CANTATAdb: A Collection of Plant Long Non-Coding RNAs

Michał W. Szczesniak*, Wojciech Rosikiewicz and Izabela Makalowska

Department of Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University in Poznan, 61-614 Poznan, Poland

*Corresponding author: E-mail, miszcz@amu.edu.pl; Fax, +48 61 829-5949.

(Received August 30, 2015; Accepted December 6, 2015)

Long non-coding RNAs (lncRNAs) represent a class of potent regulators of gene expression that are found in a wide array of eukaryotes; however, our knowledge about these molecules in plants is still very limited. In particular, a number of model plant species still lack comprehensive data sets of lncRNAs and their annotations, and very little is known about their biological roles. To meet these shortcomings, we created an online database of lncRNAs in 10 model plant species. The lncRNAs were identified computationally using dozens of publicly available RNA sequencing (RNA-Seq) libraries. Expression values, coding potential, sequence alignments as well as other types of data provide annotation for the identified lncRNAs. In order to better characterize them, we investigated their potential roles in splicing modulation and deregulation of microRNA functions. The data are freely available for searching, browsing and downloading from an online database called CANTATAdb (http://cantata.amu.edu.pl, http://yeti.amu.edu.pl/CANTATA/).

Keywords: Database • Long-non-coding RNAs • MicroRNAs • RNA–RNA interactions • Splicing.

Introduction

Long non-coding RNAs (lncRNAs) constitute untranslated RNA molecules that are at least 200 nt long. They represent a large portion of well-annotated transcripts; for instance there are 145,331 lncRNAs known in the human transcriptome (NONCODE v4, Xie et al. 2014), which is >6-fold more than the number of protein-coding transcripts in Ensembl 77 (Cunningham et al. 2014). In plants, the numbers of lncRNAs found are usually an order of magnitude lower than in animals, but they still constitute an important component of their transcriptomes. To date, however, the functions of very few lncRNAs have been analyzed in laboratories, while in silico analyses aiming at deciphering their functions have met a number of difficulties as they are highly heterogeneous in biogenesis, sequence, structure and function. The data accumulated so far have come mostly from animal studies, and they showed that lncRNAs participate in a variety of biological processes, including transcription, splicing, translation, protein localization, imprinting, the cell cycle and apoptosis. They are also implicated in human diseases, with much attention focused on their involvement in cancer progression and development. Plant lncRNAs have also been associated with a wide array of biological phenomena, including vernalization (Heo and Sung 2011), fertility (Ding et al. 2012), photomorphogenic processes (Wang et al. 2014), phosphate homeostasis (Jabnoune et al. 2013) and regulation of symbiosis between bacteria or fungi and leguminous plants (Kouchi et al. 1999). Plant lncRNAs can play these roles by affecting different steps of gene expression. First of all, some lncRNAs represent primary transcripts for small regulatory RNAs, such as micro RNAs (miRNAs) and small interfering RNAs (siRNAs) (Ben Amor et al. 2009). For instance, in Arabidopsis thaliana, 24 nt siRNAs were shown to originate from at least five long non-coding transcripts: npc34, npc351, npc375, npc520 and npc523. Another connection to the small RNA world is target mimicry, during which interactions between miRNAs and their targets are blocked by binding of the so-called sponge RNAs to miRNAs (Wu et al. 2013). A number of such competing endogenous RNAs (ceRNAs) have been identified in humans and some animals, suggesting that it is a widespread phenomenon (Salmena et al. 2011). Plant lncRNAs are also involved in alternative splicing regulation, e.g. ASCO-lncRNA and the nuclear speckle RNA-binding (NSR) protein that forms an alternative splicing regulatory module, which has been linked to the development of lateral roots (Bardou et al. 2014). Finally, lncRNAs participate in transcriptional regulation through chromatin remodeling (Bardou et al. 2014) as well as in mRNA translation modulation (Jabnoune et al. 2013).

There is a growing body of evidence showing that a large portion of lncRNAs may exert their functions by engaging in physical interactions with mate RNA molecules. After base-pairing with other RNA molecules, lncRNAs could affect their splicing by masking splice sites and splicing signals (Beltran et al. 2008). Additionally, they could mask miRNA-binding sites on mate RNAs, thus deregulating miRNA functions (Faghihi et al. 2008). lncRNA–RNA duplexes are also expected to trigger A to I editing in mRNAs (Geisler and Coller 2013). In primates, lncRNAs have been shown to guide protein-coding transcripts to the Staufen-mediated decay pathway (Gong and Maquat, 2011). Finally, mouse lncRNAs that possess SINEB2 elements were shown to increase translation efficiency in an mRNA-independent manner (Carrieri et al. 2012). These functions, relying on the formation of lncRNA–RNA duplexes, have been documented in animals. It is
highly possible that, in plants, IncRNAs could also base-pair with other RNAs to affect their expression and processing.

The evidenced functionality of IncRNAs and our limited knowledge of their diversity and biology in plants motivated us to perform a large-scale prediction of novel IncRNAs and characterize them functionally. To this end, we took advantage of publicly available sets of RNA sequencing (RNA-Seq) data, and we built a computational pipeline for their prediction, which enabled us to find 45,117 IncRNAs in 10 plant species (Table 1). We annotated them, for example by estimating their expression values and identifying their homologs. We then aimed to determine their potential functions; however, there are no available tools to perform this task. For this reason, we developed a dedicated computational pipeline, focusing on one possibility, i.e. potential consequences of IncRNAs base-pairing with mate RNA molecules. We tested two scenarios: (i) modulating alternative splicing events by masking splicing signals; and (ii) deregulating miRNA functions. Altogether, we predicted 11,896 IncRNAs (26.37%) as being putatively involved in these phenomena. The computational results were made available through an online database that we called CANTATALdb (http://cantata.amu.edu.pl, http://yeti.amu.edu.pl/CANTATA/). The resource offers browsing, searching and downloading options to its users. To the best of our knowledge, the CANTATALdb database is the largest collection of plant IncRNAs.

### Materials and Methods

#### Data download

Genome and transcriptome sequences as well as corresponding reference annotation data were retrieved from Ensembl Plants (Kersey et al. 2014) using BioMart and a download page. Fastq files associated with 107 RNA-Seq libraries were downloaded from the European Nucleotide Archive (Silvester et al. 2014), listed on the ‘RNA-Seq libraries’ page of CANTATAdb. Mature miRNA sequences came from miRBase Release 21 (Kozomara and Griffiths-Jones 2014).

#### Identification of novel transcripts and splicing isoforms using RNA-Seq data

TopHat (Trapnell et al. 2009), a splice junction mapper, was used to map RNA-Seq reads to a corresponding plant genome. Here, short reads were mapped with BowTie, a file with known transcripts from Ensembl Plants in GTF format, and the ‘mate-inner-dist’ parameter was adjusted to library characteristics. Then, StringTie (Pertea et al. 2015) was applied to assemble the obtained alignments in a BAM format into potential transcripts, using Ensembl Plants genome annotations (GTF format) as a reference. StringTie’s output files (one file per RNA-Seq library) were then processed with Cuffmerge from the Cufflinks package (Trapnell et al. 2012) to obtain a single set of genes and transcripts and compare them with reference annotations from Ensembl Plants.

#### InCRA discovery

Based on the Cuffmerge results, transcripts annotated with class code ‘=’ were removed if the reference transcript’s biotype was one of protein-coding, rRNA, tRNA, small nucleolar RNA ( snoRNA) or small nuclear RNA (snRNA). Using CNGI (Sun et al. 2013) in a plant mode (‘model’ parameter set to ‘pl’), the coding potential of the transcripts was calculated, and those predicted to be protein coding were discarded. Then, a BLASTN search against sequences from the Rfam database was performed, and all sequences with hits of E-value < 1e-5 in sense orientation were removed. Additionally, transcripts originating from plastids or mitochondria were discarded. Finally, the transcripts were filtered for length to maintain only those that were at least 200 nt long.

#### Expression estimation

To calculate transcript expression values, RNA-Seq reads were mapped with Bowtie (Langmead et al. 2009) to all transcripts in a given species (which included our predictions enhanced with Ensembl Plants annotations) with one mismatch allowed. Reads mapping to >10 positions were discarded. Then, using in-house Python scripts, reads that mapped to more than one gene were removed and RPKM (reads per kilobase of transcript per million mapped reads) values for transcripts were calculated. The requirement for gene specificity in read mapping is expected to provide increased specificity in detecting expressed IncRNAs. With this setting, however, 18.47% of CANTATAdb IncRNAs showed no expression evidence, though they were predicted from RNA-Seq data.

#### Identification of IncRNA–RNA interactions

We used ‘lastal’ from the LAST package (Kielbasa et al. 2011) to identify potential IncRNA–RNA interactions. To this end, we created a custom substitution matrix that enabled us to search for G:U (wobble) pairs. The reason why we used ‘lastal’ is that other commonly used similarity search tools, such as BLAST, do not allow user-supplied scoring matrices, although BLAST has already been used in a similar task (Chen et al. 2012). In the substitution matrix, G/C, A/T and G/T matches were scored as 4, 2 and 1, respectively. These proportions have been widely used in the field of RNA–RNA interactions, e.g. in miRNA target search algorithms (Dai and Zhao 2014). Additionally, a mismatch was scored as −6, a gap opening as −20 and a gap extension as −8. ‘Lastex’, from the same package, enabled us to estimate a threshold value for alignment scores, so that there was no more than one alignment expected to occur by chance. In the interaction search procedure, predicted IncRNAs constituted a database, while Ensembl Plants transcripts were used as a query. When looking for IncRNA base pairings with pre-miRNAs, the query transcripts contained intron sequences; any intronic sequences distant by >250 bases from 3’ or 5’ splice sites were masked with N characters using Python scripts, based on exon co-ordinates from annotation files.

#### Retrieving interactions that could be involved in regulatory processes

To identify potential cases of splicing regulation through masking splicing signals, interactions between IncRNAs and pre-miRNAs were filtered to keep only those that spanned exon–intron borders. It was also required that the exon–intron junction was subject to alternative splicing. When looking for IncRNAs that could function as modulators of miRNA-dependent regulation, miRNA target sites were superimposed with IncRNA–RNA interaction regions. The miRNA targets were obtained using psRNATarget (Dai and Zhao 2011) with default settings; in particular, the ‘maximum expectation’ was set to 3.0, the ‘length for complementarity scoring’ was set to 20 and the ‘allowed maximum energy to unpair the target site’ was set to 25.0.

---

<table>
<thead>
<tr>
<th>Species</th>
<th>No. of IncRNAs</th>
<th>No. of transcripts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amborella trichopoda</td>
<td>2,569</td>
<td>39,095</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>4,761</td>
<td>63,619</td>
</tr>
<tr>
<td>Chlamydomonas reinhardtii</td>
<td>2,214</td>
<td>39,716</td>
</tr>
<tr>
<td>Glycine max</td>
<td>3,000</td>
<td>105,001</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>8,594</td>
<td>131,870</td>
</tr>
<tr>
<td>Physcomitrella patens</td>
<td>2,711</td>
<td>9,3581</td>
</tr>
<tr>
<td>Selaginella moellendorffii</td>
<td>2,667</td>
<td>55,552</td>
</tr>
<tr>
<td>Solanum tuberosum</td>
<td>9,692</td>
<td>89,231</td>
</tr>
<tr>
<td>Vitis vinifera</td>
<td>4,506</td>
<td>65,213</td>
</tr>
<tr>
<td>Zea mays</td>
<td>4,403</td>
<td>126,734</td>
</tr>
</tbody>
</table>
Further annotation of identified lncRNAs

To provide even more information about identified lncRNAs, we performed a number of computations using available software and data. First, we used the Coding Potential Calculator (CPC) (Kong et al. 2007), a Support Vector Machine-based classifier that calculates the protein-coding potential of transcripts using different sequence features from CNCI. With default settings, as many as 40,686 lncRNAs (90.18%) were predicted to be non-coding. With TransDecoder (Haas et al. 2013), we identified putative proteins that could originate from the sense strand of lncRNAs. To this point, we used a –m 30 parameter, which enables identification of open reading frames (ORFs) that are at least 30 amino acids long (the default threshold is 100 amino acids). In the search procedure, ORFs were not required to possess start or stop codons. Having ORFs identified, we ran a BLASTP search against a non-redundant and manually curated set of proteins from the Swiss-Prot database (UniProt Consortium 2015). If complete ORFs, i.e. those with a start and stop codon, showed significant similarity against annotated proteins (defined as hits with an E-value < 1e-5), a candidate was designated as a ‘low-confidence lncRNA’. We also ran a BLASTX search against Swiss-Prot proteins, with an E-value threshold set to 1e-5, to discover any lncRNA sequence similarity to known proteins, including regions outside identified ORFs.

Database implementation and testing

The database was constructed using Hypertext Markup Language (HTML), Cascading Style Sheets (CSS), PHP 5.4 (http://www.php.net/), MySQL 5.5 (http://www.mysql.com/), and Bootstrap 3 (http://getbootstrap.com/) framework. The database layout was also enhanced with the JavaScript Highcharts library (https://www.highcharts.com/) for interactive data plotting.

The web interface and functionalities of the database were tested on Windows (XP, 7, 8 and 8.1) and Mac OS X (Snow Leopard, Lion, Mountain Lion, Mavericks and Yosemite) operating systems using Internet Explorer (versions 6, 7, 8, 9, 10 and 11), Mozilla Firefox (versions 37 and 40), Chrome (versions 12.15 and 12.16) and Safari (versions 5.1, 6.1, 7 and 8) web browsers. Only in the case of Internet Explorer versions 6 and 7 was the web interface visualized improperly without affecting the functionality of CANTATAdb.

CANTATAdb: Database Composition and Usage

Search page

To access the data stored at CANTATAdb, the user is required to select a species from the drop-down list in the main menu. A search page will appear with the following components (Fig. 1).

Search options. Here, the user can select criteria for data search and filtering: CANTATAdb lncRNA id, CPC status (coding or non-coding), length of the longest peptide found in an lncRNA, maximal expression value, potential function (miRNA-associated or splicing regulation), lncRNA confidence and options for ordering of the results.

Search summary. In this section, user-selected search parameters are provided, and the number of records found is reported. To download the filtered data into a tab-delimited text file, one needs to press the ‘Download current results’ button. Additionally, two pie charts summarize the CPC status and potential functions for filtered lncRNAs.

Search results. The search results are displayed in a table with one row per lncRNA. The presented data include: lncRNA id, species, genomic location, best hit against Swiss-Prot proteins (if any), maximal expression value across used RNA-Seq libraries, potential function (either miRNA-associated or splicing regulation), and confidence. Low-confidence records are highlighted with a pale red color. There is also a ‘Details’ button—upon clicking this, the user gets access to more information on selected IncRNA (Fig. 2A), which includes the following:

1. The lncRNA sequence.
2. BLAST search results against the Plant Non-coding RNA Database (PNRD), Swiss-Prot and A. thaliana lncRNAs from NONCODE. Additionally, most similar lncRNAs are presented there, including potential orthologs.
3. The lncRNA expression profile across analyzed RNA-Seq libraries is shown in the form of a bar chart.
4. Information about peptides detected in the lncRNAs. In the case of ‘low-confidence lncRNAs’, there is also a red button on top of the page; upon clicking, the best five peptide alignments against Swiss-Prot proteins are visualized.
5. A list of identified lncRNA–RNA interactions, including potential function predictions (either splicing regulation or miRNA-associated) with a link to detailed information on the base-pairing (Fig. 2B, C).

From this details page, the user can return to the search page without losing the search parameters by clicking a button on the top of the page.

BLAST page

Here, one can perform a sequence-based search of data stored in CANTATAdb using BLAST version 2.2.26. A user-submitted nucleotide sequence should be in FASTA format. It is possible to select between two tools, i.e. BLASTN and MEGABLAST. Except for the expectation value (E-value) and maximal number of found hits, the search parameters are left as the default, and they are provided for user reference on the same page.

Download page

Although the search results can be easily downloaded from the search page, we also enabled a bulk data download option. All lncRNAs can be downloaded in FASTA format for each species separately. There are also files with genomic co-ordinates of lncRNAs, their potential function and annotation data, including database id, species and expression values. Finally, the best BLAST hits against selected databases of lncRNAs and Swiss-Prot proteins are available there.

RNA-Seq libraries

On this page, we summarize the RNA-Seq libraries used to estimate the lncRNA expression values. The library name with a direct link to the SRA (Short Read Archive) database is provided as well as the source of the sequenced material.

Discussion and Conclusions

Confidence of found lncRNAs

Determining the protein-coding ability for a transcript is a critical point in the identification of lncRNAs, yet it represents quite a challenging task. First of all, algorithms designed for
Coding potential calculations generate a significant number of false positives and false negatives. For example, CNCI showed 97.3% accuracy when tested on human protein-coding and long non-coding transcripts (Sun et al. 2013). Additionally, lncRNAs are likely to possess ORFs purely by chance (Dinger et al. 2008), so discarding any transcripts containing ORFs would result in losing a substantial number of true lncRNAs. For example, we were able to detect peptides >30 amino acids long in 74.63% of the lncRNAs stored at CANTATAdb and in 80.24% of the PNRD lncRNAs. Regarding peptides that do not exceed 100 amino acids, the proportions are 82.53% and 76.63% in our database and PNRD, respectively, while long peptides (exceeding 200 amino acids) are contained in 8.19% of CANTATAdb lncRNAs, compared with 6.35% in PNRD. This problem might be mitigated by exploiting protein sequence conservation, with the assumption that lncRNA-derived ORFs should not be conserved unless they are translated and their products are functional. We noticed that 8.23% of our lncRNAs contained ORFs with start and stop codons and showed significant similarity to known proteins from Swiss-Prot (E-value threshold of 1e-20). The same was true for only 4.72% of the lncRNAs in PNRD and 5.40% of plant lncRNAs in NONCODE. What is more, only 43.33% were predicted by CPC as non-coding, compared with 90.18% in the case of all the CANTATAdb lncRNAs. We dubbed them ‘low-confidence lncRNAs’ as their non-coding status was questionable. To discriminate them from other candidates, they are marked with a pale red color in the database.

Comparison with existing databases of lncRNAs

There are already several databases that store plant lncRNAs, yet they are much less comprehensive and store fewer data.
Fig. 2 An example of the ‘Details page’ (A) and information regarding potential roles of a lncRNA in deregulation of miRNA functions (B) and in modulation of alternative splicing (C).
than CANTATAdb. The biggest of them, i.e. the PNRD (Yi et al. 2015), contains fasta sequences of IncRNAs from four plant species: A. thaliana (2,577 records), Oryza sativa (752 records), Populus trichocarpa (538 records) and Zea mays (1,704 records). NONCODE v4 stores IncRNAs for eight animals, yeast and 3,853 IncRNAs from A. thaliana. LncRNAdb (Quek et al. 2015) collects experimentally tested IncRNAs from nine plant species: seven from A. thaliana and between one and three from other species.

The recently released PLNcRBase (Xuan et al. 2015) contains manually collected data from nearly 200 publications, covering a total of 1,187 plant IncRNAs in 43 plant species. Finally, PlncDB (Jin et al. 2013) stores 13,230 novel transcription units from A. thaliana, including long intergenic RNAs (lncRNAs), identified using high-throughput technologies, such as TILING arrays and RNA-Seq sequencing. We found that 53.22% of CANTATAdb IncRNAs for A. thaliana are contained in PlncDB. A detailed comparison between the two resources is provided in Supplementary Table S1.

A distinguishing feature of CANTATAdb is annotation data, including predicted functions in the context of IncRNA–RNA interactions. Altogether, 11,896 IncRNAs have assigned functions, including 440 IncRNAs that are thought to take part in deregulation of miRNA functions and 11,659 IncRNAs that could function as splicing modulators through masking splicing signals. Notably, only 4.35% of them were predicted as coding by the CPC, and 2.97% of them are classified as ‘low-confidence IncRNAs’. For all CANTATAdb IncRNAs, these numbers were 9.82% and 8.23%, respectively. Moreover, in the case of 66.65% of IncRNA–RNA interactions, both RNA components are co-expressed in at least one of the libraries used in this study, which involved 73.50% of all unique IncRNAs with predicted functions. These calculations provide support for predicted functions and the non-coding status of particular IncRNAs, but extensive laboratory tests are required to check the predictions experimentally. Moreover, there are a number of other possibilities regarding roles of plant IncRNAs. With this in mind, we are planning further studies to enhance the annotation of IncRNAs stored at CANTATAdb. Still, we believe that the provided IncRNA sequences and their annotation, including bona fide functions playing a role in the context of IncRNA–RNA interactions, will contribute significantly to future efforts aimed at deciphering the biology of plant IncRNAs.

**Disclosures**

The authors have no conflicts of interest to declare.

**References**


**Supplementary data**

Supplementary data are available at PCP online.

**Funding**

This work was supported by the National Science Centre [grant No. 2014/15/D/NZ2/00525 to M.W.S.]; the Foundation for Polish Science [a START Scholarship grant editions 2014/2015 and 2015/2016 to M.W.S.]; the KNOW Poznan RNA Centre [grant No. 01/KNOW2/2014].


