNAs by fusing multiple features and using deep learning', *Molecular Biosystems*, Vol. 11, No. 3, pp.892–897, doi: 10.1039/c4mb00650j.

Fickett, J.W. (1982) 'Recognition of protein coding regions in DNA sequences', *Nucleic Acids Research*, Vol. 10, No. 17, pp.5303–5318, doi: 10.1093/nar/10.17.5303.

Fickett, J.W. and Tung, C.S. (1992) 'Assessment of protein coding measures', *Nucleic Acids Research*, Vol. 20, No. 24, pp.6441–6450, doi: 10.1093/nar/20.24.6441.

Franco-Zorrilla, J.M., Valli, A., Todesco, M., Mateos, I., Puga, M.I., Rubio-Somoza, I., Leyva, A., Weigel, D., García, J.A. and Paz-Ares, J. (2007) 'Target mimicry provides a new mechanism for regulation of microRNA activity', *Nature Genetics*, Vol. 39, No. 8, pp.1033–1037, doi: 10.1038/ng2079.

Frith, M.C., Forrest, A.R., Nourbakhsh, E., Pang, K.C., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Bailey, T.L. and Grimmond, S.M. (2006) 'The abundance of short proteins in the mammalian proteome', *PLoS Genetics*, Vol. 2, No. 4, pp.515–528, doi: 10.1371/journal. pgen.0020052.

Guo, X., Gao, L., Liao, Q., Xiao, H., Ma, X., Yang, X., Luo, H., Zhao, G., Bu, D., Jiao, F., Shao, Q., Chen, R.S. and Zhao, Y. (2013) 'Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks', *Nucleic Acids Research*, Vol. 41, No. 2, doi: 10.1093/nar/gks967.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., Cabili, M.N., Jaenisch, R., Mikkelsen, T.S., Jacks, T., Hacohen, N., Bernstein, B.E., Kellis, M., Regev, A., Rinn, J.L. and Lander, E.S. (2009) 'Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals', *Nature*, Vol. 458, No. 7235, pp.223–227, doi: 10.1038/nature07672.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J.M., Ezkurdia, I., Van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigó, R. and Hubbard, T.J. (2012) 'GENCODE: the reference human genome annotation for the ENCODE project', *Genome Research*, Vol. 22, No. 9, pp.1760–1774, doi: 10.1101/gr.135350.111.

Huang, M.L., Hung, Y.H., Lee, W.M., Li, R.K. and Jiang, B.R. (2014) 'SVM-rFE based feature selection and taguchi parameters optimization for multiclass SVM classifier', *Scientific World Journal*, doi: 10.1155/2014/795624.

Ikemura, T. (1982) 'Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes', *Journal of Molecular Biology*, Vol. 158, No. 4, pp.573–597, doi: 10.1016/0022-2836(82)90250-9 .

Jabnoune, M., Secco, D., Lecampion, C., Robaglia, C., Shu, Q. and Poirier, Y. (2013) 'A rice cis-natural antisense RNA acts as a translational enhancer for its cognate mRNA and contributes to phosphate homeostasis and plant fitness', *The Plant Cell*, Vol. 25, No. 10, pp.4166–4182, doi: 10.1105/tpc.113.116251.

Jiang, Q., Ma, R., Wang, J., Wu, X., Jin, S., Peng, J., Tan, R., Zhang, T., Li, Y. and Wang, Y. (2015) 'LncRNA2Function: A comprehensive resource for functional investigation of human lncRNAs based on RNA-seq data', *BMC Genomics*, Vol. 16, No. 3, doi: 10.1186/1471-2164-16-S3-S2.

Jin, J., Liu, J., Wang, H., Wong, L. and Chua, N.H. (2013) 'PLncDB: plant long non-coding RNA database', *Bioinformatics*, pp.1068–1071, doi: 10.1093/bioinformatics/btt107.

Kang, Y.J., Yang, D.C., Kong, L., Hou, M., Meng, Y.Q., Wei, L. and Gao, G. (2017) 'CPC 2: a fast and accurate coding potential calculator based on sequence intrinsic features', *Nucleic Acids Research*, doi: 10.1093/nar/gkx428.

Karlin, S. and Mrázek, J. (1996) 'What drives codon choices in human genes?', *Journal of Molecular Biology*, Vol. 262, No. 4, pp.459–72, doi: 10.1006/jmbi.1996.0528.

Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B.E., van Oudenaarden, A., Regev, A., Lander, E.S. and Rinn, J.L. (2009) 'Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression', *Proceedings of the National Academy of Sciences*, Vol. 106, No. 28, pp.11667–11672, doi: 10.1073/pnas.0904715106.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L. (2013) 'TopHat 2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions', *Genome Biology*, Vol. 14, No. 4, R36, doi: 10.1186/gb-2013-14-4-r36.

Klepikova, A.V., Logacheva, M.D., Dmitriev, S.E. and Penin, A.A. (2015) 'RNA-seq analysis of an apical meristem time series reveals a critical point in arabidopsis thaliana flower initiation', *BMC Genomics*, Vol. 16, doi: 10.1186/s12864-015-1688-9.

Kong, L., Zhang, Y., Ye, Z.Q., Liu, X.Q., Zhao, S.Q., Wei, L. and Gao, G. (2007) 'CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine', *Nucleic Acids Research*, Vol. 35, Suppl. 2, doi: 10.1093/nar/gkm391.

Kourmpetis, Y.A.I., Van Dijk, A.D.J., Bink, M.C.A.M., Van Ham, R.C.H.J. and Ter Braak, C.J.F. (2010) 'Bayesian markov random field analysis for protein function prediction based on network data', *PLoS ONE*, Vol. 5, No. 2, p.e9293, doi: 10.1371/journal. pone.0009293.

Lee, C. and Lee, G.G. (2006) 'Information gain and divergence-based feature selection for machine learning-based text categorization', *Information Processing and Management*, pp.155–165, doi: 10.1016/j.ipm.2004.08.006.

Li, A., Zhang, J. and Zhou, Z. (2014) 'PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme', *BMC Bioinformatics*, Vol. 15, No. 1, p.311, doi: 10.1186/1471-2105-15-311.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with burrows-wheeler transform', *Bioinformatics*, Vol. 25, No. 14, pp.1754–1760, doi: 10.1093/bioinformatics/btp.324.

Liu, C., Muchhal, U.S. and Raghothama, K.G. (1997) 'Differential expression of TPS11, a phosphate starvation-induced gene in tomato', *Plant Molecular Biology*, Vol. 33, No. 5, pp.867–874, doi: 10.1023/A: 1005729309569.

Liu, X., Hao, L., Li, D., Zhu, L. and Hu, S. (2015) 'Long non-coding RNAs and their biological roles in plants', *Genomics, Proteomics and Bioinformatics*, Vol. 13, No. 3, pp.137–147, doi: 10.1016/j. gpb.2015.02.003.

Marquardt, D.W. (1970) 'Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation', *Technometrics*, Vol. 12, No. 3, pp.591–612, doi: 10.1080/00401706. 1970.10488699.

Mercer, T.R., Dinger, M.E. and Mattick, J.S. (2009) 'Long non-coding RNAs: insights into functions', *Nature Reviews Genetics*, pp.155–159, doi: 10.1038/nrg2521.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D. and Pruitt, K.D. (2016) 'Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation', *Nucleic Acids Research*, Vol. 44, No. D1, pp.D733–d745, doi: 10.1093/nar/gkv1189.

Peng, H., Long, F. and Ding, C. (2005) 'Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy', *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 8, pp.1226–1238, doi: 10.1109/TPAMI.2005.159.

Perron, U., Provero, P. and Molineris, I. (2017) 'In silico prediction of lncRNA function using tissue specific and evolutionary conserved expression', *BMC Bioinformatics*, Vol. 18, doi: 10.1186/s12859-017-1535-x.

Pudil, P., Novovičová, J. and Kittler, J. (1994) 'Floating search methods in feature selection', *Pattern Recognition Letters*, Vol. 15, No. 11, pp.1119–1125, doi: 10.1016/0167-8655(94)90127-9.

Roth, A., Anisimova, M. and Cannarozzi, G.M. (2012) 'Measuring codon usage bias', *Codon Evolution: Mechanisms and Models*, doi: 10.1093/acprof: osobl/9780199601165.003.0013.

Roymondal, U., Das, S. and Sahoo, S. (2009) 'Predicting gene expression level from relative codon usage bias: an application to escherichia coli genome', *DNA Research*, Vol. 16, No. 1, pp.13–30, doi: 10.1093/dnares/dsn029.

Sharp, P.M., Tuohy, T.M.F. and Mosurski, K.R. (1986) 'Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes', *Nucleic Acids Research*, Vol. 14, No. 13, pp.5125–5143, doi: 10.1093/nar/14.13.5125.

Sun, L., Luo, H., Bu, D., Zhao, G., Yu, K., Zhang, C., Liu, Y., Chen, R. and Zhao, Y. (2013) 'Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts', *Nucleic Acids Research*, Vol. 41, No. 17, doi: 10.1093/nar/gkt646.

Suzuki, H., Saito, R. and Tomita, M. (2004) 'The 'weighted sum of relative entropy ': a new index for synonymous codon usage bias', *Gene*, Vol. 335, Nos. 1–2, pp.19–23, doi: 10.1016/j.gene.2004.03.001.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., Kuhn, M., Bork, P., Jensen, L.J. and Von Mering, C. (2015) 'STRING v 10: protein-protein interaction networks, integrated over the tree of life', *Nucleic Acids Research*, Vol. 43, No. D1, pp.D447–D452, doi: 10.1093/nar/gku1003.

Tibshirani, R. (1996) 'Regression selection and shrinkage via the lasso', *Journal of the Royal Statistical Society B*, pp.267–288, doi: 10.2307/2346178.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. and Pachter, L. (2012) 'Differential gene and transcript expression analysis of RNA-seq experiments with topHat and cufflinks', *Nature Protocols*, Vol. 7, No. 3, pp.562–578, doi: 10.1038/nprot.2012.016.

Wan, X-F., Xu, D., Kleinhofs, A. and Zhou, J. (2004) 'Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes', *BMC Evolutionary Biology*, Vol. 4, p.19, doi: 10.1186/1471-2148-4-19.

Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.P. and Li, W. (2013) 'CPAT: coding-potential assessment tool using an alignment-free logistic regression model', *Nucleic Acids Research*, Vol. 41, No. 6, doi: 10.1093/nar/gkt006.

Wasaki, J., Yonetani, R., Shinano, T., Kai, M. and Osaki, M. (2003) 'Expression of the osPI1 gene, cloned from rice roots using cDNA microarray, rapidly responds to phosphorus status', *New Phytologist*, Vol. 158, No. 2, pp.239–248, doi: 10.1046/j.1469-8137.2003.00748.x.

Wu, T.T. and Lange, K. (2008) 'Coordinate descent algorithms for lasso penalized regression', *Annals of Applied Statistics*, Vol. 2, No. 1, pp.224–244, doi: 10.1214/07-AOAS147.

Wunderlich, M., Groß-Hardt, R. and Schöffl, F. (2014) 'Heat shock factor HSFB2a involved in gametophyte development of arabidopsis thaliana and its expression is controlled by a heat-inducible long non-coding antisense RNA. Plant Molecular Biology, Vol. 85, No. 6, pp.541–550, doi: 10.1007/s11103-014-0202-0.

Xiao, Y., Lv, Y., Zhao, H., Gong, Y., Hu, J., Li, F., Xu, J., Bai, J., Yu, F. and Li, X. (2015) 'Predicting the functions of long noncoding RNAs using RNA-seq based on bayesian network', *Biomed Res. Int*., p.839590, doi: 10.1155/2015/839590.

Zhao, J., Song, X. and Wang, K. (2016) 'lncScore: Alignment-free identification of long noncoding\nRNA from assembled novel transcripts', *Sci. Rep.*, Vol. 6, p.34838, doi: 10.1038/srep.34838.

Zhao, J., Song, X., Wang, K., Kaikkonen, M.U., Lam, M.T., Glass, C.K., Eddy, S.R., Blignaut, M., Zeng, X., Zhang, X., Zou, Q., Liu, Y., Zeng, X., He, Z., Quan, Z., Li, G., Yu, J., Liang, T., Zou, Q., Derrien, T., Skroblin, P., Mayr, M., Mercer, T.R., Dinger, M.E., Mattick, J.S., Ponting, C.P., Oliver, P.L., Reik, W., Mercer, T.R., Mattick, J.S., Fatica, A., Bozzoni, I., Quinn, J.J., Chang, H.Y., Wang, W.M.Z.G., Snyder, M., Amaral, P.P., Clark, M.B., Gascoigne, D.K., Dinger, M.E., Mattick, J., Volders, P-J., Xie, C., Harrow, J., Yi, X., Zhang, Z., Ling, Y., Xu, W., Su, Z., Pertea, M., Sun, L., Lv, J., Lv, J., Iyer, M.K., Legeai, F., Derrien, T., Rombel, I.T., Sykes, K.F., Rayner, S., Johnston, S.A., Min, X.J., Butler, G., Storms, R., Tsang, A., Iseli, C., Jongeneel, C.V., Bucher, P., Kong, L., Arrial, R.T., Togawa, R.C., Brigido, M.M., Johnsson, P., Lipovich, L., Grandér, D., Morris, K.V., Lin, M.F., Jungreis, I., Kellis, M., Sun, K., Achawanantakun, R., Chen, J., Sun, Y., Zhang, Y., Mattick, J.S., Rinn, J.L., Zou, Q., Hu, Q., Guo, M., Wang, G., Wang, L., Li, A., Zhang, J., Zhou, Z., Fan, X-N., Zhang, S-W., Steijger, T., Howald, C., Cunningham, F., Pedregosa, F., Haerty, W., Ponting, C.P., Claverie, J-M., Bentley, J., Fickett, J.W., Fawcett, T., Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., Roberts, A., Pimentel, H., Trapnell, C., Pachter, L., Pruitt, K.D., Tatusova, T., Maglott, D.R., Chang, C-C., Lin, C-J. and Lin, C. (2016) 'lncScore: Alignment-free identification of long noncoding RNA from assembled novel transcripts', *Scientific Reports*, Vol. 6, July, pp.34838, doi: 10.1038/srep.34838.

Zhao, Y., Li, H., Fang, S., Kang, Y., Wu, W., Hao, Y., Li, Z., Bu, D., Sun, N., Zhang, M.Q. and Chen, R. (2016) 'NONCODE 2016: an informative and valuable data source of long non-coding RNAs', *Nucleic Acids Research*, Vol. 44, issue D1, pp.D203–D208, doi: 10.1093/nar/gkv1252.