D982–D985 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 13 November 2007*
doi:10.1093/nar/gkm997

Update of ASRP: the *Arabidopsis* Small RNA Project database

Tyler W. H. Backman[1,2], Christopher M. Sullivan[1,2], Jason S. Cumbie[3], Zachary A. Miller[1,2], Elisabeth J. Chapman[1,2,3], Noah Fahlgren[1,2,3], Scott A. Givan[1,2], James C. Carrington[1,2] and Kristin D. Kasschau[1,2,*]

[1]Center for Genome Research and Biocomputing, [2]Department of Botany and Plant Pathology and [3]Molecular and Cellular Biology Program, Oregon State University, Corvallis, OR 97331, USA

Received September 15, 2007; Revised October 20, 2007; Accepted October 22, 2007

ABSTRACT

Development of the *Arabidopsis* Small RNA Project (ASRP) Database, which provides information and tools for the analysis of microRNA, endogenous siRNA and other small RNA-related features, has been driven by the introduction of high-throughput sequencing technology. To accommodate the demands of increased data, numerous improvements and updates have been made to ASRP, including new ways to access data, more efficient algorithms for handling data, and increased integration with community-wide resources. New search and visualization tools have also been developed to improve access to small RNA classes and their targets. ASRP is publicly available through a web interface at http://asrp.cgrb.oregonstate.edu/db/

INTRODUCTION

High-throughput sequencing has enabled the discovery of hundreds of thousands of unique small RNA sequences from *Arabidopsis* and other plants. These small RNAs include both conserved and non-conserved microRNAs (miRNAs), which arise from self-complementary foldback structures, and several classes of siRNA that derive from long inverted duplications, bidirectional transcription or the activity of RNA-dependent RNA polymerases (1,2).

The *Arabidopsis* Small RNA Project (ASRP) database was developed to provide a public resource for genome-wide small RNA data from the model plant *Arabidopsis thaliana* (3). Here, we describe the recent updates to the ASRP database and new software algorithms for efficient searching, accessing and displaying of data on small RNAs and related features of the *Arabidopsis* genome. The ASRP database can be accessed via a website, downloaded as data files, or accessed via the BioMOBY web service (4).

DATABASE CONTENT AND DESIGN

The ASRP database contains data from several sources, and continues to grow as new small RNA libraries are sequenced. As of mid-2007, the database contained 218 585 unique small RNA sequences (663 312 reads) from wild-type and mutant plants generated by picoliter-scale pyrosequencing from the authors' group (3,5,6). Small RNA sequences are cataloged according to 30 unique small RNA libraries from various developmental stages of *Arabidopsis* (Col-0 ecotype), including inflorescence, seedling and leaf tissue, and from mutants with a range of defects in small RNA silencing pathways. Small RNA has been analyzed from mutants with defects in all four Dicer-like genes (*dcl1-7*, *dcl2-1*, *dcl3-1* and *dcl4-2* alleles), and three known functional RNA-dependent RNA-polymerase RDR genes (*rdr1-1*, *rdr2-1* and *rdr6-15* alleles) (7,8).

Small RNAs and associated features from wild-type and mutant libraries can be viewed graphically from the ASRP website using the Generic Model Organism Database Project genome browser (9). A user can select genome coordinates, small RNA names or other locus-defining codes to view small RNA locus positions that are color-coded by size. The viewer also displays transcripts, miRNA precursors, repeat elements and other annotation features. Information pages about small RNAs and other features can be accessed by mouse-over clicks. In addition to small RNA sequences derived in-house, additional small RNA data from other groups, including those of the David Bartel group (10,11), are viewable in distinct tracks. Data from other groups are currently not included in

*To whom correspondence should be addressed. Tel: 541 737 3679; Fax: 541 737 3045; Email: kasschau@cgrb.oregonstate.edu
Present address:
Elisabeth J. Chapman, Department of Biology, Indiana University, Bloomington, IN 47405, USA

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

ASRP database searches, but will be added in the future. Also, genome-wide DNA methylation data from the Steven Henikoff group (12), chromosome data from NCBI, transcript data from TAIR and repeat element data from Repbase at GIRI can be displayed (13–15). Clicking the image for a gene locus in the browser directs the user to an information page containing notes, exon/intron coordinates and small RNAs that derive from the gene. Importantly, the relative transcript accumulation profiles, based on microarray data from a series of RNA silencing defective mutants (5,16), are shown graphically.

The hardware and software used to maintain and update the database has been designed to accommodate large quantities of data from millions of small RNA sequences. The addition of new ASRP data is expedited with the use of software automation, and efficient database population algorithms written specifically for the ASRP database. For example, new small RNA library results are processed through a data-mining pipeline developed with hashes to parse data from raw reads, and then through an XOR string-matching algorithm to find sequences with identity to the *Arabidopsis* genome. Custom algorithms have been developed to minimize iterative interactions with the MySQL (http://www.mysql.com) database during the introduction of new data. These algorithms employ large hash table data structures residing in computer RAM to rapidly identify new sequences and to update existing sequences. Prior to updating the database tables, the large hash tables are used to create MySQL dump files. The dump files are used to update the database, which is much more efficient than iteratively inserting each row. All software have been developed in Perl (http://www.perl.com) and C++ and are available upon request. Frequently requested data and downloads are cached or pre-rendered on the server, enabling the site to serve data simultaneously to multiple users while automatically updating the caches when the data are revised.

The ASRP database utilizes a multistep system of checks and balances to ensure data quality. Known datasets ranging in size from 10 to 1 million data points are generated with built-in false discovery and false positive data. The known datasets can be used to debug and fix changes to the parsing algorithms, and to test timing and throughput of systems and data-mining operations. Known datasets are reviewed regularly, and when changes are made, all code are re-tested and new base lines are established. Data-handling algorithms are also analyzed for logical flaws, and then audited by multiple individuals after being converted to code.

DATABASE ACCESS AND WEB INTERFACE

There are three primary means for outside researchers to access ASRP data. These include the ASRP homepage, direct data file downloads and the BioMoby web service. The ASRP database homepage itself provides multiple ways to access information including a variety of search tools, the genome browser and hyperlinked lists of small RNA families and classes (Figure 1).

The primary search tool allows users to find sequences, genes and small RNA classes by searching with genome coordinates, keywords (such as 'miR171') or sequences. Multiple queries of the same or different category can be done simultaneously and results are displayed quickly using AJAX technology (http://www.adaptivepath.com/ideas/essays/archives/000385.php). Results are viewed one category at a time. Hyperlinks allow users to view additional information or to download search results as a .csv spreadsheet file.

A quick search tool is provided on the navigation toolbar of each page. This tool facilitates rapid access to data for a particular miRNA, tasiRNA, database entry, sequence or gene locus by entering a locus name, label or sequence. A third search tool is provided on the ASRP homepage through a visual map of the five *Arabidopsis* chromosomes. Mouse-over clicks of any region on the chromosome diagrams forwards the user to the corresponding region in the ASRP genome browser.

Detailed miRNA data are provided through hyperlinked lists sorted by miRNA family and target gene family. The main list displays numbers of locus-specific and miRNA family-specific reads from different small RNA libraries. A page for each miRNA family shows sequence variants associated with the family, reads associated with each sequence, miRNA gene transcript data, miRNA target genes and foldback structures. Reads are displayed both in raw counts/library and in library size-normalized form to allow for comparison between small RNA libraries. A hyperlink is provided to link each miRNA to miRBase (17). Data on individual tasiRNAs and tasiRNA families are presented in the same format as for miRNAs.

All data in the ASRP database are provided both as downloadable files accessible from the homepage and as a large MySQL dump file that contains the entire database. Downloads of small RNA data are provided for specific small RNA libraries, for miRNAs and tasiRNAs. Each dataset is provided in three separate formats where appropriate. Genome coordinates are provided in a general feature format (GFF) file, sequences are provided in a FASTA format and both sequences and coordinates are provided as a comma separated spreadsheet (.csv) that can be accessed with a wide array of spreadsheet applications.

The ASRP database can also be accessed via the BioMOBY (http://biomoby.org) web service using a BioMOBY client, such as Taverna workbench (http://taverna.sourceforge.net)(4). The BioMOBY web service allows other researchers to integrate data from the ASRP database into their databases and programs automatically without manually downloading and converting files. One of the main advantages to this approach is that updated data can quickly and automatically propagate to users without manually downloading and re-processing files for each update.

FUTURE DIRECTIONS

Recent sequencing technology, such as sequencing by synthesis (18) will yield millions of small RNA reads in

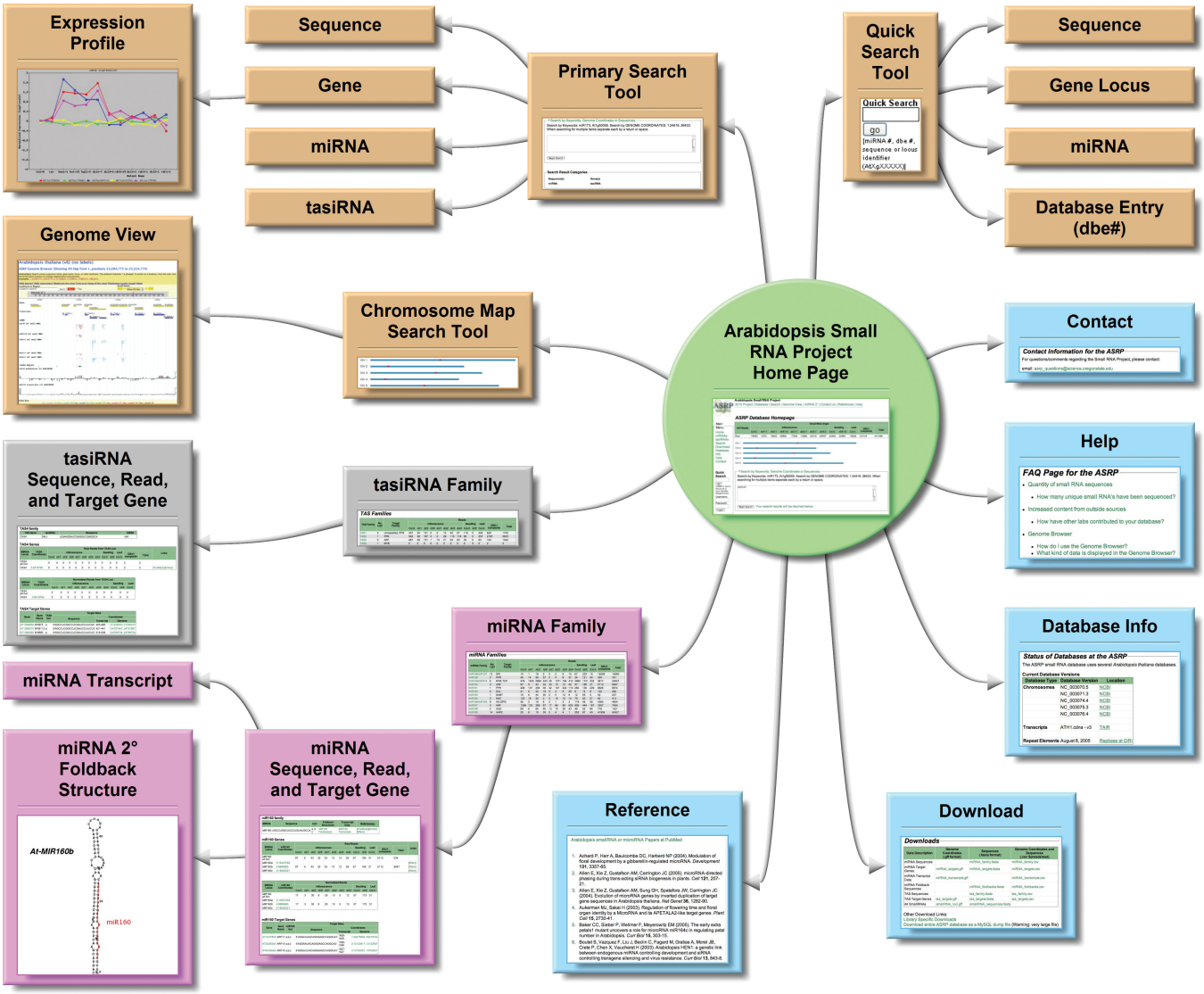

Figure 1. Sitemap of functions available using the ASRP website.

single runs, soon resulting in orders of magnitude increases in the ASRP database content. New algorithms and data-displaying techniques are being integrated into the ASRP database to accommodate the increased quantity and resolution of data from our group as well as other groups. As other databases begin to offer their data via web services, such as BioMOBY, the ASRP database will be modified to integrate more closely with external resources. We envision that the small RNA component of genetic and epigenetic regulation in *Arabidopsis* will become much more apparent and better understood as more systems data are integrated.

ACKNOWLEDGEMENTS

We thank the members of the Carrington laboratory for their input into the development of the ASRP database and website. We thank Mark Dasenko for sequencing small RNA libraries and Amy Shatswell for laboratory management. The *Arabidopsis* small RNA project database is supported by a 2010 project grant from the National Science Foundation (MCB-0618433). Funding to pay the Open Access publication charges for this article was provided by the National Science Foundation (MCB-0618433).

Conflict of interest statement. None declared.

REFERENCES

1. Baulcombe,D. (2004) RNA silencing in plants. *Nature*, **431**, 356–363.
2. Mello,C.C. and Conte,D.Jr (2004) Revealing the world of RNA interference. *Nature*, **431**, 338–342.
3. Gustafson,A.M., Allen,E., Givan,S., Smith,D., Carrington,J.C. and Kasschau,K.D. (2005) ASRP: the Arabidopsis Small RNA Project Database. *Nucleic Acids Res.*, **33**, D637–640.
4. Kawas,E., Senger,M. and Wilkinson,M.D. (2006) BioMoby extensions to the Taverna workflow management and enactment software. *BMC Bioinformatics*, **7**, 523.

5. Kasschau,K.D., Fahlgren,N., Chapman,E.J., Sullivan,C.M., Cumbie,J.S., Givan,S.A. and Carrington,J.C. (2007) Genome-wide profiling and analysis of Arabidopsis siRNAs. *PLoS Biol*., **5**, e57.
6. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J. *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
7. Xie,Z., Allen,E., Wilken,A. and Carrington,J.C. (2005) DICER-LIKE 4 functions in trans-acting small interfering RNA biogenesis and vegetative phase change in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA*, **102**, 12984–12989.
8. Xie,Z., Johansen,L.K., Gustafson,A.M., Kasschau,K.D., Lellis,A.D., Zilberman,D., Jacobsen,S.E. and Carrington,J.C. (2004) Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol*., **2**, E104.
9. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W. *et al*. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res*., **12**, 1599–1610.
10. Axtell,M.J., Jan,C., Rajagopalan,R. and Bartel,D.P. (2006) A two-hit trigger for siRNA biogenesis in plants. *Cell*, **127**, 565–577.
11. Rajagopalan,R., Vaucheret,H., Trejo,J. and Bartel,D.P. (2006) A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. *Genes Dev*., **20**, 3407–3425.
12. Zilberman,D., Gehring,M., Tran,R.K., Ballinger,T. and Henikoff,S. (2007) Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nat. Genet*., **39**, 61–69.
13. Jurka,J., Kapitonov,V.V., Pavlicek,A., Klonowski,P., Kohany,O. and Walichiewicz,J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res*., **110**, 462–467.
14. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R. *et al*. (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **35**, D5–D12.
15. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G. *et al*. (2003) The Arabidopsis information resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res*., **31**, 224–228.
16. Allen,E., Xie,Z., Gustafson,A.M. and Carrington,J.C. (2005) microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell*, **121**, 207–221.
17. Griffiths-Jones,S. (2006) miRBase: the microRNA sequence database. *Methods Mol. Biol*., **342**, 129–138.
18. Mikkelsen,T.S., Ku,M., Jaffe,D.B., Issac,B., Lieberman,E., Giannoukos,G., Alvarez,P., Brockman,W., Kim,T.K. *et al*. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.