D174–D182 *Nucleic Acids Research, 2022, Vol. 50, Database issue* *Published online 13 October 2021*
https://doi.org/10.1093/nar/gkab912

LIRBase: a comprehensive database of long inverted repeats in eukaryotic genomes

Lihua Jia[1,2,†], Yang Li[1,†], Fangfang Huang[1], Yingru Jiang[1], Haoran Li[1], Zhizhan Wang[1], Tiantian Chen[1], Jiaming Li[1], Zhang Zhang[3,4,5,6,*] and Wen Yao[1,*]

[1]National Key Laboratory of Wheat and Maize Crop Science, College of Life Sciences, Henan Agricultural University, Zhengzhou 450002, China, [2]National Key Laboratory of Wheat and Maize Crop Science, College of Agronomy, Henan Agricultural University, Zhengzhou 450002, China, [3]China National Center for Bioinformation, Beijing 100101, China, [4]National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China, [5]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China and [6]University of Chinese Academy of Sciences, Beijing 100101, China

Received July 31, 2021; Revised September 20, 2021; Editorial Decision September 22, 2021; Accepted September 25, 2021

ABSTRACT

Small RNAs (sRNAs) constitute a large portion of functional elements in eukaryotic genomes. Long inverted repeats (LIRs) can be transcribed into long hairpin RNAs (hpRNAs), which can further be processed into small interfering RNAs (siRNAs) with vital biological roles. In this study, we systematically identified a total of 6 619 473 LIRs in 424 eukaryotic genomes and developed LIRBase (https://venyao.xyz/lirbase/), a specialized database of LIRs across different eukaryotic genomes aiming to facilitate the annotation and identification of LIRs encoding long hpRNAs and siRNAs. LIRBase houses a comprehensive collection of LIRs identified in a wide range of eukaryotic genomes. In addition, LIRBase not only allows users to browse and search the identified LIRs in any eukaryotic genome(s) of interest available in GenBank, but also provides friendly web functionalities to facilitate users to identify LIRs in user-uploaded sequences, align sRNA sequencing data to LIRs, perform differential expression analysis of LIRs, predict mRNA targets for LIR-derived siRNAs, and visualize the secondary structure of candidate long hpRNAs encoded by LIRs. As demonstrated by two case studies, collectively, LIRBase bears the great utility for systematic investigation and characterization of LIRs and functional exploration of potential roles of LIRs and their derived siRNAs in diverse species.

INTRODUCTION

Small RNAs (sRNAs) are short non-coding RNAs with regulatory roles in almost every biological process of plants and animals by modulating gene expression levels (1). The biological functions of the majority of small interfering RNAs (siRNAs), which are a sort of sRNA with the largest quantity and species, remain elusive (1). In 2006, Henderson et al. reported a new biogenesis pathway of siRNAs from long inverted repeats (LIRs) in *Arabidopsis thaliana*, which was soon verified and characterized in *Drosophila* (2–4). An inverted repeat is a single stranded DNA sequence followed by its reverse complement at the downstream, designated as the left arm and the right arm of the inverted repeat, respectively. The two arms are intervened by zero or more nucleotides, termed as the loop of the inverted repeat. It was found that LIRs can form secondary stem-loop structures and can be transcribed into long hairpin RNAs (hpRNAs), which are much longer than typical animal or plant pre-miRNAs (3,4). The long hpRNAs can be further processed into 21-nt and 22-nt siRNAs by canonical RNA interference factors including Dicer-2, Hen1 and Argonaute 2 (3).

In the past decades, LIRs as well as their derived siRNAs have been extensively reported to play vital biological roles in diverse species (3–11). LIRs were deemed to be able to lead to tumorigenesis in human by causing DNA rearrangements in somatic cells (12). In 2018, Lin *et al.* identified two long hpRNAs encoded by LIRs with ∼2.8- and ∼3.1-kb in length in *Drosophila simulans*, which were processed into 21-nt siRNAs (5,13,14). These siRNAs can repress the expression of the *Dox* and *MDox* genes, thus promoting X chromosome transmission by suppressing Y-bearing sperms. As a result, the two LIRs and the derived

*To whom correspondence should be addressed. Tel: +86 371 6355 5790; Email: yaowen@henau.edu.cn
Correspondence may also be addressed to Zhang Zhang. Tel: +86 10 8409 7261; Fax: +86 10 8409 7720; Email: zhangzhang@big.ac.cn
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

siRNAs are critical to the maintenance of balanced sex ratio in the offspring of *D. simulans*. In mouse, siRNAs derived from LIRs were reported to regulate gene expression in oocytes (8,11). In apple, a 1489 bp LIR and the generated 21-nt siRNAs contributed to the resistance of apple to leaf spot disease caused by the *Alternaria alternata* f. sp *mali* (ALT1) fungus, by targeting five resistance genes (6). In soybean, 22-nt siRNAs derived from an intronic LIR with ∼5-kb, formed by the chalcone synthase (CHS) genes *CHS1* and *CHS3*, targeted other *CHS* genes to regulate the seed coat color of soybean (7,15,16). Additionally, in our previous study, we conducted sRNA sequencing of the flag leaves of a rice F_2 population derived from a self-cross of an elite rice hybrid (10,17) and found that the presence/absence variations of two LIRs (with 15 012-bp and 3013-bp in length) were associated with expression variations of siRNAs derived from the LIRs in the F_2 population (10).

Despite the vital biological roles as identified above, functionally studying LIRs and their derived siRNAs is still a great challenge, since there lacks a database that houses a comprehensive collection of LIRs across diverse species. Equally important, it is in urgent need for a database to facilitate the identification of LIRs underlying phenotype variations in confidence intervals delimited by QTL mapping and GWAS (genome-wide association study), and to allow the alignment of user-uploaded sRNA sequencing data to all collected LIRs, with the aim to identify LIRs encoding long hpRNAs and siRNAs. Although valuable efforts from different research groups were made in the identification of LIRs and long hpRNAs in different organisms (18,19), they only focused on very limited species (20–22). To make it short, there is no database up to now designed for genome-wide identification and annotation of LIRs and long hpRNAs across a wide range of species.

In this study, we identified a total of 6 619 473 LIRs in the whole genomes of 424 eukaryotes and constructed LIRBase (https://venyao.xyz/lirbase/), a comprehensive collection of LIRs across diverse eukaryotic genomes, with the aim to facilitate the annotation and identification of LIRs encoding long hpRNAs and siRNAs. Users can browse and search LIRs in LIRBase by genomic locations, LIR identifiers, or sequence similarities. LIRBase also provides friendly functionalities to allow users to upload and align sRNA sequencing data to all collected LIRs, perform differentially expression analysis of LIRs or sRNAs, and visualize the predicted secondary structure of long hpRNAs encoded by LIRs.

DATA COLLECTION AND METHODS

Collection of eukaryotic genomes

The whole-genome sequences and genome annotations of 77 invertebrate metazoa genomes and 208 vertebrate genomes were downloaded from Ensembl (https://www.ensembl.org) (23). The genome sequences and annotations of 44 plant genomes were extracted from Gramene (http://www.gramene.org/) (24). The other 95 plant genomes were obtained by searching PubMed and NCBI for featured plant species. As a result, a total of 424 eukaryotic genomes were collected (Supplementary Table S1).

Identification of LIRs in eukaryotic genomes

We tested several tools including detectIR (25), findIR (26), Inverted Repeats Finder (IRF) (20) and Lirex (27) to identify LIRs, and found that IRF is the only tool that was able to identify LIRs with imperfectly matched arms longer than 600 bp. On the contrary, detectIR and findIR were used to identify inverted repeats shorter than 100 nt, developed based on the commercial software MATLAB (25,26). Lirex was able to identify LIRs with long internal loops, but failed to identify LIRs with long arms harboring multiple mismatches and indels between each other (27). In reality, growing evidence shows that the 3.1-kb LIR reported in *Drosophila* (5) as well as the 15 012-bp and 3013-bp LIRs reported in rice (10) can only be identified by IRF. As a result, IRF (http://tandem.bu.edu/irf/irf.download.html) was utilized to identify inverted repeats in the genomes of 424 eukaryotes with recommended parameters ('2 3 5 80 10 40 500000 100000 -d -f 500') (20). The first three numbers are the match weight, the penalty for mismatches and indels used to perform Smith–Waterman local alignment when searching for complementary matched sequence pairs. With the recommended parameters, a minimum of 80% sequence matches, a maximum of 10% indels, and a minimum alignment score of 40 were required for the alignment between the two arms of identified LIRs. A maximum of 500 000-nt arm length and a maximum of 100 000-nt loop length were allowed in the identification. Two 500-nt flanking sequences at the upstream and downstream of the identified LIR were also reported. To remove short inverted repeats, potential miniature inverted-repeat transposable elements (MITEs) and Alu elements, the results of IRF were further filtered by requiring both arms of an inverted repeat longer than 400 nt, which is the arm length of typical RNAi transgenes depending on canonical RNAi factors to generate siRNAs resembling LIR-derived siRNAs (28). The output of IRF included a plain text file, reporting the genomic locations and sequences of the two arms for all LIRs, and the complementary mismatches and indels between the two arms of all LIRs. The hairpin structures of all LIRs identified by IRF were reported in a HTML file, which was further processed by an in-house R script. Finally, genomic locations and other information of all LIRs, sequences of both arms and the loop of all LIRs and hairpin structures of all LIRs, were stored in three separate files.

Construction of the LIRBase database

The R/Shiny framework was employed to build the LIRBase database (29,30). R is a prominent programming language for data science, which is extensively used in biological data analysis and bioinformatics (29). Shiny, an R package for building powerful web applications (30), has been widely used to construct biological databases and data analysis platforms (31,32). In brief, an R/Shiny application is mainly composed of two R scripts, viz., server.R and ui.R. Specifically, ui.R is used to define the graphical interface of LIRBase to receive input from the users, including data uploading and parameters settings. By compiling R code into HTML, CSS, and JavaScript code, ui.R was programmed to design the appearance and page layout of LIRBase. Similarly, ui.R was utilized to design HTML widgets such as

radio buttons, action buttons, download buttons, checkboxes, and slider input, which function as basic building blocks of LIRBase. server.R monitors all user inputs from ui.R, performs the calculation, and then displays the results in the graphical interface of LIRBase. The source code of LIRBase is freely available at https://github.com/venyao/LIRBase.

Annotation of transposable elements in the genome of *Oryza sativa* L. cv. Minghui 63

A manually curated transposon library of the rice genome was provided by the EDTA package (https://github.com/oushujun/EDTA), which is a comprehensive pipeline for *de-novo* annotation of transposons and generation of high-quality non-redundant TE libraries (33). RepeatMasker (https://www.repeatmasker.org/) was then utilized to annotate the transposons in the genome of *Oryza sativa* L. cv. Minghui 63 (MH63) with the transposon library provided by EDTA (34).

DATABASE CONTENTS AND FEATURES

Identification and characterization of LIRs in eukaryotes

We conducted the LIR identification in 424 eukaryotic genomes across 374 species by utilizing Inverted Repeats Finder (IRF) (see details in Data Collection and Methods; Figure 1A) (20). As a result, a total of 6 619 473 LIRs, including 297 317 LIRs in 77 invertebrate metazoa genomes, 1 731 978 LIRs in 139 plant genomes and 4,590,178 LIRs in 208 vertebrate genomes, were identified and deposited in LIRBase. The median size of LIRs was 3320-nt in invertebrate metazoans, 7910-nt in plants, and 7426-nt in vertebrates, respectively (Supplementary Figure S1). In general, the number of LIRs identified in different genomes was positively correlated with the size and complexity of the genome (Figure 1B). By comparing and clustering 30 423 LIRs identified in five different rice genomes, we found that only 6616 LIRs (∼30.0%) were shared by the *indica* and *japonica* subspecies (Supplementary Notes and Figure S2). Collectively, further characterizations of LIRs in seven different species (*Arabidopsis thaliana*, *Brachypodium distraction*, *Brassica napus*, cucumber, soybean, maize, as well as rice) revealed that LIRs are highly divergent between different species (Supplementary Notes). A thorough comparison between 6325 LIRs and transposable elements in a rice genome MH63 disclosed that a large proportion of LIRs were formed by adjacent transposons with identical or very similar sequences located in complementary strands (Supplementary Figure S3 and Table S2) (34). Nevertheless, many LIRs were found to be overlapped with protein-coding genes in the genome (Figure 1C, Supplementary Table S3). For about half of all 139 plant genomes, >25% of the LIRs in these genomes were found to be overlapped with genes (Figure 1C). For example, the arms of 877 LIRs in the genome of MH63 were covered by 1113 genes with ≥70% sequence coverage. A further comparison revealed that a total of 7918 splicing sites of 2340 mRNAs were covered by the arms of 2076 LIRs (Supplementary Table S4). More than half of the 2076 LIRs were overlapped with at least two splicing sites of mRNAs.

To make all identified LIRs publicly available, we constructed LIRBase, housing a comprehensive collection of 6 619 473 LIRs in 424 eukaryotic genomes. In LIRBase, thus, each LIR has a wealth of relevant information, including its identifier, nucleotide sequence, hairpin structure, genomic location, genomic coordinates of the left/right arm, the left/right arm length, the loop length, the percentage of sequence matches between the two arms, the percentage of indels between the two arms, and the overlaps between the LIR and genes. All identified LIRs across 424 eukaryotic genomes can be searched by the genomic locations or the IDs of LIRs (Figure 2A), which are publicly accessible and downloadable at https://venyao.xyz/lirbase/.

LIRBase also features online identification and annotation of LIRs for user-input sequences, which is achieved by equipping with IRF that is a command-line tool to detect inverted repeats in user-input DNA sequences (20). LIRBase provides a user-friendly interface to accept user-uploaded sequences of interest (Figure 2B). Various widgets are designed to allow users to set the parameters, including the penalty of complementary mismatches and indels between the two arms of the LIR, the maximum arm length, the maximum loop length allowed for the LIR, and so on. The results of IRF, including hairpin structures, genomic locations, and sequences of all annotated LIRs, are displayed on the web page in a table format, which can also be downloaded as plain text files. We further implemented a graphical interface to enable users to search LIRs by sequence similarities using BLASTN in LIRBase (35). Individual BLASTN database was built for LIRs identified in each of the 424 eukaryotic genomes. Searches between the input DNA sequences and one or multiple BLASTN databases can be conducted. The BLASTN results were displayed in a data table, with each row representing a BLASTN hit and containing the alignment details and the relevant LIR.

Detection of candidate LIRs encoding long hpRNAs and siRNAs

Large amounts of siRNAs can be generated from LIRs. However, an LIR can not necessarily be transcribed into a long hpRNA and siRNAs. Alignment of high-throughput sRNA sequencing data to LIRs of a specific genome is the best approach to identification of candidate LIRs encoding long hpRNAs and siRNAs, by detecting the origination of siRNAs from LIRs. Towards this end, LIRBase is also capable to detect LIRs-encoding siRNAs; it accepts high-throughput sRNA sequencing data and then aligns the data to LIRs of a given genome utilizing Bowtie (Figure 2C) (36). For efficiency, the input data should be read count of distinct sRNAs rather than the raw sequencing data. For each sRNA, we allow a maximum of 100 alignment hits, which can be easily adjusted in LIRBase. By default, no mismatches are allowed for each alignment, which can also be adjusted. By allowing a single mismatch in the alignment, we observed a higher A-to-G mismatch rate in the alignment results of three independent sRNA sequencing datasets from *Oryza sativa* (17), *Arabidopsis thaliana* (GEO accession number GSM1533527), and *Drosophila melanogaster* (GEO accession number GSM4990861). This was probably caused by potential A-to-I RNA editing in

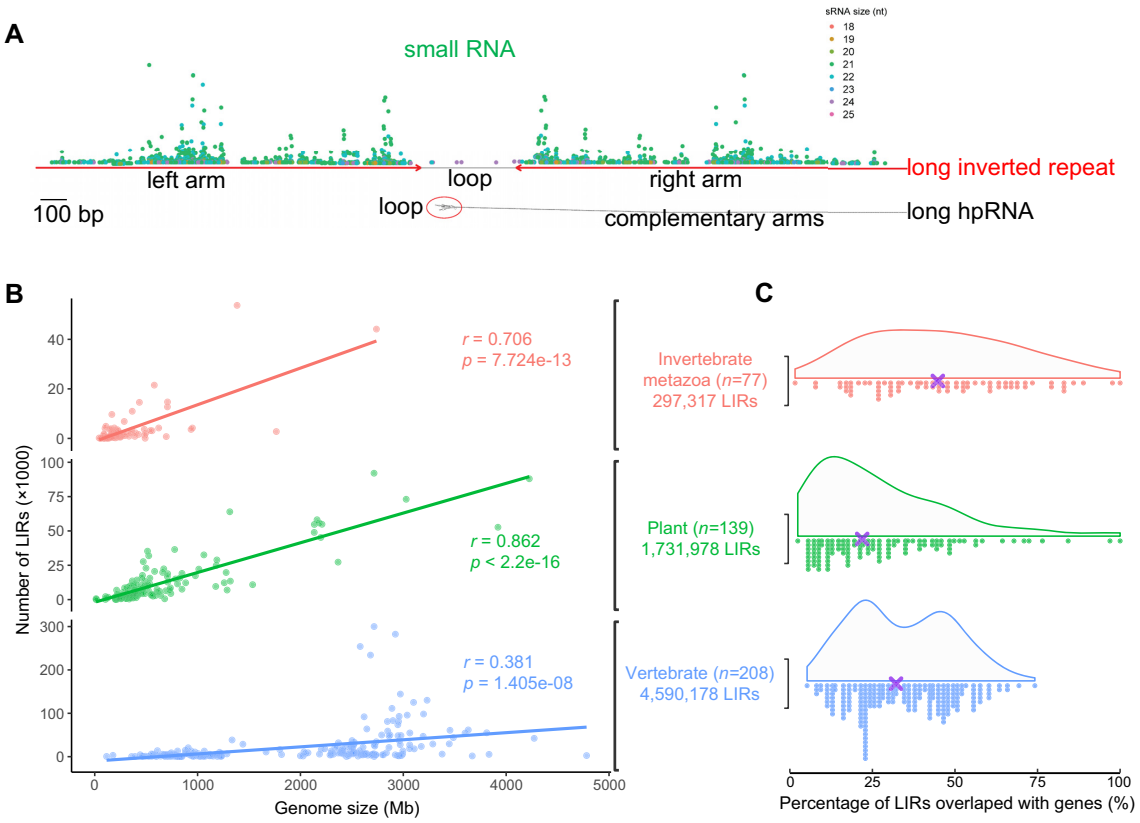

Figure 1. Overview of LIRs deposited in LIRBase. (**A**) Depiction of the biogenesis pathway of siRNAs from an LIR. The structure of the LIR is represented by two red arrows with an internal grey loop. siRNAs derived from the LIR are displayed as points with different colors along the arms of the LIR. The expression levels of siRNAs are indicated by their heights along the vertical axis. (**B**) Correlation between the number of LIRs and the genome size in invertebrate metazoa ($n = 77$), plant ($n = 139$) and vertebrate ($n = 208$). The correlation coefficient (r) and the P-value (P) of correlation test are indicated in each plot. (C) Percentage of LIRs overlapped with genes for genomes of different categories are displayed as violin plot and dot points. The purple cross indicates the median value in each plot.

the double-stranded RNA precursors of siRNAs derived from LIRs (Supplementary Figure S4) (37,38). The whole alignment by LIRBase can be downloaded as a plain text file for further processing. For a specific LIR, the alignment summaries are displayed in a tabular manner, including the number of sRNAs, the number of sRNA reads, the percentage of 21-22-nt sRNAs, the percentage of 24-nt sRNAs, and the total read count of sRNAs aligned to each LIR that can be used to quantify the expression level of LIRs. LIRBase also provides visualization functionalities to plot the alignment of all the siRNAs (Figure 2C), the expression level (TPM) of each siRNA and the structure of LIR. In addition, the percentage of sRNAs aligned to different parts of the LIR (arms, loop, and the flanking sequences) are also reported. All these results can be filtered in a user-customized manner, with the purpose to help users identify candidate LIRs encoding long hpRNAs and siRNAs.

Differential expression analysis of LIRs or sRNAs

Conditioning of gene expression level is an important approach regulating various biological processes to respond to the changing environment (39). Differential expression analysis of protein-coding genes or non-coding RNAs has been extensively used to identify candidate genes underlying various phenotype variations (40). We implemented a graphical interface in LIRBase for users to conduct differential expression analysis of LIRs or sRNAs between different biological samples/tissues utilizing DESeq2 (Figure 3A) (41). A read count matrix of all expressed LIRs/sRNAs in different samples and a sample information table should be prepared as the input data for DESeq2. The result of differentially expressed LIRs/sRNAs was listed in a data table, which was also downloadable as a plain text file. A volcano plot was generated to visualize and identify differentially expressed genes between different conditions (Supplementary Figure S5). In addition, a heatmap of sample-to-sample distances was created to demonstrate the similarities between different samples regarding whole-genome gene expression levels.

Prediction of mRNA targets of siRNAs encoded by an LIR

It was reported that siRNAs derived from LIRs can trigger the cleavage of their complementary mRNA targets to repress their expressions (5–7). As a result, we developed an analysis module in LIRBase to identify the mRNA targets of siRNAs generated from a single LIR through detecting the complementary matches between siRNAs and

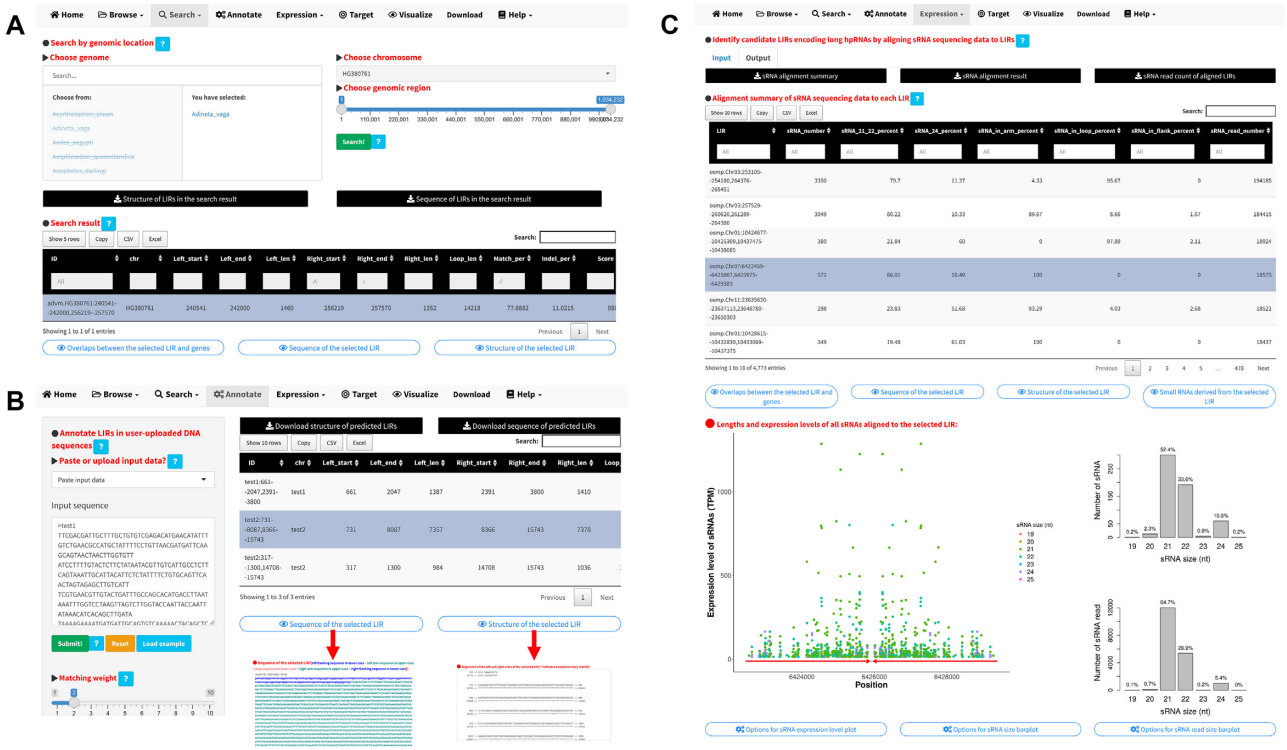

Figure 2. Features and web interfaces for LIR search, annotation and identification. (**A**) Search LIRBase by genomic locations. (**B**) Annotate LIRs in user-input DNA sequences. (**C**) Identify candidate LIRs encoding hpRNAs and siRNAs by aligning sRNA sequencing data to LIRs.

the cDNA sequence of protein-coding genes (Figure 3B). Bowtie (36) was utilized to perform the alignment of siRNAs to the cDNA sequences of a specified genome. The alignment result was further processed to report the number of all species of siRNAs, 21-nt siRNAs, and 22-nt siRNAs that were complementary aligned to each mRNA target, separately. The functional annotation of each mRNA was also extracted and displayed in the result.

Visualization of secondary structure of LIR-encoded hpRNAs

RNAs can form complex secondary structures, which are critical to their biological functions. The secondary structure of candidate long hpRNA encoded by an LIR is critical to the biogenesis of siRNAs from the LIR. Therefore, LIRBase was also implemented with RNAfold (42), a widely used tool to predict and visualize the secondary structure of RNAs (Figure 3C). LIRBase accepts an LIR's DNA sequence as input and outputs its predicted secondary structure as a plain text file of dot-bracket notation. The plain text file of dot-bracket notation and the centroid secondary structure in PDF format can be freely downloaded for downstream analysis.

CASE STUDIES

Analysis of the sRNA sequencing dataset of a rice F_2 population

In our previous study (17), we reported the sRNA sequencing of an immortalized F_2 (IMF2) population, derived from a cross between *Oryza sativa* L. cv. Zhenshan 97 (ZS97) and MH63, which are the parents of an elite rice hybrid Shanyou 63 (SY63). Using the 'Quantify' and 'DESeq' features of LIRBase, we identified 159 candidate LIRs encoding long hpRNAs and siRNAs in MH63, as well as 201 differentially expressed LIRs between MH63 and SY63 (Supplementary Tables S5–S6 and Figures S5–S7). We next aligned the sequences of 1,805,909 sRNA expression traits identified in the previous study to the LIRs of ZS97 and MH63 (10,17). We identified 1055 candidate LIRs encoding 106 744 sRNAs in the IMF2 population, by requesting each LIR covered by ≥100 sRNAs and ≥80% of the sRNAs derived from the arms of the LIR. Based on the QTL analysis results of the 1 805 909 sRNAs, we retrieved 37 489 QTLs regulating the expression of 34 492 sRNAs derived from LIRs. It was found that more than 300 sRNAs were regulated by each of the eight QTL hotspots represented by molecular markers Bin359, Bin795, Bin903, Bin1556, Bin454 (*OsDCL2a*), Bin827, Bin1139 (*OsDCL2b*) and Bin1325 (Figure 4A). The majority of sRNAs regulated by the QTL hotspots except for Bin454 and Bin1139 were aligned to an LIR located in the corresponding QTL hotspot (Figure 4B and C, Supplementary Figures S8–S13). Notably, variations were detected between these LIRs in the parental genome, leading to significant differences in the species and expression levels of siRNAs derived from the LIRs between the parental genomes. For example, the LIR present in Bin827 of ZS97 genome was totally absent from the genome of MH63. As a result, sRNAs regulated by Bin827 were highly expressed in ZS97 but were missing or lowly expressed in MH63 (Figure 4B and C).

Figure 3. Features and web interfaces for LIR analysis, visualization and download. (**A**) Differentially expression analysis of LIRs between different samples. (**B**) Predict mRNA targets of siRNAs encoded by an LIR. (**C**) Visualize secondary structure of LIR-encoded hpRNAs. (**D**) Download datasets from LIRBase.

Identification of a candidate LIR involved in disease resistance in rice

To investigate the expression profiles of LIRs and identify candidate LIRs involved in disease resistance in rice, we re-analyzed a sRNA sequencing dataset published in a previous study (43). sRNA sequencing was conducted with the pathogen (*Xoo* PXO99)-infected and mock leaves (treated by water) of *Oryza sativa* L. cv. Nipponbare at 2, 6, 12 and 24 h postinfection (hpi) (44). Processed sRNAs of all eight libraries were aligned to the LIRs of Nipponbare in LIRBase, separately. Based on similar criteria used in the analysis of sRNA sequencing data of MH63, we identified an average of 187 LIRs (ranging from 129 to 230) encoding long hpRNAs and siRNAs in the eight libraries. A total of 280 LIRs were identified by taking together the results of all eight libraries. We observed an LIR that was highly expressed in pathogen-infected leaves at 6 hpi but was barely expressed in the other seven libraries (Figure 4D). This LIR was formed by two adjacent germin-like protein (*GLP*) genes (LOC_Os02g29000 and LOC_Os02g29010) (45). A total of 1,278 siRNAs were derived from this LIR, 56.3% of which were 21-22 nt (Figure 4E). The majority of the 1278 siRNAs were aligned to the left arm of this LIR, harboring more than 110 SNPs against the right arm of the LIR (Supplementary Figure S14). Using the 'Target' feature of LIRBase, we found that a cluster of *GLP* genes on chromosome 8 of the Nipponbare genome were targeted by siRNAs derived from this LIR (Supplementary Table S7). This gene cluster was reported to function as a major-effect QTL conferring broad-spectrum disease resistance in rice (46). Powered by LIRBase, these results demonstrated that this identified LIR was most likely involved in the regulation of disease resistance in rice.

DISCUSSION AND FUTURE DIRECTIONS

Evidence has accumulated that LIRs as well as their derived functional siRNAs exert important biological roles in plants and animals (5–9). In this study, we built the first database hosting a comprehensive collection of LIRs identified in 424 eukaryotic genomes. We observed considerable discrepancy in the sequences of LIRs between different species/subspecies, which in turn resulted in significant variations in species and expression levels of sRNAs among different individuals. These results implied that LIRs contributed greatly to sequence diversity between different genomes, which may play vital roles in genome evolution (19,47). We chose to remove LIRs related with Alu elements or MITEs from LIRBase by requiring a minimum length of 400 nt for both arms of an LIR, as LIRs and their derived sRNAs related with Alu elements or MITEs are distinct from typical LIRs focused by LIRBase (3,22,48,49). However, this may also mistakenly remove a genuine LIR from LIRBase. Nevertheless, LIRBase provides a great resource

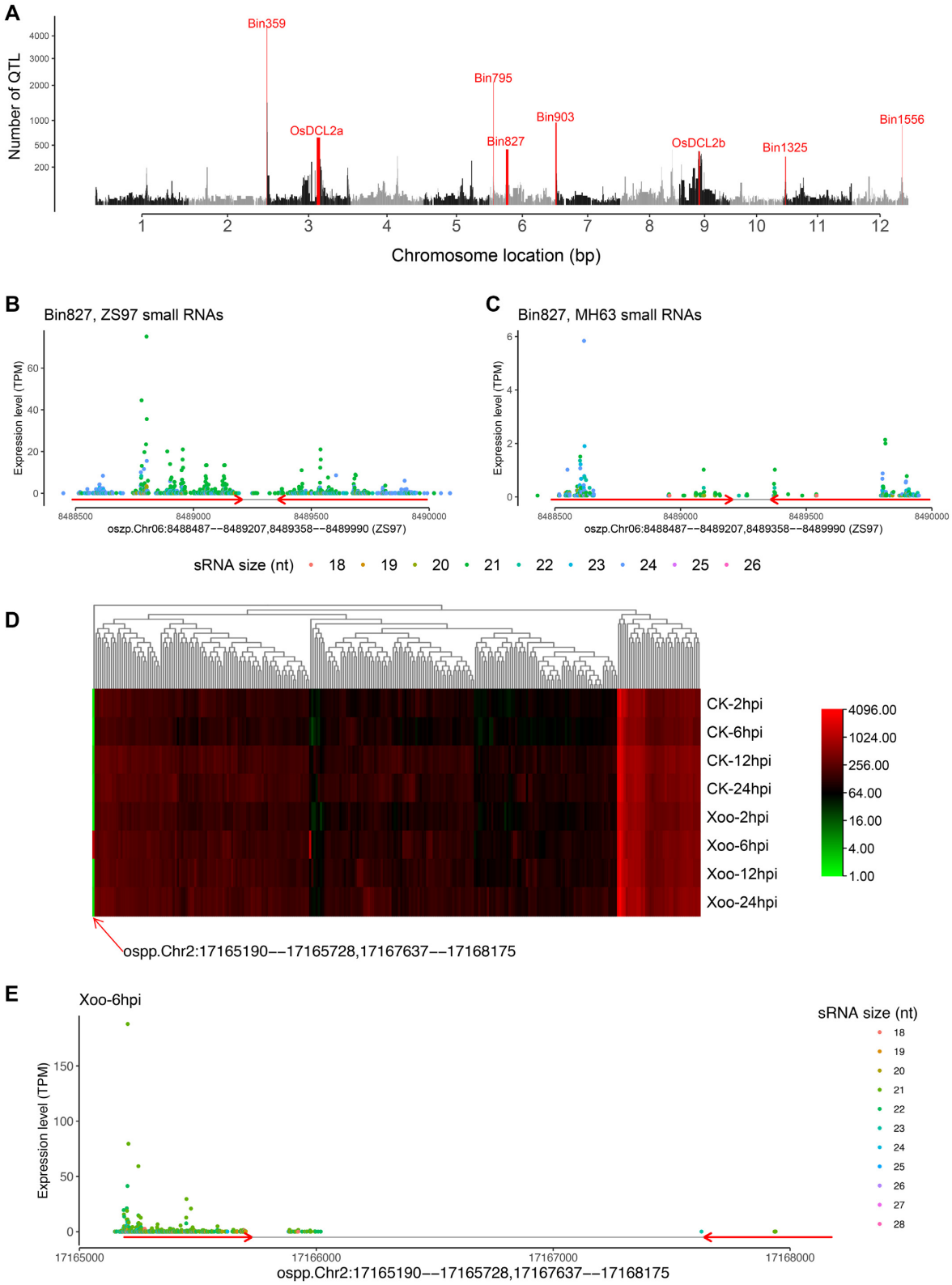

Figure 4. Two case studies of LIRBase. (**A**) Number of QTL represented by the 1567 markers of the rice F_2 population (17). Each marker corresponds to a genomic region denoted as a filled bar. The width of the bar represents the size of the genomic region. The y axis is in square root scale. Adjacent chromosomes are represented with different colors. QTL hotspots regulating the expressions of more than 300 sRNAs are denoted in red color. (**B**) Expression of sRNAs from *Oryza sativa* L. cv. ZS97 aligned to LIR oszp.Chr06:8488487–8489207,8489358–8489990 in Bin827. (**C**) Expression of sRNAs from *Oryza sativa* L. cv. MH63 aligned to LIR oszp.Chr06:8488487–8489207,8489358–8489990 in Bin827. (**D**) Expression profile of 280 LIRs in pathogen-infected and mock leaves (treated by water) of *Oryza sativa* L. cv. Nipponbare. CK, control (treated by water). Xoo, infected by *Xoo*. hpi, hours post infection. (**E**) Expression of sRNAs derived from ospp.Chr2:17165190–17165728,17167637–17168175 in pathogen-infected leaves of Nipponbare at 6 hours post infection.

for comparative investigation of LIRs in different genomes and in-depth exploration of potential roles of LIRs in evolution, adaption, and phenotype variation (Figure 3D). Importantly, LIRBase not only allows users to browse and search LIRs in any of the 424 eukaryotic genomes, but also provides friendly web functionalities to facilitate users to perform various analyses on LIRs and their derived siRNAs. Clearly, as testified by the two case studies, LIRBase bears the great utility for in-depth investigation and characterization of LIRs in diverse species. Therefore, future directions of LIRBase include the identification and integration of LIRs in a larger number of genomes. In addition, we plan to use a shorter arm length threshold for LIRs and remove potential Alu repeats and MITEs by thorough identification of Alu repeats in primate genomes and MITEs in all eukaryote genomes. We also plan to characterize the expression profiles of LIRs and the origination of siRNAs from LIRs by aligning publicly available sRNA sequencing data to all collected LIRs in LIRBase. In addition, more user-friendly interfaces and tools will be frequently updated and enhanced, with the purpose to provide users with more straightforward and interactive functionalities. We believe these features of LIRBase will greatly facilitate the annotation and functional studies of LIRs and the derived siRNAs.

DATA AVAILABILITY

LIRBase is a comprehensive database of LIRs in a wide range of eukaryotic genomes (https://venyao.xyz/lirbase/). Detailed information of all collected genomes is available under the Download menu of LIRBase.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Key Research & Development Program of China [2017YFC0907502]; National Natural Science Foundation of China [31900451, 31871328, 32030021]; Research start-up fund to topnotch talents of Henan Agricultural University [30500581]; Scientific and Technological Research Project of Henan Province [202102110015, 212102110243]; Strategic Priority Research Program of the Chinese Academy of Sciences [XDA19050302].
Conflict of interest statement. None declared.

REFERENCES

1. Yu,Y., Zhang,Y., Chen,X. and Chen,Y. (2019) Plant noncoding RNAs: hidden players in development and stress responses. *Annu. Rev. Cell Dev. Biol.*, **35**, 407–431.
2. Henderson,I.R., Zhang,X., Lu,C., Johnson,L., Meyers,B.C., Green,P.J. and Jacobsen,S.E. (2006) Dissecting *Arabidopsis thaliana* DICER function in small RNA processing, gene silencing and DNA methylation patterning. *Nat. Genet.*, **38**, 721–725.
3. Okamura,K., Chung,W.-J., Ruby,J.G., Guo,H., Bartel,D.P. and Lai,E.C. (2008) The *Drosophila* hairpin RNA pathway generates endogenous short interfering RNAs. *Nature*, **453**, 803–806.
4. Czech,B., Malone,C.D., Zhou,R., Stark,A., Schlingeheyde,C., Dus,M., Perrimon,N., Kellis,M., Wohlschlegel,J.A., Sachidanandam,R. *et al.* (2008) An endogenous small interfering RNA pathway in *Drosophila*. *Nature*, **453**, 798–802.
5. Lin,C.-J., Hu,F., Dubruille,R., Vedanayagam,J., Wen,J., Smibert,P., Loppin,B. and Lai,E.C. (2018) The hpRNA/RNAi pathway is essential to resolve intragenomic conflict in the *Drosophila* male germline. *Dev. Cell*, **46**, 316–326.
6. Zhang,Q., Ma,C., Zhang,Y., Gu,Z., Li,W., Duan,X., Wang,S., Hao,L., Wang,Y., Wang,S. *et al.* (2018) A single-nucleotide polymorphism in the promoter of a hairpin RNA contributes to *Alternariaalternata* leaf spot resistance in apple (*Malus* × *domestica*). *Plant Cell*, **30**, 1924–1942.
7. Jia,J., Ji,R., Li,Z., Yu,Y., Nakano,M., Long,Y., Feng,L., Qin,C., Lu,D., Zhan,J. *et al.* (2020) Soybean DICER-LIKE2 regulates seed coat color via production of primary 22-nucleotide small interfering RNAs from long inverted repeats. *Plant Cell*, **32**, 3662–3673.
8. Tam,O.H., Aravin,A.A., Stein,P., Girard,A., Murchison,E.P., Cheloufi,S., Hodges,E., Anger,M., Sachidanandam,R., Schultz,R.M. *et al.* (2008) Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, **453**, 534–538.
9. Wen,J., Duan,H., Bejarano,F., Okamura,K., Fabian,L., Brill,J.A., Bortolamiol-Becet,D., Martin,R., Ruby,J.G. and Lai,E.C. (2015) Adaptive regulation of testis gene expression and control of male fertility by the *Drosophila* hairpin RNA pathway. *Mol. Cell*, **57**, 165–178.
10. Yao,W., Li,Y., Xie,W. and Wang,L. (2020) Features of sRNA biogenesis in rice revealed by genetic dissection of sRNA expression level. *Comput. Struct. Biotechnol. J.*, **18**, 3207–3216.
11. Watanabe,T., Totoki,Y., Toyoda,A., Kaneda,M., Kuramochi-Miyagawa,S., Obata,Y., Chiba,H., Kohara,Y., Kono,T., Nakano,T. *et al.* (2008) Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature*, **453**, 539–543.
12. Tanaka,H. and Yao,M.-C. (2009) Palindromic gene amplification — an evolutionarily conserved role for DNA inverted repeats in the genome. *Nat. Rev. Cancer*, **9**, 216–224.
13. Tao,Y., Masly,J.P., Araripe,L., Ke,Y. and Hartl,D.L. (2007) A sex-ratio meiotic drive system in *Drosophilasimulans*. I: an autosomal suppressor. *PLoS Biol.*, **5**, e292.
14. Tao,Y., Araripe,L., Kingan,S.B., Ke,Y., Xiao,H. and Hartl,D.L. (2007) A sex-ratio meiotic drive system in *Drosophilasimulans*. II: an X-linked distorter. *PLoS Biol.*, **5**, e293.
15. Tuteja,J.H., Zabala,G., Varala,K., Hudson,M. and Vodkin,L.O. (2009) Endogenous, tissue-specific short interfering RNAs silence the chalcone synthase gene family in *Glycine max* seed coats. *Plant Cell*, **21**, 3063–3077.
16. Cho,Y.B., Jones,S.I. and Vodkin,L. (2013) The transition from primary siRNAs to amplified secondary siRNAs that regulate chalcone synthase during development of *Glycine max* seed coats. *PLoS One*, **8**, e76954.
17. Wang,J., Yao,W., Zhu,D., Xie,W. and Zhang,Q. (2015) Genetic basis of sRNA quantitative variation analyzed using an experimental population derived from an elite rice hybrid. *Elife*, **4**, e04250.
18. Axtell,M.J. (2013) Classification and comparison of small RNAs from plants. *Annu. Rev. Plant Biol.*, **64**, 137–159.
19. Wang,Y. and Leung,F.C.C. (2006) Long inverted repeats in eukaryotic genomes: Recombinogenic motifs determine genomic plasticity. *FEBS Lett.*, **580**, 1277–1284.
20. Warburton,P.E., Giordano,J., Cheung,F., Gelfand,Y. and Benson,G. (2004) Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Genome Res.*, **14**, 1861–1869.
21. Tschudi,C., Shi,H., Franklin,J.B. and Ullu,E. (2012) Small interfering RNA-producing loci in the ancient parasitic eukaryote *Trypanosomabrucei*. *BMC Genomics*, **13**, 427.
22. Aygun,N. (2015) Correlations between long inverted repeat (LIR) features, deletion size and distance from breakpoint in human gross gene deletions. *Sci. Rep.*, **5**, 8300.
23. Yates,A.D., Achuthan,P., Akanni,W., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R. *et al.* (2020) Ensembl 2020. *Nucleic Acids Res.*, **48**, D682–D688.

24. Tello-Ruiz,M.K., Naithani,S., Stein,J.C., Gupta,P., Campbell,M., Olson,A., Wei,S., Preece,J., Geniza,M.J., Jiao,Y. *et al.* (2018) Gramene 2018: unifying comparative genomics and pathway resources for plant research. *Nucleic Acids Res.*, **46**, D1181–D1189.
25. Ye,C., Ji,G., Li,L. and Liang,C. (2014) detectIR: a novel program for detecting perfect and imperfect inverted repeats using complex numbers and vector calculation. *PLoS One*, **9**, e113349.
26. Sreeskandarajan,S., Flowers,M.M., Karro,J.E. and Liang,C. (2014) A MATLAB-based tool for accurate detection of perfect overlapping and nested inverted repeats in DNA sequences. *Bioinformatics*, **30**, 887–888.
27. Wang,Y. and Huang,J.-M. (2017) Lirex: a package for identification of long inverted repeats in genomes. *Genomics Proteomics Bioinformatics*, **15**, 141–146.
28. Okamura,K., Chung,W.-J. and Lai,E.C. (2008) The long and short of inverted repeat genes in animals: MicroRNAs, mirtrons and hairpin RNAs. *Cell Cycle*, **7**, 2840–2845.
29. R Core Team. (2020) In: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
30. Chang,W., Cheng,J., Allaire,J., Xie,Y. and McPherson,J. (2021) In: *shiny: web application framework for R. R package version 1.6.0*.
31. Yu,Y., Yao,W., Wang,Y. and Huang,F. (2019) shinyChromosome: an R/Shiny application for interactive creation of non-circular plots of whole genomes. *Genomics Proteomics Bioinformatics*, **17**, 535–539.
32. Zhou,W., Wang,L., Zheng,W. and Yao,W. (2019) MaizeSNPDB: a comprehensive database for efficient retrieve and analysis of SNPs among 1210 maize lines. *Comput. Struct Biotechnol. J.*, **17**, 1377–1383.
33. Ou,S., Su,W., Liao,Y., Chougule,K., Agda,J.R.A., Hellinga,A.J., Lugo,C.S.B., Elliott,T.A., Ware,D., Peterson,T. *et al.* (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.*, **20**, 275.
34. Zhang,J., Chen,L.-L., Xing,F., Kudrna,D.A., Yao,W., Copetti,D., Mu,T., Li,W., Song,J.-M., Xie,W. *et al.* (2016) Extensive sequence divergence between the reference genomes of two elite *indica* rice varieties Zhenshan 97 and Minghui 63. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E5163–E5171.
35. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
36. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
37. Porath,H.T., Knisbacher,B.A., Eisenberg,E. and Levanon,E.Y. (2017) Massive A-to-I RNA editing is common across the Metazoa and correlates with dsRNA abundance. *Genome Biol.*, **18**, 185.
38. Li,M., Xia,L., Zhang,Y., Niu,G., Li,M., Wang,P., Zhang,Y., Sang,J., Zou,D., Hu,S. *et al.* (2019) Plant editosome database: a curated database of RNA editosome in plants. *Nucleic Acids Res.*, **47**, D170–D174.
39. Groen,S.C., Ćalić,I., Joly-Lopez,Z., Platts,A.E., Choi,J.Y., Natividad,M., Dorph,K., Mauck,W.M., Bracken,B., Cabral,C.L.U. *et al.* (2020) The strength and pattern of natural selection on gene expression in rice. *Nature*, **578**, 572–576.
40. Van den Berge,K., Hembach,K.M., Soneson,C., Tiberi,S., Clement,L., Love,M.I., Patro,R. and Robinson,M.D. (2019) RNA sequencing data: hitchhiker's guide to expression analysis. *Annu. Rev. Biomed Data. Sci.*, **2**, 139–173.
41. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
42. Lorenz,R., Bernhart,S.H., Höner zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
43. Zhao,Y.-T., Wang,M., Wang,Z.-M., Fang,R.-X., Wang,X.-J. and Jia,Y.-T. (2015) Dynamic and coordinated expression changes of rice small RNAs in response to *Xanthomonasoryzae* pv. *oryzae*. *J. Genet. Genomics.*, **42**, 625–637.
44. Kawahara,Y., de la Bastide,M., Hamilton,J.P., Kanamori,H., McCombie,W.R., Ouyang,S., Schwartz,D.C., Tanaka,T., Wu,J., Zhou,S. *et al.* (2013) Improvement of the *Oryzasativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice*, **6**, 4.
45. Li,L., Xu,X., Chen,C. and Shen,Z. (2016) Genome-wide characterization and expression analysis of the germin-like protein family in rice and *Arabidopsis*. *Int. J. Mol. Sci.*, **17**, 1622.
46. Manosalva,P.M., Davidson,R.M., Liu,B., Zhu,X., Hulbert,S.H., Leung,H. and Leach,J.E. (2009) A germin-like protein gene family functions as a complex quantitative trait locus conferring broad-spectrum disease resistance in rice. *Plant Physiol.*, **149**, 286–296.
47. Thybert,D., Roller,M., Navarro,F.C.P., Fiddes,I., Streeter,I., Feig,C., Martin-Galvez,D., Kolmogorov,M., Janoušek,V., Akanni,W. *et al.* (2018) Repeat associated mechanisms of genome evolution and function revealed by the *Muscaroli* and *Muspahari* genomes. *Genome Res.*, **28**, 448–459.
48. Lu,C., Chen,J., Zhang,Y., Hu,Q., Su,W. and Kuang,H. (2012) Miniature Inverted–Repeat Transposable Elements (MITEs) have been accumulated through amplification bursts and play important roles in gene expression and species diversity in *Oryzasativa*. *Mol. Biol. Evol.*, **29**, 1005–1017.
49. Jiang,N., Feschotte,C., Zhang,X. and Wessler,S.R. (2004) Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). *Curr. Opin. Plant Biol.*, **7**, 115–119.