An Online Database for Exploring Over 2,000 Arabidopsis Small RNA Libraries[1[OPEN]]

Li Feng,[a,b,2] Fei Zhang,[b,2] Hong Zhang,[b] Yan Zhao,[b] Blake C. Meyers,[c,d] and Jixian Zhai[b,3,4]

[a]Harbin Institute of Technology, Harbin, Heilongjiang 150001, China
[b]Department of Biology and Institute of Plant and Food Science, Southern University of Science and Technology, Shenzhen, Guangdong 518055, China
[c]Donald Danforth Plant Science Center, St. Louis, Missouri 63132
[d]University of Missouri–Columbia, Division of Plant Sciences, Columbia, Missouri 65211

ORCID IDs: 0000-0001-9780-431X (H.Z.); 0000-0002-3087-4767 (Y.Z.); 0000-0003-3436-6097 (B.C.M.); 0000-0002-0217-0666 (J.Z.).

Small RNAs (sRNAs) play a wide range of important roles in plants, from maintaining genome stability and enhancing disease resistance to regulating developmental processes. Over the past decade, next-generation sequencing technologies have allowed us to explore the sRNA populations with unprecedented depth and accuracy. The community has accumulated a tremendous amount of sRNA sequencing (sRNA-seq) data from various genotypes, tissues, and treatments. However, it has become increasingly challenging to access these "big data" and extract useful information, particularly for researchers lacking sophisticated bioinformatics tools and expensive computational resources. Here, we constructed an online website, Arabidopsis Small RNA Database (ASRD, http://ipf.sustech.edu.cn/pub/asrd), that allows users to easily explore the information from publicly available Arabidopsis (*Arabidopsis thaliana*) sRNA libraries. Our database contains ~2.3 billion sRNA reads, representing ~250 million unique sequences from 2,024 sRNA-seq libraries. We downloaded the raw data for all libraries and reprocessed them with a unified pipeline so that the normalized abundance of any particular sRNA or the sum of abundances of sRNAs from a genic or transposable element region can be compared across all libraries. We also integrated an online Integrative Genomics Viewer browser into our Web site for convenient visualization. ASRD is a free, web-accessible, and user-friendly database that supports the direct query of over 2,000 Arabidopsis sRNA-seq libraries. We believe this resource will help plant researchers take advantage of the vast next-generation sequencing datasets available in the public domain.

Small RNAs (sRNAs) in plants, generally 20–24 nucleotides (nt) in size, are regulatory RNA molecules functioning in growth and developmental processes, as well as in protecting plants against viruses, transgenes, and transposable elements (TEs; Baulcombe, 2004; Carthew and Sontheimer, 2009; Chen, 2009; Ruiz-Ferrer and Voinnet, 2009; Bologna and Voinnet, 2014; Borges and Martienssen, 2015; Meyers and Axtell, 2019). There are two major classes of sRNAs: microRNAs (miRNAs) and small interfering RNAs (siRNAs; Axtell, 2013). For miRNA biogenesis, RNA polymerase II transcribes miRNA genes into long, single-stranded primary miRNAs, then these single-stranded primary miRNAs form imperfectly matched foldback structures and can be further processed by Dicer-like1 (DCL1) protein into mature miRNA/miRNA* duplexes (Carthew and Sontheimer, 2009; Voinnet, 2009; Axtell, 2013). In contrast, siRNAs are processed by DCLs other than DCL1, with their perfectly matched double-stranded precursors produced by RNA-dependent RNA polymerases (RDRs; Carthew and Sontheimer, 2009; Axtell, 2013). Transacting siRNAs (ta-siRNAs) are usually 21 nt in length and their biogenesis depends on RDR6 and DCL4 activity for their generation, triggered by either "two-hits" (Allen et al., 2005; Axtell et al., 2006) or a single 22-nt miRNA/siRNA (Fei et al., 2013). The production of 24-nt siRNAs is generally initiated by RNA polymerase IV (Pol IV) transcription, followed by RDR2 activity to form a complementary strand that is subsequently cleaved by DCL3. These 24-nt siRNAs can mediate transcriptional gene silencing by targeting DNA methylation and repressive histone modifications (Matzke and Mosher, 2014; Cuerda-Gil and Slotkin, 2016).

The development of next-generation sequencing (NGS) technologies has drastically improved sensitivity and accuracy in detecting sRNA molecules. A large number

[1]This work was supported by the National Key R&D Program of China Grant (2019YFA0903903), the National Natural Science Foundation of China (31871234 to J.Z.), the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (2016ZT06S172), and the Shenzhen Sci-Tech Fund (KYTDPT20181011104005).

[2]These authors contributed equally to this article.

[3]Author for contact: zhaijx@sustech.edu.cn.

[4]Senior author.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Jixian Zhai (zhaijx@sustech.edu.cn).

J.Z. and B.C.M. conceived the original research plan; J.Z. supervised the analysis; L.F. collected and processed publicly available sRNA-seq data; L.F., H.Z. and Y.Z. analyzed sRNA-seq data; L.F. and F.Z. built the database and website; L.F. wrote the article; all authors revised the article.

[OPEN]Articles can be viewed without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.19.00959

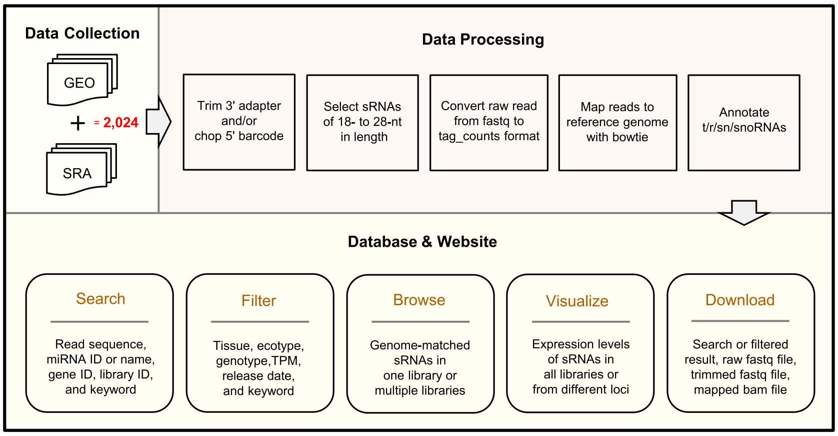

Data Collection
GEO
+ = 2,024
SRA
Data Processing
Trim 3' adapter and/or chop 5' barcode
Select sRNAs of 18- to 28-nt in length
Convert raw read from fastq to tag_counts format
Map reads to reference genome with bowtie
Annotate t/r/sn/snoRNAs
Database & Website
Search
Read sequence, miRNA ID or name, gene ID, library ID, and keyword
Filter
Tissue, ecotype, genotype,TPM, release date, and keyword
Browse
Genome-matched sRNAs in one library or multiple libraries
Visualize
Expression levels of sRNAs in all libraries or from different loci
Download
Search or filtered result, raw fastq file, trimmed fastq file, mapped bam file

Figure 1. Overflow of the ASRD. A total of 2,024 publicly available Arabidopsis sRNA-seq libraries were collected from GEO and SRA databases, and processed with a unified pipeline for cross-library comparisons, and all the sRNA-related information can be accessed via keyword-based searching on the ASRD Web site (http://ipf.sustech.edu.cn/pub/asrd/).

of sRNA sequencing (sRNA-seq) datasets have been stored in public databases, such as the Gene Expression Omnibus (GEO; Clough and Barrett, 2016) and the Sequence Read Archive (SRA; Leinonen et al., 2011), which include samples from various genotypes, tissues, and treatments. However, it is difficult to directly compare sRNA data across publications using the original results because different pipelines and methods were usually used to analyze these sRNA-seq libraries in each study. Thus, it is difficult for researchers, especially those with a wet-lab background, to take advantage of the vast collection of public NGS datasets.

RESULTS

To address this challenge, here we present the Arabidopsis Small RNA Database (ASRD), an online database with integrated, multifaceted functions for exploring published Arabidopsis (*Arabidopsis thaliana*) sRNA-seq libraries (Fig. 1). ASRD currently hosts 2,024 sRNA-seq libraries collected from GEO and SRA databases. The raw sequencing data of these libraries were downloaded and processed with a unified pipeline so that we can compare the normalized abundances of sRNAs across all libraries. ASRD provides easy-to-access links to download the raw, trimmed, and mapped reads so that researchers can apply their favorite tools to these published libraries with ease. ASRD supports a "Google-like" search through querying of a single sRNA sequence, miRNA ID, miRNA name, miRNA sequence, gene ID, library ID, or library-related keyword (Fig. 2, A and B). ASRD also supports "Advanced options" for narrowing down the search area. All search results can be downloaded via a simple click and be further analyzed using Microsoft Excel. We also added a built-in online Integrative Genomics Viewer (IGV) interface (Robinson et al., 2011) to visualize and browse sRNA alignments (Fig. 2C). Functions of ASRD are described in the following sections (please see the online video tutorial for a step-by-step instruction on how to use the website):

An "All libraries" Table

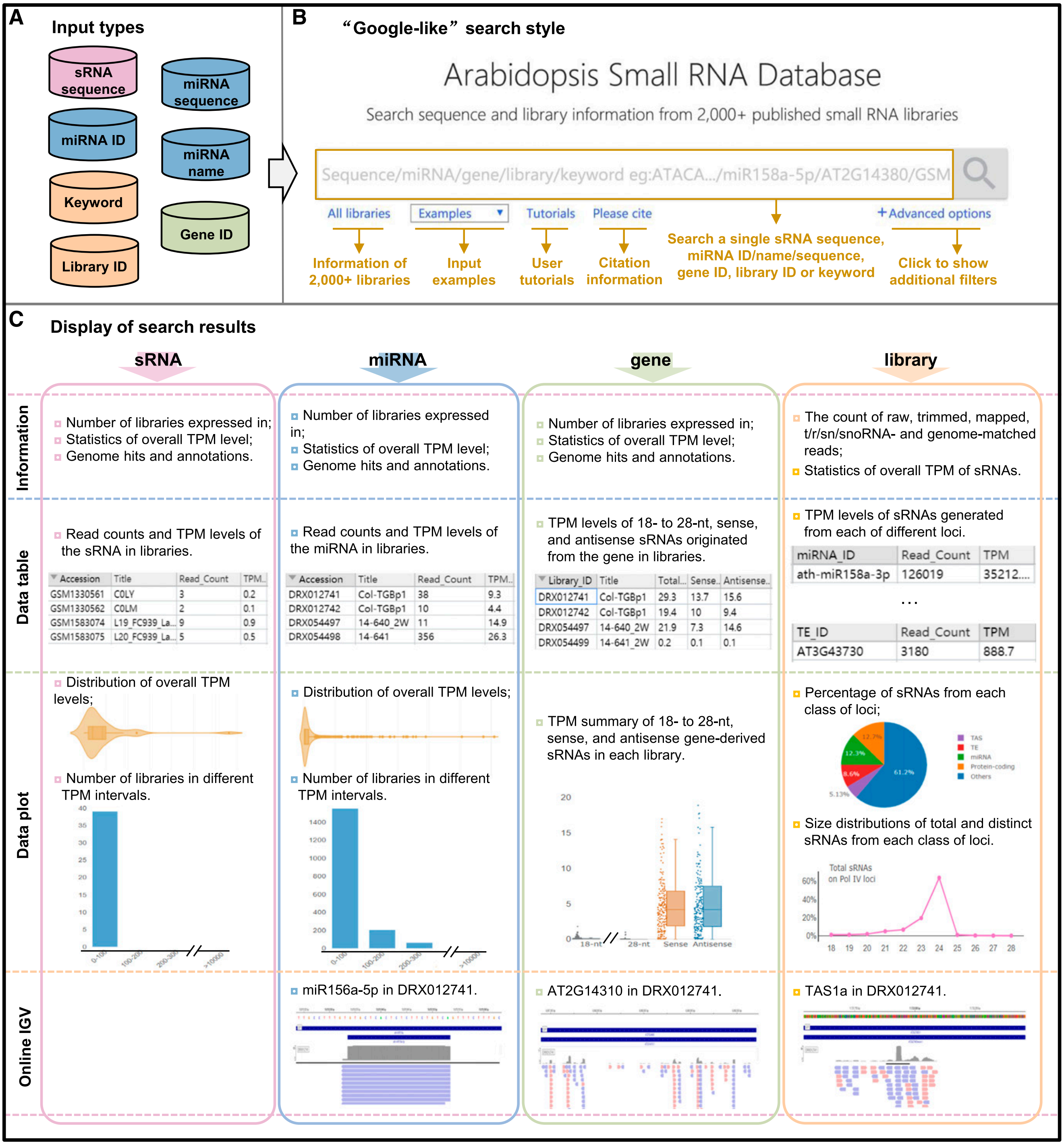
To give an overview of all 2,024 sRNA-seq libraries in ASRD (Supplemental Table S1), we provided a table with detailed information of each library, including description, genotype, ecotype, tissue, the counts of raw, transfer/ribosomal/small nuclear/small nucleolar RNAs (t/r/sn/snoRNA)-matched, genome-matched total and distinct reads, the transcripts per million (TPM) levels of sRNAs produced from miRNA, protein-coding (PC), and Pol IV loci. A table with more information (such as size distribution of total sRNAs and different classes of sRNAs in each library) can be downloaded via the link on the "All libraries" page.

Search by sRNA Sequence or miRNA ID/Name

For querying the information of a single miRNA or sRNA, ASRD supports searches by a single sRNA sequence, miRNA ID, miRNA name, or miRNA sequence, and it will return the statistics of expression levels across all libraries. Taking the query of miR158a-5p as an example, Figure 3A shows the maximum, median, mean, and minimum TPM levels of miR158a-5p in all libraries. The violin plot shows the overall TPM distribution of miR158a-5p, and the bar diagram displays the number of miR158a-5p expressed libraries grouped by different TPM intervals (Fig. 3B). The result table with the raw count, TPM, TP5M, and TP10M of miR158a-5p in each library is also downloadable. Furthermore, users can narrow down the query results via the "Advanced options" by applying a combination of filters including tissue, ecotype, genotype, TPM level, keyword, and release date (Fig. 3C). The integrated online IGV interface can be used to browse the mapped miR158a-5p sequences on the genome in one or more libraries (Fig. 3D).

Search Library Using ID or Keyword

This function supports querying of a single library ID or library-related keyword. ASRD shows not only the



sRNA

- Number of libraries expressed in;
- Statistics of overall TPM level;
- Genome hits and annotations.
- Read counts and TPM levels of the sRNA in libraries.

Accession	Title	Read_Count	TPM..
GSM1330561	C0LY	3	0.2
GSM1330562	C0LM	2	0.1
GSM1583074	L19_FC939_La...	9	0.9
GSM1583075	L20_FC939_La...	5	0.5

- Distribution of overall TPM levels;
- Number of libraries in different TPM intervals.

miRNA

- Number of libraries expressed in;
- Statistics of overall TPM level;
- Genome hits and annotations.
- Read counts and TPM levels of the miRNA in libraries.

Accession	Title	Read_Count	TPM..
DRX012741	Col-TGBp1	38	9.3
DRX012742	Col-TGBp1	10	4.4
DRX054497	14-640_2W	11	14.9
DRX054498	14-641	356	26.3

- Distribution of overall TPM levels;
- Number of libraries in different TPM intervals.
- miR156a-5p in DRX012741.

gene

- Number of libraries expressed in;
- Statistics of overall TPM level;
- Genome hits and annotations.
- TPM levels of 18- to 28-nt, sense, and antisense sRNAs originated from the gene in libraries.

Library_ID	Title	Total...	Sense..	Antisense..
DRX012741	Col-TGBp1	29.3	13.7	15.6
DRX012742	Col-TGBp1	19.4	10	9.4
DRX054497	14-640_2W	21.9	7.3	14.6
DRX054499	14-641_2W	0.2	0.1	0.1

- TPM summary of 18- to 28-nt, sense, and antisense gene-derived sRNAs in each library.
- AT2G14310 in DRX012741.

library

- The count of raw, trimmed, mapped, t/r/sn/snoRNA- and genome-matched reads;
- Statistics of overall TPM of sRNAs.
- TPM levels of sRNAs generated from each of different loci.

miRNA_ID	Read_Count	TPM
ath-miR158a-3p	126019	35212....

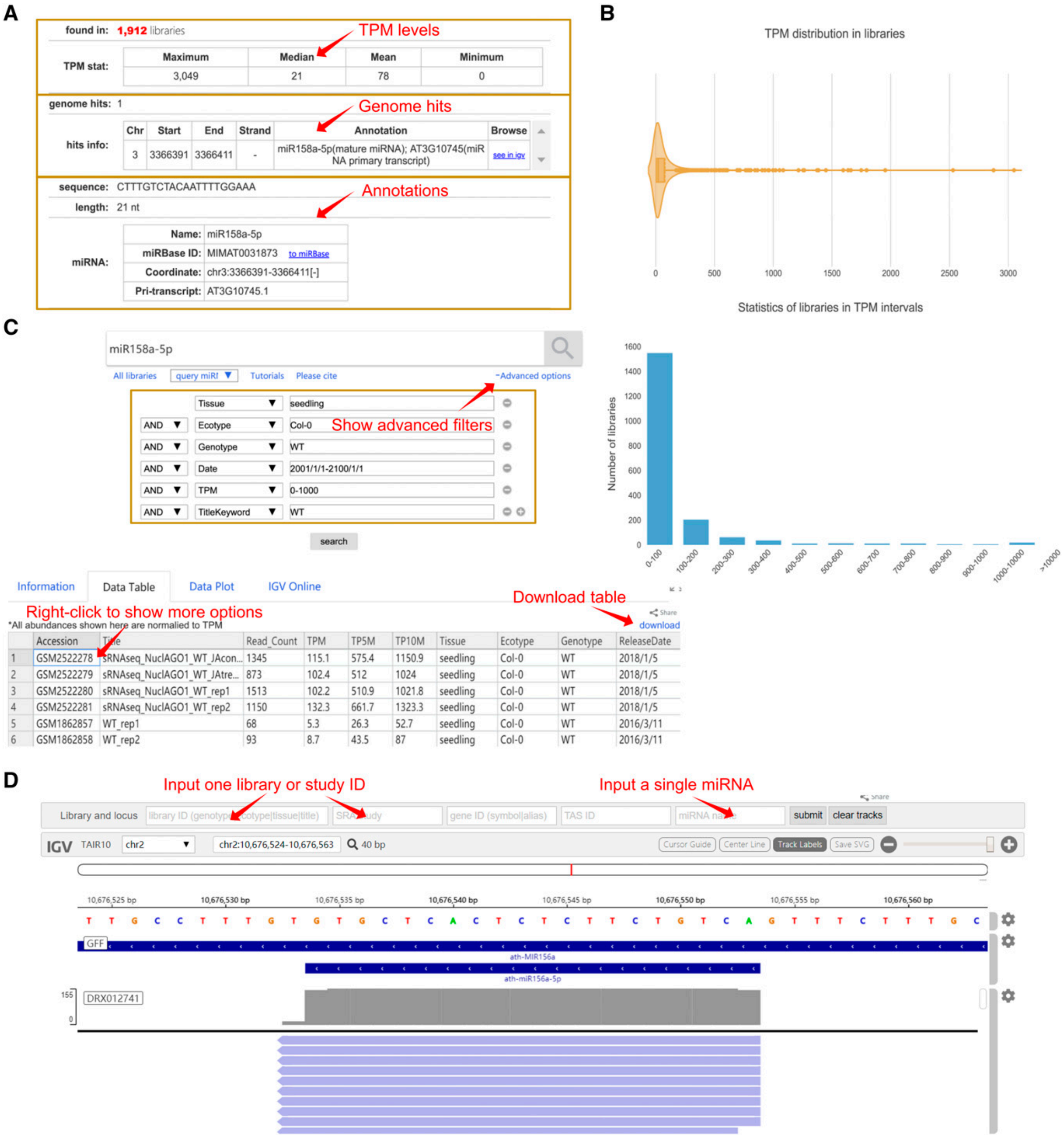
...

TE_ID	Read_Count	TPM
AT3G43730	3180	888.7

- Percentage of sRNAs from each class of loci;
- Size distributions of total and distinct sRNAs from each class of loci.
- TAS1a in DRX012741.

Figure 2. Basic functions of ASRD. A, Input can be an sRNA sequence, an miRNA sequence, an miRNA ID, an miRNA name, a keyword, a library ID, or a gene ID. B, The introduction of ASRD main page functions. ASRD supports a "Google-like" search that allows the users to search each item in one single input box. "All libraries" shows the detailed information of libraries; "Examples" presents different types of queries; "Tutorials" links to instructions to users; "Advanced options" describes various additional filters. C, Display of search results, including the information, data table, data plot, and online IGV, based on different queries of "sRNA," "miRNA," "gene," and "library." All results can be downloaded to a local computer.

library-related information gathered from the GEO or SRA database but also the statistics of sRNA expression levels across libraries. Read counts include raw, trimmed, mapped, t/r/sn/snoRNA-matched, genome-matched total and distinct reads. TPM levels are for sRNAs generated from miRNA, PC, Pol IV, TE, ta-siRNA loci



found in: 1,912 libraries

TPM stat:

Maximum	Median	Mean	Minimum
3,049	21	78	0

genome hits: 1

hits info:

Chr	Start	End	Strand	Annotation	Browse
3	3366391	3366411	-	miR158a-5p(mature miRNA); AT3G10745(miRNA primary transcript)	see in igv

sequence: CTTTGTCTACAATTTTGGAAA

length: 21 nt

miRNA:

Name:	miR158a-5p
miRBase ID:	MIMAT0031873 to miRBase
Coordinate:	chr3:3366391-3366411[-]
Pri-transcript:	AT3G10745.1

miR158a-5p

All libraries | query miRI ▼ | Tutorials | Please cite | -Advanced options

	Tissue ▼	seedling
AND ▼	Ecotype ▼	Col-0
AND ▼	Genotype ▼	WT
AND ▼	Date ▼	2001/1/1-2100/1/1
AND ▼	TPM ▼	0-1000
AND ▼	TitleKeyword ▼	WT

search

Information | Data Table | Data Plot | IGV Online

Share
download

*All abundances shown here are normalied to TPM

	Accession	Title	Read_Count	TPM	TP5M	TP10M	Tissue	Ecotype	Genotype	ReleaseDate
1	GSM2522278	sRNAseq_NuclAGO1_WT_JAcon...	1345	115.1	575.4	1150.9	seedling	Col-0	WT	2018/1/5
2	GSM2522279	sRNAseq_NuclAGO1_WT_JAtre...	873	102.4	512	1024	seedling	Col-0	WT	2018/1/5
3	GSM2522280	sRNAseq_NuclAGO1_WT_rep1	1513	102.2	510.9	1021.8	seedling	Col-0	WT	2018/1/5
4	GSM2522281	sRNAseq_NuclAGO1_WT_rep2	1150	132.3	661.7	1323.3	seedling	Col-0	WT	2018/1/5
5	GSM1862857	WT_rep1	68	5.3	26.3	52.7	seedling	Col-0	WT	2016/3/11
6	GSM1862858	WT_rep2	93	8.7	43.5	87	seedling	Col-0	WT	2016/3/11

Share

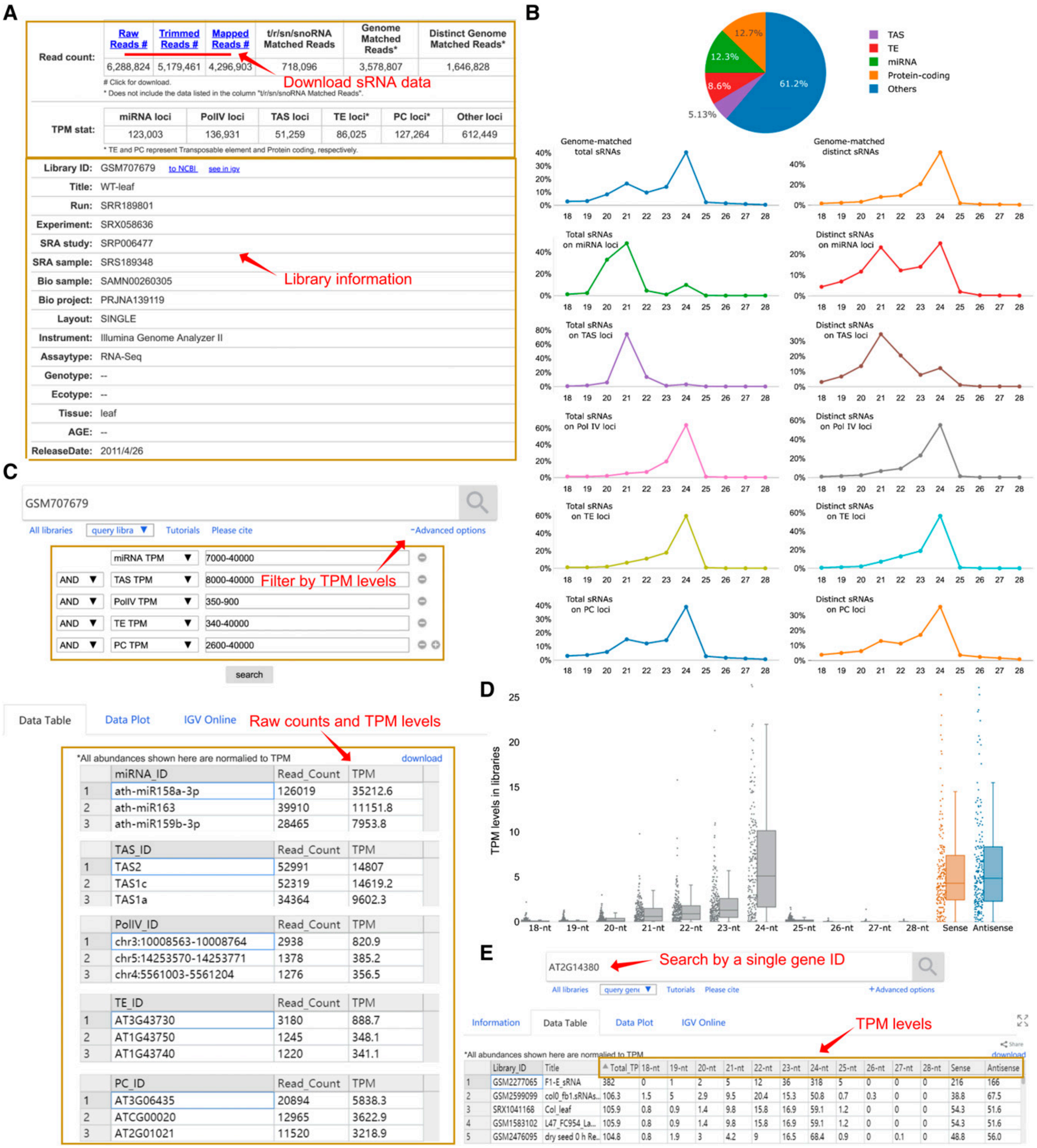
Library and locus | library ID (genotype|ecotype|tissue|title) | SRA study | gene ID (symbol|alias) | TAS ID | miRNA name | submit | clear tracks

IGV TAIR10 | chr2 ▼ | chr2:10,676,524-10,676,563 | 40 bp | Cursor Guide | Center Line | Track Labels | Save SVG

10,676,525 bp | 10,676,530 bp | 10,676,535 bp | 10,676,540 bp | 10,676,545 bp | 10,676,550 bp | 10,676,555 bp | 10,676,560 bp

T T G C C T T T G T G T G C T C A C T C T C T T C T G T C A G T T T C T T T G C

Figure 3. Example of a query using miR158a-5p. A, Example data for miR158a-5p, such as the statistics of maximum, median, mean, and minimum TPM levels in all libraries, as well as genome hits and their annotations on miRBase. B, The violin plot shows the overall TPM distribution, and each point represents the TPM level of miR158a-5p in a library. The bar diagram displays the number of libraries in different TPM intervals. C, The result table displays the read count, TPM, TP5M, and TP10M levels of miR158a-5p in all libraries. The advanced options can be used to filter the results by tissue, ecotype, genotype, release date, TPM level, or keyword. Right-clicking on each column of this table shows more operations, such as adding, removing, or sorting a column, linking a library to the National Center for Biotechnology Information, or adding a library to online IGV. D, The online IGV browses the mapped miR158a-5p sequences in the DRX012741 library.



A
Download sRNA data
Library information
B
TAS
TE
miRNA
Protein-coding
Others
12.7%
12.3%
8.6%
5.13%
61.2%
Genome-matched total sRNAs
Genome-matched distinct sRNAs
Total sRNAs on miRNA loci
Distinct sRNAs on miRNA loci
Total sRNAs on TAS loci
Distinct sRNAs on TAS loci
Total sRNAs on Pol IV loci
Distinct sRNAs on Pol IV loci
Total sRNAs on TE loci
Distinct sRNAs on TE loci
Total sRNAs on PC loci
Distinct sRNAs on PC loci
0% 10% 20% 30% 40% 60% 80%
18 19 20 21 22 23 24 25 26 27 28
C
Filter by TPM levels
Raw counts and TPM levels
D
TPM levels in libraries
0 5 10 15 20 25
18-nt 19-nt 20-nt 21-nt 22-nt 23-nt 24-nt 25-nt 26-nt 27-nt 28-nt Sense Antisense
E
Search by a single gene ID
TPM levels

Read count:

Raw Reads #	Trimmed Reads #	Mapped Reads #	t/r/sn/snoRNA Matched Reads	Genome Matched Reads*	Distinct Genome Matched Reads*
6,288,824	5,179,461	4,296,903	718,096	3,578,807	1,646,828

Click for download.
* Does not include the data listed in the column "t/r/sn/snoRNA Matched Reads".

TPM stat:

miRNA loci	PolIV loci	TAS loci	TE loci*	PC loci*	Other loci
123,003	136,931	51,259	86,025	127,264	612,449

* TE and PC represent Transposable element and Protein coding, respectively.

Library ID:	GSM707679 to NCBI see in igv
Title:	WT-leaf
Run:	SRR189801
Experiment:	SRX058636
SRA study:	SRP006477
SRA sample:	SRS189348
Bio sample:	SAMN00260305
Bio project:	PRJNA139119
Layout:	SINGLE
Instrument:	Illumina Genome Analyzer II
Assaytype:	RNA-Seq
Genotype:	--
Ecotype:	--
Tissue:	leaf
AGE:	--
ReleaseDate:	2011/4/26

GSM707679

All libraries | query libra ▼ | Tutorials | Please cite | −Advanced options

	miRNA TPM ▼	7000-40000
AND ▼	TAS TPM ▼	8000-40000
AND ▼	PolIV TPM ▼	350-900
AND ▼	TE TPM ▼	340-40000
AND ▼	PC TPM ▼	2600-40000

search

Data Table | Data Plot | IGV Online

*All abundances shown here are normalied to TPM download

	miRNA_ID	Read_Count	TPM
1	ath-miR158a-3p	126019	35212.6
2	ath-miR163	39910	11151.8
3	ath-miR159b-3p	28465	7953.8

	TAS_ID	Read_Count	TPM
1	TAS2	52991	14807
2	TAS1c	52319	14619.2
3	TAS1a	34364	9602.3

	PolIV_ID	Read_Count	TPM
1	chr3:10008563-10008764	2938	820.9
2	chr5:14253570-14253771	1378	385.2
3	chr4:5561003-5561204	1276	356.5

	TE_ID	Read_Count	TPM
1	AT3G43730	3180	888.7
2	AT1G43750	1245	348.1
3	AT1G43740	1220	341.1

	PC_ID	Read_Count	TPM
1	AT3G06435	20894	5838.3
2	ATCG00020	12965	3622.9
3	AT2G01021	11520	3218.9

AT2G14380

All libraries | query gene ▼ | Tutorials | Please cite | +Advanced options

Information | Data Table | Data Plot | IGV Online

*All abundances shown here are normalied to TPM Share download

	Library_ID	Title	▲Total_TP	18-nt	19-nt	20-nt	21-nt	22-nt	23-nt	24-nt	25-nt	26-nt	27-nt	28-nt	Sense	Antisense
1	GSM2277065	F1-E_sRNA	382	0	1	2	5	12	36	318	5	0	0	0	216	166
2	GSM2599099	col0_fb1.sRNAs...	106.3	1.5	5	2.9	9.5	20.4	15.3	50.8	0.7	0.3	0	0	38.8	67.5
3	SRX1041168	Col_leaf	105.9	0.8	0.9	1.4	9.8	15.8	16.9	59.1	1.2	0	0	0	54.3	51.6
4	GSM1583102	L47_FC954_La...	105.9	0.8	0.9	1.4	9.8	15.8	16.9	59.1	1.2	0	0	0	54.3	51.6
5	GSM2476095	dry seed 0 h Re...	104.8	0.8	1.9	3	4.2	9	16.5	68.4	0.9	0	0.1	0	48.8	56.0

Figure 4. Examples of queries using library ID and gene ID. A to C, Search by library ID. A, The library information includes statistics of the raw, trimmed, mapped, t/r/sn/snoRNA-matched, genome-matched total and distinct reads, and the TPM levels of sRNAs generated from miRNA, TAS, Pol IV, TE, PC, and other classes of loci. B, The pie diagram exhibits the percentage of each class of sRNAs in the library, and the line diagrams show the size distributions of genome-matched total and distinct sRNAs from each class, with *x* axis indicating sRNA size (nt), and *y* axis showing the percentage of sRNAs. C, The tables display the read counts and TPM levels of sRNAs derived from each of miRNAs, TAS, Pol IV, TE, and the Top-100 abundant PCs. The advanced options additionally allow the users to filter the results by the TPM levels of sRNAs produced from different classes of loci. D and E, Search by gene ID. D, The diagram with scatters and boxes describes the TPM levels of sRNAs on the queried locus across all libraries. The gray, orange, and blue colors represent 18–28-nt sRNAs, sense, and antisense sRNAs, respectively. E, The table shows the TPM levels of 18–28-nt, sense, and antisense sRNAs generated from that locus in each library.

(TAS), and others. The files of raw, trimmed, and mapped reads can be downloaded via a single click and fed into other analysis tools that the users prefer (Fig. 4A). For characterizing different classes of sRNAs, the pie diagram shows their corresponding percentage, and the line diagrams describe the size distributions of genome-matched total and distinct sRNAs (Fig. 4B). For a quick glance at each library, ASRD also provides the read counts and TPM levels of sRNAs from each of the miRNA, TAS, and "Top-100" PC, Pol IV, and TE loci ranked by TPM (Fig. 4C). The "Advanced options" can also be used to filter the search results by the TPM levels of sRNAs produced from different classes of loci.

Search by Gene/TE ID

ASRD supports searches by any of the 38,621 gene/TE IDs annotated in the most recent Araport11 release (Cheng et al., 2017; Fig. 4A). To give an overview of sRNAs generated from a given locus, the "Data Plot" page shows the TPM levels of 18–28 nt sRNAs across all libraries (Fig. 4D), and the "Data table" page details the size distribution, sense, and antisense abundances in each library (Fig. 4E). This table can be downloaded for local use.

IGV Visualization

ASRD integrates an online IGV interface to browse genome-matched sRNAs in any number of libraries. Besides the links on the "Information" pages, the users can use several input boxes and a submit button to easily explore sRNAs at any genomic region in one or more libraries. We extended the IGV function to allow the users to submit a single library ID; genotype-, ecotype-, or tissue-related keyword; SRA study ID (to simultaneously select multiple libraries included in the same study); gene ID, gene symbol, or alias; TAS family members' name and ID; and miRNA name.

DISCUSSION

Many excellent web-based resources have been developed for hosting sRNA data, such as the MPSS Web site (Nakano et al., 2006), the UCSC Genome Browser (Kent et al., 2002), the Anno-J Browser (Lister et al., 2008), and the EPIC-CoGe Browser (Nelson et al., 2018). Compared to these existing resources that are mostly designed for a single project or multiple projects, ASRD can easily and quickly extract the information from more than 2,000 Arabidopsis sRNA-seq libraries using a simple "Google-like" search and enables the direct comparison of sRNA abundances across different libraries by reprocessing all the raw data from scratch. In the future, we plan to update ASRD on a regular basis by adding more recently published sRNA-seq libraries, and we also intend to experiment with new functions that allow the users to upload their own sRNA-seq libraries and host them at ASRD.

MATERIALS AND METHODS

Data Collection, Processing, and Analysis

We collected Arabidopsis (*Arabidopsis thaliana*) sRNA-seq data published until July 2019 from GEO and SRA databases by searching with the following combinations of keywords: "([sRNA] OR [sRNAs] OR siRNA OR smallRNA OR smallRNAs OR miRNA OR sRNA OR sRNAs OR siRNAs OR miRNAs) and Arabidopsis." We obtained a total of 2,024 nonredundant libraries from the NGS platform (Illumina) with raw sequencing data. Figure 1 describes the data collection, processing, and database construction pipeline. The raw datasets in SRA format were downloaded, processed, and analyzed by in-house scripts (all of our scripts are available upon request). In brief, we used the fastq-dump from the SRA Toolkit (v2.8.2; https://www.ncbi.nlm.nih.gov/books/NBK158900/) to convert raw data from sra to fastq format; if a 3′ adapter sequence was not provided, we would predict and trim it with the software tools DNApi (Tsuji and Weng, 2016) and Cutadapt v1.16 (Martin, 2011); if 5′ barcoding was used, we would also chopped the 5′ barcode sequence off; we then processed the remaining 18–28 nt reads in the fasta file to tag_count format. To annotate sRNA features, we mapped these reads to the Arabidopsis reference genome (The Arabidopsis Information Resource 10) using the program BowTie v1.2.1.1 (Langmead et al., 2009), allowing zero mismatches (−v 0) and multiple hits (−a). We used Araport11-annotated t/r/sn/snoRNAs to flag corresponding types of sRNAs in each bam file. After importing all sRNAs, our database contains 2,357,941,025 genome-matched sRNAs representing 254,678,199 distinct sRNAs.

Analysis of sRNA Abundance

The annotation of 426 mature Arabidopsis miRNAs, including both miRNA and miRNA* sequences (named as miRNA-5p and miRNA-3p), was obtained from the program miRbase (v22.1; Kozomara et al., 2019). Eight TAS loci (*TAS1a*, *TAS1b*, *TAS1c*, *TAS2*, *TAS3*, *TAS3b*, *TAS3c*, and *TAS4*) were used for calculating TAS abundance. The annotations of 38,621 genes were from Araport11. The list of 7,632 P4-siRNA loci was the same as described in Zhai et al. (2015). The 27,655 PC genes and 3,901 TEs annotated in Araport11 were used to calculate the abundance of PC gene-generating siRNAs and TE-generating siRNAs, respectively. The abundance of sRNAs in each library was calculated as TPM by normalizing to the total number of genome-matched reads excluding t/r/sn/snoRNA-derived ones. The TPM of sRNAs at a given locus was the sum of genome hits-normalized TPMs for all mapped reads at that locus.

Supplemental Data

The following supplemental materials are available.

Supplemental Table S1. The information of 2,024 publicly available libraries in ASRD.

ACKNOWLEDGMENTS

We thank all the research groups that contributed sRNA-seq data to the public domain, and we apologize for not being able to cite all the related articles in the main text due to limited space. References for all libraries that we used are listed in Supplemental Table S1.

Received August 2, 2019; accepted December 2, 2019; published December 16, 2019.

LITERATURE CITED

Allen E, Xie Z, Gustafson AM, Carrington JC (2005) microRNA-directed phasing during trans-acting siRNA biogenesis in plants. Cell **121:** 207–221

Axtell MJ (2013) Classification and comparison of small RNAs from plants. Annu Rev Plant Biol **64:** 137–159

Axtell MJ, Jan C, Rajagopalan R, Bartel DP (2006) A two-hit trigger for siRNA biogenesis in plants. Cell **127:** 565–577

Baulcombe D (2004) RNA silencing in plants. Nature **431:** 356–363

Bologna NG, Voinnet O (2014) The diversity, biogenesis, and activities of endogenous silencing small RNAs in Arabidopsis. Annu Rev Plant Biol **65:** 473–503

Borges F, Martienssen RA (2015) The expanding world of small RNAs in plants. Nat Rev Mol Cell Biol **16:** 727–741

Carthew RW, Sontheimer EJ (2009) Origins and mechanisms of miRNAs and siRNAs. Cell **136:** 642–655

Chen X (2009) Small RNAs and their roles in plant development. Annu Rev Cell Dev Biol **25:** 21–44

Cheng CY, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD (2017) Araport11: A complete reannotation of the *Arabidopsis thaliana* reference genome. Plant J **89:** 789–804

Clough E, Barrett T (2016) The Gene Expression Omnibus Database. Methods Mol Biol **1418:** 93–110

Cuerda-Gil D, Slotkin RK (2016) Non-canonical RNA-directed DNA methylation. Nat Plants **2:** 16163

Fei Q, Xia R, Meyers BC (2013) Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. Plant Cell **25:** 2400–2415

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome Res **12:** 996–1006

Kozomara A, Birgaoanu M, Griffiths-Jones S (2019) miRBase: From microRNA sequences to function. Nucleic Acids Res **47**(D1): D155–D162

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol **10:** R25

Leinonen R, Sugawara H, Shumway M; International Nucleotide Sequence Database Collaboration (2011) The sequence read archive. Nucleic Acids Res **39:** D19–D21

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell **133:** 523–536

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J **17:** 10–12

Matzke MA, Mosher RA (2014) RNA-directed DNA methylation: An epigenetic pathway of increasing complexity. Nat Rev Genet **15:** 394–408

Meyers BC, Axtell MJ (2019) MicroRNAs in plants: Key findings from the early years. Plant Cell **31:** 1206–1207

Nakano M, Nobuta K, Vemaraju K, Tej SS, Skogen JW, Meyers BC (2006) Plant MPSS databases: Signature-based transcriptional resources for analyses of mRNA and small RNA. Nucleic Acids Res **34:** D731–D735

Nelson ADL, Haug-Baltzell AK, Davey S, Gregory BD, Lyons E (2018) EPIC-CoGe: Managing and analyzing genomic data. Bioinformatics **34:** 2651–2653

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer. Nat Biotechnol **29:** 24–26

Ruiz-Ferrer V, Voinnet O (2009) Roles of plant small RNAs in biotic stress responses. Annu Rev Plant Biol **60:** 485–510

Tsuji J, Weng Z (2016) DNApi: A De Novo Adapter Prediction Algorithm for small RNA sequencing data. PLoS One **11:** e0164228

Voinnet O (2009) Origin, biogenesis, and activity of plant microRNAs. Cell **136:** 669–687

Zhai J, Bischof S, Wang H, Feng S, Lee TF, Teng C, Chen X, Park SY, Liu L, Gallego-Bartolome J, et al (2015) A one precursor one siRNA model for Pol IV-dependent siRNA biogenesis. Cell **163:** 445–455