BMC Bioinformatics

DATABASE

Open Access

Lantern: an integrative repository of functional annotations for lncRNAs in the human genome

Swapna Vidhur Daulatabad[1], Rajneesh Srivastava[2] and Sarath Chandra Janga[1,3,4*]

*Correspondence: scjanga@iupui.edu
[1] Department of BioHealth Informatics, School of Informatics and Computing, Indiana University Purdue University, Informatics and Communications Technology Complex, 535 W Michigan St., IT 475H, Indianapolis, IN 46202, USA
Full list of author information is available at the end of the article

Abstract

Background: With advancements in omics technologies, the range of biological processes where long non-coding RNAs (lncRNAs) are involved, is expanding extensively, thereby generating the need to develop lncRNA annotation resources. Although, there are a plethora of resources for annotating genes, despite the extensive corpus of lncRNA literature, the available resources with lncRNA ontology annotations are rare.

Results: We present a lncRNA annotation extractor and repository (Lantern), developed using PubMed's abstract retrieval engine and NCBO's recommender annotation system. Lantern's annotations were benchmarked against lncRNAdb's manually curated free text. Benchmarking analysis suggested that Lantern has a recall of 0.62 against lncRNAdb for 182 lncRNAs and precision of 0.8. Additionally, we also annotated lncRNAs with multiple omics annotations, including predicted cis-regulatory TFs, interactions with RBPs, tissue-specific expression profiles, protein co-expression networks, coding potential, sub-cellular localization, and SNPs for ~ 11,000 lncRNAs in the human genome, providing a one-stop dynamic visualization platform.

Conclusions: Lantern integrates a novel, accurate semi-automatic ontology annotation engine derived annotations combined with a variety of multi-omics annotations for lncRNAs, to provide a central web resource for dissecting the functional dynamics of long non-coding RNAs and to facilitate future hypothesis-driven experiments. The annotation pipeline and a web resource with current annotations for human lncRNAs are freely available on sysbio.lab.iupui.edu/lantern.

Keywords: LncRNA, Ontology, Expression, Interaction, eCLIP, Coding potential, Snps, eQTL

Background

Non-coding RNAs across the genome have been associated with a variety of biological processes [1–4], ranging from regulation of splicing to remodeling of chromatin [5, 6]. Amongst the repertoire of non-coding sequences, lies a critical species of RNAs called long non-coding RNAs (lncRNAs) [7]. An increasing number of studies suggest that lncRNAs significantly contribute to a large spectrum of human phenotypes including cancers [8–10], neurological disorders such as Alzheimer's disease [11], Heart failure

[12], and Diabetes [13]. However, the complete functional role of lncRNAs is not fully characterized [14–17]. Hence, uncovering the role of lncRNAs in disease phenotypes is critical for targeting them as potential therapeutic agents [18, 19].

This rapidly emerging field, therefore, requires the development of a robust framework for comprehensive annotations of lncRNAs. Several approaches have been published so far to annotate and characterize lncRNAs [20–23]. The GENCODE project [24] has identified thousands of lncRNAs across the human genome out of which only a small percentage are functionally understood, leaving a significant part of the functional non-coding genome unexplored.

Currently, available lncRNA annotation resources attempt to bridge the gap between lncRNAs and their biological functions. Although resources like Noncode [25], lncRNAdb [26], lncRNome [27], LNCipedia [28] and lncRNADisease [29] have functionally annotated lncRNAs, they are either limited to a small set of lncRNAs or do not provide an extensive understanding of their biological functions.

In addition to controlled annotations, the available resources have not elaborated other key parameters that elucidate the role of lncRNAs like their tissue specificity, sub-cellular compartment localization, and cross-talk with various cellular components.

In this study, we propose a lncRNA annotation extractor and repository, Lantern, which renders users annotate lncRNA with multiple ontologies along with multi-omic lncRNA annotations. Lantern integrates lncRNA functional ontology information extracted from natural language processing of abstracts from PubMed, protein -RNA interactions obtained from CLIP experiments, expression profiles across human tissues, and lncRNA characteristics like protein-coding potential, SNPs, and sub-cellular localization into one resource.

Availability of such functional annotations for lncRNAs can aid experimental scientists to design hypothesis-driven research based on annotations from Lantern. Therefore, Lantern will significantly aid our understanding of the non-coding transcriptome and its role in human disease phenotypes.

Overview of Lantern

Lantern hosts multi-omic annotation information for 11,290 lncRNAs, amongst which, 769 lncRNAs are annotated with multiple functional ontology information extracted from 6942 abstracts of lncRNA pertaining literature. A total of 9982 lncRNAs were annotated with tissue-specific expression across 53 tissues from the GTEx data portal, and 6714 lncRNAs were annotated with SNPs across 48 tissues based on eQTL and SNP information from GTEx and GWAS studies. Additionally, the protein-coding potential of 9898 lncRNAs was computed and recorded. Sub-cellular localization of 11,290 lncRNAs in 10 cellular compartments across 15 cell lines was also obtained and hosted on Lantern. A total of 617,074 interactions across 161 transcription factors and 10,727 lncRNAs were also predicted and mounted onto the Lantern interface. RNA binding protein (RBP) interactions of 7942 lncRNAs across 18 cell lines were computed from ENCODE eCLIP data, along with protein co-expression information, extracted from analysis based on information from Human Proteome Map and GTEx, for 5331 lncRNAs are available on Lantern (see Construction and Content section). Table 1 summarizes the various datasets employed in constructing Lantern. The functional ontology annotations

Table 1 Various annotation components available on Lantern for lncRNAs along with the number of annotated lncRNAs with respective annotations along with their coverage

Annotation	Number of lncRNAs with respective annotation	Features
Ontology	769	Gene ontology, human phenotype ontology, disease ontology, and SNOMEDCT ontology extracted from 6942 abstracts
RBP-interactions	7942	Across 18 cell lines
Tissue-specific expression	9982	Across 53 tissues
Protein co-expression	5331	Across 14 tissues
Transcription factor interaction	10,727	Potential 617,074 regulatory interactions with 161 Transcription Factors
Coding Potential	9898	For 27,907 transcripts
Sub-cellular localization	11,290	For 10 cellular compartments across 15 cell lines
GTEx eQTL SNPs	6714	Across 48 tissues
GWAS SNPs	2569	1421 phenotypic characteristics

extracted by the NLP-based literature mining framework, the Mobius pipeline, were benchmarked against the manually curated gold standard annotations from lncRNAdb. From the benchmarking analysis, we observe that our pipeline can recall 62% of the original annotations, and the extracted annotations had a precision of 80% (see "Construction and Content" section). Around 26% of all the lncRNAs with gene ontology annotations extracted using our novel Mobius pipeline could be benchmarked based on available annotations from prior resources. Across these Mobius pipeline-based annotations of lncRNAs, we observed that 'NEOPLASM'/'TUMOR' (292 lncRNAs associated), 'TYPE I PROGRAMMED CELL DEATH' (240 lncRNAs associated) ontologies were associated with the most number of lncRNAs. Users can search and visualize various levels of lncRNA annotations, information including where a lncRNA localizes in the cell, which tissue is more representative of a specific lncRNA, which RBPs interact with which lncRNAs, SNPs prevalent on lncRNA specific to tissue and phenotype, and most importantly controlled ontology annotations. Lantern enables users to understand the interactive nature of lncRNAs and visualize characteristics of lncRNAs to design studies that help further explore the functional scope of the non-coding transcriptome. As conceptualized in Fig. 1, Lantern is a resource generated by the integration of multiple pipelines extracting and presenting several key levels of lncRNA annotation information.

Construction and content

In this study, we integrated multiple pipelines to extract functional annotations and cellular interactions of lncRNAs as illustrated in Fig. 1. The crux of Lantern is a semi-automated ontology annotation pipeline, hereby referred to as the "Mobius" pipeline, which uses concepts of Natural Language Processing (NLP) to mine lncRNA relevant literature to extract accurate non-coding transcriptome ontology annotations. The Mobius pipeline annotates lncRNAs with 4 ontologies: gene ontology (GO), human phenotype ontology (HPO), disease ontology (DO), and SNOMEDCT ontology. The annotations extracted by the Mobius pipeline were benchmarked against an established gold standard repository, and the pipeline was deployed to overall GENCODE registered lncRNAs

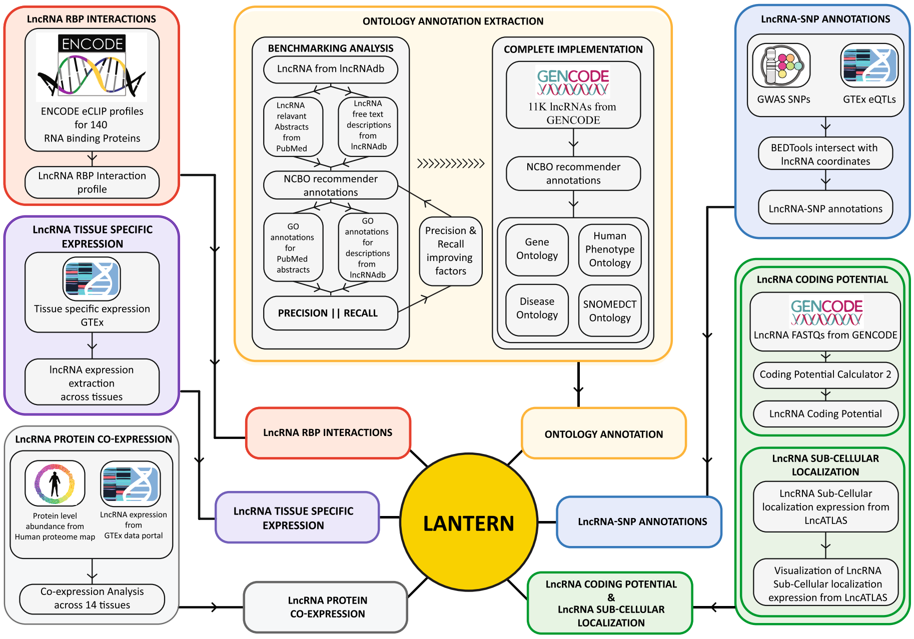

LncRNA RBP INTERACTIONS
ENCODE
ENCODE eCLIP profiles for 140 RNA Binding Proteins
LncRNA RBP Interaction profile
LncRNA TISSUE SPECIFIC EXPRESSION
Tissue specific expression GTEx
lncRNA expression extraction across tissues
LncRNA PROTEIN CO-EXPRESSION
Protein level abundance from Human proteome map
LncRNA expression from GTEx data portal
Co-expression Analysis across 14 tissues
ONTOLOGY ANNOTATION EXTRACTION
BENCHMARKING ANALYSIS
LncRNA from lncRNAdb
LncRNA relavant Abstracts from PubMed
LncRNA free text descriptions from lncRNAdb
NCBO recommender annotations
GO annotations for PubMed abstracts
GO annotations for descriptions from lncRNAdb
PRECISION || RECALL
Precision & Recall improving factors
COMPLETE IMPLEMENTATION
GENCODE
11K lncRNAs from GENCODE
NCBO recommender annotations
Gene Ontology
Human Phenotype Ontology
Disease Ontology
SNOMEDCT Ontology
LncRNA-SNP ANNOTATIONS
GWAS SNPs
GTEx eQTLs
BEDTools intersect with lncRNA coordinates
LncRNA-SNP annotations
LncRNA CODING POTENTIAL
GENCODE
LncRNA FASTQs from GENCODE
Coding Potential Calculator 2
LncRNA Coding Potential
LncRNA SUB-CELLULAR LOCALIZATION
LncRNA Sub-Cellular localization expression from LncATLAS
Visualization of LncRNA Sub-Cellular localization expression from LncATLAS
LncRNA RBP INTERACTIONS
ONTOLOGY ANNOTATION
LncRNA TISSUE SPECIFIC EXPRESSION
LANTERN
LncRNA-SNP ANNOTATIONS
LncRNA PROTEIN CO-EXPRESSION
LncRNA CODING POTENTIAL & LncRNA SUB-CELLULAR LOCALIZATION

Fig. 1 Lantern workflow overview. The flow chart shows various NGS and ontology extraction pipelines merging to annotate lncRNAs with multi-level omics and ontology information. All the annotation pipelines were integrated, and the extracted annotations are hosted on the online resource. The ontology annotation extraction pipeline showing a two-part workflow, the first part benchmarks the pipeline developed, and the annotations extracted, second part shows the implementation of the developed pipeline over all the human lncRNAs from GENCODE. Respective flowcharts on either side showing data resources and steps involved in extracting a diverse set of omics data for annotating lncRNAs

and the data was deposited onto an online resource. Additionally, we used established Next-Generation Sequencing (NGS) data analysis approaches to annotate lncRNAs with diverse omics information such as lncRNA-RBP interactions, tissue-specific lncRNA expression, lncRNA-protein co-expression, coding potential, sub-cellular localization, and SNPs in lncRNA.

Ontology annotation for long non-coding RNAs (lncRNAs)

The lncRNA ontology annotation extraction pipeline, the Mobius pipeline, contains two parts as seen in Fig. 1. In the first part of this implementation, the pipeline is deployed over a subset of lncRNAs and benchmarked against an established manually curated lncRNA annotation resource. In the second part, the benchmarked pipeline is deployed over human genome-wide lncRNAs from GENCODE, to extract annotations for all the lncRNAs that have been studied so far. The individual steps involved in the Mobius pipeline are as follows:

Abstract extraction for lncRNA relevant literature from PubMed

Our gold standard for this study, lncRNAdb [26] annotates lncRNAs with function, localization, characteristics, and conservation information by manual curation of literature and recording information for each lncRNA, thereby making it an appropriate gold standard. We extracted all the 298 lncRNA names from lncRNAdb which had characteristic information. These extracted lncRNA names were then plugged into PubMed's search engine via an application programming interface (API) using python, to retrieve abstracts relevant to the respective lncRNA. To make the results precise, additional key terms such as 'RNA', 'lncRNA', 'long non-coding RNA', and 'non-coding RNA' were added to the lncRNA names before performing the search. The resulting hits were obtained as XML of PMIDs. These PMIDs were then used to obtain the abstracts using PubMed eUtils [30]. The output abstracts were also extracted in XML and were parsed using python to extract specific information like abstract, and PMID.

Ontology extraction using NCBO recommender system

The National Center for Biomedical Ontology (NCBO) [31] recommender parses a given input text to identify and suggest the most representative ontology based on the keywords contained in the text. To identify the functional ontology terms, the extracted abstracts were pushed to the NCBO ontology recommender endpoint using REST API [32]. The NCBO ontology recommender then identified and returned the annotations in the posted text with respect to 4 ontologies; Gene ontology [33], Human phenotype ontology [34], Disease ontology [35], and SNOMEDCT [36]. The returned ontology associations were annotated with the respective lncRNAs. Along with lncRNA-specific ontology annotation, the corresponding ontology IDs were extracted and recorded.

Benchmarking predicted lncRNA ontology annotations against a gold standard

To measure the accuracy of the extracted ontology information, we compared our annotations with those from lncRNAdb, across 182 lncRNAs that had annotation from both sources. We extracted all the characteristic descriptions of each lncRNA from lncRNAdb as free text. The extracted free text was then parsed through NCBO ontology recommender as previously performed with lncRNA abstracts from PubMed. Thereby obtaining annotations for gold-standard information, which was now comparable to the annotations extracted from PubMed abstracts. The number of ontology IDs common across annotations from the Mobius pipeline and gold standard free text from lncRNAdb was computed for each lncRNA, based on which the recall was calculated using the following equation (1).

$$\text{Recall} = \frac{\text{Number of shared annotations across extracted and gold standard annotations}}{\text{Total number of annotations extracted}} \tag{1}$$

The lncRNA functional ontology annotations extracted by the Mobius pipeline demonstrated a recall of 62% across the annotations extracted from the gold standard, lncRNAdb.

Since the gold standard lacks updated annotations and the Mobius pipeline extracts annotations from the literature until December 2018, precision was calculated in a non-traditional approach. Precision was manually computed by randomly selecting 50 lncRNAs, reading through each of the annotated abstracts to identify which annotations from the Mobius pipeline were accurate and which were misannotated. Subsequently, the precision of the Mobius pipeline was computed as 80% using equation (2). The record of this curation step is provided in Additional file 1: Table S1.

$$\text{Precision} = \frac{\text{Number of true annotations extracted}}{\text{Total number of annotations extracted}} \quad (2)$$

Genome-scale ontology annotation for lncRNAs using Mobius pipeline

As seen in Fig. 1, after the benchmarking analysis, the Mobius pipeline was deployed overall 11,405 human lncRNAs recorded in GENCODE, to annotate the non-coding transcriptome as much as possible with functional ontology terms. The PubMed abstract retrieval system was employed over these 11,405 lncRNAs, out of which 769 lncRNAs had at least one abstract mentioning the respective lncRNA's name in the abstract or title of the article. A total of 6942 unique abstracts were extracted for 769 lncRNAs. These abstracts were then parsed by the NCBO ontology recommender to mine ontology annotations present in the free text. The PMIDs and their respective ontology annotations were retrieved and reannotated to the respective lncRNAs, to obtain 698 lncRNAs with Gene Ontology annotations, 433 lncRNAs with Human Phenotype ontology annotations, 239 lncRNAs with Disease ontology annotations and 357 lncRNAs with SNOMEDCT ontology annotations. These annotations were formatted into tables and fed into the database to be visualized as interaction network of lncRNAs and respective ontology annotations on Lantern, the same information can also be visualized in form of tables.

LncRNAdb is a resource that was generated by manually curating literature, which enables capturing of deeper functional annotations which contemporary natural language processing approaches cannot retrieve, which is why there is a variation in the number of lncRNAs that are recorded on lncRNAdb and the number of lncRNAs which had PMIDs annotated. Also, our approach extracts annotations that are solely based on the abstract retrieved, as we observed that annotations from other sections of the articles gave rise to substantial noise in the annotations. While extracting the abstracts for respective lncRNA, we observed that the recall spikes as the number of annotated abstracts increases, however the same depletes as the number of abstracts exceeds 150 as seen in Fig. 2. Therefore, we administered an upper limit of 150 abstracts, to reduce the noise or non lncRNA specific abstracts. We put forth the rationale for selecting the four ontologies annotated as gene ontology, human phenotype ontology, disease ontology, and SNOMEDCT because these were the most representative of lncRNA function and are tangible for translational observations. While benchmarking the retrieved ontology annotations, the number of annotations retrieved for human phenotype ontology, disease ontology, and SNOMEDCT were very few from the gold standard to make any comparisons. NCBO returns 5 scores with respect to annotations for each ontology, however, these scores represent the most appropriate annotated ontology for a given text, which in the case of Lantern are auxiliary, as Lantern tries to extract annotations rather than compare ontologies. Therefore, to filter the extracted annotations we employed a frequency of occurrence-based approach. In the post-search result page of the Lantern, annotated ontology IDs are sorted and displayed by their frequency of occurrence across the source. Another filter we employed was filtering out the non-informative, non-specific annotations like "gene", "cell", "protein". This filtration step was performed by manually mining through all the annotations and identifying 1478 non-informative annotations and removing these annotations from the database tables. Lantern

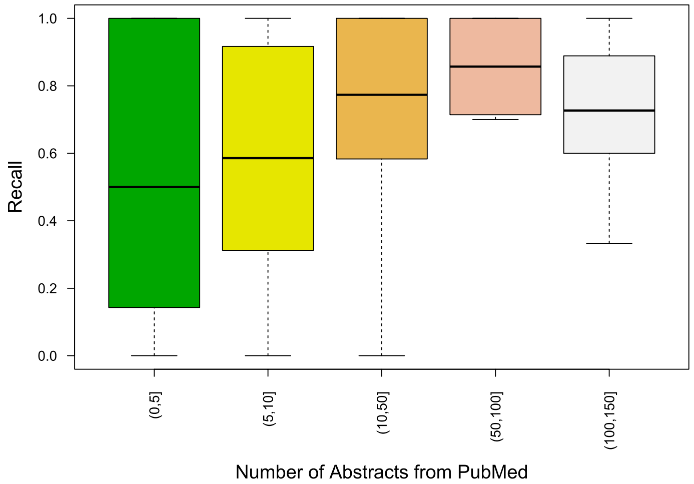


Fig. 2 Boxplots showing the distribution of recall values across lncRNAs categorized based on the number of abstracts retrieved per lncRNA from PubMed. This analysis suggests that the more the number of abstracts available for a lncRNA, the higher is the recall for annotations extracted by Lantern. Our data also shows an optimal peak of recall at 100 to 150 abstracts extracted per lncRNA

also provides additional features and integrated multi-omics resources for the LncRNAs as follows:

Mapping lncRNA-RBP interactions by mining publicly available CLIP-seq datasets

Long non-coding RNAs are well known to interact with RNA binding proteins (RBP) in a function-specific and cell-type-specific manner [37]. To understand the functional role of the non-coding transcriptome, it is crucial to dissect and record RBP–lncRNA interaction profiles. Therefore, we extracted lncRNA-RBP interactions from UV cross-linking protocol-based protein-RNA interaction (Crosslinking and Immunoprecipitation;CLIP) experiments from ENCODE [38] project, for 7942 lncRNAs and 139 RBPs in 18 cell lines. The downloaded bed file containing the cell line specific binding site coordinates of 139 RBPs were parsed thoroughly over the annotated lncRNA loci from GENCODE using BEDTools 'intersect' option [39]. The extracted 468,348 interactions of RBPs and lncRNAs were tabulated and mounted onto the Lantern database for efficient retrieval and visualization in the form of a network. Additionally, we extracted the Pfam domain information of interacting RBPs from the biomart [40] and incorporated it in Lantern to display the Pfam ID and description along with the cell lines. Also, we further analyzed CLIP-seq based lncRNA-RBP interactions to identify the RBP families most frequently interacting with the long non-coding transcriptome. From our examination, we observed that the RBPs belonging to HNRNP and CSTF2 groups were the most interactive with 5538 lncRNAs associated on average.

Comparative visualization of lncRNA expression profiles across human tissues from GTEx RNA-seq samples

To understand how each lncRNA abundance varies across human tissues, we downloaded the transcript-level expression profile as a quantification matrix across 53 tissues from the GTEx data portal [41]. RNA-seq derived transcript abundance data were extracted for each lncRNA. After extracting the expression levels across all the available tissues from the GTEx portal, python package 'pandas' and 'pylab' [42] was used to generate dynamic boxplots presenting the queried lncRNA expression. This segment of Lantern can help the user to visually interpret the nature of 9982 lncRNAs across multiple tissues.

Generating lncRNA: protein co-expression associations by integrating tissue-specific expression profiles

Several studies show the role of LncRNAs in multiple biological processes including transcriptional/ post-transcriptional regulation [43–46]. A majority of them are characterized for their conservation across multiple species [47, 48]. Therefore, it is imperative to infer the potential crosstalk between protein(s) and lncRNAs. To establish such molecular interaction between LncRNAs and proteins, we downloaded the protein level abundance of 9983 proteins from the human proteome map (HPM) [49] and lncRNA expression data from the GTEx portal [41] as described previously. We considered the 14 tissues which are common between the two databases, to deploy the co-expression analysis and established a rank-based association network between protein and lncRNA expression levels using spearman rank correlation [50]. From the co-expression analysis, each instance of protein-lncRNA co-expression was annotated with r-value, p-value, and FDR to gauze the intensity and significance of the interaction. The resulting high confidence (5% FDR) association network with a total of 15,314 nodes and 5,038,058 edges across all 5331 lncRNA was visualized for each lncRNA on Lantern using Cytoscape JavaScript framework [51].

Prediction of potential upstream regulatory transcription factors of lncRNAs

LncRNAs are known to express in a tissue-specific manner and hence require an investigation of upstream regulators such as transcription factors genome-wide. We made an in silico attempt to navigate the possible TFs regulating the lncRNAs. First, we downloaded the Position Specific Weight Matrices (PWMs) of ~ 2100 transcription factors from TRANSFAC [52] and MEME [53] databases. Also, we used bedtools [54] to extract the 2 kb upstream + 0.5 kb instream sequence from the TSS (Transcript Start Site) of all lncRNAs (based on lncRNA genomic coordinates from Ensembl hg38.p84) in fasta format. Next, PWMs of these TFs were scanned onto the extracted regulatory region of lncRNAs using Find Individual Motif Occurrences (FIMO) [55] with default parameters. The resulting motif-based binding sites were filtered at a 5% FDR cutoff. A total of 617,074 interactions across 161 transcription factors and 10,727 lncRNAs were identified and mounted onto the Lantern interface.

Estimating the protein-coding potential of lncRNAs at genome-scale

The coding potential is a useful metric that helps to differentiate non-coding transcripts from other biotypes. Coding potential can not only help to dissect and elucidate the functionality of lncRNAs but also help identify novel lncRNAs [56]. For computing the coding potential of lncRNAs, we extracted the genomic coordinates of lncRNA transcripts from GENCODE (hg38) in GTF (Gene Transfer Format) and deployed a robust machine learning-based approach, Coding Potential Calculator 2 (CPC2) [57], which considers sequence features to estimate the coding potential of a transcript. CPC2 was deployed across human lncRNA coordinates, acquired in the form of GTF from GENCODE, to generate the coding potential estimates for all lncRNA. The coding potential of 27,907 lncRNA transcripts, mapped to 9898 lncRNAs can be seen on Lantern along with intrinsic information such as peptide length or length of the open reading frame, Fickett score [58] for the transcript, isoelectric point, and open reading frame integrity, which depicts if the ORF starts and stops with the appropriate codons.

Annotating and visualizing lncRNAs' preferential sub-cellular localization

A crucial element that directly affects the function of a given transcript is its cellular localization [59]. Similarly, the function of lncRNAs' is also influenced by where the lncRNA is localized in the cell [60, 61]. Therefore, it is critical to understand the compartment-specific localization of lncRNA. To investigate this, we obtained the lncRNA localization information from lncATLAS [62], which was originally computed from ENCODE consortium data [38]. The lncRNA expression localization information of around 11 K lncRNAs across 15 cell lines and 10 sub-cellular compartments was processed and mounted onto the database, to be visualized as a heatmap with cell lines and sub-cellular compartments as axes. The heatmap can be viewed in two normalization formats, row normalized (i.e. normalized with respect to compartments across each cell line), which is the default, and column normalized (i.e. normalized with respect to cell lines across each compartment). The heat map visualization was enabled using the high charts JavaScript framework.

Transcriptome-wide identification of eQTL and GWAS SNPs occurring on lncRNAs

Another key factor that can influence the functions of lncRNAs in a cellular context is the single nucleotide polymorphism (SNP) [63–65]. It is reported that more than 85% of SNPs that are associated with disease occur in the non-coding regions [66]. This makes it imperative to explore various SNPs associated with lncRNAs. Therefore, to identify SNPs pertaining to lncRNAs, and their effect on the cellular and tissue level behavior, we mapped SNPs from the GWAS catalog and GTEx eQTL analysis onto lncRNAs. The latest version of the GWAS SNP association was obtained from the GWAS catalog [67]. To examine for SNPs across lncRNA, SNP information from GWAS was processed and converted to BED format. Using BEDTools [54], 5897 GWAS SNPs whose coordinates overlapped with lncRNA coordinates, obtained from GENCODE, were extracted along with their physical trait information. Each GWAS SNP on lncRNA is annotated with the dbSNP 'rsid' [68], position, and phenotypic trait.

The Expression quantitative trait locus (eQTL) analysis from the GTEx project [41] has mapped SNPs to altered gene expression across 48 tissues. This expression dysregulating tissue-specific 39,741 SNPs were mapped onto lncRNAs. The obtained eQTLs were recorded along with information such as dbSNP 'rsid' for the respective SNP, tissue in which the SNP was observed, genomic coordinates, and an r-value. Thereby recording 45,530 SNPs in 6714 lncRNAs, across 48 tissues, along with GWAS annotated phenotypic trait and GTEx annotated eQTL information. This extensive SNP-lncRNA annotation can help understand and explore the phenotypic nature of SNPs across lncRNAs.

Database and web interface

A web resource was developed for hosting all the extracted and benchmarked annotations. In the back-end, a MySQL database was used to create and connect tables with lncRNA-specific annotations. The database schema was a straightforward connection across each annotation component as a separate table and lncRNA names as a unique table, associated with the lncRNA name. The front-end is a 'php' based, Twitter bootstrapped webpage, developed to generate a user-friendly and intuitive interface. The search boxes are enabled with auto-suggestions based on user input using jQuery.

Lantern can be browsed for annotations via two search routes, user can search by lncRNA, using lncRNA names and Ensembl gene ID (ENSG), by toggling between search buttons above the search bar, users can also search for ontologies on the adjacent search tab, to retrieve all the lncRNAs annotated to that ontology. Lantern also provides network visualizations of lncRNA and its annotations across ontology annotation, lncRNA-RBP interactions, and protein co-expression. These network visualizations were enabled using the Cytoscape JavaScript framework. All the tables retrieved after search on Lantern can be downloaded. Some columns in the tables retrieved after search on Lantern are hyperlinked to various resources, showing additional information about that respective tuple. For instance, the PMID column in the ontology annotation retrieved table redirects the user to all the literature that supports the annotation. Similarly, the ontology ID column, gene ID columns, Pfam ID column, and dbSNP rsid columns are as well hyperlinked. In the tissue-specific lncRNA expression section, we embedded a hyperlink to FireBrowse [69] which redirects the user to the specific lncRNA's expression across multiple cancers. The tissue-specific expression of lncRNA is depicted as boxplots across the tissues, this image can be downloaded using the embedded buttons. The sub-cellular localization of lncRNA can be visualized as a heatmap showing intensity of a lncRNA's abundance across compartments and cell lines, this image can be exported in multiple formats. The downloads page on Lantern enables the user to retrieve and download all available information for that lncRNA on one page. The documentation page briefly describes the implementations on Lantern and the user guide page enables users to get an overview of all the functionalities on Lantern.

Utility and discussion

Lantern facilitates analysis and annotation of lncRNAs on multiple levels using established computational pipelines. It employs a swift integration of Application Programming Interface (API), established tools, and resources to extract the most recent and relevant lncRNA functional annotations. The ontology annotations available on Lantern were extracted by a semi-automated pipeline which leverages the potential of Natural Language Processing (NLP), by using it to mine recent lncRNA literature. The extracted ontology annotations are benchmarked against manually curated gold standard lncRNA annotation resources. Lantern also houses lncRNA annotations across a range of omics data, which provides extensive insights into lncRNA's role from a fundamental level, as cellular functions, to an empirical scale such as tissue level. Information such as a list of RBPs that have been experimentally identified to interact with lncRNAs, proteins that co-express with lncRNAs, tissues that are more representative of a specific set of lncRNAs, sub-cellular expression of lncRNAs, and Single Nucleotide Polymorphisms in lncRNAs can essentially enable the exploration of the functional role of lncRNAs. Lantern attempts to bridge this knowledge gap with accurate and up-to-date lncRNA annotations.

Utility case

LncRNAs have been observed to be involved in a spectrum of regulatory processes, including remodeling of chromatin, regulation of splicing, translation of mRNA, and stability of mRNA, and protein [70]. Despite having a broad range of regulatory roles, lncRNAs are known to have tissue or phenotype exclusive regulatory facets [42, 47, 71, 72]. Thereby it is important to, not only understand lncRNAs and cellular elements interacting with them but also to investigate the role of lncRNAs in the context of tissue or phenotype exclusivity. Lantern precisely captures such annotations of lncRNAs and projects them via dynamic visualizations. For instance, metastasis-associated lung adenocarcinoma transcript 1 (MALAT1) is one of the extensively studied lncRNAs. It is known to play a key role in various disorders like lung adenocarcinoma [73–76], hepatocellular carcinoma [77], ovarian cancer [78], osteosarcoma [73, 79], and atherosclerosis [80], all of which Lantern captures and portrays as an ontology network as seen in Fig. 3A. MALAT1 is observed to be upregulated in various types of cancer and acts as a proto-oncogene in hepatocellular carcinoma [81, 82]. It is widely known to interact with the oncogenic splicing factor SRSF1[83]. Lantern captures the phenotypic annotations attributed to hepatocellular carcinoma that were documented previously along with its known tissue-specific biological interaction with SRSF1 observed in HepG2 [84]. In addition to that, Lantern also provides other accurately annotated RBP interactions as shown in Fig. 3B. These interactions highlight the underlying mechanism of RNA binding proteins and lncRNAs as co-mediators of cellular functions and phenotypes, as previously seen in literature [85–89]. Therefore, such interactions available on Lantern facilitates research groups to further understand the potential lncRNA-centric regulome and associated co-complexing RBPs.

To further portray and understand lncRNAs' expression pattern in multiple tissues, we downloaded the processed expression profile of lncRNAs across 53 tissues from the GTEx portal [41]. We integrated the dataset in Lantern as a boxplot to show the comparative expression plot that helps the user to visually interpret the nature of 9982 lncRNAs. MALAT1, as described previously is observed with a global expression profile across multiple tissues, as seen in Fig. 3C, in coherence with the range of phenotypes and diseases it is associated with. In contrast, Lantern also reports certain highly tissue-specific lncRNAs. For instance, lncRNA HULC, which is known to be highly expressed in hepatocellular carcinoma [90–93], demonstrates a distinct signal of upregulation in the liver, as seen in Fig. 3D.

LncRNAs are recruited based on their functional role and abundance in a given sub-cellular compartment [94, 95]. Therefore, it is essential to understand and explore the respective sub-cellular compartment abundance of lncRNAs. Lantern provides an interactive visual platform to investigate the lncRNA abundance across cellular compartments in 15 cell lines, available as a heatmap. MALAT1 for instance has extensively been identified as a nuclear lncRNA [96–99].In agreement with previous

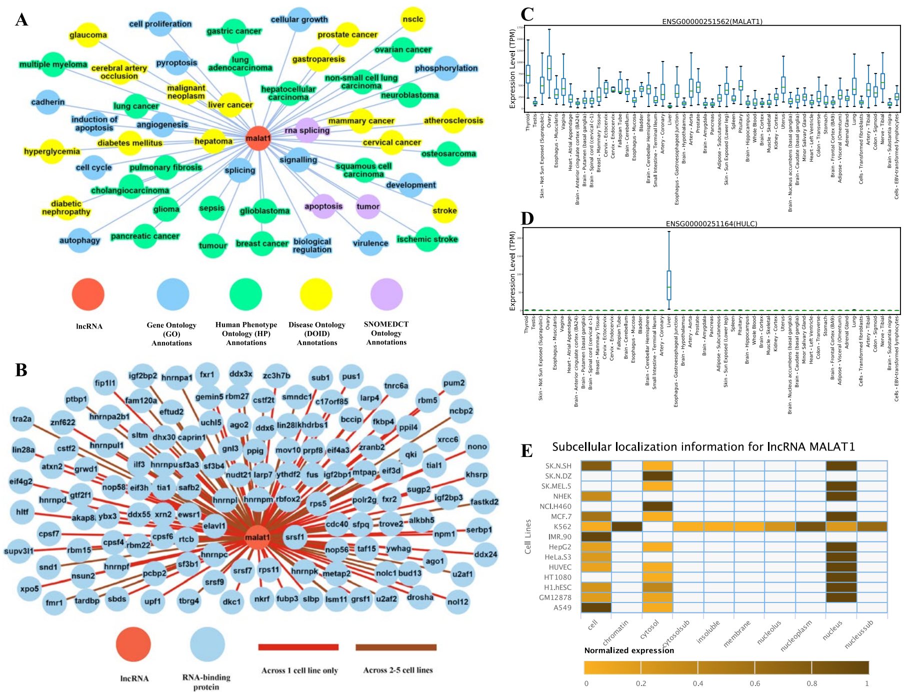
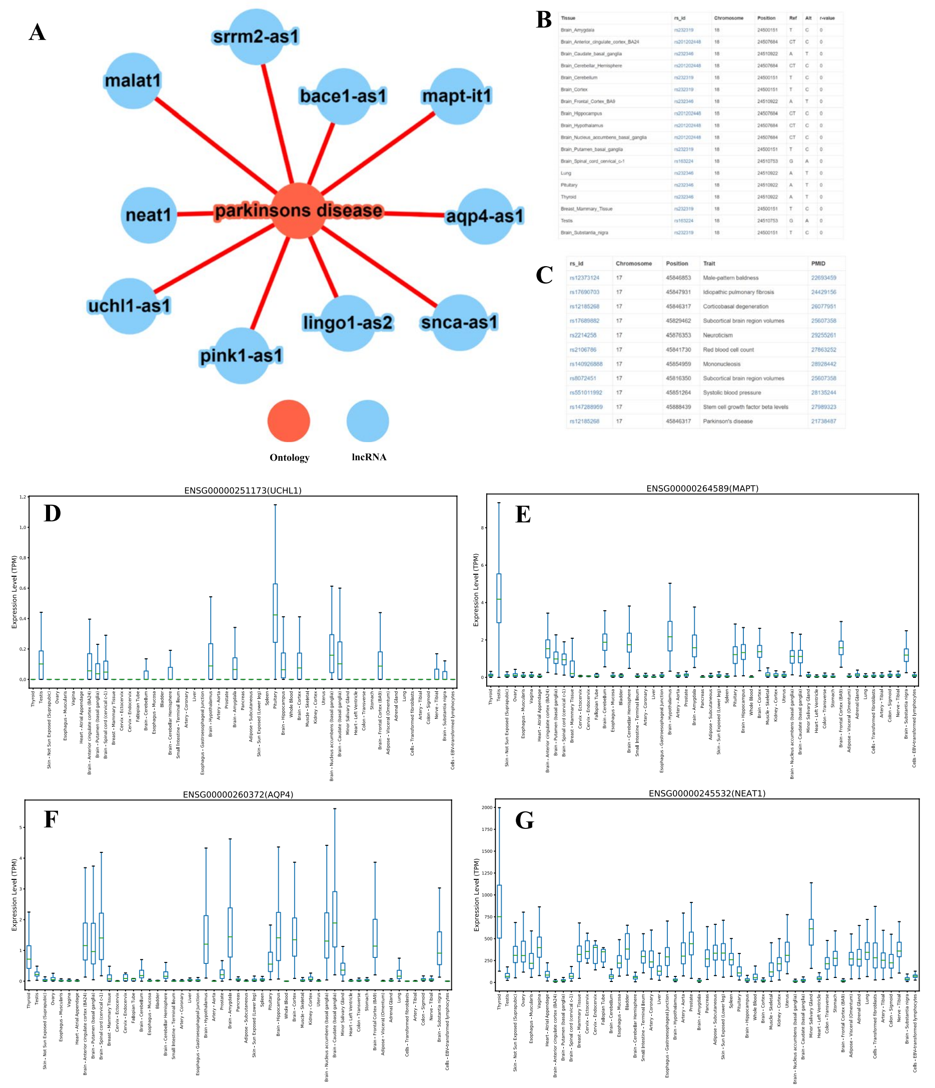


Fig. 3 Selection of multi-level genomic and ontology annotations for lncRNA MALAT1 as seen on Lantern. **A** ontology annotation interaction network showing accurate ontology annotations of lncRNA MALAT1, extracted from literature mining. **B** RNA binding protein interaction profile of MALAT1 across multiple cell lines. **C** Boxplot showing global expression of lncRNA MALAT1 across tissues. **D** Boxplot showing liver-specific expression of lncRNA HULC indicative of upregulation in liver tissue, emphasizing Lantern's ability to capture tissue-specific expression profile of lncRNAs. **E** Heatmap showing the sub-cellular localization of MALAT1 in various compartments across multiple tissues. In agreement with established observations, MALAT1 was found to exhibit enriched abundance in the nucleus over cytosol

literature [84], localization of MALAT1 was observed predominantly in the nuclear compartment of the cell (Fig. 3E).

Lantern also provides lncRNA disease phenotype associations. Our understanding of a disease phenotype can improve by recognizing the various cellular elements involved. To identify and understand lncRNAs' functional role in various disease phenotypes, an ontology-based search was integrated onto Lantern. Users can search for a phenotype and Lantern will return all the associated lncRNAs, based on the functional annotations extracted from the literature by the Mobius pipeline (see "Construction and Content" section). A sample phenotype-centric network is visualized as Fig. 4A, wherein the human disease ontology 'Parkinson's disease has been annotated with several lncRNAs by the Mobius pipeline, including BACE1-AS1[100], MAPT-IT1[101, 102], AQP4-AS1[103] and SNCA-AS1[104], etc., along with the peer-reviewed literature from which this phenotype-lncRNA association was extracted. Lantern also integrates phenotype-specific SNPs across lncRNAs. Reinforcing the annotations for lncRNAs like AQP4-AS1 and MAPT-IT1, which were annotated to 'Parkinson's disease by the Mobius pipeline, the tissue-specific SNPs in AQP4-AS1 predominantly occur in brain tissue and GWAS SNP annotation for MAPT-IT1 can be seen as 'Parkinson's Disease' from Fig. 4B, C.



Tissue	rs_id	Chromosome	Position	Ref	Alt	r-value
Brain_Amygdala	rs232319	18	24500151	T	C	0
Brain_Anterior_cingulate_cortex_BA24	rs201202448	18	24507684	CT	C	0
Brain_Caudate_basal_ganglia	rs232346	18	24510922	A	T	0
Brain_Cerebellar_Hemisphere	rs201202448	18	24507684	CT	C	0
Brain_Cerebellum	rs232319	18	24500151	T	C	0
Brain_Cortex	rs232319	18	24500151	T	C	0
Brain_Frontal_Cortex_BA9	rs232346	18	24510922	A	T	0
Brain_Hippocampus	rs201202448	18	24507684	CT	C	0
Brain_Hypothalamus	rs201202448	18	24507684	CT	C	0
Brain_Nucleus_accumbens_basal_ganglia	rs201202448	18	24507684	CT	C	0
Brain_Putamen_basal_ganglia	rs232319	18	24500151	T	C	0
Brain_Spinal_cord_cervical_c-1	rs163224	18	24510753	G	A	0
Lung	rs232346	18	24510922	A	T	0
Pituitary	rs232346	18	24510922	A	T	0
Thyroid	rs232346	18	24510922	A	T	0
Breast_Mammary_Tissue	rs232319	18	24500151	T	C	0
Testis	rs163224	18	24510753	G	A	0
Brain_Substantia_nigra	rs232319	18	24500151	T	C	0

rs_id	Chromosome	Position	Trait	PMID
rs12373124	17	45846853	Male-pattern baldness	22693459
rs17690703	17	45847931	Idiopathic pulmonary fibrosis	24429156
rs12185268	17	45846317	Corticobasal degeneration	26077951
rs17689882	17	45829462	Subcortical brain region volumes	25607358
rs2214258	17	45876353	Neuroticism	29255261
rs2106786	17	45841730	Red blood cell count	27863252
rs140926888	17	45854959	Mononucleosis	28928442
rs8072451	17	45816350	Subcortical brain region volumes	25607358
rs551011992	17	45851264	Systolic blood pressure	28135244
rs147288959	17	45888439	Stem cell growth factor beta levels	27989323
rs12185268	17	45846317	Parkinson's disease	21738487

Fig. 4 lncRNAs annotated to Parkinson's disease and their pertaining information as seen on Lantern. **A** Network depicting the lncRNAs associated with the ontology 'Parkinson's disease', illustrating the preciseness of annotations. **B** Screenshot of tissue-specific SNPs in lncRNA AQP4-AS1 from GTEx eQTL. An enriched abundance of SNPs in brain tissue reinforcing the association of AQP4-AS1 and Parkinson's disease. **C** Screenshot of GWAS SNP phenotype annotations for lncRNA MAPT-IT1 showing its established association with Parkinson's disease, corroborating Lantern's annotations. **D**–**G** boxplots showing brain-exclusive varied expression of lncRNAs annotated to Parkinson's disease as seen on Lantern. Panel **D**, **E**, and **f** showing elevated expression levels in brain tissue for UCHL1-AS1, MAPT-IT1, and AQP4-AS1 respectively. While in panel **G** lncRNA NEAT1 with exclusive modest abundance across all brain tissues supporting its established association as identified by Lantern

As evident from Lantern's tissue-specific expression panel, these lncRNAs viz. UCHL-AS1, MAPT-IT1, and AQP4-AS1 were observed to be relatively elevated in brain tissue (Fig. 4D–F), signifying their regulatory role in the normal functioning of the brain, in coherence with the previously reported observations [105–107]. In contrast, lncRNA NEAT1 was found to be down-regulated across all the GTEx brain regions (Fig. 4E), which corroborates with our current understanding that NEAT1 has been extensively observed to be upregulated in the brain and neurodegenerative disorders [108, 109]. Since the expression on Lantern is extracted from GTEx based normal tissue abundance, which is why NEAT1 expression across normal brain tissue is highly downregulated.

Lantern can not only help explore the regulatory role of extensively studied lncRNAs but can also recognize novel associations and features of non-putative lncRNAs. LncRNAs have been widely recognized for varying functions across different tissues and phenotypes. For instance lncRNA XISTis recognized as an oncogene in human colorectal cancer [110] but also acts as a tumor suppressor in prostate cancer [111]. Increasing evidence for the variable role of lncRNAs generates the necessity for an understanding of the novel as well as well-studied lncRNAs in a cell line or tissue-specific manner. For instance, the functional role of lncRNA MHENCR is not extensively explored. Although, it is known to be upregulated in melanoma [112] and is implicated with a poor survival rate in melanoma patients. This oncogene MHENCR is known to be influence cell proliferation and apoptosis by attenuating miR-425 and miR-489, thus upregulating their target genes IGF1, SPIN1, and activating the PI3K-Akt pathway. In corroboration to this evidence, MHENCR expression is recorded to be low in GTEx extracted tissue expression. However, we observed a 30-fold upregulation of MHENCR expression in thyroid tissue, as seen in Fig. 5A. The role of MHENCR has not been elucidated previously in the context of thyroid or related phenotypes. This upregulation of this lncRNA in the thyroid could potentially mean that it is essential for the normal function of the thyroid. Interestingly, we also observed that a GTEx eQTL analysis-based SNP was also identified in the locus of this lncRNA.

With these observations, Lantern not only provides a platform to recognize the role of lncRNA in multiple diseases but can also facilitate the identification of tissue-specific intrinsic features of lncRNA which is indispensable for understanding their molecular functions. Additionally, annotations from resources like TRlnc [113] combined with ontology and omic-annotations from Lantern can enable us to develop a better understanding of transcriptional regulatory mechanisms of lncRNAs.

Another interesting novel lncRNA was ELOVL2-AS1, unlike its protein-coding counterpart ELOVL2-AS1 which is not fully studied. However, Lantern registers its interactions with various RNA binding and cellular proteins like SRSF1, RBFOX2, and PTBP1 which have a significant role in cellular processes like splicing and development especially in the context of the brain [114]. We observe a specific expression pattern of ELOVL2-AS1 from Lantern, even though it has higher expression in testis, from Fig. 5B, we observed that ELOVL2-AS1 has a relatively higher abundance across all of the brain tissues. Moreover, we also observed that 5 out of 6 SNPs (rs3945981, rs397954604, rs4713170, rs7774711, rs2147041) annotated by GTEx eQTL analysis, identified in and around its locus were predominantly from brain tissues, seen in Fig. 5C. With such extensive multi-level lncRNA annotations, Lantern has the potential to aid the identification of several other novel lncRNAs, as well as facilitate and develop a hypothesis around well-studied lncRNAs to understand their regulatory role in diseases.

Conclusion

This study describes the development of a semi-automated pipeline for annotating lncRNA with ontology information using literature mining. The extracted ontology annotations were benchmarked against a manually curated lncRNA information resource. All the

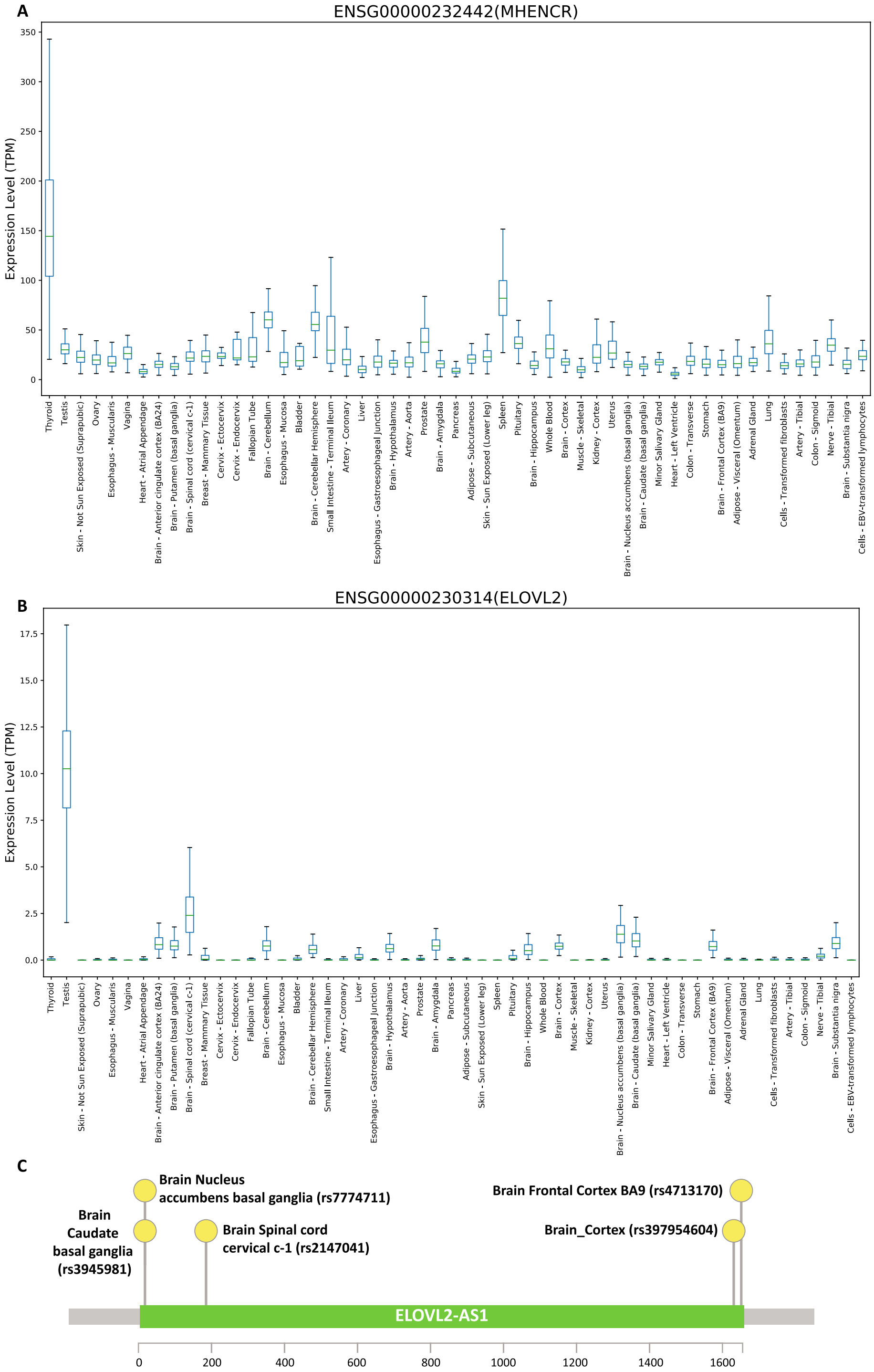


Fig. 5 Understanding the functional role of novel and not extensively studied lncRNAs. **A** Plot showing tissue-specific expression of lncRNA MHENCR. **B** Tissue-specific expression plot showing brain-specific expression for novel lncRNA ELOVL2-AS1. **C** Lollipop plot showing the predominant occurrence of ELOVL2-AS1 SNPs in the brain

extracted ontology information is deposited onto a web interface to easily navigate and retrieve ontology annotations for lncRNA. A molecular-level interaction profile of lncRNAs was put together, along with tissue-specific and SNP information utilizing contemporary NGS data analysis pipelines. Subsequently, a public resource with high-quality-controlled ontology annotations and comprehensive omics annotations was developed for improving the annotation of the non-coding transcriptome.

Abbreviations

LncRNAs: Long non-coding RNAs; SNP: Single nucleotide polymorphism; GTEx: Genotype-tissue expression; GWAS: Genome-wide association studies; RBP: RNA binding protein; CLIP: Crosslinking and immunoprecipitation; NLP: Natural Language Processing; GO: Gene ontology; HPO: Human phenotype ontology; DO: Disease ontology; SNOMEDCT: Systematized nomenclature of medicine—clinical terms; NGS: Next-generation sequencing; API: Application programming interface; NCBO: National Center for Biomedical Ontology; HPM: Human proteome map; TSS: Transcript start site; FIMO: Find Individual Motif Occurrences; GTF: Gene transfer format; ORF: Open reading frame; ENSG: Ensembl gene ID.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04207-3.

Additional file 1: Table S1. Manual precision calculation for benchmarking extracted ontology annotations.

Acknowledgements

We thank members of the Janga lab for their valuable suggestions over the course of this project. We are also thankful to colleagues and lab members who reviewed and provided feedback on previous versions of this manuscript.

Authors' contributions

SVD and SCJ designed the study, SVD implemented the computational approaches and performed the benchmarking of the resource. SVD and RS processed multi OMICs data and formatted it for data integration into Lantern. SVD, RS, and SCJ wrote the manuscript. All authors read and approved the final manuscript.

Funding

Research reported in this publication was in part supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM123314 (SCJ). Funding body did not have any role in the design of the study and collection, analysis, and interpretation of data or in writing the manuscript.

Availability of data and materials

Lantern database is hosted at sysbio.lab.iupui.edu/lantern, which provides downloadable tables and figures. Source codes are available at GitHub (at https://github.com/VidhurDS/Lantern) [115].

Declarations

Ethics approval and consent to participate

Ethics approval does not apply to this work.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing financial interests.

Author details

[1]Department of BioHealth Informatics, School of Informatics and Computing, Indiana University Purdue University, Informatics and Communications Technology Complex, 535 W Michigan St., IT 475H, Indianapolis, IN 46202, USA. [2]Department of Surgery, Indiana Center for Regenerative Medicine and Engineering (ICRME), Indiana University School of Medicine, Indianapolis, IN 46202, USA. [3]Department of Medical and Molecular Genetics, Medical Research and Library Building, Indiana University School of Medicine, 975 West Walnut Street, Indianapolis, IN 46202, USA. [4]Centre for Computational Biology and Bioinformatics, Indiana University School of Medicine, 5021 Health Information and Translational Sciences (HITS), 410 West 10th Street, Indianapolis, IN 46202, USA.

Received: 21 February 2021 Accepted: 18 May 2021
Published online: 26 May 2021

References

1. Diamantopoulos MA, Tsiakanikas P, Scorilas A. Non-coding RNAs: the riddle of the transcriptome and their perspectives in cancer. Ann Transl Med. 2018;6(12):241.
2. Geisler S, Coller J. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. Nat Rev Mol Cell Biol. 2013;14(11):699–712.
3. Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. Nat Rev Genet. 2009;10(3):155–9.
4. Patrushev LI, Kovalenko TF. Functions of noncoding sequences in mammalian genomes. Biochem Mosc. 2014;79(13):1442–69.
5. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annu Rev Biochem. 2012;81:145–66.
6. Kugel JF, Goodrich JA. Non-coding RNAs: key regulators of mammalian transcription. Trends Biochem Sci. 2012;37(4):144–51.
7. Gil N, Ulitsky I. Regulation of gene expression by cis-acting long non-coding RNAs. Nat Rev Genet. 2020;21(2):102–17.
8. Li S, et al. Exploring functions of long noncoding RNAs across multiple cancers through co-expression network. Sci Rep. 2017;7(1):754.
9. Cabanski CR, et al. Pan-cancer transcriptome analysis reveals long noncoding RNAs with conserved function. RNA Biol. 2015;12(6):628–42.
10. Verma A, et al. Transcriptome sequencing reveals thousands of novel long non-coding RNAs in B cell lymphoma. Genome Med. 2015;7:110.
11. Faghihi MA, et al. Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. Nat Med. 2008;14(7):723–30.
12. Grote P, et al. The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. Dev Cell. 2013;24(2):206–14.
13. Wallace C, et al. The imprinted *DLK1-MEG3* gene region on chromosome *14q32.2* alters susceptibility to type 1 diabetes. Nat Genet. 2010;42(1):68–71.
14. DiStefano JK. The emerging role of long noncoding RNAs in human disease. Methods Mol Biol. 2018;1706:91–110.
15. Wapinski O, Chang HY. Long noncoding RNAs and human disease. Trends Cell Biol. 2011;21(6):354–61.
16. Schmitt AM, Chang HY. Long noncoding RNAs in cancer pathways. Cancer Cell. 2016;29(4):452–63.
17. Li L, et al. Long non-coding RNA in neuronal development and neurological disorders. Front Genet. 2018;9:744.
18. Hu X, et al. A functional genomic approach identifies FAL1 as an oncogenic long noncoding RNA that associates with BMI1 and represses p21 expression in cancer. Cancer Cell. 2014;26(3):344–57.
19. Prensner JR, et al. The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. Nat Genet. 2013;45(11):1392–8.
20. Fritah S, Niclou SP, Azuaje F. Databases for lncRNAs: a comparative evaluation of emerging tools. RNA. 2014;20(11):1655–65.
21. Volders PJ, et al. An update on LNCipedia: a database for annotated human lncRNA sequences. Nucleic Acids Res. 2015;43(8):4363–4.
22. Chakraborty S, et al. LncRBase: an enriched resource for lncRNA information. PLoS ONE. 2014;9(9):e108010.
23. Ma L, et al. LncRNAWiki: harnessing community knowledge in collaborative curation of human long non-coding RNAs. Nucleic Acids Res. 2015;43(Database issue):D187–92.
24. Harrow J, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22(9):1760–74.
25. Zhao Y, et al. NONCODE 2016: an informative and valuable data source of long non-coding RNAs. Nucleic Acids Res. 2016;44(D1):D203–8.
26. Quek XC, et al. lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. Nucleic Acids Res. 2015;43(Database issue):D168–73.
27. Bhartiya D, et al. lncRNome: a comprehensive knowledgebase of human long noncoding RNAs. Database (Oxford). 2013;2013:bat034.
28. Volders PJ, et al. LNCipedia: a database for annotated human lncRNA transcript sequences and structures. Nucleic Acids Res. 2013;41(Database issue):D246–51.
29. Chen G, et al. LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res. 2013;41(Database issue):D983–6.
30. Sayers E. A general introduction to the E-utilities. In: Entrez programming utilities help [Internet]. Bethesda, MD: National Center for Biotechnology Information (US); 2010.
31. Musen MA, et al. The National Center for Biomedical Ontology. J Am Med Inform Assoc. 2012;19(2):190–5.
32. Martinez-Romero M, et al. NCBO Ontology Recommender 2.0: an enhanced approach for biomedical ontology recommendation. J Biomed Semantics. 2017;8(1):21.
33. Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res. 2015;43(Database issue):D1049–56.
34. Kohler S, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res. 2014;42(Database issue):D966–74.
35. Kibbe WA, et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res. 2015;43(Database issue):D1071–8.
36. Spackman KA, Campbell KE, Cote RA. SNOMED RT: a reference terminology for health care. In: Proceedings on AMIA Annual Fall Symposium; 1997. p. 640–4.
37. Porto FW, Daulatabad SV, Janga SC. Long non-coding RNA expression levels modulate cell-type-specific splicing patterns by altering their interaction landscape with RNA-Binding Proteins. Genes (Basel). 2019;10(8):593.
38. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74.
39. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.

40. Zerbino DR, et al. Ensembl 2018. Nucleic Acids Res. 2018;46(D1):D754-d761.
41. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, Foster B. The genotype-tissue expression (GTEx) project. Nat Genet. 2013;45(6):580–5.
42. Misawa A, Orimo H. lncRNA HOTAIR inhibits mineralization in osteoblastic osteosarcoma cells by epigenetically repressing ALPL. Calcif Tissue Int. 2018;103(4):422–30.
43. Hung T, Chang HY. Long noncoding RNA in genome regulation: prospects and mechanisms. RNA Biol. 2010;7(5):582–5.
44. Sola I, et al. RNA-RNA and RNA-protein interactions in coronavirus replication and transcription. RNA Biol. 2011;8(2):237–48.
45. Kishore S, Luber S, Zavolan M. Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression. Brief Funct Genomics. 2010;9(5–6):391–404.
46. Hanly DJ, Esteller M, Berdasco M. Interplay between long non-coding RNAs and epigenetic machinery: emerging targets in cancer? Philos Trans R Soc Lond B Biol Sci. 2018;373(1748):20170074.
47. Bush SJ, et al. Cross-species inference of long non-coding RNAs greatly expands the ruminant transcriptome. Genet Sel Evol. 2018;50(1):20.
48. Johnsson P, et al. Evolutionary conservation of long non-coding RNAs; sequence, structure, function. Biochim Biophys Acta. 2014;1840(3):1063–71.
49. Kim MS, et al. A draft map of the human proteome. Nature. 2014;509(7502):575–81.
50. Zar JH. Spearman rank correlation. Encyclopedia of Biostatistics, vol. 7. Hoboken: Wiley; 2005.
51. Franz M, et al. Cytoscape.js: a graph theory library for visualisation and analysis. Bioinformatics. 2016;32(2):309–11.
52. Matys V, et al. TRANSFAC®: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. 2003;31(1):374–8.
53. Bailey TL, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37(suppl_2):W202–8.
54. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics. 2014;47:11.12.1-11.12.34.
55. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27(7):1017–8.
56. Iyer MK, et al. The landscape of long noncoding RNAs in the human transcriptome. Nat Genet. 2015;47(3):199–208.
57. Kang YJ, et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res. 2017;45(W1):W12-w16.
58. Fickett JW. Recognition of protein coding regions in DNA sequences. Nucleic Acids Res. 1982;10(17):5303–18.
59. Buxbaum AR, Haimovich G, Singer RH. In the right place at the right time: visualizing and understanding mRNA localization. Nat Rev Mol Cell Biol. 2015;16(2):95–109.
60. Chen LL. Linking long noncoding RNA localization and function. Trends Biochem Sci. 2016;41(9):761–72.
61. Carlevaro-Fita J, Johnson R. Global Positioning System: understanding long noncoding RNAs through subcellular localization. Mol Cell. 2019;73(5):869–83.
62. Mas-Ponte D, et al. LncATLAS database for subcellular localization of long noncoding RNAs. RNA. 2017;23(7):1080–7.
63. Tang JY, et al. Long noncoding RNAs-related diseases, cancers, and drugs. Sci World J. 2013;2013:943539.
64. Li J, Xuan Z, Liu C. Long non-coding RNAs and complex human diseases. Int J Mol Sci. 2013;14(9):18790–808.
65. Ling H, et al. CCAT2, a novel noncoding RNA mapping to 8q24, underlies metastatic progression and chromosomal instability in colon cancer. Genome Res. 2013;23(9):1446–61.
66. Freedman ML, et al. Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet. 2011;43(6):513–8.
67. MacArthur J, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017;45(D1):D896-d901.
68. Sherry ST, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308–11.
69. Deng M, et al. FirebrowseR: an R client to the Broad Institute's Firehose Pipeline. Database. 2017;2017:1–6.
70. Yang L, et al. lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs. Nature. 2013;500(7464):598–602.
71. Sun X, et al. Down-regulation of lncRNA MEG3 promotes endothelial differentiation of bone marrow derived mesenchymal stem cells in repairing erectile dysfunction. Life Sci. 2018;208:246–52.
72. Tao H, et al. Emerging role of long noncoding RNAs in lung cancer: current status and future prospects. Respir Med. 2016;110:12–9.
73. Li Q, et al. Long noncoding RNA MALAT1 promotes cell proliferation through suppressing miR-205 and promoting SMAD4 expression in osteosarcoma. Oncotarget. 2017;8(63):106648–60.
74. Li H, et al. Long non-coding RNA MALAT1 decreases the sensitivity of resistant glioblastoma cell lines to temozolomide. Cell Physiol Biochem. 2017;42(3):1192–201.
75. Bi S, et al. LncRNA-MALAT1-mediated Axl promotes cell invasion and migration in human neuroblastoma. Tumour Biol. 2017;39(5):1010428317699796.
76. Zhang X, et al. Long noncoding RNA Malat1 regulates cerebrovascular pathologies in ischemic stroke. J Neurosci. 2017;37(7):1797–806.
77. Li C, et al. Down-regulation of miR-146b-5p by long noncoding RNA MALAT1 in hepatocellular carcinoma promotes cancer growth and metastasis. Oncotarget. 2017;8(17):28683–95.
78. Lei R, et al. Long noncoding RNA MALAT1-regulated microRNA 506 modulates ovarian cancer growth by targeting iASPP. Onco Targets Ther. 2017;10:35–46.
79. Gao KT, Lian D. Long non-coding RNA MALAT1 is an independent prognostic factor of osteosarcoma. Eur Rev Med Pharmacol Sci. 2016;20(17):3561–5.
80. Han Y, et al. Low-dose sinapic acid abates the pyroptosis of macrophages by downregulation of lncRNA-MALAT1 in rats with diabetic atherosclerosis. J Cardiovasc Pharmacol. 2018;71(2):104–12.
81. Munoz U, et al. Hepatocyte growth factor enhances alternative splicing of the Kruppel-like factor 6 (KLF6) tumor suppressor to promote growth through SRSF1. Mol Cancer Res. 2012;10(9):1216–27.
82. Wang J, et al. Mutual inhibition between YAP and SRSF1 maintains long non-coding RNA, Malat1-induced tumourigenesis in liver cancer. Cell Signal. 2014;26(5):1048–59.

83. Malakar P, et al. Long noncoding RNA MALAT1 promotes hepatocellular carcinoma development by SRSF1 upregulation and mTOR activation. Cancer Res. 2017;77(5):1155–67.
84. Muñoz Ú, et al. Hepatocyte growth factor enhances alternative splicing of the Kruppel-like factor 6 (KLF6) tumor suppressor to promote growth through SRSF1. Mol Cancer Res. 2012;10(9):1216–27.
85. Clark MB, et al. Genome-wide analysis of long noncoding RNA stability. Genome Res. 2012;22(5):885–98.
86. Ayupe AC, et al. Global analysis of biogenesis, stability and sub-cellular localization of lncRNAs mapping to intragenic regions of the human genome. RNA Biol. 2015;12(8):877–92.
87. Noh JH, et al. Cytoplasmic functions of long noncoding RNAs. Wiley Interdiscip Rev RNA. 2018;9(3):e1471.
88. Lakhotia SC. From heterochromatin to long noncoding RNAs in Drosophila: expanding the arena of gene function and regulation. Adv Exp Med Biol. 2017;1008:75–118.
89. Hu WL, et al. GUARDIN is a p53-responsive long non-coding RNA that is essential for genomic stability. Nat Cell Biol. 2018;20(4):492–502.
90. Sonohara F, et al. Prognostic value of long non-coding RNA HULC and MALAT1 following the curative resection of hepatocellular carcinoma. Sci Rep. 2017;7(1):16142.
91. Ma Z, et al. Current advances of long non-coding RNA highly upregulated in liver cancer in human tumors. Onco Targets Ther. 2017;10:4711–7.
92. Wang Y, et al. The long noncoding RNA HULC promotes liver cancer by increasing the expression of the HMGA2 oncogene via sequestration of the microRNA-186. J Biol Chem. 2017;292(37):15395–407.
93. Xiong H, et al. lncRNA HULC promotes the growth of hepatocellular carcinoma cells via stabilizing COX-2 protein. Biochem Biophys Res Commun. 2017;490(3):693–9.
94. Su ZD, et al. iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. Bioinformatics. 2018;34(24):4196–204.
95. Hacisuleyman E, et al. Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. Nat Struct Mol Biol. 2014;21(2):198–206.
96. Ip JY, Nakagawa S. Long non-coding RNAs in nuclear bodies. Dev Growth Differ. 2012;54(1):44–54.
97. Cabili MN, et al. Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. Genome Biol. 2015;16:20.
98. Sone M, et al. The mRNA-like noncoding RNA Gomafu constitutes a novel nuclear domain in a subset of neurons. J Cell Sci. 2007;120(Pt 15):2498–506.
99. Hutchinson JN, et al. A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. BMC Genomics. 2007;8:39.
100. Wu P, et al. Roles of long noncoding RNAs in brain development, functional diversification and neurodegenerative diseases. Brain Res Bull. 2013;97:69–80.
101. Valenca GT, et al. The role of MAPT haplotype H2 and isoform 1N/4R in Parkinsonism of older adults. PLoS ONE. 2016;11(7):e0157452.
102. Coupland KG, et al. Role of the long non-coding RNA MAPT-AS1 in regulation of Microtubule Associated Protein Tau (MAPT) expression in Parkinson's disease. PLoS ONE. 2016;11(6):e0157924.
103. Thenral ST, Vanisree AJ. Peripheral assessment of the genes AQP4, PBP and TH in patients with Parkinson's disease. Neurochem Res. 2012;37(3):512–5.
104. Yan Y, et al. SNCA is a functionally low-expressed gene in lung adenocarcinoma. Genes (Basel). 2018;9(1):16.
105. Carrieri C, et al. Expression analysis of the long non-coding RNA antisense to Uchl1 (AS Uchl1) during dopaminergic cells' differentiation in vitro and in neurochemical models of Parkinson's disease. Front Cell Neurosci. 2015;9:114.
106. Qureshi IA, Mattick JS, Mehler MF. Long non-coding RNAs in nervous system function and disease. Brain Res. 2010;1338:20–35.
107. Quan Z, Zheng D, Qing H. Regulatory roles of long non-coding RNAs in the central nervous system and associated neurodegenerative diseases. Front Cell Neurosci. 2017;11:175.
108. Yang X, et al. Silencing of the long non-coding RNA NEAT1 suppresses glioma stem-like properties through modulation of the miR-107/CDK6 pathway. Oncol Rep. 2017;37(1):555–62.
109. Yan W, et al. LncRNA NEAT1 promotes autophagy in MPTP-induced Parkinson's disease through stabilizing PINK1 protein. Biochem Biophys Res Commun. 2018;496(4):1019–24.
110. Song H, et al. Long non-coding RNA XIST functions as an oncogene in human colorectal cancer by targeting miR-132-3p. J Buon. 2017;22(3):696–703.
111. Du Y, et al. LncRNA XIST acts as a tumor suppressor in prostate cancer through sponging miR-23a to modulate RKIP expression. Oncotarget. 2017;8(55):94358–70.
112. Chen X, et al. Long noncoding RNA MHENCR promotes melanoma progression via regulating miR-425/489-mediated PI3K-Akt pathway. Am J Transl Res. 2017;9(1):90–102.
113. Li Y, et al. TRlnc: a comprehensive database for human transcriptional regulatory information of lncRNAs. Brief Bioinform. 2020;22(2):1929–39.
114. Li YI, et al. RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. Genome Res. 2015;25(1):1–13.
115. Daulatabad V. Lantern source codes in GitHub. 2018.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.