Published online 3 November 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D1187–D1193***
doi:10.1093/nar/gkr823

PlantNATsDB: a comprehensive database of plant natural antisense transcripts

Dijun Chen[1,2], Chunhui Yuan[1], Jian Zhang[2], Zhao Zhang[1], Lin Bai[1], Yijun Meng[1], Ling-Ling Chen[2,*] and Ming Chen[1,*]

[1]Department of Bioinformatics, State Key Laboratory of Plant Physiology and Biochemistry, College of Life Sciences, Zhejiang University, Hangzhou, 310058 and [2]State Key Laboratory of Crop Genetic Improvement, College of Life Science and Technology, Huazhong Agricultural University, Wuhan, 430070, P.R. China

Received June 10, 2011; Revised August 29, 2011; Accepted September 18, 2011

ABSTRACT

Natural antisense transcripts (NATs), as one type of regulatory RNAs, occur prevalently in plant genomes and play significant roles in physiological and pathological processes. Although their important biological functions have been reported widely, a comprehensive database is lacking up to now. Consequently, we constructed a plant NAT database (PlantNATsDB) involving approximately 2 million NAT pairs in 69 plant species. GO annotation and high-throughput small RNA sequencing data currently available were integrated to investigate the biological function of NATs. PlantNATsDB provides various user-friendly web interfaces to facilitate the presentation of NATs and an integrated, graphical network browser to display the complex networks formed by different NATs. Moreover, a 'Gene Set Analysis' module based on GO annotation was designed to dig out the statistical significantly overrepresented GO categories from the specific NAT network. PlantNATsDB is currently the most comprehensive resource of NATs in the plant kingdom, which can serve as a reference database to investigate the regulatory function of NATs. The PlantNATsDB is freely available at http://bis.zju.edu.cn/pnatdb/.

INTRODUCTION

Gene regulation at RNA level has been progressively shown to be more important and prevalent than previously presumed (1,2). With the advances of high-throughput experimental technologies and bioinformatics methods, an explosion of recent findings underscores both the predominance and complexity of regulatory RNA molecules in eukaryotes, including the discovery of ubiquitous regulatory short non-coding RNAs (ncRNAs) (3), including microRNAs (miRNAs), endogenous short interfering RNAs (siRNAs) and Piwi-interacting RNAs (piRNAs), and the functional long ncRNAs (1,4). Natural antisense transcripts (NATs), as a new member of regulatory RNAs, occur prevalently in prokaryote and eukaryote genomes, and play significant roles in physiological and/or pathological processes (5). NATs are a group of endogenous RNA molecules containing sequences that are complementary to other transcripts (5–7). This class of RNAs includes both protein- and non-coding transcripts. NATs can be grouped into two categories, *cis*-NATs and *trans*-NATs, based on whether they act in *cis* or *trans*. *Cis*-NAT pairs are transcribed from opposing DNA strands at the same genomic locus and have a variety of orientations and differing lengths of overlap between the perfect sequence complementary regions, whereas *trans*-NAT pairs are transcribed from different loci and form partial complementarily (5). Although underlying mechanistic insights are largely unknown, NATs have been implicated in many aspects of gene regulation including genomic imprinting, transcriptional interference, RNA masking, RNA editing, RNA interference (RNAi) and translational regulation (5,7,8). However, since the discovery of the founder example of *cis*-NATs, SRO5 and P5CDH, involving in the regulation of salt tolerance through RNAi pathway in *Arabidopsis* (*Arabidopsis thaliana*) (9), more and more examples of NATs have been shown to act together with endogenous siRNAs (nat–siRNAs) from the overlapping regions in both plant and animal species (10–16). Moreover, deep sequencing of small RNAs (sRNAs) together with bioinformatics analysis reveals that the overlap portions of NATs are the hotspots for siRNA

*To whom correspondence should be addressed. Tel/Fax: +86 27 87280877; Email: llchen@mail.hzau.edu.cn
Correspondence may also be addressed to Ming Chen. Tel/Fax: +86 571 88206612; Email: mchen@zju.edu.cn

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

generation (12,13,17,18), further indicating that NATs are an important biogenesis mechanism of endogenous siRNAs. These recent discoveries revealed the unexpected complexity of the regulatory networks formed by NATs (17).

Whole-genome searches based on computational analysis have identified thousands of NAT pairs in multiple eukaryotes. Thus, standardized applications or databases are required for data description, deposition, organization, parsing and analysis, and also allowing for functional discovery by integrating other biological data. To date, there are just a few free available NAT databases, one of which, NATsDB (19), comprises 10 animal species. However, the existing databases mainly focus on *cis*-NATs and none of them expand to any plant species, although both *cis*-NATs and *trans*-NATs have been reported in several plant species including two model plants, the monocot rice (*Oryza sativa*) (17,18,20) and the eudicot *Arabidopsis* (18,21–23). Furthermore, the functional annotation and graphical visualization of the NATs is limited.

In the current analysis, we developed a genome-scale computational pipeline to identify NATs in plant species. A convenient database of plant NATs (PlantNATsDB) was constructed, which contains 69 plant species and provides the most comprehensive data set to date. PlantNATsDB serves the plant research community by providing facilitated access to a huge amount of resources regarding the NATs as well as a variety of specific analysis tools including browsing, searching, viewing, downloading and so on. In addition, it integrates Gene Ontology (GO) annotation (24) and sRNA high-throughput sequencing data sets to evaluate and investigate the function of NATs. Moreover, a 'Gene Set Analysis' module based on GO annotation was implemented to excavate the statistical significantly overrepresented GO categories from the complex network formed by different NATs. PlantNATsDB provides an information rich and user-friendly interface and an integrated, graphical network browser to facilitate mining-specific functional NAT pairs (Figure 1). Detailed

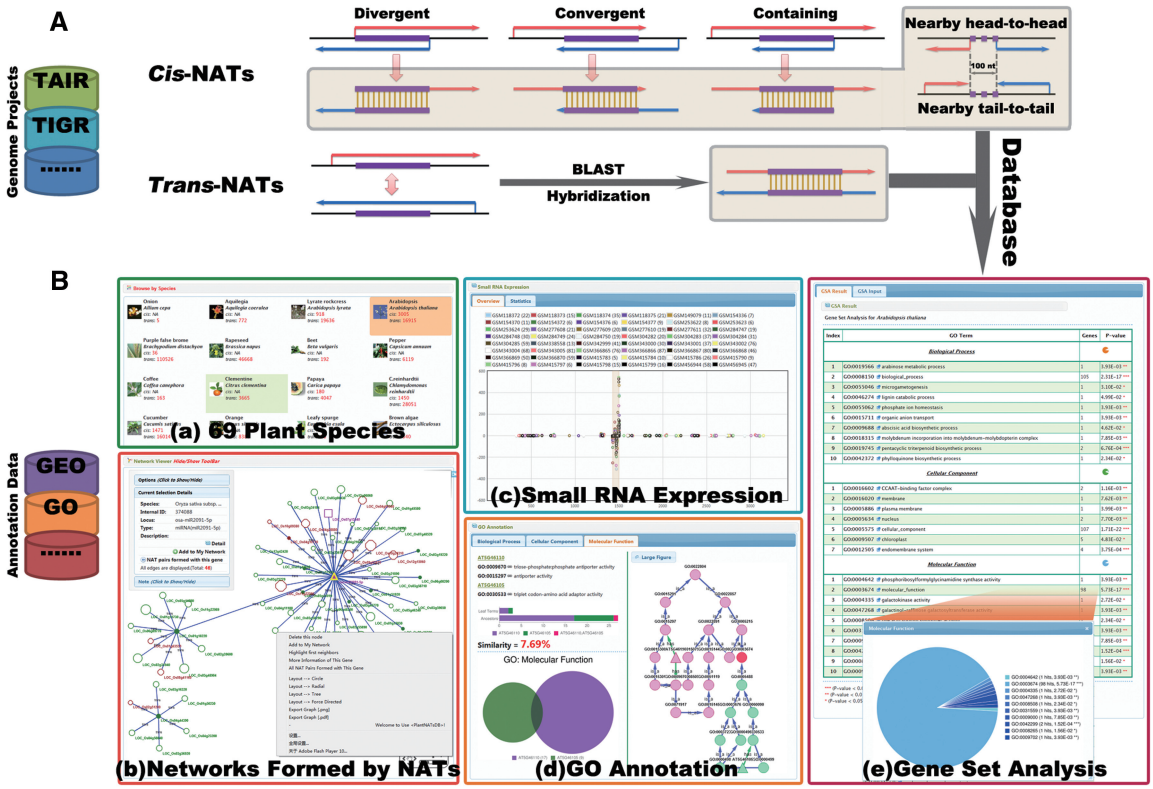

Figure 1. System overview of PlantNATsDB core framework. (**A**) Schematic presentation of the five different types of *cis*-NATs (natural antisense transcripts) (i.e. Divergent, Convergent, Containing, Nearby head-to-head and Nearby tail-to-tail) and the *trans*-NATs predicted by PlantNATsDB. The complementary regions are highlighted and linked with vertical lines. Sequences used for NAT prediction were retrieved from variously public databases, as detailed in the website page. All NATs predicted by PlantNATsDB were deposited in MySQL relational databases. (**B**) Highlighted features of PlantNATsDB, which integrates various data to evaluate the function of NATs. [**B(a)**] The 69 plant species currently available in PlantNATsDB. [**B(b)**] Network formed by different NATs displayed in the integrated network browser, which is based on Cytoscape Web program (31). Note that this network can be edited and used for further analysis, such as 'Gene Set Analysis'. An example of the output for 'Small RNA Expression' of a NAT pair is shown in [**B(c)**] and 'GO Annotation' in [**B(d)**]. Please note that small RNAs are enriched in the overlapped region and the two genes of the NAT pair share very similar GO annotation. [**B(e)**] An example of the output for 'Gene Set Analysis' based on GO annotation. The enriched GO categories are listed in the table and the *P*-value indicating the significance of enrichment. The number of genes in each GO category is indicated and shown in the pie chart. Additional functional modules, such as 'Browser', 'Searcher' and 'Viewer' as detailed in the PlantNATsDB website.

information is provided at the PlantNATsDB website (http://bis.zju.edu.cn/pnatdb/).

DATABASE CONSTRUCTION

Data source

Of the 69 plant species, 27 have genomic information. For these 27 genomically sequenced species, the annotated transcription units (TUs) used for NAT prediction and other annotation information were downloaded from the specific genome-sequencing projects. Based on the fact that pseudogenes and transposons can form NATs with protein-coding genes (10,11,17), all the pseudogenes and transposons were retained for NAT prediction. For the remaining 42 plant species, the tentative consensus sequences (TCs), which can be used to provide putative genes with functional annotation similar to TUs were used for NAT prediction, and their related information were downloaded from The Gene Index Project (25).

SRNA high-throughput sequencing data sets of each species were obtained from the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/) (26). All the sRNA data sets retrieved for this study were summarized in Table 1.

Prediction of NAT pairs

Prediction of NAT pairs was performed as previously described (17,18,22). Specifically, the following criteria were used to identify *cis*-NATs and *trans*-NATs, respectively.

Table 1. Summary statistics of small RNA data sets in PlantNATsDB

No.	Species	GEO Data sets[a]	
		Series	Samples
1	*Arabidopsis thaliana*	15	80
2	*Arabidopsis lyrata*	3	8
3	*Brachypodium distachyon*	2	4
4	*Chlamydomonas reinhardtii*	3	6
5	*Citrus sinensis*	1	2
6	*Gossypium hirsutum*	2	6
7	*Glycine max*	2	5
8	*Medicago truncatula*	2	5
9	*Nicotiana benthamiana*	2	6
10	*Oryza sativa subsp. indica*	1	2
11	*Oryza sativa subsp. japonica*	6	38
12	*Physcomitrella patens*	3	10
13	*Prunus persica*	1	2
14	*Solanum lycopersicum*	1	2
15	*Selaginella moellendorffii*	1	1
16	*Triticum aestivum*	2	2
17	*Vitis vinifera*	2	5
18	*Zea mays*	4	12
Total		54	196

[a]Number of GEO Series or GEO Samples in each species, including biological and technical replicates. Detailed information of the data sets in each species is provided at the PlantNATsDB website.
Note that all small RNA data sets in this study were downloaded from the GEO database (http://www.ncbi.nlm.nih.gov/geo/) (26).

For *cis*-NATs, they can be grouped into five categories, namely: (i) Divergent (head to head or 5′ to 5′ overlap); (ii) Convergent (tail-to-tail or 3′ to 3′ overlap); (iii) Containing (full overlap); (iv) Nearby head-to-head (5′ close to 5′) and (v) Nearby tail-to-tail (3′ close to 3′) according to their relative orientation and degree of overlap (Figure 1A) (27). If a pair of transcripts is located in opposite strands at adjacent genomic loci and has at least 1 nt overlapping, or their distance on the chromosome is no >100 nt, then they were considered as a *cis*-NAT pair. In total, 27 plant species were subjected to *cis*-NAT prediction.

For *trans*-NATs, BLASTN (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/, Release 2.2.20) (28) was used to search for transcript pairs with high sequence complementary to each other and the following criteria should be satisfied for each transcript pair: (i) If the complementary region identified by BLAST covered more than half the length of either transcript, this transcript pair was designated to be a 'high-coverage' (HC) *trans*-NAT pair; (ii) If the two transcripts had a continuous complementary region >100 nt, they were classified as a '100-nt' pair. Functional *trans*-NATs should form RNA–RNA duplexes *in vivo*. We therefore used DINAMelt (29) to verify whether the transcript pairs could melt into RNA–RNA duplexes in the complementary regions *in silico*. All the *trans*-NAT pairs based on BLAST search were further used to DINAMelt hybridization validation. The *trans*-NAT pair was retained if it satisfied: (i) the paired region indentified by DINAMelt should be coincident with the BLAST-based search; (ii) any bubble in the paired region predicted by DINAMelt should be no longer than 10% of the region. For the BLAST-based *trans*-NAT pairs that contain transcripts >10 kb, they were not applied to DINAMelt validation due to the heavy computational work. Instead, if the paired region identified by BLAST was >10% of its longer transcript, it was considered as verified *trans*-NAT.

All the NAT pairs predicted in this study were summarized in Table 2.

Small RNA analysis

SRNA sequences containing incomplete information (such as containing 'N') with length <18 or >28 were removed for further analysis. For each data set, the filtered sRNA sequences were mapped to all the gene models of the related plant species. All mapping steps were performed using the Bowtie algorithm (30) allowing no mismatch. Besides, for comparison, the normalized abundance of sRNAs from each data set was calculated as RPMs (reads per million), which divided the read number of each sRNA by the total reads from this data set, and multiplied by 10^6.

For each NAT, an enrichment score was calculated to evaluate whether sRNAs were enriched in the overlapping region (17,18). The enrichment score E was calculated using the following formula:

$$E = \frac{S_o/L_o}{S_a/L_a}$$

Table 2. Statistical result of NATs predicted in this study

No.	Species	Genes[a]	NATs (*cis*, *trans*)[b]	No.	Species	Genes[a]	NATs (*cis*, *trans*)[b]
1	*Allium cepa*	4063 (10)	5 (*NA*, 5)	36	*Nicotiana benthamiana*	7712 (429)	564 (*NA*, 564)
2	*Aquilegia coerulea*	13 556 (655)	141 (*NA*, 141)	37	*Nicotiana tabacum*	45 554 (3962)	3521 (*NA*, 3521)
3	*Arabidopsis lyrata*	32 670 (4841)	6757 (918, 5839)	38	*Oryza sativa subsp. indica*	40 745 (10 153)	144 088 (387, 143 701)
4	*Arabidopsis thaliana*	33 239 (8049)	7788 (3005, 4783)	39	*Oryza sativa subsp. japonica*	57 624 (30 799)	409 789 (1186, 408 603)
5	*Beta vulgaris*	4785 (249)	192 (*NA*, 192)	40	*Ostreococcus lucimarinus CCE9901*	7805 (2773)	1498 (1482, 16)
6	*Brachypodium distachyon*	25 532 (5363)	107 933 (36, 107 897)	41	*Ostreococcus tauri*	7725 (3030)	1790 (1620, 170)
7	*Brassica napus*	50 542 (20771)	45 930 (*NA*, 45 930)	42	*Panicum virgatum*	52 936 (4631)	4802 (*NA*, 4802)
8	*Capsicum annuum*	14 727 (2138)	6119 (*NA*, 6119)	43	*Petunia hybrida*	2259 (39)	25 (*NA*, 25)
9	*Carica papaya*	25 536 (3991)	7302 (180, 7122)	44	*Phaseolus coccineus*	22 518 (1063)	754 (*NA*, 754)
10	*Chlamydomonas reinhardtii*	15 935 (6549)	26 879 (1450, 25 429)	45	*Phaseolus vulgaris*	11 954 (638)	433 (*NA*, 433)
11	*Citrus clementina*	32 287 (2243)	3554 (*NA*, 3554)	46	*Physcomitrella patens*	35 938 (3976)	24 396 (195, 24 201)
12	*Citrus sinensis*	26 081 (3451)	7492 (*NA*, 7492)	47	*Picea abies*	42 746 (22360)	43 535 (*NA*, 43 535)
13	*Coffea canephora*	7511 (202)	163 (*NA*, 163)	48	*Pinus taeda*	39 798 (10897)	14 298 (*NA*, 14 298)
14	*Cucumis sativus*	32 775 (6104)	23 373 (1471, 21 902)	49	*Populus trichocarpa*	41 377 (5001)	13 107 (744, 12 363)
15	*Ectocarpus siliculosus*	9122 (387)	340 (*NA*, 340)	50	*Prunus persica*	27 852 (4642)	26 163 (298, 25 865)
16	*Euphorbia esula*	10 727 (103)	96 (*NA*, 96)	51	*Quercus robur*	17 804 (2138)	2142 (*NA*, 2142)
17	*Festuca arundinacea*	10 617 (309)	151 (*NA*, 151)	52	*Raphanus sativus*	17 939 (356)	233 (*NA*, 233)
18	*Festuca pratensis*	12 248 (156)	96 (*NA*, 96)	53	*Ricinus communis*	31 221 (2570)	3348 (495, 2853)
19	*Fragaria vesca*	34 809 (10 622)	117 786 (574, 117 212)	54	*Saccharum officinarum*	42 377 (5311)	7210 (*NA*, 7210)
20	*Glycine max*	46 367 (11 352)	78 339 (436, 77 903)	55	*Secale cereale*	1471 (52)	32 (*NA*, 32)
21	*Gossypium hirsutum*	50 081 (31 296)	80 835 (*NA*, 80 835)	56	*Selaginella moellendorffii*	22 285 (2399)	1558 (669, 889)
22	*Gossypium raimondii*	9508 (667)	426 (*NA*, 426)	57	*Solanum lycopersicum*	28 167 (2039)	1793 (*NA*, 1793)
23	*Helianthus annuus*	20 130 (2460)	2255 (*NA*, 2255)	58	*Solanum melongena*	14 512 (219)	336 (*NA*, 336)
24	*Hordeum vulgare*	43 306 (6993)	8503 (*NA*, 8503)	59	*Solanum tuberosum*	31 972 (2849)	2866 (*NA*, 2866)
25	*Ipomoea nil*	11 754 (57)	31 (*NA*, 31)	60	*Sorghum bicolor*	34 496 (8231)	145 374 (241, 145 133)
26	*Lactuca sativa*	12 505 (347)	263 (*NA*, 263)	61	*Striga hermonthica*	9275 (178)	128 (*NA*, 128)
27	*Lactuca serriola*	8047 (215)	140 (*NA*, 140)	62	*Theobroma cacao*	14 724 (889)	1593 (*NA*, 1593)
28	*Lotus japonicus*	40 504 (7783)	29 575 (126, 29 449)	63	*Triphysaria eriantha*	17 442 (1491)	1224 (*NA*, 1224)
29	*Malus x domestica*	34 945 (3631)	4356 (*NA*, 4356)	64	*Triphysaria versicolor*	7165 (672)	539 (*NA*, 539)
30	*Manihot esculenta*	47 443 (14 342)	30 308 (4454, 25 854)	65	*Triticum aestivum*	93 508 (32 258)	120 316 (*NA*, 120 316)
31	*Medicago truncatula*	50 962 (18 083)	164 686 (1151, 163 535)	66	*Vigna unguiculata*	19 333 (592)	405 (*NA*, 405)
32	*Mesembryanthemum crystallinum*	3627 (207)	156 (*NA*, 156)	67	*Vitis vinifera*	26 346 (11 898)	108 392 (685, 107 707)
33	*Micromonas pusilla CCMP1545*	10 547 (4717)	11 881 (1573, 10 308)	68	*Volvox carteri*	15 669 (7438)	90 222 (273, 89 949)
34	*Micromonas sp. RCC299*	10 108 (4321)	2338 (2189, 149)	69	*Zea mays*	32 540 (6944)	25 726 (1528, 24 198)
35	*Mimulus guttatus*	27 501 (8885)	160 109 (1032, 159 077)	Total[c]		1 746 886 (384 466)	2 138 498 (28 398, 2 110 100)

[a]Number of genes used for NAT prediction in each species. The number of genes formed at least one NAT pair with other genes is shown in parenthesis.
[b]Number of predicted NAT (*cis*- and *trans*-NAT) pairs in each species.
[c]The total number in all species belonging to each categories.

where S_o = the total normalized abundance of the sRNAs generated from the overlapping region, L_o = the total length of the paired region of the two transcripts of the NATs, S_a = the total normalized abundance of the sRNAs generated from these two transcripts and L_a = the total length of the two transcripts. Furthermore, a standard χ^2 test (Pearson's chi-square test) was performed to test the significance of the enrichment.

Database implementation

All the predicted NATs and processed sRNAs were organized and stored in the MySQL database (http://www.mysql.com/). Besides, the gene sequence information, annotated gene models and their functional annotations, including GO annotations, were collected and stored in the database. These genes can also be linked to external genome browsers. PlantNATsDB was implemented in JSP language and deployed on the Apache Tomcat web server (http://tomcat.apache.org/). The integrated network browser is created by Cytoscape Web program (http://cytoscapeweb.cytoscape.org/) (31). JavaScript and adobe flash player are required in order to use the full functionality of PlantNATsDB. PlantNATsDB can be accessed through IE 6.0 or higher, Netscape 7.0 or higher, Safari, Opera, Chrome and Firefox from multiple platforms.

WEB INTERFACE AND DATABASE USAGE

Search modules

PlantNATsDB provides various query interfaces and graphical visualization tools to facilitate the retrieve and demonstration of NAT data. Four major search modules for retrieving NATs are designed: 'Simple Searcher', 'Batched Searcher', 'Advanced Searcher' and 'BLAST Searcher'. Alternatively, users can get the entire NAT list by species in the 'Browser' module. The 'Simple Searcher' module allows users to enter any keyword in all fields for all data entries, including gene locus identifiers (IDs), gene aliases or any words in their annotation texts. The 'Batched Searcher' module supports gene set search, which allows users to enter a list of gene locus IDs or gene aliases. The 'Advanced Searcher' was designed to facilitate users to access any NAT data according multiple options such as the plant species, the types of NATs, the length of overlapping regions and the GO annotation. In addition, users can perform a BLAST sequence search to retrieve NAT data in the fourth module, 'BLAST Searcher'. All the search results performed by the above search modules can be further used for functional investigation (see below).

NAT information page

For each NAT pair, PlantNATsDB provides rich annotation according to the relationships between the related two genes. The result page largely comprises four main parts, i.e. NAT summarization, gene information, GO annotation and sRNA expression. Generally, all parts are displayed vividly in the graphical fashion. Figure 2 shows the example of SRO5 and P5CDH *cis*-NAT pair (9). The first part is the summary of NAT information and the overlapping region is highlighted (Figure 2A). The second part shows the detailed annotation of the two genes (Figure 2B). The third part displays the GO functional assessment of this NAT pair based on the GO annotation of the two genes (Figure 2C). Functional NAT pairs are expected to have similar 'Molecular Function', involve in the related 'Biological Process' and/or locate in the same 'Cellular Component'. Therefore, the same GO terms shared by the two genes are highlighted in the GO network graph. The information provided in this part is very useful for evaluating the function of NAT. The last part provides sRNA expression derived from the NAT pair (Figure 2D). Based on the finding that sRNAs were the important component in the NAT regulatory pathway (9), most, if not all, of the sRNA data sets currently available were collected and further processed and organized into the database (Table 1). Thus, these invaluable data sources will be of much help to users to inquire the function of NATs. Furthermore, a user-friendly interface is provided that allows users to add or remove data sets for analysis and to highlight different regions of the NAT.

GO functional analysis module

Gene set analysis based on GO annotation (24) and statistical test is widely used to identify enriched GO categories and to explore the most important biological terms associated with the given gene set. A 'Gene Set Analysis' module (Figure 1B) has been developed for organizing a set of genes based on GO annotation, where the set of genes can be found by the search modules (see above) or collected in the network formed by NAT pairs (see below). Here, we used the combination of the χ^2 test and Fisher's exact test to evaluate the significance of enrichment for GO category. Detailed methods can be referred from the PlantNATsDB website.

Graphical interaction network visualization

One gene may forms multiple NAT pairs with other antisense transcription partners, just as multiple paralogous genes may form RNA duplexes with the same antisense transcripts. Different NAT pairs might form complex regulatory networks in the related process (17). To this end, a graphical browser based on Cytoscape Web program (31) was developed to display the network formed by different NAT pairs (Figure 1B). Different types of nodes (genes) and relationships (NAT pairs) are colored distinctly. Moreover, the network graph can be edited (such as, to click/double-click/right-click the nodes/edges, to delete the nodes/edges, to apply distinct layouts and to export the graph in various formats) and all the genes contained in the network can be further subject to gene set analysis based on GO annotation (see above). In addition, users can use the toolkit of 'My Network', where genes or NAT pairs of interest may be stored temporarily on the server side during the session period and later retrieved in the 'My Network' page. There is a button to add selected genes or NAT pairs to 'My Network' in

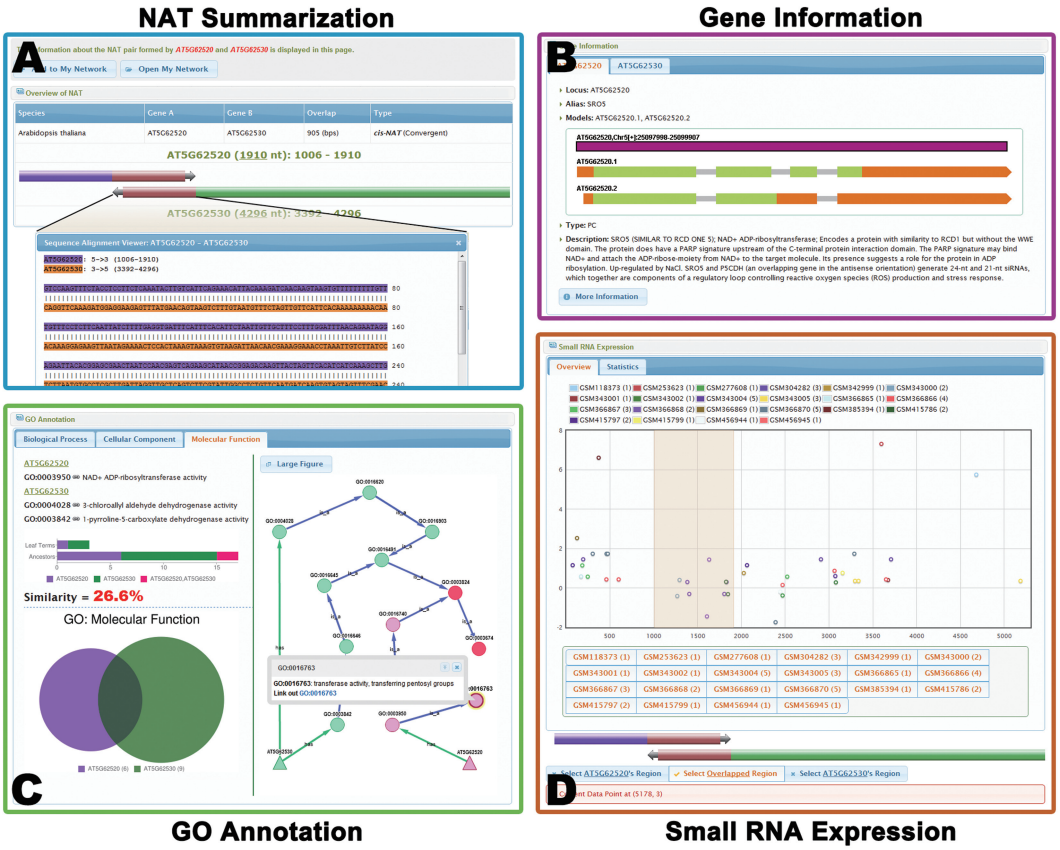

Figure 2. The information page of the NAT pair formed by SRO5 (*AT5G62520*) and P5CDH (*AT5G62530*) (9). (**A**) Summary of the NAT information, including the type, sequences and length of overlapped region. The sequence of the overlapped region is highlighted below. (**B**) Detailed annotation of the two genes of this *cis*-NAT pair. (**C**) GO functional annotation of the two genes. The annotated GO terms are displayed in Venn chart and GO network graph. The GO network graph contains two types of nodes: those that represent the NAT pairs (triangle nodes) and those that represent GO hierarchical terms (circle nodes). The shared GO terms (red color) and specific GO terms (purple and green colors) are shown in different colors. Functional similarity of these two genes is represented by the percent of shared GO terms. (**D**) The expression of the small RNAs derived from the NAT pair. Small RNAs from different data sets are indicated by dots in different colors. The overlapped region is highlighted in the chart. Small RNA data sets can be added or removed for demonstration in the chart by clicking the buttons below. The enrichment score for small RNA generated from the overlapped region is calculated based on the specific data sets. Please note that there is no observation of enriched small RNA derived from the overlapped region because this NAT pair is specially formed in the salt stress condition and PlantNATsDB lacks such data sets.

many pages of the website, which will greatly facilitate users' digging out specific biological network formed by related NAT pairs involved in regulation of the interrelated process.

SUMMARY AND FUTURE DIRECTIONS

This work presents a comprehensive collection of plant NATs, which are organized and deposited in an online database named PlantNATsDB. The biological function of NAT pairs can be elucidated from the variously integrated data currently available. Moreover, vivid web interfaces are also designed to facilitate the presentation of NATs. PlantNATsDB serves the plant research community by providing a reference database to investigate the functions of NATs.

In the near future, PlantNATsDB will collect and include more experimentally validated data and plan to make distinction between experimentally determined and predicted NATs. In addition, more useful and precise algorithms or tools will be designed to evaluate the functions of NAT pairs or to dig out functional NAT pairs based on GO network graphs and NATs-formed regulatory network. For example, it would be helpful to put such a regulatory subnetwork graph to the context of a larger network. Besides, some NAT pairs or subnetworks formed by NATs may be conserved between species. PlantNATsDB intends to allow users to select a specific family and to make comparisons within the family members.

As new and improved high-throughput technologies are applied to a broader set of species, cell lines, tissues and conditions, more and more data sets will be generated, PlantNATsDB will be continuously maintained and timely updated to keep up with these improvements. In addition, gene expression data, such as ESTs (expression sequence tags), microarray and RNA-Seq data and degradome-sequencing data (32,33) will be integrated into PlantNATsDB to improve our understanding of the regulatory networks formed by NATs.

ACKNOWLEDGEMENTS

The authors thank the Joint Genome Institute (http://www.jgi.doe.gov/) for the availability of the draft genome assemblies and the genomic annotation of *Arabidopsis lyrata*, *Cucumis sativus*, *Manihot esculenta*, *Mimulus guttatus* and *Selaginella moellendorffii*. The authors thank Dr Christian Klukas for his kind discussions. The authors also thank Dr Michael Galperin and the three anonymous referees for their constructive and helpful suggestions.

FUNDING

The National Natural Sciences Foundation of China (30971743, 31050110121, 31071659); The Ministry of Science and Technology of China (2009DFA32030); Program for New Century Excellent Talents in University of China (NCET-07-0740); Huazhong Agricultural University Scientific & Technological Self-innovation Foundation (2010SC07). Funding for Open Access charge: Partial waiver by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Ponting,C.P., Oliver,P.L. and Reik,W. (2009) Evolution and functions of long noncoding RNAs. *Cell*, **136**, 629–641.
2. Brosnan,C.A. and Voinnet,O. (2009) The long and the short of noncoding RNAs. *Curr. Opin. Cell Biol.*, **21**, 416–425.
3. Ghildiyal,M. and Zamore,P.D. (2009) Small silencing RNAs: an expanding universe. *Nat. Rev. Genet.*, **10**, 94–108.
4. Mercer,T.R., Dinger,M.E. and Mattick,J.S. (2009) Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.*, **10**, 155–159.
5. Lapidot,M. and Pilpel,Y. (2006) Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. *EMBO Rep.*, **7**, 1216–1222.
6. Vanhee-Brossollet,C. and Vaquero,C. (1998) Do natural antisense transcripts make sense in eukaryotes? *Gene*, **211**, 1–9.
7. Lavorgna,G., Dahary,D., Lehner,B., Sorek,R., Sanderson,C.M. and Casari,G. (2004) In search of antisense. *Trends Biochem. Sci.*, **29**, 88–94.
8. Faghihi,M.A. and Wahlestedt,C. (2009) Regulatory roles of natural antisense transcripts. *Nat. Rev. Mol. Cell. Biol.*, **10**, 637–643.
9. Borsani,O., Zhu,J., Verslues,P.E., Sunkar,R. and Zhu,J.K. (2005) Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis. *Cell*, **123**, 1279–1291.
10. Watanabe,T., Totoki,Y., Toyoda,A., Kaneda,M., Kuramochi-Miyagawa,S., Obata,Y., Chiba,H., Kohara,Y., Kono,T., Nakano,T. *et al.* (2008) Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature*, **453**, 539–543.
11. Tam,O.H., Aravin,A.A., Stein,P., Girard,A., Murchison,E.P., Cheloufi,S., Hodges,E., Anger,M., Sachidanandam,R., Schultz,R.M. *et al.* (2008) Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, **453**, 534–538.
12. Okamura,K., Balla,S., Martin,R., Liu,N. and Lai,E.C. (2008) Two distinct mechanisms generate endogenous siRNAs from bidirectional transcription in Drosophila melanogaster. *Nat. Struct. Mol. Biol.*, **15**, 998.
13. Czech,B., Malone,C.D., Zhou,R., Stark,A., Schlingeheyde,C., Dus,M., Perrimon,N., Kellis,M., Wohlschlegel,J.A., Sachidanandam,R. *et al.* (2008) An endogenous small interfering RNA pathway in Drosophila. *Nature*, **453**, 798–802.
14. Ghildiyal,M., Seitz,H., Horwich,M.D., Li,C., Du,T., Lee,S., Xu,J., Kittler,E.L., Zapp,M.L., Weng,Z. *et al.* (2008) Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. *Science*, **320**, 1077–1081.
15. Ron,M., Alandete Saez,M., Eshed Williams,L., Fletcher,J.C. and McCormick,S. (2010) Proper regulation of a sperm-specific cis-nat-siRNA is essential for double fertilization in Arabidopsis. *Genes Dev.*, **24**, 1010–1021.
16. Katiyar-Agarwal,S., Morgan,R., Dahlbeck,D., Borsani,O., Villegas,A. Jr, Zhu,J.K., Staskawicz,B.J. and Jin,H. (2006) A pathogen-inducible endogenous siRNA in plant immunity. *Proc. Natl Acad. Sci. USA*, **103**, 18002–18007.
17. Zhou,X., Sunkar,R., Jin,H., Zhu,J.K. and Zhang,W. (2009) Genome-wide identification and analysis of small RNAs originated from natural antisense transcripts in Oryza sativa. *Genome Res.*, **19**, 70–78.
18. Chen,D., Meng,Y., Ma,X., Mao,C., Bai,Y., Cao,J., Gu,H., Wu,P. and Chen,M. (2010) Small RNAs in angiosperms: sequence characteristics, distribution and generation. *Bioinformatics*, **26**, 1391–1394.
19. Zhang,Y., Li,J., Kong,L., Gao,G., Liu,Q.R. and Wei,L. (2007) NATsDB: Natural Antisense Transcripts DataBase. *Nucleic Acids Res.*, **35**, D156–D161.
20. Osato,N., Yamada,H., Satoh,K., Ooka,H., Yamamoto,M., Suzuki,K., Kawai,J., Carninci,P., Ohtomo,Y., Murakami,K. *et al.* (2003) Antisense transcripts with rice full-length cDNAs. *Genome Biol.*, **5**, R5.
21. Wang,X.J., Gaasterland,T. and Chua,N.H. (2005) Genome-wide prediction and identification of cis-natural antisense transcripts in Arabidopsis thaliana. *Genome Biol.*, **6**, R30.
22. Wang,H., Chua,N.H. and Wang,X.J. (2006) Prediction of trans-antisense transcripts in Arabidopsis thaliana. *Genome Biol.*, **7**, R92.
23. Jin,H., Vacic,V., Girke,T., Lonardi,S. and Zhu,J.K. (2008) Small RNAs and the regulation of cis-natural antisense transcripts in Arabidopsis. *BMC Mol. Biol.*, **9**, 6.
24. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
25. Lee,Y., Tsai,J., Sunkara,S., Karamycheva,S., Pertea,G., Sultana,R., Antonescu,V., Chan,A., Cheung,F. and Quackenbush,J. (2005) The TIGR gene indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Res.*, **33**, D71–D74.
26. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
27. Osato,N., Suzuki,Y., Ikeo,K. and Gojobori,T. (2007) Transcriptional interferences in cis natural antisense transcripts of humans and mice. *Genetics*, **176**, 1299–1306.
28. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
29. Markham,N.R. and Zuker,M. (2005) DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res.*, **33**, W577–W581.
30. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
31. Lopes,C.T., Franz,M., Kazi,F., Donaldson,S.L., Morris,Q. and Bader,G.D. (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347–2348.
32. German,M.A., Pillay,M., Jeong,D.H., Hetawal,A., Luo,S., Janardhanan,P., Kannan,V., Rymarquis,L.A., Nobuta,K., German,R. *et al.* (2008) Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nat. Biotechnol.*, **26**, 941–946.
33. Addo-Quaye,C., Eshoo,T.W., Bartel,D.P. and Axtell,M.J. (2008) Endogenous siRNA and miRNA targets identified by sequencing of the Arabidopsis degradome. *Curr. Biol.*, **18**, 758–762.