Breakthrough Technologies

Next-Generation Sequence Databases: RNA and Genomic Informatics Resources for Plants[1][OPEN]

Mayumi Nakano,[a,b] Kevin McCormick,[b,2] Caghan Demirci,[b,3] Feray Demirci,[a,b,4] Sai Guna Ranjan Gurazada,[a,b] Deepti Ramachandruni,[a,b] Ayush Dusia,[b,c] Joshua A. Rothhaupt,[a,b] and Blake C. Meyers[a,b,d,5,6]

[a]Donald Danforth Plant Science Center, St. Louis, Missouri 63132
[b]Delaware Biotechnology Institute, University of Delaware, Newark, Delaware 19711
[c]Department of Computer and Information Sciences, University of Delaware, Newark, Delaware 19716
[d]University of Missouri, Division of Plant Sciences, Columbia, Missouri 65211

ORCID IDs: 0000-0001-9317-4509 (M.N.); 0000-0002-3084-5683 (S.G.R.G.); 0000-0003-1310-8523 (A.D.); 0000-0003-3436-6097 (B.C.M.).

We developed public web sites and resources for data access, display, and analysis of plant small RNAs. These web sites are interconnected with related data types. The current generation of these informatics tools was developed for Illumina data, evolving over more than 15 years of improvements. Our online databases have customized web interfaces to uniquely handle and display RNA-derived data from diverse plant species, ranging from Arabidopsis (*Arabidopsis thaliana*) to wheat (*Triticum* spp.), including many crop and model species. The web interface displays the abundance and genomic context of data for small RNAs, parallel analysis of RNA ends/degradome reads, RNA sequencing, and even chromatin immunoprecipitation sequencing data; it also provides information about potentially novel transcripts (antisense transcripts, alternative splice isoforms, and regulatory intergenic transcripts). Numerous options are included for downloading data as tables or via web services. Interpretation of these data is facilitated by the inclusion of extensive repeat or transposon data in our genome viewer. We have developed graphical and analytical tools, including a new viewer and a query page for the analysis of phased small RNAs; these are particularly useful for understanding the complex small RNA pathways of plants. These public databases are accessible at https://mpss.danforthcenter.org.

Plant small RNAs and their RNA targets function in diverse pathways ranging from development to stress responses and from posttranscriptional control to genome defense. In parallel to advances in our understanding of these RNA activities, sequencing technologies have improved, yielding exponential increases in small RNA data from thousands of libraries. These data are often mapped onto sequenced plant genomes, which are also growing in number. For more than 15 years, our group has developed public web resources for the analysis of plant gene expression data. The purpose of these resources has been to facilitate gene- and genome-based analysis of RNA data, allowing users to browse, analyze, and visualize results for loci of interest. Our experience in this field started with the first next-generation sequencing technology called massively parallel signature sequencing (MPSS; Brenner et al., 2000; Meyers et al., 2004c) and continued with the development of small RNA sequencing (Lu et al., 2005). The sequencing technologies have substantially advanced, as have our databases, web sites, visualization tools, and applications for transcriptional and chromatin analyses.

The original prototype of our web-based database for next-generation sequence data was developed in 2003, when we launched our first public web site for a model plant, Arabidopsis (*Arabidopsis thaliana*), based on mRNA MPSS data (Meyers et al., 2004a, 2004b; Nakano et al., 2006). Although that first generation of our web resources was already equipped with many of the major analysis tools we offer today, in the last 13+ years, all our tools have gone through a series of overhauls and improvements to adapt to ever-increasing sizes and types of data sets. On the back end, our database structures have been continuously reexamined and

[1] This work was supported by numerous grants, primarily from the National Science Foundation (award nos. IOS 1649424, IOS 1754097, and IOS 1842698).

[2] Present address: University of Pennsylvania, 424 Johnson Pavilion, 3610 Hamilton Walk, Philadelphia, PA 19104.

[3] Present address: Microsoft, One Microsoft Way, Redmond, WA 98052.

[4] Present address: FiDoSoft Software Consulting, Redmond, WA 98052.

[5] Author for contact: bmeyers@danforthcenter.org.

[6] 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: Blake C. Meyers (bmeyers@danforthcenter.org).

M.N., K.M., C.D., F.D., and S.G.R.G. designed and wrote code and developed tools and applications; D.R., F.D., and A.D. built databases and processed data; J.A.R. provided technical assistance; M.N. and B.C.M. wrote the article with major contributions by K.M., C.D., and S.G.R.G.; B.C.M. conceived the project and supervised and completed the writing.

[OPEN] Articles can be viewed without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.19.00957

reorganized to store and process ever-growing amounts of data more efficiently and more rapidly. The web interface has been completely rewritten or substantially modified several times to improve performance, to accommodate changes made to the database structure, and to introduce more advanced features for analysis and display of the data. These improvements were necessary as we expanded our resources by adding more plant species (and even organisms from other kingdoms) and more data types while adapting to changing sequencing technologies. For historical reasons, we retained the MPSS name of our sites, although the data we display are almost exclusively from Illumina's sequencing-by-synthesis technology.

The data on our web sites include small RNA, parallel analysis of RNA ends (PARE)/degradome, RNA sequencing, and chromatin immunoprecipitation sequencing. As of September 2019, we have launched 70 public web sites featuring 31 different organisms (see "Genomic Databases" below for the complete listing). Our web pages are unique because they provide extensive information about individual reads or RNAs, with those data embedded in the images and easily accessible. Taking small RNAs as a primary focus of our site, the accessibility of individual reads (representing a single, distinct small RNA) is important because plants have a number of biogenesis pathways that are absent from animals. For example, heterochromatic small interfering RNAs (siRNAs) functioning in RNA-directed DNA methylation and phased, trans-acting siRNAs are plant specific. This complex set of small RNAs is best analyzed using information-rich visualization tools.

Here, we describe our process of database building/loading and our web interfaces, focusing on the developments we have made for analyses of various types of sequencing-by-synthesis data from different organisms since our last web site-centric publication (Nakano et al., 2006). Because we have developed a pipeline to process genomic and expression data and a common set of web scripts to display the processed data, only minimum customization is now made to our databases and interfaces when we launch a new web site. This report also covers our new phasing analysis viewer and tool for making rapid visual assessments of small RNAs, because identification of phased small RNAs is an important aspect of work in our lab. All our public resources are accessible at https://mpss.danforthcenter.org.

OVERVIEW OF OUR DATABASES

Each of our web sites is connected to one of each of two types of databases, which we call the genomic database and the expression database. For each type of database, we developed a standardized schema and an automated building process to perform all of the necessary steps of data loading. The pipeline for each data-loading function is a combination of Perl, Python, and MySQL scripts that takes data files as input and uses them to construct databases from which data are later extracted to build web pages. We have continuously revamped the method used to load raw sequences into our expression databases, allowing large libraries to be added quickly. We now use Bowtie, an extremely fast program capable of mapping tens of millions of reads in just a few minutes. We continue to use databases because their relational structure facilitates complex queries, while developments to maintain their speed and capacity over the last 15 years have kept up with the growth in the size of our tables and data sets.

Genomic Databases

Our public web sites are all based on sequenced genomes for which we can acquire the assembled and annotated sequences from public sources. Thus, as a research group, we are a consumer of genomic data rather than a generator. Genomes that we maintain for our sites include Arabidopsis, rice (*Oryza sativa* and *Oryza glaberrima*), maize (*Zea mays*), *Medicago truncatula*, soybean (*Glycine max*), wheat (*Triticum aestivum*), *Aegilops tauschii*, asparagus (*Asparagus officinalis*), common bean (*Phaseolus vulgaris*), litchi (*Litchi chinensis*), strawberry (*Fragaria ananassa*), apple (*Malus domestica*), cassava (*Manihot esculenta*), grape (*Vitis vinifera*), lettuce (*Lactuca sativa*), orange (*Citrus sinensis*), papaya (*Carica papaya*), peach (*Prunus persica*), potato (*Solanum tuberosum*), tomato (*Solanum lycopersicum*), *Amborella trichopoda*, *Mimulus guttatus*, petunia (*Petunia axillaris* and *Petunia inflata*), poplar (*Populus trichocarpa*), *Brachypodium distachyon*, *Spirodela polyrhiza*, *Ustilago maydis*, *Magnaporthe grisea* (the rice blast fungus), and chicken (*Gallus gallus*). We have a separate genomic database for each organism, which contains data obtained from its genomic sequence and annotation. Any analysis that requires chromosomal, genic, or related information of that organism must consult the genomic database. No expression data (libraries and read sequences) are stored in genomic databases, and thus the genome databases are utilized across many sets of transcriptional or related data. Each genomic database contains pieces of information such as chromosomes, genes and noncoding RNAs, k-mer frequencies, and unannotated repeats.

Expression Data Preprocessing

Our expression database-building pipeline can be run to load large libraries into databases, but in efforts to standardize the process, we require that a short preprocessing protocol is followed prior to running the actual pipeline (Fig. 1). This phase is called data preprocessing, and it consists of three steps: parsing/trimming, filtering/chopping/alignment, and read mapping. The basic function of this preprocessing protocol is to convert various raw file formats into a standard format, trim off any ligated library or sequencing adapters, and store a copy of

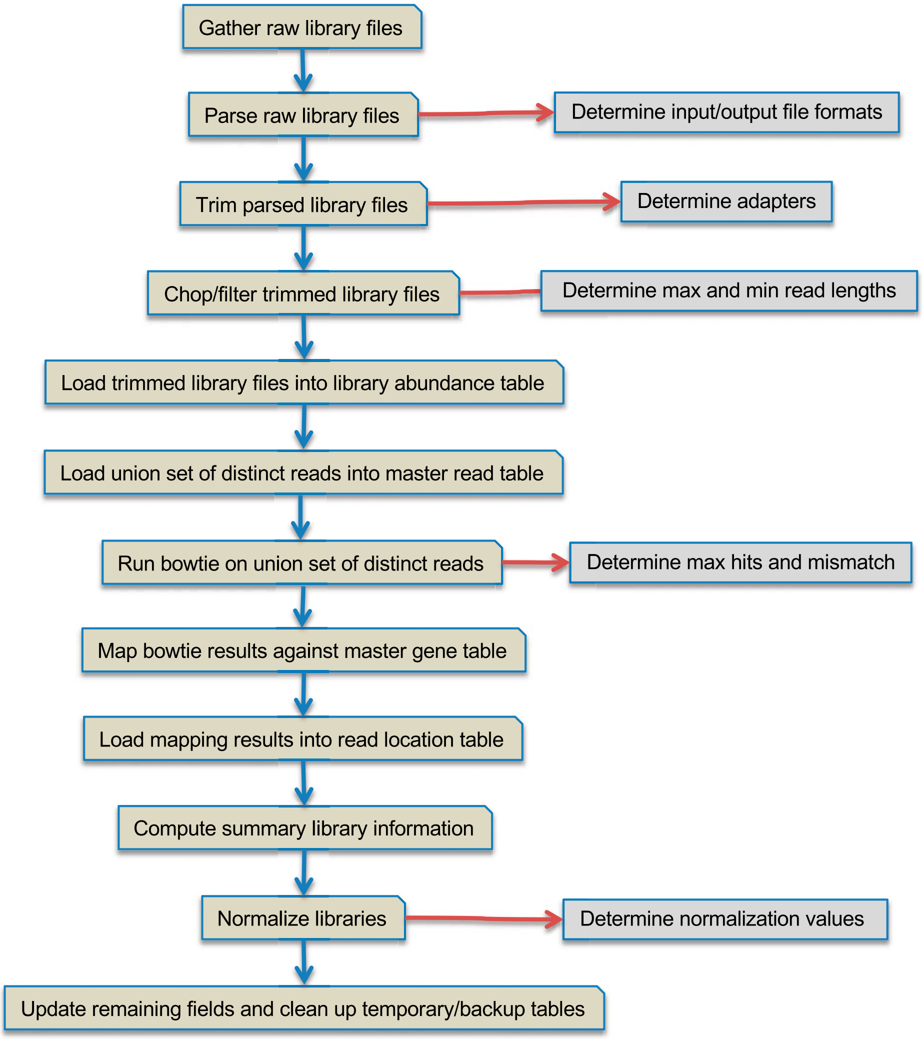

Gather raw library files
Parse raw library files
Determine input/output file formats
Trim parsed library files
Determine adapters
Chop/filter trimmed library files
Determine max and min read lengths
Load trimmed library files into library abundance table
Load union set of distinct reads into master read table
Run bowtie on union set of distinct reads
Determine max hits and mismatch
Map bowtie results against master gene table
Load mapping results into read location table
Compute summary library information
Normalize libraries
Determine normalization values
Update remaining fields and clean up temporary/backup tables

Figure 1. Expression data preprocessing pipeline.

the expression data (as flat files) in a central repository, which is then used as the standard input to the database-building scripts. But with time, as the number of available libraries increased, before loading them into databases, a preliminarily analysis to assess the quality of the libraries became a pivotal and integral part of this phase too. The preliminary analysis yields graphs of the size distribution of the sequenced molecules (particularly useful to assess small RNA and PARE data) and the percentage of reads with genome matches (useful to assess data quality).

Expression Databases

Expression databases store quantitative RNA information for various tissue samples. They connect to our web sites to provide visualization tools for the expression data, and they also allow our researchers to run complex MySQL queries directly via the web interface. We build an expression database for each set of libraries we want to analyze, and the main tables in the database record pieces of information such as library abundances, genomic locations, and library summaries.

WEB INTERFACES: DATA ACCESSIBILITY

After each expression database is built, its data, along with those from its related genomic database, can be accessed and viewed via a series of dynamic web pages. A majority of our web scripts are written in PHP (an open-source general-purpose programming language widely used for web development) embedded into HTML with JavaScript and Cascading Style Sheet, and the graphical outputs included in query results are generated on the fly by the GD library (an image-drawing tool used by PHP) or HTML5 Canvas (an HTML element that contains images drawn via JavaScript). It is the latter version of those graphical outputs, which we call Jiffy View, that we have added for faster page loading (as it excludes slower-to-load mouse events or links) and more flexible viewing options (e.g. changing the bp per pixel, image scaling, etc., which require redrawing the image). We have kept the original version of the web pages, which we call Interactive View, as they allow the user to mouse over or click every element to see more information or go to the related detailed pages for analysis of individual elements like single RNAs. Those two versions (Jiffy and Interactive) are interchangeable and interlinked on our pages as embedded genome viewers. Another improvement we have made to offer customized viewing of the data and images is a Control Panel (CP), which is a utility allowing the user to select and adjust libraries and/or sequences to display in order to facilitate analyses, and some of the CP options are mentioned in the following sections that briefly explain how to use our web sites.

Quick Start: Basic Queries and Chromosomal Navigation

The home page of each Next-Gen Sequence DB Web site is a starting point for both experienced and inexperienced users, since most of our major analysis tools and resources are accessible from this page (Fig. 2). For those who are new to our web-based databases, various example links on this page provide quick access. For instance, the home page of our Arabidopsis Small RNA Database lists example links to our image viewer output pages for known microRNAs (miRNAs), inverted/tandem repeats, pericentromeric regions, weakly predicted transposons, and trans-acting siRNAs (Fig. 2A).

For those who are ready to run analyses, the home page offers input boxes for a number of Basic Queries based on a protein-entry code for the Gene Analysis tool (GA; for an analysis of a specific gene or intergenic region) or the Library Abundances tool (LA; for an analysis of individual libraries and their abundance data in a specific region) with single or multiple gene identifiers, a read sequence for the Signature Analysis tool (for an analysis of a specific read) with single or multiple sequences, or a keyword used for predicted protein function or repeat classes for the Keyword Search tool to list a group of genes or repeats in a certain category (Fig. 2B). Each query's input section also includes an example input linked to its output page, so the user can simply click on the link to preview the related tool.

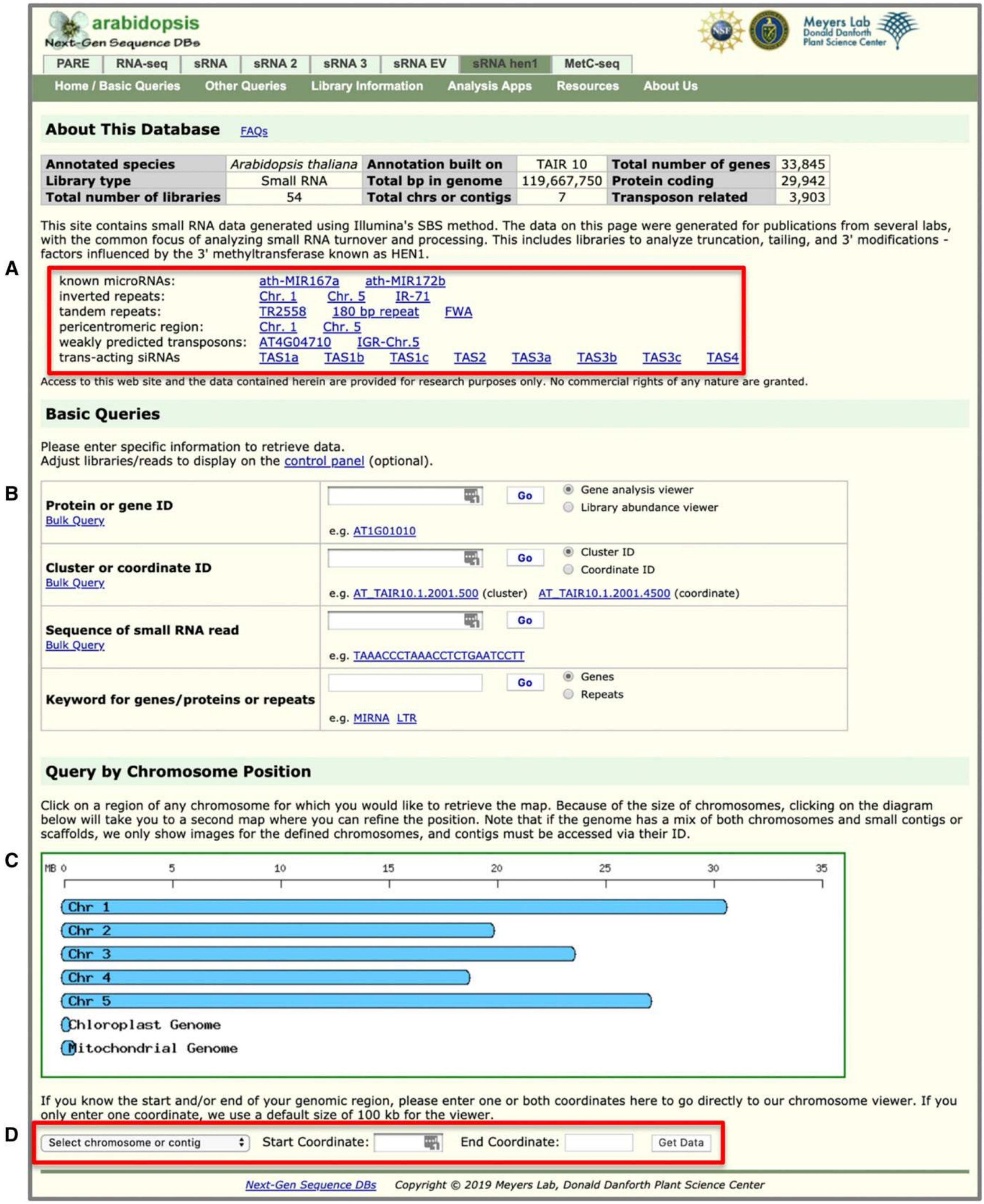
The home page also offers two options for the Query by Chromosome Position. The first option is by clicking on the image map of the featured genome (Fig. 2C). Each chromosome bar appearing on this map is divided into a series of hidden rectangles, or hot spots, which represent different chromosomal regions (the coordinates pop up on a mouse over), and these regions are linked to a new page containing the zoomed-in map of the corresponding locus. We call this linked page the level 2 (or intermediate) version of Chromosome Viewer (CV), as it provides a second, general view of the selected region. This page too consists of image elements linked to the final version of CV, and thus it can be used to pinpoint the selected region more precisely and further zoom into that region. The other option, if the user knows exactly which chromosomal region to look up, is to select the chromosome from the drop-down list and enter the start and/or end coordinates in the input boxes just below the image map (Fig. 2D). This will take the user directly to the final CV page, bypassing the intermediate version.

Image Viewers: Visualizing Genomic Information and Sequences

We have four different levels at which to visualize the featured genome, starting from the image map for all the chromosomes on the home page, moving to intermediate and then final versions of CV to narrow the view to a specific chromosomal region, and ending with the viewer on the GA or LA page to focus on each specific gene or region (Fig. 3). The last two levels are what we mainly refer to as viewers because of their embedded genome browsers. The legend that helps the user to understand each image is linked from those pages displaying the viewers.

Although the CV (final version) and the GA/LA viewers are based on different scales and sizes, both viewers share common elements and display options, some of which can be customized via the CP linked from their host pages. They display genes represented as a series of colored rectangles that include exons, untranslated regions, and/or introns. The color of each gene is usually determined by the strand to which it belongs (i.e. red for top and blue for bottom), but the tRNA/rRNA/small nuclear RNA/small nucleolar RNA genes are colored differently. Another displayed genomic element is the repeat data (retrotransposon/transposon related, tandem, inverted, etc.), repetitive sequences represented as boxes of different, mainly pastel, colors and heights in the viewers. We have identified those generally unannotated repeats by using such programs as RepeatMasker, einverted, and etandem/Tandem Repeats Finder; we use a low stringency to identify low-homology repeats to ensure maximal identification of repeats. The repeat data may be particularly useful for interpreting small RNA data, since all types of these repeats are rich sources of small RNAs in plant genomes. One other genomic element displayed in the viewers is the k-mer line plot represented as a squared-off, purple line graph. This represents the average degree of repetitiveness in the genome determined by a 20-nucleotide window of the moving average of 20-nucleotide fragments spanning the genome; the k-mer line plot is shown by default, but it can be hidden by changing the option in the CP.

Our viewers also display sequences of the featured data types (small RNA, PARE, RNA sequencing, chromatin immunoprecipitation, etc.) that are associated with the region and have been sequenced in any of the libraries. Sequences can be displayed with their individual abundances represented as small dots or with the sum of abundances represented as bars. This display option can be interchanged via the CP, or a third view showing the sum of hits-normalized abundances, represented by bars. Each dot for individual abundances



About This Database

FAQs

Annotated species	*Arabidopsis thaliana*	**Annotation built on**	TAIR 10	**Total number of genes**	33,845
Library type	Small RNA	**Total bp in genome**	119,667,750	**Protein coding**	29,942
Total number of libraries	54	**Total chrs or contigs**	7	**Transposon related**	3,903

This site contains small RNA data generated using Illumina's SBS method. The data on this page were generated for publications from several labs, with the common focus of analyzing small RNA turnover and processing. This includes libraries to analyze truncation, tailing, and 3' modifications - factors influenced by the 3' methyltransferase known as HEN1.

known microRNAs:	ath-MIR167a	ath-MIR172b						
inverted repeats:	Chr. 1	Chr. 5	IR-71					
tandem repeats:	TR2558	180 bp repeat	FWA					
pericentromeric region:	Chr. 1	Chr. 5						
weakly predicted transposons:	AT4G04710	IGR-Chr.5						
trans-acting siRNAs	TAS1a	TAS1b	TAS1c	TAS2	TAS3a	TAS3b	TAS3c	TAS4

Access to this web site and the data contained herein are provided for research purposes only. No commercial rights of any nature are granted.

Basic Queries

Please enter specific information to retrieve data.
Adjust libraries/reads to display on the control panel (optional).

Protein or gene ID
Bulk Query
● Gene analysis viewer
○ Library abundance viewer
e.g. AT1G01010

Cluster or coordinate ID
Bulk Query
● Cluster ID
○ Coordinate ID
e.g. AT_TAIR10.1.2001.500 (cluster) AT_TAIR10.1.2001.4500 (coordinate)

Sequence of small RNA read
Bulk Query
e.g. TAAACCCTAAACCTCTGAATCCTT

Keyword for genes/proteins or repeats
● Genes
○ Repeats
e.g. MIRNA LTR

Query by Chromosome Position

Click on a region of any chromosome for which you would like to retrieve the map. Because of the size of chromosomes, clicking on the diagram below will take you to a second map where you can refine the position. Note that if the genome has a mix of both chromosomes and small contigs or scaffolds, we only show images for the defined chromosomes, and contigs must be accessed via their ID.

If you know the start and/or end of your genomic region, please enter one or both coordinates here to go directly to our chromosome viewer. If you only enter one coordinate, we use a default size of 100 kb for the viewer.

Select chromosome or contig Start Coordinate: End Coordinate:

Next-Gen Sequence DBs Copyright © 2019 Meyers Lab, Donald Danforth Plant Science Center

Figure 2. The home page of one of our Arabidopsis small RNA web sites. A, Example links provide quick access to various outputs of our image viewer. B, The Basic Query section allows the user to start most of our major analyses. C, The user can click on the image map linked to the intermediate version of the CV. D, The user can also enter a specific genomic location to access the final version of the CV.

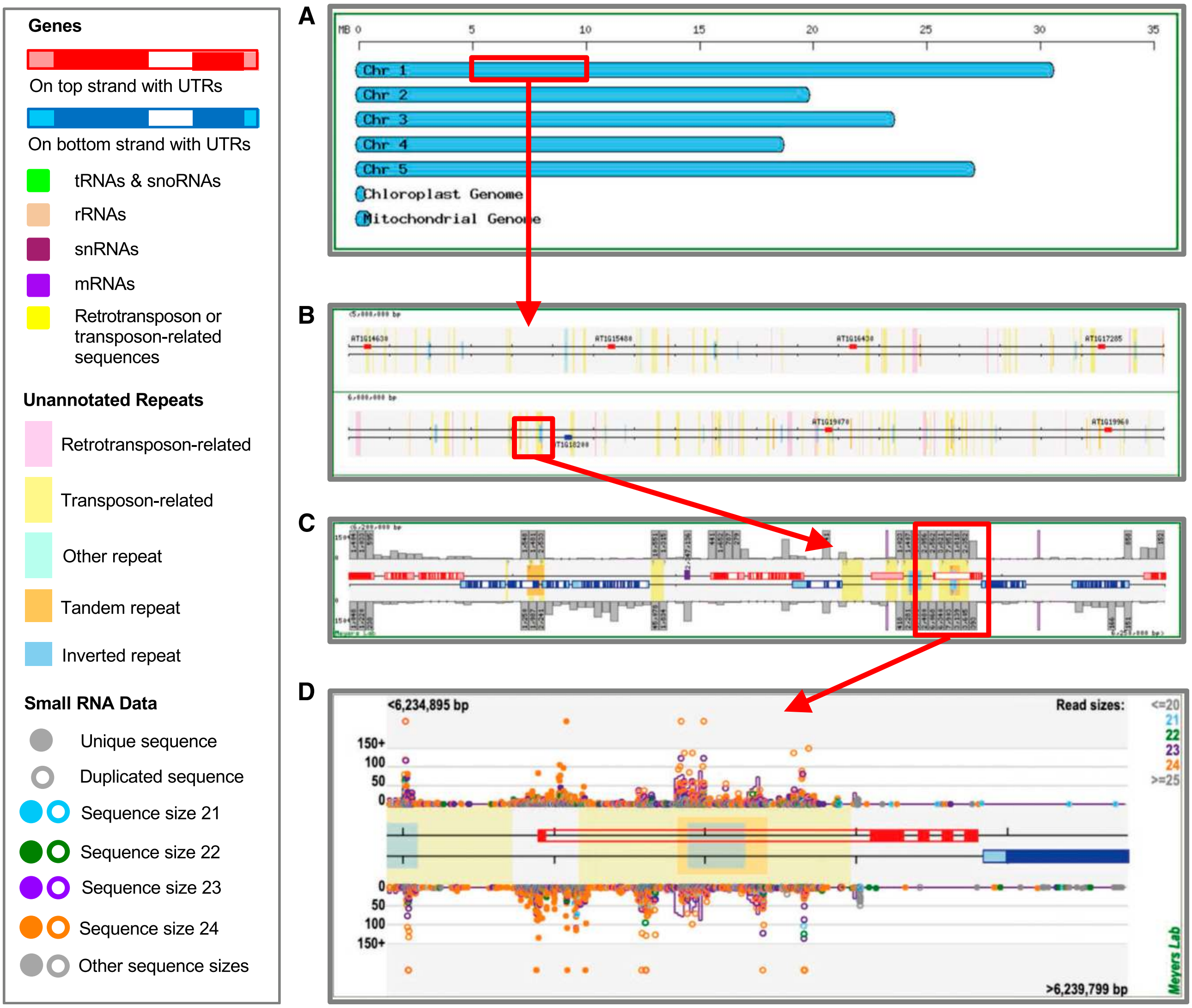


Figure 3. Four different levels of viewing the featured genome. A, The image map for all the Arabidopsis chromosomes on the home page. The red box indicates a clickable link to the intermediate version of the CV for the selected region. B, The intermediate version of the CV. The red box indicates a clickable link to the final version of the CV for the selected region. C, The final version of the CV, which displays genes on each strand as a series of red or blue narrow rectangles, repeat data (shadow boxes of different colors and heights), the k-mer line plot as a purple line graph, and the sum of sequence abundances as bars. The red box indicates a clickable link to the GA tool for the selected gene. D, The viewer on the GA or LA page for each specific gene or region. In this viewer, individual small RNA sequences are displayed as small dots in different colors according to their sizes (unique sequences as filled dots and duplicated sequences as hollow dots). snoRNAs, Small nucleolar RNAs; snRNAs, small nuclear RNAs; UTRs, untranslated regions.

in the Interactive View of the final CV and the GA/LA viewer is linked to the Signature Analysis page for that particular sequence (the dot is filled for a unique sequence and hollow for a duplicated one). For small RNA sequences, the user also has options to display individual abundances in different colors according to their sizes and to display only some of the sizes (Fig. 3D).

JavaScript Object Notation Web Services

JavaScript Object Notation (JSON) is a data-interchange format representing data in an easy-to-access way, and it has become a preferred choice for data interchange standard today. As a text-based system using name-value pairs, JSON is lightweight and compatible with most modern programming languages. It is easy not only for humans to read and write but also for computers to parse and generate. We have been offering the user to download query results from our analysis pages as JSON in addition to a comma-separated values file, and we recently set up a new JSON web services page where the user can choose from a wide range of data (expression, genomic, or metadata), send a query, and then receive the requested data. This facilitates data accessibility. The resulting data can be viewed as a web page, and it also can be saved as a file that the user can later edit, convert, decode, or parse for their own analyses. A variety of JSON editors/converters are available, and so are JSON decoders/parsers for different programming languages.

SMALL RNA PHASING ANALYSIS

Heterochromatic siRNAs, miRNAs, and small RNA decay products are generated in a stochastic manner from their precursors. However, there are quite a few locations where the small RNAs have a very precise spacing (i.e. phasing) of 21 or 24 nucleotides, exemplified by the 21-nucleotide Arabidopsis trans-acting siRNAs (Vazquez et al., 2004). This pattern describes secondary siRNAs generated by miRNA triggers in a phased pattern (Fei et al., 2013). In many angiosperms, particularly including the grasses, there are numerous loci at which the small RNAs are spaced by exactly 24 nucleotides (Zhai et al., 2015; Xia et al., 2019). In the phasing analysis, the web site analyzes all the small RNA sequences existing in a given region by making each sequence as the start of a window of 10 cycles of the selected phasing length to see if sequences in that window are spaced by that length. We call those sequences spaced precisely in-phase and others out-of-phase. Based on the numbers of these sequences, the phasing score of each window is calculated, somewhat similar to a previously described algorithm (De Paoli et al., 2009). We have made this analysis available at our public web sites. The following sections describe how we obtain and display phasing analysis data on our web pages.

List of Qualified Small RNAs

First, we obtain a list of small RNA sequences, which can belong to all the possible phasing windows existing between the start and end coordinates specified by the user, from an expression database we have built for a specific set of libraries (the user may select a subset of the libraries via the CP). We previously performed this analysis only on the top strand (sequences on the bottom strand were converted to the top strand with their positions offset by two nucleotides with the overhang at the 3′ end), but now the user may view the result on both strands (each window on its own strand: default), top or bottom strand (windows on the opposite strand converted to each strand), or top and bottom strands together. Thus, we keep two sets of coordinates for each sequence by converting the original coordinates to the opposite strand, and then one set of those strand-specific coordinates (original and/or converted) is used for the analysis on each strand. The formulas we use are: [original coordinate] − ([phasing length] − 3) for the bottom-to-top conversion and [original coordinate] + ([phasing length] − 3) for the top-to-bottom conversion.

The small RNA list must consist of not only those sequences for which each can be the start of a window on either of the strands in the region but also those sequences that can belong to any of the cycles in a window, especially when the window starts around the start or end coordinate. Thus, we adjust the start and end coordinates for the query by using the end of the possible last window on the bottom strand ([original start coordinate] − [phase length] × 10 + 1) as the smallest coordinate and the end of the possible last window on the top strand ([original end coordinate] + [phase length] × 10 – 1) as the biggest coordinate. Among those sequences that the query result includes for the adjusted coordinates, as empirically defined filters for Pol II-derived phased siRNAs (phasiRNAs), we exclude those small RNAs that hit to the featured genome more than 20 times, whose types are rRNAs, small nuclear RNAs, small nucleolar RNAs, or tRNAs, or whose sizes are not equal to the featured length. For each qualified small RNA, we record its start position, strand, W- and C-strand positions, sequence, number of genomic hits, and normalized abundance sum, and the final list is available as JSON data via our web services.

List of Phasing Windows

Next, from the set of prescreened small RNAs obtained above, we generate a list of phasing windows existing in the specified region for the featured strand (top or bottom). Depending on the user's selection of the strand(s) for viewing the analysis result, this process may be repeated for the opposite strand as well. By using the strand-specific coordinates, which may or may not have been converted from opposite-strand coordinates, we make each sequence from the small RNA list the start of a phasing window. Those sequences whose coordinates are outside the region are excluded. According to the orientation of each strand, we sort the small RNA list by the featured-strand coordinates in ascending order for the top strand and in descending order for the bottom strand. With each version of the sorted list, we move from one window to another, one cycle to another, in the same orientation.

For each window, all the small RNAs existing in the window are examined to assess whether they are in-phase. We consider a sequence in-phase if its position is equal to the start coordinate of one of the window's 10 cycles while allowing −1 and +1 mismatches. While going through sequences for the window, we collect the following values to calculate its phasing score: [A] the total abundance sum of in-phase sequences; [B] the total abundance sum of out-of-phase sequences; and [C] the number of cycles occupied by distinct, occupied phase positions. The formula we use to calculate the score is log(pow((1 + 10 × [A]/(1 + [B])), ([C] − 2))) (De Paoli et al., 2009), but the score should be 0 if the number of the occupied cycles is less than 3. During the score calculation, we also track the highest phasing-score position (HPSP) in the entire region so that we can later flag those windows that are in-phase with the HPSP (−1/0/+1 positional differences are allowed).

In addition to the phasing score and its related values mentioned above, each phasing window's entry in the final list consists of its start and end coordinates, phasing length, the original strand and original position of the sequence starting the window, whether the window is the HPSP, and whether it is exactly or almost in-phase with the HPSP. Like the small RNA list, this list can be obtained from our JSON web services, and a specific region's JSON data are directly linked from those pages displaying phasing analysis data.

Visualization of Phasing Windows

We have developed a new viewer customized to visualize our phasing analysis result, and it can be displayed along with our regular viewer on two of our most popular analysis pages: the GA and LA pages (Fig. 4A). The Phasing Analysis (PA) viewer looks very similar to the regular viewer at a glance, since they both consist of the same set of genomic elements. However, in the new viewer, each of the small dots (colored or gray and filled or hollow) represents a window of 10 cycles of the phasing length and the score of small RNAs positioned in-phase with this, instead of individual abundances represented in the regular viewer. By using this viewer, the user can see where miRNAs have triggered secondary siRNA production at their targets.

The PA viewer can be turned on for small RNA data (it is disabled by default) in the CP, the utility allowing the user to select and adjust libraries and/or sequences to display in order to facilitate analyses. The CP offers a few more display options for the user to customize this viewer. The user can adjust the strand(s) to show phasing windows. The default value is both strands (each window on its original strand), but this can be changed to either top or bottom strand only (the opposite strand window is converted to the featured strand) or top and bottom strands together (all windows on each strand). The user can also change each phasing cycle's length (between 19 and 25 nucleotides). The default value is 21, which is the length generally found in the model plant Arabidopsis, but other lengths may be more appropriate for other species. Finally, the user can optionally define a certain score (between 0 and 100) to highlight all positions exceeding that score.

By using the phasing window list(s) discussed above, the PA viewer draws, on each strand, windows whose strand-specific coordinates exist in the specified region. By default, W-strand windows appear on the top portion while C-strand windows appear on the bottom, but the user can select another option via the CP. The red-filled dot indicates the window that has the HPSP. Color-filled dots (the color is determined by the phasing length) indicate those windows that are in-phase with the HPSP (exactly in-phase ones as filled dots and almost in-phase ones as hollow dots), while gray hollow dots indicate those windows that are out of phase. If the user has defined a certain score via the CP, all windows exceeding that score will be colored, whether they are in-phase or not (the exactly in-phase windows are still drawn as filled dots).

The 20× image scale (magnification), which can be selected in the New Search section on the GA/LA page, provides the best view for phasing analysis, since each window can be separated from the others. The Interactive View of this image scale marks the HPSP along the red vertical line, and then it shows 10 in-phase positions to its left and right by lines of another color (Fig. 4B). The interactive view also shows each window's information (start and end coordinates, score, sum of in-phase or out-of-phase abundances, and number of occupied phase cycle positions) as a mouse-over popup text, and the window's image is linked to the page that shows this phasing window's in-depth analysis.

In-Depth Analysis of a Phasing Window

The Phasing Analysis Information (PA Info) page displays detailed information on each phasing window for the start coordinate and the strand specified by the user (or passed from the linking page; Fig. 5). Each phasing window consists of 10 cycles, and each cycle's length depends on the featured length (the user can change this via the CP). The orientation of the window and its cycles depends on the specified strand: for the W strand, the window moves from left to right with coordinates increasing from the specified start, while for the C strand, the window moves from right to left with coordinates decreasing from the specified start. The tabular data discussed below can be viewed and downloaded from the JSON links on the page.

The first table found on this page lists the positions of 10 cycles on both strands (for simplicity, the same order of cycle numbers is used for both strands; Fig. 5A). The user-specified start of the window is highlighted for the specified strand (it is on the far left for the W strand and on the far right for the C strand), and those cycles occupied by any in-phase sequence are linked to the corresponding rows in the sequence tables below.

The sequence table for each strand consists of small RNAs and their information such as number of genomic hits, location, and library-specific normalized abundance values (Fig. 5B). To filter qualified small RNA sequences, this table mostly uses the same criteria as the small RNA list described above, but sequences in a range of lengths are included here. In the JSON data linked from the top of the table, different-length sequences are included and flagged as in-phase as long as their positions are within −1/0/+1 differences. However, the sequence table highlights only those sequences in the featured length as exact or almost in-phase, and the PA viewer appearing at the bottom of this page considers only exact-length sequences.

The PA Info page is equipped with the regular viewer and its special version of the PA viewer (both interactive 20× image scale). In this PA viewer, which is

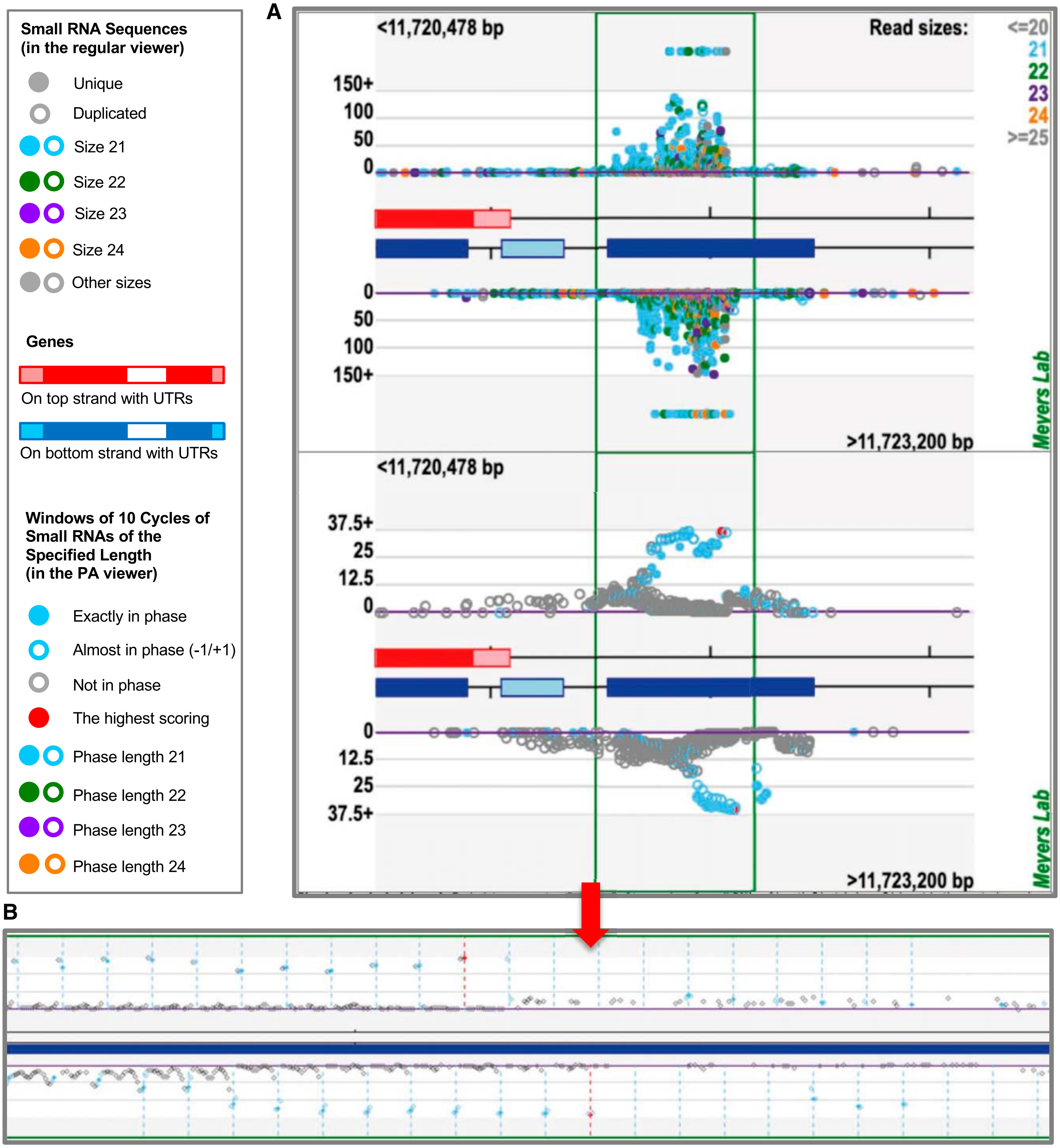
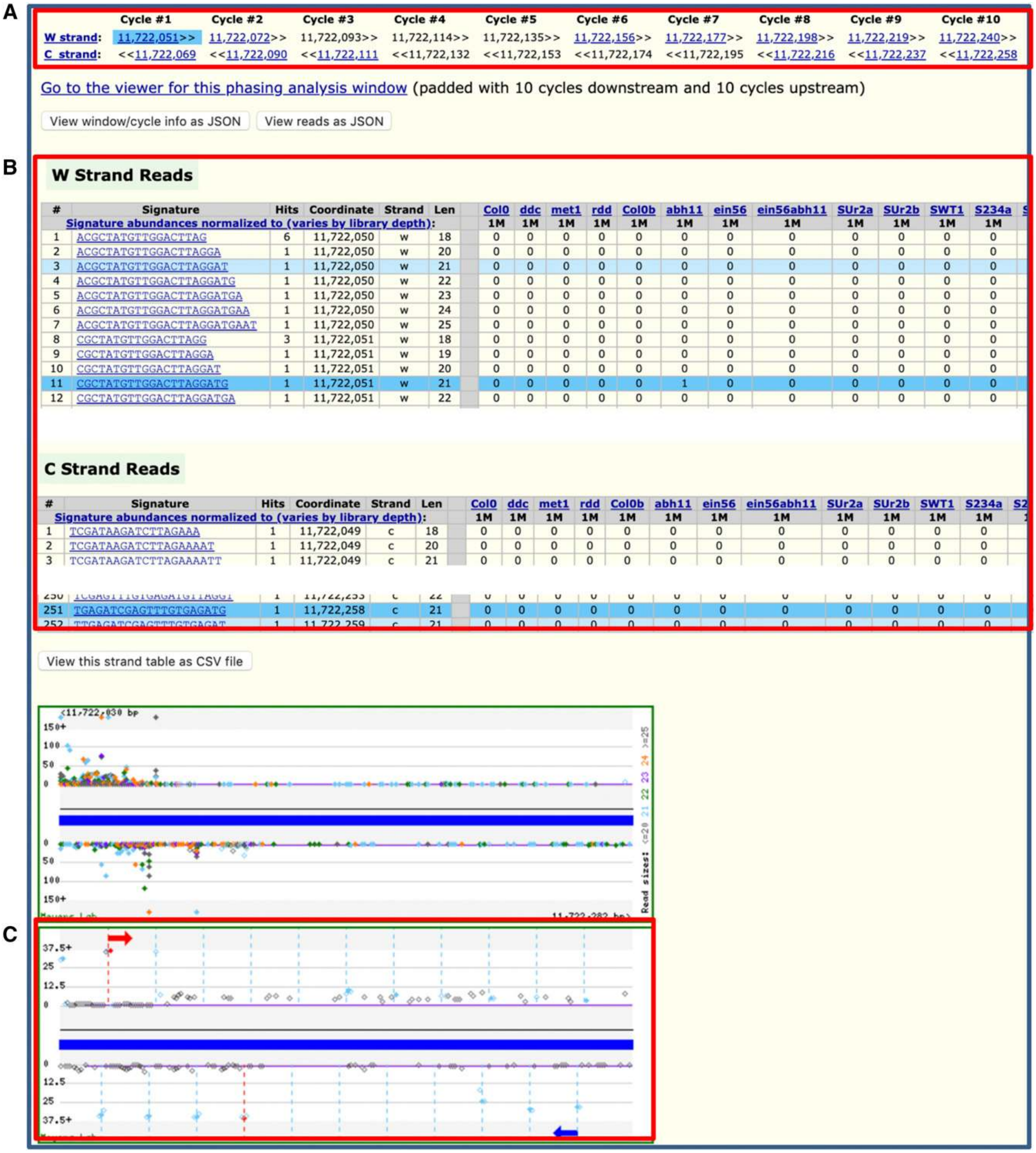


Figure 4. The customized PA viewer as an add-on to the small RNA browser. A, The PA viewer (bottom) displayed along with the regular viewer (top). In the bottom section, each dot represents a window of 10 cycles of small RNAs of the specified length in the top section (21 nucleotides in this example). The phasing score (*y* axis) indicates the degree of phasing. The red dot is the window with the best score in this region, and other colored dots are windows that are in-phase with it (perfectly phased windows as filled dots and those with −1/+1 mismatches as hollow dots). Gray hollow dots are out-of-phase windows. B, The Interactive View of the PA viewer from the bottom part of A in 20× image scale. UTRs, Untranslated regions.

independent from the user's selection of the strand to display windows via the CP, each qualified small RNA shows up as a PA window itself always on its original strand, and its score is based on those sequences whose strand-specific coordinates (opposite-strand coordinates get converted) belong to that window (with −1/+1 positional differences allowed; Fig. 5C). This viewer covers the selected window with an additional cycle



A

	Cycle #1	Cycle #2	Cycle #3	Cycle #4	Cycle #5	Cycle #6	Cycle #7	Cycle #8	Cycle #9	Cycle #10
W strand:	11,722,051>>	11,722,072>>	11,722,093>>	11,722,114>>	11,722,135>>	11,722,156>>	11,722,177>>	11,722,198>>	11,722,219>>	11,722,240>>
C strand:	<<11,722,069	<<11,722,090	<<11,722,111	<<11,722,132	<<11,722,153	<<11,722,174	<<11,722,195	<<11,722,216	<<11,722,237	<<11,722,258

Go to the viewer for this phasing analysis window (padded with 10 cycles downstream and 10 cycles upstream)

B

W Strand Reads

#	Signature	Hits	Coordinate	Strand	Len		Col0	ddc	met1	rdd	Col0b	abh11	ein56	ein56abh11	SUr2a	SUr2b	SWT1	S234a	S
	Signature abundances normalized to (varies by library depth):						1M	1M	1M	1M	1M	1M	1M	1M	1M	1M	1M	1M	
1	ACGCTATGTTGGACTTAG	6	11,722,050	w	18		0	0	0	0	0	0	0	0	0	0	0	0	
2	ACGCTATGTTGGACTTAGGA	1	11,722,050	w	20		0	0	0	0	0	0	0	0	0	0	0	0	
3	ACGCTATGTTGGACTTAGGAT	1	11,722,050	w	21		0	0	0	0	0	0	0	0	0	0	0	0	
4	ACGCTATGTTGGACTTAGGATG	1	11,722,050	w	22		0	0	0	0	0	0	0	0	0	0	0	0	
5	ACGCTATGTTGGACTTAGGATGA	1	11,722,050	w	23		0	0	0	0	0	0	0	0	0	0	0	0	
6	ACGCTATGTTGGACTTAGGATGAA	1	11,722,050	w	24		0	0	0	0	0	0	0	0	0	0	0	0	
7	ACGCTATGTTGGACTTAGGATGAAT	1	11,722,050	w	25		0	0	0	0	0	0	0	0	0	0	0	0	
8	CGCTATGTTGGACTTAGG	3	11,722,051	w	18		0	0	0	0	0	0	0	0	0	0	0	0	
9	CGCTATGTTGGACTTAGGA	1	11,722,051	w	19		0	0	0	0	0	0	0	0	0	0	0	0	
10	CGCTATGTTGGACTTAGGAT	1	11,722,051	w	20		0	0	0	0	0	0	0	0	0	0	0	0	
11	CGCTATGTTGGACTTAGGATG	1	11,722,051	w	21		0	0	0	0	0	1	0	0	0	0	0	0	
12	CGCTATGTTGGACTTAGGATGA	1	11,722,051	w	22		0	0	0	0	0	0	0	0	0	0	0	0	

C Strand Reads

#	Signature	Hits	Coordinate	Strand	Len		Col0	ddc	met1	rdd	Col0b	abh11	ein56	ein56abh11	SUr2a	SUr2b	SWT1	S234a	S2
	Signature abundances normalized to (varies by library depth):						1M	1M	1M	1M	1M	1M	1M	1M	1M	1M	1M	1M	1
1	TCGATAAGATCTTAGAAA	1	11,722,049	c	18		0	0	0	0	0	0	0	0	0	0	0	0	
2	TCGATAAGATCTTAGAAAAT	1	11,722,049	c	20		0	0	0	0	0	0	0	0	0	0	0	0	
3	TCGATAAGATCTTAGAAAATT	1	11,722,049	c	21		0	0	0	0	0	0	0	0	0	0	0	0	
250	TCGAGTTTGTGAGATGTTAGGT	1	11,722,253	c	22		0	0	0	0	0	0	0	0	0	0	0	0	
251	TGAGATCGAGTTTGTGAGATG	1	11,722,258	c	21		0	0	0	0	0	0	0	0	0	0	0	0	
252	TTGAGATCGAGTTTGTGAGAT	1	11,722,259	c	21		0	0	0	0	0	0	0	0	0	0	0	0	

Figure 5. The PA Info page, displaying detailed information on each phasing window with the start coordinate and the strand chosen by the user. A, The list of the positions of 10 cycles on both strands. B, The sequence table for each strand highlights exactly in-phase positions in a darker shade and almost in-phase positions in a lighter shade. C, The PA viewer (interactive 20× image scale) displays the selected phasing window with an additional cycle added to both 5′ and 3′ ends.

added to both 5′ and 3′ ends by showing 11 vertical lines on each strand with a red arrow by the first line on the W strand and a blue arrow by the last line on the C strand, which both represent the start of each strand-specific, first cycle window. The orientation for W-strand windows is from left to right with increasing coordinates, while that for C-strand windows is from right to left with decreasing coordinates. By clicking on each window's image, the user can directly go to the PA Info page for that locus.

FUTURE DEVELOPMENTS

The analysis of phasiRNAs has been a key project in our lab for the last decade (Zhai et al., 2011; Xia et al., 2019). The identification of loci yielding phasiRNAs (*PHAS* loci) is important because their biogenesis requires multiple factors. Thus, a union set of the *PHAS* loci we identify for multiple expression databases may be considered permanent features of a genome and could be added to its annotation data. This is perhaps comparable to the annotation of miRNAs in miRBase (Griffiths-Jones et al., 2008). Our *PHAS* loci identification work is currently separate from our genome and expression database-building pipelines, and thus it is not easy to systematically obtain *PHAS* data for all of our data sets, although we have developed a mechanism of storing and displaying *PHAS* loci data at our web sites. At the moment, we only have one public Arabidopsis web site and some of the rice web sites showing a genome-specific list of *PHAS* loci we have identified and displaying images of loci in the genome viewer.

In terms of future developments, as the first step, we are planning to establish a new mechanism to run the phasiRNA analysis on each expression database and generate a genome-specific list of distinct *PHAS* loci from the analysis outputs, and then integrate this mechanism into our existing expression and genome database-building pipelines. While adding *PHAS* data to more web sites, we will develop a new web tool for the user to search and compare *PHAS* loci from different organisms by type, phase length, title, keyword, and so on. In addition, we want to start a new registry/repository for *PHAS* locus data consolidated from different organisms, hoping that we can contribute to genome annotation projects and provide a repository for information on these important loci.

ACKNOWLEDGMENTS

We thank Jixian Zhai for assistance during the development of the PA viewer and Atul Kakrana for work on the preprocessing script. We thank Jason Miller and other Araport team members for their suggestions on our web services. We also thank the many users who have provided input over the years on bugs and desired features, and the many labs that have provided both the genomic data we depend on and the public libraries that we include on our sites.

Received August 6, 2019; accepted October 18, 2019; published November 4, 2019.

LITERATURE CITED

Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, et al (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol **18:** 630–634

De Paoli E, Dorantes-Acosta A, Zhai J, Accerbi M, Jeong DH, Park S, Meyers BC, Jorgensen RA, Green PJ (2009) Distinct extremely abundant siRNAs associated with cosuppression in petunia. RNA **11:** 1965–1970

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

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: Tools for microRNA genomics. Nucleic Acids Res **36:** D154–D158

Lu C, Tej SS, Luo S, Haudenschild CD, Meyers BC, Green PJ (2005) Elucidation of the small RNA component of the transcriptome. Science **309:** 1567–1569

Meyers BC, Lee DK, Vu TH, Tej SS, Edberg SB, Matvienko M, Tindell LD (2004a) Arabidopsis MPSS: An online resource for quantitative expression analysis. Plant Physiol **135:** 801–813

Meyers BC, Tej SS, Vu TH, Haudenschild CD, Agrawal V, Edberg SB, Ghazal H, Decola S (2004b) The use of MPSS for whole-genome transcriptional analysis in Arabidopsis. Genome Res **14:** 1641–1653

Meyers BC, Vu TH, Tej SS, Ghazal H, Matvienko M, Agrawal V, Ning J, Haudenschild CD (2004c) Analysis of the transcriptional complexity of *Arabidopsis thaliana* by massively parallel signature sequencing. Nat Biotechnol **22:** 1006–1011

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

Vazquez F, Vaucheret H, Rajagopalan R, Lepers C, Gasciolli V, Mallory AC, Hilbert JL, Bartel DP, Crété P (2004) Endogenous trans-acting siRNAs regulate the accumulation of Arabidopsis mRNAs. Mol Cell **16:** 69–79

Xia R, Chen C, Pokhrel S, Ma W, Huang K, Patel P, Wang F, Xu J, Liu Z, Li J, et al (2019) 24-nt reproductive phasiRNAs are broadly present in angiosperms. Nat Commun **10:** 627

Zhai J, Jeong DH, De Paoli E, Park S, Rosen BD, Li Y, González AJ, Yan Z, Kitto SL, Grusak MA, et al (2011) MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased, trans-acting siRNAs. Genes Dev **25:** 2540–2553

Zhai J, Zhang H, Arikit S, Huang K, Nan GL, Walbot V, Meyers BC (2015) Spatiotemporally dynamic, cell-type-dependent premeiotic and meiotic phasiRNAs in maize anthers. Proc Natl Acad Sci USA **112:** 3146–3151