aBIOTECH (2022) 3:169–177
https://doi.org/10.1007/s42994-022-00084-3

aBIOTECH

BRIEF COMMUNICATION

LncPheDB: a genome-wide lncRNAs regulated phenotypes database in plants

Danjing Lou[1], Fei Li[1], Jinyue Ge[1], Weiya Fan[1], Ziran Liu[2], Yanyan Wang[1], Jingfen Huang[1], Meng Xing[1], Wenlong Guo[1], Shizhuang Wang[1], Weihua Qiao[1,3], Zhenyun Han[1], Qian Qian[1,4,5], Qingwen Yang[1,3]✉, Xiaoming Zheng[1,3,6]✉

[1] National Key Facility for Crop Gene Resources and Genetic Improvement, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China
[2] College of Life Science, Shenyang Normal University, Shenyang 110034, China
[3] National Nanfan Research Institute (Sanya), Chinese Academy of Agricultural Sciences, Sanya 572000, China
[4] State Key Laboratory of Rice Biology, China National Rice Research Institute, Chinese Academy of Agricultural Sciences, Hangzhou 310006, China
[5] Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China
[6] International Rice Research Institute, DAPO box 7777 Metro Manila, The Philippines

Received: 30 June 2022 / Accepted: 12 September 2022 / Published online: 5 October 2022

Abstract LncPheDB (https://www.lncphedb.com/) is a systematic resource of genome-wide long non-coding RNAs (lncRNAs)-phenotypes associations for multiple species. It was established to display the genome-wide lncRNA annotations, target genes prediction, variant-trait associations, gene-phenotype correlations, lncRNA-phenotype correlations, and the similar non-coding regions of the queried sequence in multiple species. LncPheDB sorted out a total of 203,391 lncRNA sequences, 2000 phenotypes, and 120,271 variants of nine species (*Zea mays* L., *Gossypium barbadense* L., *Triticum aestivum* L., *Lycopersicon esculentum* Mille, *Oryza sativa* L., *Hordeum vulgare* L., *Sorghum bicolor* L., *Glycine max* L., and *Cucumis sativus* L.). By exploring the relationship between lncRNAs and the genomic position of variants in genome-wide association analysis, a total of 68,862 lncRNAs were found to be related to the diversity of agronomic traits. More importantly, to facilitate the study of the functions of lncRNAs, we analyzed the possible target genes of lncRNAs, constructed a blast tool for performing similar fragmentation studies in all species, linked the pages of phenotypic studies related to lncRNAs that possess similar fragments and constructed their regulatory networks. In addition, LncPheDB also provides a user-friendly interface, a genome visualization platform, and multi-level and multi-modal convenient data search engine. We believe that LncPheDB plays a crucial role in mining lncRNA-related plant data.

Keywords LncRNA, GWAS, Phenotype, SNP, Plants

INTRODUCTION

LncRNAs are a class of non-coding RNAs that are more than 200 nucleotides in length. Initially, this type of RNA was once considered to be "junk" material in the genome. However, as the research continues, there is

✉ Correspondence: yangqingwen@caas.cn (Q. Yang), zhengxiaoming@caas.cn (X. Zheng)

growing evidence that lncRNAs are key players in growth and development, metabolism and regulatory processes in a variety of organisms, particularly in mammals and humans (Kopp and Mendell 2018; Kung et al. 2013; Morris and Mattick 2014; Sun et al. 2018; Uchida and Dimmeler 2015; Wu et al. 2017). However, the study of lncRNAs in plants remains in its infancy. Currently, it has been found in plants that lncRNAs not only play an important role in regulating growth and developmental processes such as growth hormone transport and signal transduction in plants. It also plays an important role in improving crop yield (Wang et al. 2018), leaf distortion (Liu et al. 2018), plant fertility (Fang et al. 2019; Zhao et al. 2018), fruit fertility (Fan et al. 2016) and other important agronomic traits. But the vast majority of lncRNA regulatory explorations with clear mechanisms are nowadays performed in *Arabidopsis thaliana*. Our understanding of the mechanisms regulating lncRNAs in crop species remains limited. In addition, in recent years, transcriptome data have been used to carry out a large number of lncRNAs-related studies (Katayama et al. 2005; Osato et al. 2003; Terryn and Rouzé 2000; Wang et al. 2005; Zhang et al. 2006, 2014; Zhu and Deng 2012). Studies have shown that there are 32,397 lncRNAs in maize, 11,565 lncRNAs in rice, and 12,577 lncRNAs in soybean (Jin et al. 2021). It has also been revealed that lncRNAs are generally characterized by low expression, poor conservativeness among different species, and tissue specificity (Derrien et al. 2012; Cabili et al. 2011). These characteristics make the study of lncRNAs functions a herculean task. At present, although a large number of lncRNAs have been identified through transcriptome research, the lncRNAs whose functions have been further verified are less than 1% (Quek et al. 2015). Furthermore, the genome-wide association study (GWAS) of multiple species revealed that 84% of trait-related variation loci are located in non-coding sequences (Cheetham et al. 2013). However, the non-coding regions in the genome lack annotations and other relevant information. This hinders our further research on the non-coding regions.

The lncRNAs database is a very good tool to facilitate a detailed and accurate study of lncRNAs. In recent years, a total of 20 plant-related lncRNA databases have been established. They have averaged a whopping 530 citations since publication. But most of these databases provide the basic information of lncRNAs in species and target gene prediction according to transcriptome data. For instance, the PLncDB database (Jin et al. 2021) can provide basic information about various plants, such as lncRNA genome position, sequence, and structure, the expression in tissues, and the query and visual display of gene regulation networks. However, the database can only perform a Basic Local Alignment Search Tool (BLAST) analysis of single species. The CANTATAdb 2.0 database (Szcześniak et al. 2019), which contains lncRNAs of plants and algae, leverages on JBrowse, eFP Browser, EPexplorer, and other analysis tools to search for the maximum peptide length, maximum expression level, number of lncRNA exons, and other information of lncRNAs in species. The GreeNC database (Gallart et al. 2016) can extract the position, sequence, coding potential, folding energy, and other information of lncRNAs in various species; it can be used to perform a BLAST analysis of one or more species. Most of the databases constructed by researchers in the early days focused on some basic annotation information about the sequence and position of lncRNAs. However, they lacked comprehensive annotation information. In addition, very few databases could provide information about the correlation between lncRNAs and phenotypes, the similarity of lncRNAs among multiple species and display the possible correlation between these similar fragments and phenotypes. The RiceLncPedia database (Zhang et al. 2021), a newly built database, has comprehensive annotation information of lncRNAs. For instance, the database collects multi-omics information, such as quantitative trait locus, GWAS, transposons, and variant sites (SNPs). However, it only shows the lncRNAs of rice, but no blast tool is available to study the similarity of lncRNAs among different species. Therefore, it is necessary to build a database that explores the similarity of lncRNAs in multiple species and combines lncRNAs with GWAS.

In this study, we built a database containing the lncRNAs information of nine common crops, including *Zea mays* L., *Gossypium barbadense* L., *Triticum aestivum* L., *Lycopersicon esculentum* Mille, *Oryza sativa* L., *Hordeum vulgare* L., *Sorghum bicolor* L., *Glycine max* L., and *Cucumis sativus* L. The database provides information about the sequence and position of lncRNAs, the distribution of lncRNAs in the genome, the population variation of lncRNAs, and the phenotypic traits that may be regulated, among others. In addition, the database can also use the BLAST tool to investigate the conservativeness of target gene sequences in various species and the phenotypic conditions that may be regulated. Our database is designed to further improve the annotation information of lncRNAs in plants to further explore the possible functions of lncRNAs.

MATERIALS AND METHODS

Data collection and sorting

For the LncPheDB database, we selected nine important model plants (including *Zea mays* L., *Gossypium barbadense* L., *Triticum aestivum* L., *Lycopersicon esculentum* Mille, *Oryza sativa* L., *Hordeum vulgare* L., *Sorghum bicolor* L., *Glycine max* L., and *Cucumis sativus* L.) with great economic value and a high-quality reference genome. According to the data sequencing method and data sequencing depth, we extracted a total of 2324 RNA sequencing (RNA-Seq) datasets from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database (https://www.ncbi.nlm.nih.gov/sra/) (Supplemental Table S1). Using the SRA toolkit (Version 2.8) under the Linux system, we first converted the extracted SRA file into Fastq format and trimmed the adapter sequences using Trim Galore (version 0.50) (https://www.bioinformatics.babraham.ac.uk/projects/trim galore/) to obtain clean data. HIAST2 (Kim et al. 2015) was used to make a comparison between the clean data and the reference genome; afterward, the clean data were assembled with StringTie (Pertea et al. 2015). StringTie-merge was used to obtain the transcript set of each species. The transcripts were filtered out according to the following criteria: transcript length less than 200 base pairs and open reading frame greater than 120 amino acids. Finally, BLASTx was used to search the SWISS-PROT database to filtered out the transcripts that may encode small peptides with the parameters -e 1.0e-4-S 1. A comparison between the database and the Rfam database was performed to filter out tRNAs, rRNAs, sRNAs, and miRNAs. The transcripts were collected after the filtering. The CPC (Kong et al. 2007), CREMA (Simopoulos et al. 2018), PLEK (Li et al. 2014), and RNAplonc (Negri et al. 2019) programs were used to calculate the protein-coding ability of transcripts, and the non-protein-coding transcripts detected in at least two software were used as candidate lncRNAs (Fig. 1B). In addition, to enrich lncRNAs types, we sorted out the lncRNAs sequences of the nine species mentioned above in the RNAcentral Database (The et al. 2017) and the EVLncRNAs Databases (Zhou et al. 2018).

To extract comprehensive and high-quality information from published GWAS articles, we used the keywords "species" and "GWAS" to search for articles published in PubMed and we obtained 2227 relevant research articles that were published after 2009. Afterward, Articles were selected if there were a large number of candidates for significant SNP-phenotype correlation analysis data, while articles with segmental and phenotypic correlation data or no SNP-phenotype correlation analysis data were removed. We found 497 articles with data that are significantly related to genome-wide variation loci and phenotypic traits. Finally, 421 articles were further screened according to the *P*-value ($P < 10^{-3}$) of significant GWAS data. In addition, the basic information of these articles is listed in Supplemental Table S2.

To link the lncRNAs data with the GWAS result data, we used the BWA tool (version 0.7.17) to unify the SNPs from GWAS data in each species and the reference genome from lncRNAs data in the same species into the same reference genome. Afterward, we first mapped the long segments according to the distance between SNPs (The distance between variant sites was shorter than the length of the region of linkage disequilibrium (LD)) (Supplemental Table S3), and then amplified the mapped long segments according to the LD of each species, if the lncRNAs and genes are within the incremental region, these lncRNAs are considered to regulate the corresponding phenotype and are associated with genes. At the same time, we also amplified a single site in the GWAS results based on the length of the region of LD of each species, and based on the positional relationship between the gene or lncRNA and the amplified segment, to determine the phenotypes that lncRNAs or genes may regulate (Guttman and Rinn 2012; Guttman et al. 2011; Huarte et al. 2010; Lee 2009; Martianov et al. 2007; Nagano et al. 2008; Rinn and Chang 2012; Sleutels et al. 2002).

Implementation

LncPheDB was implemented using PostgreSQL (https://www.postgresql.org; a powerful, open-source object-relational database system with over 30 years of active development that has earned it a strong reputation for reliability, feature robustness, and performance) and Django development server (https://docs.djangoproject.com/en/2.2/intro/tutorial01/#the-development-server; a lightweight web server written purely in Python). Web user interfaces were developed using Django (https://www.djangoproject.com; a high-level Python web framework that encourages rapid development and clean, pragmatic design), HTML5, CSS3, AJAX (Asynchronous JavaScript and XML; a set of web development techniques used to create asynchronous applications without interfering with the display and behavior of the existing page), JQuery (a cross-platform and feature-rich JavaScript library; http://jquery.com, version 1.10.2), Vue (https://vuejs.org; the Progressive JavaScript Framework, version 2.6.14), layui (https://github.com/sentsin/layui/; a classic modular front-end UI framework), and Boot-Strap (an open-source toolkit

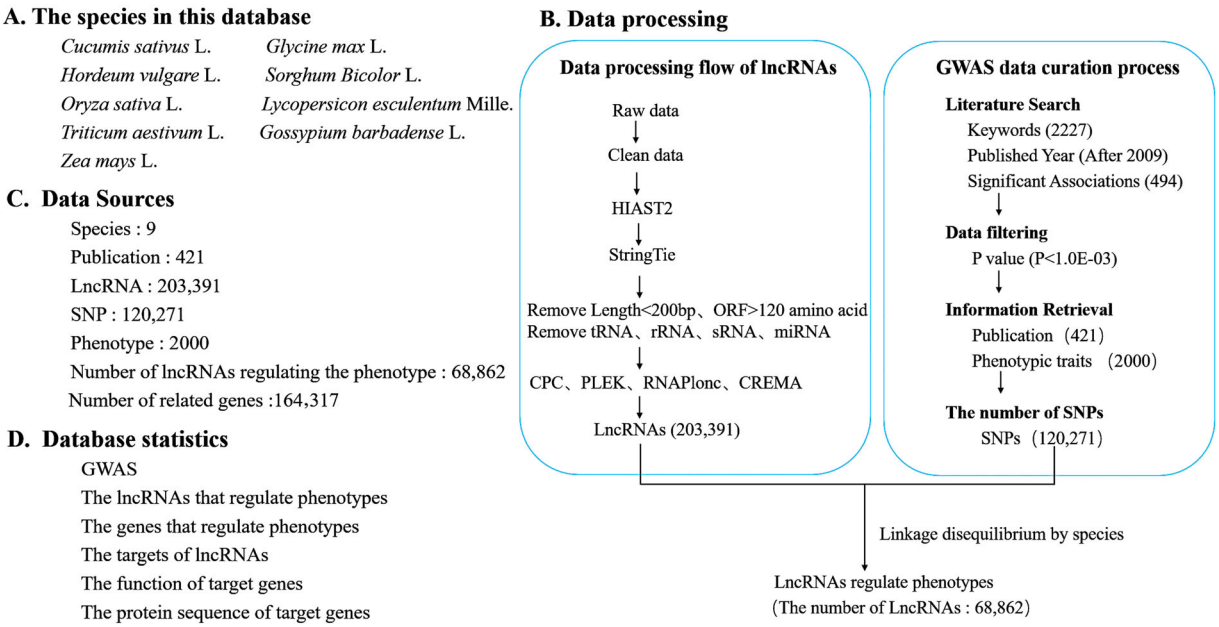

Fig. 1 Data processing workflow and outcomes of LncPheDB. **A** The nine species included in the database. **B** The data processing workflow of lncRNA and the curation process adopted by the GWAS is on the right. **C** Summary of the data contained in LncPheDB. **D** Database statistics in this study

for developing web projects with HTML, CSS, and JS; https://getbootstrap.com, version4.6.0). For dynamic genome visualization and analysis, JBrowse Genome Browser (a fast, scalable genome browser built completely with JavaScript and HTML5; https://jbrowse.org/jbrowse1.html, version 1.16.11) was adopted to generate interactive charts.

RESULTS

GWAS revealed many genetic variants associated with phenotypes. Thousands of GWAS studies have revealed that 93% of common genetic variants associated with specific traits or diseases are located in non-coding regions (Finucane et al. 2015; Schaid et al. 2018). Of these, more than 90% of the variants were SNPs. In addition, the density of SNPs in lncRNA regions is similar to that in protein-coding regions. Some lncRNA intervals even have higher SNP densities than the genomic mean (Jin et al. 2011). SNP variants in lncRNA can affect mRNA expression through variable shear, localization, and stability of mRNA. Therefore, the association between lncRNA SNPs and phenotypes needs to be studied in depth. It has been shown that lncRNAs can influence complex traits at multiple levels of epigenetic regulation, transcriptional regulation, and post-transcriptional regulation (Zhang et al. 2018). To provide a comprehensive resource for linking lncRNAs to phenotypes. First, by carrying out RNA-seq analysis and sorting out the data of various non-coding region databases in RNAcentral and EVLncRNAs, we obtained a total of 203,391 LncRNA sequences. Precisely, 32,397, 32,192, 43,659, 8,741, 11,565, 25,884, 27,623, 12,577, 8,753 lncRNAs were obtained for *Zea mays* L., *Gossypium barbadense* L., *Triticum aestivum* L., *Lycopersicon esculentum* Mille, *Oryza sativa* L., *Hordeum vulgare* L., *Sorghum Bicolor* L., *Glycine max* L., and *Cucumis sativus* L., respectively. And based on the standard screening process, we integrated 2,000 important agronomic traits and 120,271 SNPs that have a significant effect on the phenotype of the nine species from the 421 articles. Among them, *Oryza sativa* L. and *Zea mays* L. have 764 and 573 traits, respectively, which account for 66.85% of all traits, while *Gossypium barbadense* L. has the least traits, which account for 0.5%. Meanwhile, 68,862 lncRNA sequences that can regulate important agronomic traits were predicted (Table 1).

In addition, to make it easier and more efficient for users to use the data. We provide a web service interface-LncPheDB. LncPheDB provides a user-friendly interface, a visual platform and a variety of search options. The LncPheDB database mainly provides the reference genome information of nine species (the size of the reference genome, number of chromosomes, and number of protein-coding genes). Basic information regarding all lncRNAs and phenotype-related lncRNAs (e. g. species, lncRNA identity (ID), chromosome, start

Table1 Detail information about LncPheDB

Species	Phenotype	Var	Publications	LncRNAs	lncRNAs (Phenotype)	Version
Zea mays L.	573	71,058	151	32,397	28,164	B73_RefGen_v4
Gossypium barbadense L.	10	111	2	32,192	813	GCA_008761655.1
Triticum aestivum L.	50	755	11	43,659	4773	refseqv1.0
Lycopersicon esculentum Mille	132	787	9	8741	1212	ITAG4.0
Oryza sativa L.	764	23,690	117	11,565	8384	MSU_osa1r7
Hordeum vulgare L.	17	750	6	25,884	5508	version.1.0
Sorghum Bicolor L.	250	17,855	57	27,623	16,431	GCF_000003195.3
Glycine max L.	193	5129	66	12,577	3273	GCF_000004515.5
Cucumis sativus L.	11	136	2	8753	304	GCF_000004075.3

site, termination site, and positive and negative chain), as well as basic information of GWAS results (e. g. GWAS phenotypic traits, location of peak in genome, and *P*-value) is provided. Furthermore, LncPheDB also provides functional information on genes associated with lncRNAs and protein sequence information of genes in various species (by searching the SWISS-PROT database), and the regulatory network information of lncRNAs related to phenotypes (Fig. 2).

LncPheDB provides two search engines: the lncRNA search engine and the GWAS search engine. The lncRNA module provides comprehensive lncRNA-phenotype correlation data in each species, which are created in the form of columns into tables. Each correlation data mainly includes phenotype-related lncRNA ID, species, chromosome position, lncRNA initiation and termination sites, Positive and negative chains, regulated phenotype, Peak Position, *P*-value of phenotype-SNP correlation, mapped genes, and sequence of mapped genes. In this module, we merge adjacent significant SNPs whose distance is less than the species LD into a single association signal based on the LD decay of each species. The SNP with the minimum *P* value in a signal region was considered to be the lead SNP. Finally, the related lncRNA and mRNA were predicted according to the LD of each species. This module focuses on exploring the linkage among SNPs and the linkage between SNPs and lncRNA or mRNA. There are also more

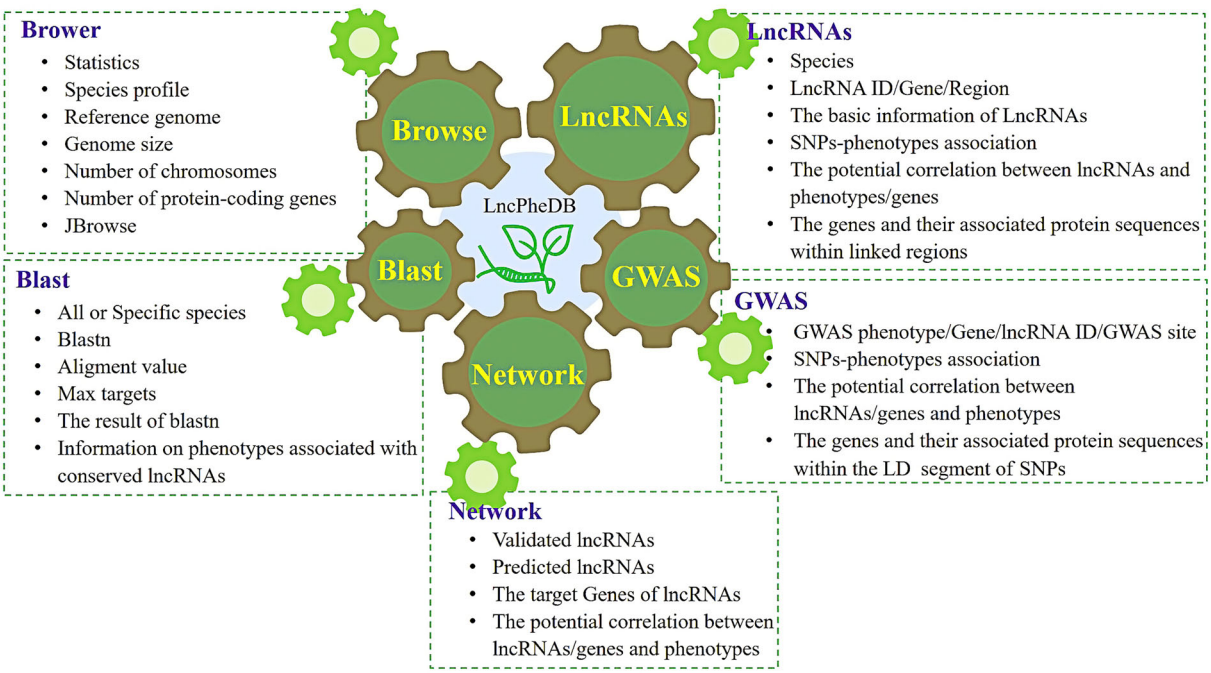

Fig. 2 Database contents and functions of LncPheDB

phenotypes highlighted in this module, such as: the SNPs 201,770,002 (P = 3.65E-59), 201,770,047 (P = 4.97E-07), and 201,770,048 (P = 3.65E-59) located on chromosome 2 are significantly associated with maize leaves, and the SNPs is located within the lncRNA URS0000D75A41_4577.4871 (201,769,823–201,770,124). So we speculate that lncRNA URS0000D75A41_4577.4871 may be associated with maize leaves. In addition, for lncRNAs of interest, users can use our database for in-depth exploration. For instance, for maize lncRNA EL0549, after selecting the maize species, if you enter lncRNA EL0549 and click "search", you can easily find information regarding the position of lncRNA EL0549, relevant GWAS information, and the information that EL0549 regulates maize's flour fiber content, proline content, breakdown viscosity, flour fiber content, flour protein content, ear infructescence position, and maize kernels. To further determine the biological processes between lncRNA and traits, such as maize entrainment, protein content, and fiber concentration, among others, Users can click "Function" to view the functional information of genes associated with lncRNAs. Meanwhile, users can also click "Sequence" to view the protein sequence of genes (Supplemental Fig. S1). By phenotype, lncRNA/Gene ID or GWAS locus input, the GWAS module can be used to obtain phenotype-associated genes or lncRNAs for each species, genome-wide variant loci significantly associated with phenotypes, correlation P values, etc. The correlation data for this module are mainly obtained based on the amplification of individual variant loci, emphasizing the relative position between the variant loci and the lncRNA or gene. In the GWAS module, users can explore the phenotypes of their interest. For instance, the keyword "100 grain weight" can be used for maize (Supplemental Fig. S2). All search results can be downloaded in the form of a list. The combination of this lncRNA module and the GWAS module allows for a more comprehensive genome-wide prediction of phenotypic traits that may be regulated by lncRNAs or gene. Meanwhile, we also added the JBrowse genome browser, which allows users to intuitively search for the relative position distribution of lncRNAs and genes on chromosomes.

To study the sequence similarity, we designed a Blast tool (version 2.12). By searching specific species in the whole database, the BLAST service enables users to search for similar lncRNA sequences. In the BLAST results, users can directly view the phenotypic traits related to lncRNAs with similar fragments by clicking the "Click here to search LncRNA: lncRNA ID" tab. To enable users to view LncRNA and its regulated target genes clearly and concisely, we predicted the target genes of known and predicted lncRNAs by psRobot (Wu et al. 2012), psMimic (Wu et al. 2013) and IntaRNA (Mann et al. 2017), which were presented in the form of regulatory networks, marked them with different colors, and set three buttons, which allow users to hide corresponding genes by clicking the corresponding buttons. In addition to downloading the information from the corresponding search page, users can also download the reference genome information for each species, the lncRNA fasta sequence files, lncRNA Potential Encoding File, lncRNA Expression File and the GFF files for database construction via the download page. Moreover, users can also download the GWAS information file (such as associated phenotypic information, SNP, p-value, and information about studies) and the gene GFF file of each species.

DISCUSSION

With the development of sequencing technology in the past few years, a large number of lncRNAs have been identified and great progress has been made in the study of lncRNAs in plants. However, compared with the lncRNAs in animals and humans, there is a very limited understanding of lncRNAs in plants, especially in terms of the mechanism of lncRNAs in regulating important agronomic traits and affecting the yield and quality of model plants (Heo et al. 2013; Liu et al. 2012; Mann et al. 2017; Xiao et al. 2009; Yang et al. 2014). With the deepening of research, some well-annotated databases, such as PLncDB V2.0 (Jin et al., 2011) and GREENC (Gallart et al. 2016), have given comprehensive annotations to some basic information of lncRNAs, such as the position and sequence. Researchers have shifted their focus from identifying new lncRNAs to the functional research of lncRNAs. In recent years, researchers have investigated the functions of lncRNAs in plants. However, at present, the identified lncRNAs whose regulatory mechanism has been clarified are less than 1% (Quek et al. 2015). In addition, the research results of some lncRNAs provide a low reference value for the study of other lncRNAs due to the differences in types and functions of lncRNAs, which affect gene expression in a wide range at different levels. Therefore, researchers' understanding and research on lncRNAs are limited. At present, it is imperative to use a genome-wide database to investigate the relationship between lncRNAs and phenotypes and explore the potential regulatory mechanism of lncRNAs.

Compared with other plant lncRNA databases, LncPheDB focuses on exploring data resources about lncRNA-regulated phenotypes. Using standardized

screening criteria, LncPheDB manually sorted a total of 203,391 lncRNA sequences, 2000 phenotypes, and 120,271 SNPs. Finally, it listed 68,862 lncRNA sequences that are associated with agronomic traits. And according to the study. The lncRNA osa-eTM160 (Osa-eTM160 is a 688 bp long lncRNA transcribed between LOC_Os03g12815 and LOC_Os03g12820 of rice chromosome 3) in rice has a role in regulating rice fertility and seed size by competitively binding OsmiR160 with OsARF18. However, the potential regulatory significance of lncRNA URS00008EDDE3_39947.4350 (also known as osa-eTM160) on rice seed fertility, days to flowering, seed weight, arsenic accumulation, germination rate and grain Mn concentration is predicted in our database, which further confirms the significance of our database. Moreover, users can use the lncRNA sequences they are investigating to conduct a BLAST comparison with all species in the data resource to identify the conservative lncRNA-regulated phenotypes. Furthermore, LncPheDB also provides users with convenient browsing and search services. Thus, users can search lncRNAs correlation from various aspects, such as Gene ID, LncRNA ID, genome position, SNP, and phenotype. To help users explore the potential molecular regulatory mechanism of lncRNAs in complex traits, we summarized and sorted out the target gene prediction of lncRNAs and visually displayed it in the form of a regulation network. Users can hide or display the corresponding data by clicking different buttons.

As a future perspective, by focusing on the study of data resources regarding lncRNA-regulated phenotypes, we will add more lncRNA-related phenotypes for more species. In addition, since we found that the number of relevant studies was unexpectedly large when collecting and sorting out data, we will sort out more data regarding lncRNA-regulated phenotypes with clear regulatory mechanisms and predictions from existing studies and timely update the data resources. To further clarify the regulatory mechanism of lncRNAs, we will add more sequence information of miRNAs that are complementary to lncRNAs and increase the tissue-specific expression information of lncRNAs. Meanwhile, to enrich the transcriptome information of rice, we will add relevant transcriptome data in our research to facilitate scientific research and utilization. Notwithstanding, we also encourage all researchers to submit their relevant studies via the contact page. We believe that LncPheDB will provide assistance for the study of the functions of lncRNAs.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42994-022-00084-3.

Acknowledgements We thank all the members who participated in the construction of this database. Thanks for the support of the Key Laboratory of Grain Crop Genetic Resources Evaluation and Utilization.

Authors' contributions Danjing Lou is involved in conceptualizing, writing and editing this manuscript. Xiaoming Zheng, Qingwen Yang, Qian Qian conceived the project. Danjing Lou, Fei Li, Jinyue Ge, Weiya Fan, Ziran Liu, Yanyan Wang, Jingfen Huang, Meng Xing, Wenlong Guo, Shizhuang Wang, Weihua Qiao and Zhenyun Han analysed the data.

Funding This work was supported by the National Key Research and Development Program of China (2021YFD1200101 to Z.X.M.), the National Natural Science Foundation of China (31670211 and 31970237 to Z.X.M.), Sanya Yazhou Bay Science and Technology City (SKJC-2020–02-001 to Z.X.M.), the Central Public-interest Scientific Institution Basal Research Fund (S2021ZD01 to Z.X.M.).

Data availability LncPheDB is freely available at https://www.lncphedb.com/.

Code availability This study involves database building code.

Declarations

Conflict of interest Author declares no conflicts of interests.

Ethical approval This manuscript is not involved in any animal experiments.

Consent to participate Necessary approval is obtained.

Consent for publication Necessary approval is obtained.

References

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Gene Dev 25:1915–1927. https://doi.org/10.1101/gad.17446611

Cheetham SW, Gruhl F, Mattick JS, Dinger ME (2013) Long noncoding RNAs and the genetics of cancer. Brit J Cancer 108:2419–2425. https://doi.org/10.1038/bjc.2013.233

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J,

Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigó R (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res 22:1775–1789. https://doi.org/10.1101/gr.132159.111

Fan Y, Yang J, Mathioni SM, Yu J, Shen J, Yang X, Wang L, Zhang Q, Cai Z, Xu C, Li X, Xiao J, Meyers BC, Zhang Q (2016) PMS1T, producing phased small-interfering RNAs, regulates photoperiod-sensitive male sterility in rice. PNAS 113:15144–15149. https://doi.org/10.1073/pnas.1619159114

Fang J, Zhang F, Wang H, Wang W, Zhao F, Li Z, Sun C, Chen F, Xu F, Chang S, Wu L, Bu Q, Wang P, Xie J, Chen F, Huang X, Zhang Y, Zhu X, Han B, Deng X, Chu C (2019) Ef-cd locus shortens rice maturity duration without yield penalty. PNAS 116:18717–18722. https://doi.org/10.1073/pnas.1815030116

Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P, Anttila V, Xu H, Zang C, Farh K, Pipke S, Day FR, Consortium R, Purcell S, Stahl E, Lindstrom S, Perry JRB, Okada Y, Raychaudhuri S, Daly MJ, Patterson N, Neale BM, Price AL (2015) Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet 47:1228–1235. https://doi.org/10.1038/ng.3404

Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Young G, Lucas AB, Ach R, Bruhn L, Yang X, Amit I, Meissner A, Regev A, Rinn JL, Root DE, Lander ES (2011) lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 477:295–300. https://doi.org/10.1038/nature10398

Guttman M, Rinn JL (2012) Modular regulatory principles of large non-coding RNAs. Nature 482:339–346. https://doi.org/10.1038/nature10887

Heo JB, Lee Y, Sung S (2013) Epigenetic regulation by long noncoding RNAs in plants. Chromosome Res 21:685–693. https://doi.org/10.1007/s10577-013-9392-6

Huarte M, Guttman M, Feldser D, Garber M, Koziol MJ, Kenzelmann-Broz D, Khalil AM, Zuk O, Amit I, Rabani M, Attardi LD, Regev A, Lander ES, Jacks T, Rinn JL (2010) A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell 142:409–419. https://doi.org/10.3410/f.5523957.5491055

Jin G, Sun J, Isaacs SD, Wiley KE, Kim ST, Chu LW, Zhang Z, Zhao H, Zheng SL, Isaacs WB, Xu J (2011) Human polymorphisms at long non-coding RNAs (lncRNAs) and association with prostate cancer risk. Carcinogenesis 32:1655–1659. https://doi.org/10.1093/carcin/bgr187

Jin J, Lu P, Xu Y, Li Z, Yu S, Liu J, Wang H, Chua N, Cao P (2021) PLncDB V2.0: a comprehensive encyclopedia of plant long noncoding RNAs. Nucleic Acids Res 49:D1489–D1495. https://doi.org/10.1093/nar/gkaa910

Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, Suzuki H, Carninci P, Hayashizaki Y, Wells C, Frith M, Ravasi T, Pang KC, Hallinan J, Mattick J, Hume DA, Lipovich L, Batalov S, Engström PG, Mizuno Y, Faghihi MA, Sandelin A, Chalk AM, Mottagui-Tabar S, Liang Z, Lenhard B, Wahlestedt C (2005) Antisense transcription in the mammalian transcriptome. Science 309:1564–1566. https://doi.org/10.1126/science.1112009

Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12:357–360. https://doi.org/10.1038/nmeth.3317

Kong L, Zhang Y, Ye Z, Liu X, Zhao S, Wei L, Gao G (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res 35:W345–W349. https://doi.org/10.1093/nar/gkm391

Kopp F, Mendell JT (2018) Functional classification and experimental dissection of long noncoding RNAs. Cell 172:393–407. https://doi.org/10.1016/j.cell.2018.01.011

Kung JTY, Colognori D, Lee JT (2013) Long noncoding RNAs: past, present, and future. Genetics 193:651–669. https://doi.org/10.1534/genetics.112.146704

Lee JT (2009) Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. Gene Dev 23:1831–1842. https://doi.org/10.1101/gad.1811209

Li A, Zhang J, Zhou Z (2014) PLEK: a tool for predicting long noncoding RNAs and messenger RNAs based on an improved k-mer scheme. BMC Bioinformatics 15:311. https://doi.org/10.1186/1471-2105-15-311

Liu J, Jung C, Xu J, Wang H, Deng S, Bernad L, Arenas-Huertero C, Chua N (2012) Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in *Arabidopsis*. Plant Cell 24:4333–4345. https://doi.org/10.1105/tpc.112.102855

Liu X, Li D, Zhang D, Yin D, Zhao Y, Ji C, Zhao X, Li X, He Q, Chen R, Hu S, Zhu L (2018) A novel antisense long noncoding RNA, TWISTED LEAF, maintains leaf blade flattening by regulating its associated sense R2R3-MYB gene in rice. New Phytol 218:774–788. https://doi.org/10.1111/nph.15023

Mann M, Wright PR, Backofen R (2017) IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions. Nucleic Acids Res 45:W435–W439. https://doi.org/10.1093/nar/gkx279

Martianov I, Ramadass A, Serra Barros A, Chow N, Akoulitchev A (2007) Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. Nature 445:666–670. https://doi.org/10.1038/nature05519

Morris KV, Mattick JS (2014) The rise of regulatory RNA. Nat Rev Genet 15:423–437. https://doi.org/10.1038/nrg3722

Nagano T, Mitchell JA, Sanz LA, Pauler FM, Ferguson-Smith AC, Feil R, Fraser P (2008) The air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. Science 322:1717–1720. https://doi.org/10.1126/science.1163802

Negri TDC, Alves WAL, Bugatti PH, Saito PTM, Domingues DS, Paschoal AR (2019) Pattern recognition analysis on long noncoding RNAs: a tool for prediction in plants. Brief Bioinform 20:682–689. https://doi.org/10.1093/bib/bby034

Osato N, Yamada H, Satoh K, Ooka H, Yamamoto M, Suzuki K, Kawai J, Carninci P, Ohtomo Y, Murakami K, Matsubara K, Kikuchi S, Hayashizaki Y (2003) Antisense transcripts with rice full-length cDNAs. Genome Biol 5:R5. https://doi.org/10.1186/gb-2003-5-1-r5

Paytuví Gallart A, Hermoso Pulido A, Lagrán AMD, I, Sanseverino W, Aiese Cigliano R, (2016) GREENC: a Wiki-based database of plant lncRNAs. Nucleic Acids Res 44:D1161–D1166. https://doi.org/10.1093/nar/gkv1215

Pertea M, Pertea GM, Antonescu CM, Chang T, Mendell JT, Salzberg SL (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33:290–295. https://doi.org/10.1038/nbt.3122

Quek XC, Thomson DW, Maag JLV, Bartonicek N, Signal B, Clark MB, Gloss BS, Dinger ME (2015) lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. Nucleic Acids Res 43:D168–D173. https://doi.org/10.1093/nar/gku988

Rinn JL, Chang HY (2012) Genome regulation by long noncoding RNAs. Annu Rev Biochem 81:145–166. https://doi.org/10.1146/annurev-biochem-051410-092902

Schaid DJ, Chen W, Larson NB (2018) From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet 19:491–504. https://doi.org/10.1038/s41576-018-0016-z

Simopoulos CMA, Weretilnyk EA, Golding GB (2018) Prediction of plant lncRNA by ensemble machine learning classifiers. BMC Genomics 19:316. https://doi.org/10.1186/s12864-018-4665-2

Sleutels F, Zwart R, Barlow DP (2002) The non-coding Air RNA is required for silencing autosomal imprinted genes. Nature 415:810–813. https://doi.org/10.1038/415810a

Sun X, Zheng H, Sui N (2018) Regulation mechanism of long non-coding RNA in plant response to stress. Biochem Bioph Res Co 503:402–407. https://doi.org/10.1016/j.bbrc.2018.07.072

Szcześniak MW, Bryzghalov O, Ciomborowska-Basheer J, Makałowska I (2019) CANTATAdb 2.0: Expanding the Collection of Plant Long Noncoding RNAs. In: Chekanova JA, Wang HV (eds) Plant Long Non-Coding RNAs: Methods and Protocols. New York, NY, Springer, New York, pp 415–429

Terryn N, Rouzé P (2000) The sense of naturally transcribed antisense RNAs in plants. Trends Plant Sci 5:394–396. https://doi.org/10.1016/S1360-1385(00)01696-4

The RC, Petrov AI, Kay SJE, Kalvari I, Howe KL, Gray KA, Bruford EA, Kersey PJ, Cochrane G, Finn RD, Bateman A, Kozomara A, Griffiths-Jones S, Frankish A, Zwieb CW, Lau BY, Williams KP, Chan PP, Lowe TM, Cannone JJ, Gutell R, Machnicka MA, Bujnicki JM, Yoshihama M, Kenmochi N, Chai B, Cole JR, Szymanski M, Karlowski WM, Wood V, Huala E, Berardini TZ, Zhao Y, Chen R, Zhu W, Paraskevopoulou MD, Vlachos IS, Hatzigeorgiou AG, Ma L, Zhang Z, Puetz J, Stadler PF, McDonald D, Basu S, Fey P, Engel SR, Cherry JM, Volders P, Mestdagh P, Wower J, Clark MB, Quek XC, Dinger ME (2017) RNAcentral: a comprehensive database of non-coding RNA sequences. Nucleic Acids Res 45:D128–D134. https://doi.org/10.1093/nar/gkw1008

Uchida S, Dimmeler S (2015) Long noncoding RNAs in cardiovascular diseases. Circ Res 116:737–750. https://doi.org/10.1161/CIRCRESAHA.116.302521

Wang X, Gaasterland T, Chua N (2005) Genome-wide prediction and identification of cis-natural antisense transcripts in *Arabidopsis thaliana*. Genome Biol 6:R30. https://doi.org/10.1186/gb-2005-6-4-r30

Wang Y, Luo X, Sun F, Hu J, Zha X, Su W, Yang J (2018) Overexpressing lncRNA LAIR increases grain yield and regulates neighbouring gene cluster expression in rice. NC 9:1–9

Wu H, Ma Y, Chen T, Wang M, Wang X (2012) PsRobot: a web-based plant small RNA meta-analysis toolbox. Nucleic Acids Res 40:W22–W28. https://doi.org/10.1093/nar/gks554

Wu H, Wang Z, Wang M, Wang X (2013) Widespread long noncoding RNAs as endogenous target mimics for microRNAs in plants. Plant Physiol 161:1875–1884. https://doi.org/10.1104/pp.113.215962

Wu H, Yang L, Chen L (2017) The diversity of long noncoding RNAs and their generation. Trends Genet 33:540–552. https://doi.org/10.1016/j.tig.2017.05.004

Xiao B, Zhang X, Li Y, Tang Z, Yang S, Mu Y, Cui W, Ao H, Li K (2009) Identification, bioinformatic analysis and expression profiling of candidate mRNA-like non-coding RNAs in Sus scrofa. J Genet Genomics 36:695–702. https://doi.org/10.1016/S1673-8527(08)60162-9

Xu S, Dong Q, Deng M, Lin D, Xiao J, Cheng P, Xing L, Niu Y, Gao C, Zhang W, Xu Y, Chong K (2021) The vernalization-induced long non-coding RNA VAS functions with the transcription factor TaRF2b to promote TaVRN1 expression for flowering in hexaploid wheat. Mol Plant 14:1525–1538. https://doi.org/10.1016/j.molp.2021.05.026

Yang G, Lu X, Yuan L (2014) LncRNA: a link between RNA and cancer. Biochim Biophys Acta Gene Regul Mech 1839:1097–1109. https://doi.org/10.1016/j.bbagrm.2014.08.012

Zhang Y, Liu XS, Liu Q, Wei L (2006) Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis -NATs) in ten species. Nucleic Acids Res 34:3465–3475. https://doi.org/10.1093/nar/gkl473

Zhang Y, Liao J, Li Z, Yu Y, Zhang J, Li Q, Qu L, Shu W, Chen Y (2014) Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice. Genome Biol 15:512. https://doi.org/10.1186/s13059-014-0512-1

Zhang Z, Xu Y, Yang F, Xiao B, Li G (2021) RiceLncPedia: a comprehensive database of rice long non-coding RNAs. Plant Biotechnol J 19:1492–1494. https://doi.org/10.1111/pbi.13639

Zhang Y, Tao Y, Liao Q (2018) Long noncoding RNA: a crosslink in biological regulatory network. Brief Bioinformatics 19:930–945. https://doi.org/10.1093/bib/bbx042

Zhao X, Li J, Lian B, Gu H, Li Y, Qi Y (2018) Global identification of *Arabidopsis* lncRNAs reveals the regulation of MAF4 by a natural antisense RNA. NC 9:1–12

Zhou B, Zhao H, Yu J, Guo C, Dou X, Song F, Hu G, Cao Z, Qu Y, Yang Y, Zhou Y, Wang J (2018) EVLncRNAs: a manually curated database for long non-coding RNAs validated by low-throughput experiments. Nucleic Acids Res 46:D100–D105. https://doi.org/10.1093/nar/gkx677

Zhu D, Deng XW (2012) A non-coding RNA locus mediates environment-conditioned male sterility in rice. Cell Res 22:791–792. https://doi.org/10.1038/cr.2012.43