PCP
PLANT & CELL PHYSIOLOGY

Special Issue – Databases

ARTADE2DB: Improved Statistical Inferences for Arabidopsis Gene Functions and Structure Predictions by Dynamic Structure-Based Dynamic Expression (DSDE) Analyses

Kei Iida[1,4], Shuji Kawaguchi[1,4], Norio Kobayashi[1], Yuko Yoshida[1], Manabu Ishii[1], Erimi Harada[1], Kousuke Hanada[1,2], Akihiro Matsui[2], Masanori Okamoto[2,3], Junko Ishida[2], Maho Tanaka[2], Taeko Morosawa[2], Motoaki Seki[2] and Tetsuro Toyoda[1,*]

[1]RIKEN BASE (Bioinformatics And Systems Engineering) Division, Yokohama, Kanagawa, 230-0045 Japan
[2]RIKEN Plant Science Center, Yokohama, Kanagawa, 230-0045 Japan
[3]Present address: Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA
[4]These authors contributed equally to this work
*Corresponding author: E-mail, toyoda@base.riken.jp; Fax, +81-45-503-9553
(Received October 29, 2010; Accepted December 20, 2010)

Recent advances in technologies for observing high-resolution genomic activities, such as whole-genome tiling arrays and high-throughput sequencers, provide detailed information for understanding genome functions. However, the functions of 50% of known *Arabidopsis thaliana* genes remain unknown or are annotated only on the basis of static analyses such as protein motifs or similarities. In this paper, we describe dynamic structure-based dynamic expression (DSDE) analysis, which sequentially predicts both structural and functional features of transcripts. We show that DSDE analysis inferred gene functions 12% more precisely than static structure-based dynamic expression (SSDE) analysis or conventional co-expression analysis based on previously determined gene structures of *A. thaliana*. This result suggests that more precise structural information than the fixed conventional annotated structures is crucial for co-expression analysis in systems biology of transcriptional regulation and dynamics. Our DSDE method, ARabidopsis Tiling-Array-based Detection of Exons version 2 and over-representation analysis (ARTADE2-ORA), precisely predicts each gene structure by combining two statistical analyses: a probe-wise co-expression analysis of multiple transcriptome measurements and a Markov model analysis of genome sequences. ARTADE2-ORA successfully identified the true functions of about 90% of functionally annotated genes, inferred the functions of 98% of functionally unknown genes and predicted 1,489 new gene structures and functions. We developed a database ARTADE2DB that integrates not only the information predicted by ARTADE2-ORA but also annotations and other functional information, such as phenotypes and literature citations, and is expected to contribute to the study of the functional genomics of *A. thaliana*. URL: http://artade.org.

Keywords: *Arabidopsis thaliana* • Database • Function prediction • Genome tiling array • Unknown genes.

Abbreviations: DSDE, dynamic structure-based dynamic expression; GO, gene ontology; miRNA, microRNA; ORA, over-representation analysis; PCC, Pearson's correlation coefficient; PO, plant ontology; snoRNA, small nucleolar RNA; SSDE, static structure-based dynamic expression.

Introduction

Arabidopsis thaliana is one of the most studied model plants. Its genome sequence has been determined (Arabidopsis Genome Initiative 2000) and several genome-wide functional genomic projects were completed or are in progress. (Its status is well reviewed in 'The Multinational Coordinated *Arabidopsis thaliana* Functional Genomics Project, Annual Report 2010' at http://www.arabidopsis.org/portals/masc/.) Despite these efforts, only a small fraction of its genes are well characterized today. The main reason is that detailed studies of a gene often require considerable time, human resources and finances. Therefore, it is important to support experimental efforts properly with computational approaches. Function prediction by motif prediction methods or similar searches, such as Pfam and BLAST, are popular (Altschul et al. 1997, Finn et al. 2010). We call such analyses 'static analyses', because they do not include transcription dynamics such as gene expression

Plant Cell Physiol. 52(2): 254–264 (2011) doi:10.1093/pcp/pcq202, available online at www.pcp.oxfordjournals.org

Table 1 Types of analyses for studying gene functions computationally

Types of analysis[a]	Structure[b]	Expression[c]	Analysis	Reliability of the coding sequence on the gene[d]	Reliability of dynamic expression analysis[d]	Ability to find novel genes[e]	Tools/Databases[f]
SSSA	Static	Static	Homology/motif search based on reference gene structures	⊙	Not applicable	−	BLAST (1), Pfam (2)
SSDE	Static	Dynamic	Co-expression analysis based on reference gene structures	⊙	◯	−	ATTED-II (3), CressExpress (4)
DSDE	Dynamic	Dynamic	Simultaneous elucidation of gene structures and dynamism of expression	◯	⊙	+	ARTADE2-ORA

[a] SSSA, static structure-based static analysis; SSDE, static structure-based dynamic expression; DSDE, dynamic structure-based dynamic expression.
[b] Static structures are pre-defined gene structures such as annotated genes, whereas dynamic structures are constructed gene models depending on the studied transcriptome.
[c] Static expression indicates gene expression analyses in which conditional changes are ignored. An example is a cDNA collection study for correcting gene structures. Dynamic expression indicates gene expression changes observed under multiple conditions.
[d] ⊙, very good; ◯, good.
[e] +, positive; −, negative.
[f] References: 1, Altschul et al. (1997); 2 Gunasekaran et al. (2010); 3, Obayashi et al. (2009); 4, Srinivasasainagendra et al. (2008).

changes (SSSA; static structure-based static analysis in **Table 1**). While static expression analysis is based on sequence information, systems biology approaches have attempted to characterize gene and genome functions on the basis of the dynamism of gene expression observed in multiple experimental conditions that include various perturbations such as stresses. We call such analyses based on multiple transcriptomes 'dynamic expression analyses'. Co-expression analysis is a dynamic expression approach. Gene co-expression relationships are known to provide gene function information (van Noort et al. 2003). Thus, we can predict gene functions by analyzing the enrichment of specific gene functions among co-expressed genes (Alexa et al. 2006, Grossmann et al. 2007), which is called over-representation analysis (ORA). Several databases for co-expression analysis exist, such as ATTED-II (*Arabidopsis thaliana* trans-factor and cis-element prediction database) (Obayashi et al. 2009) and CressExpress (Srinivasasainagendra et al. 2008). These co-expression analyses use microarray outputs that are designed on the basis of annotated gene sets. We call such approaches 'static structure-based dynamic expression (SSDE)' analyses (**Table 1**). Although SSDE is a useful approach, it has theoretical limitations. Under a specified set of studied conditions, not all annotated genes will necessarily be expressed. Moreover, the shapes of the transcripts change dynamically according to the conditions through the regulation of transcription start sites, alternative splicing events, alternative polyadenylation events, mRNA degradation, and so on. Differences between the static gene structures and the real transcriptome may negatively affect dynamic analyses including co-expression analyses.

We might obtain very precise results from co-expression analyses if we can construct gene structures that directly reflect the real transcriptome of the studied RNA samples. We call such gene structures 'dynamic gene structures', in contrast to 'static gene structures'. In this study, we propose a new approach called 'dynamic structure-based dynamic expression (DSDE)' analysis (**Table 1**). With this approach, we predict specific dynamic gene structures that appear in the studied transcriptome and use them to predict gene functions. New technologies such as genome tiling arrays and RNA-Seq have progressed recently, enabling us to choose the DSDE approach. Our particular DSDE approach contains two modules: a gene model construction method, ARabidopsis Tiling-Array-based Detection of Exons version 2 (ARTADE2), and a function prediction method, ORA; thus, this approach is named as 'ARTADE2-ORA'.

Here we show that our DSDE analysis provides much more precise function predictions than those provided by SSDE analysis. Moreover, the method provides function predictions for most of the functional unknown genes of *A. thaliana*. Importantly, our dynamic gene structures contain more than a thousand novel gene candidates, and DSDE can predict their functions. The constructed dynamic gene structures and their predicted functions are now served from our database ARTADE2DB. In this paper, we describe the way to browse ARTADE2DB and show that the contents found in this database may contribute to functional genomics.

Results and Discussion

Gene models constructed with tiling arrays and mathematical methods

To realize DSDE approaches, we had to construct gene models that reflected transcriptome dynamism. In a previous study, we developed the program ARTADE that constructed gene models from tiling array results derived from RNA samples taken under a single set of conditions (Toyoda and Shinozaki 2005). The program constructs gene structures with the significance of gene expression signals and transition of a genome sequence on a Markov model. Although we were moderately successful in

constructing gene models, such models based on a single transcriptome are not suitable for studying multiple transcriptomes. To construct a basic set of gene models for dynamic expression analysis, we developed a second version of the program, ARTADE2 (S. Kawaguchi et al. in preparation, and announced at the 20th International Conference on Arabidopsis Research). This program uses probe-wise correlation values but does not use a directory of tiling array expression values. The probe-wise correlation values are calculated for each combination of tiling array probes by using vectors consisting of expression values of multiple tiling array experiments. With these correlation values and Markov model analysis, ARTADE2 constructed a single set of gene models from multiple transcriptomes that reflected transcriptional dynamism.

In this study, each tiling array sample has at least three times the number of biological repeats (see the Materials and Methods section). We obtained 55 sets of expression profiles with Affymetrix genome tiling arrays (Yamada et al. 2003, Zhang et al. 2006). For each experiment, a set of genome tiling arrays including R-chip and F-chip were used, which made it possible to obtain strand-specific expression profiles (Zhang et al. 2006). Using ARTADE2, we obtained 17,591 gene models. Of these, 16,120 genes have overlapping regions with annotated genes consisting of The Arabidopsis Information Resource version 9 (TAIR9) genes (Swarbreck et al. 2008), micro-RNAs (miRNAs) described in miRBase (Griffiths-Jones et al. 2008) or small nucleolar RNAs (snoRNAs) described in the plant snoRNA database (Brown et al. 2003). Therefore, we obtained 1,471 genes that are not described in the Arabidopsis gene databases including TAIR, miRBase and the plant snoRNA database, i.e. they were novel gene candidates. We note that another 18 genes were described only in the plant snoRNA database and not in the TAIR9 annotation set (**Supplementary Table S1**). The ARTADE2 gene models are named using the annotation OMAT*x*P*yzzzzz*, where OMAT indicates 'Omics-studies-based gene model of *Arabidopsis thaliana*', *x* is a number of a chromosome, *y* is an identifier of the gene direction (0, plus; 1, minus) and *z* is a number specifying gene models.

We suggest that tiling array-based analyses are useful even when RNA-Seq results are available. Although future progress of sequencing technologies will provide a complete set of full-length sequences of every mRNA, we can use only a set of fragmented cDNA reads at present. Several efforts have been made to observe the transcriptome with RNA-Seq (Lister et al. 2008, Filichkin et al. 2010). However, it is still difficult to determine reliable gene models from RNA-Seq. Tiling array analysis and our mathematical method ARTADE2 make it possible to predict gene models with a certain level of reliability (S. Kawaguchi et al. in preparation).

Dynamic structures improve ORA

We tried to select annotation terms that were enriched among the annotations of the co-expressed genes, i.e. a method called ORA; a similar approach was found in Alexa et al. (2006) and Grossmann et al. (2007). In ORA, we list 200 genes with the highest correlation values of gene expression against each tested gene and then count the genes that have terms from a gene ontology (GO) (Ashburner et al. 2000), plant ontology (PO) (Avraham et al. 2008) or gene definition described as a TAIR annotation (see Materials and Methods for details). We compared the fraction of genes with each term in the co-expressed genes and that in the entire gene set. When the fraction was significantly higher than that in the entire gene set, the term was described as a function prediction for the tested gene (described in detail in the Materials and Methods section). As a result, functional annotation terms are mapped to the selected genes. We performed ORA using both annotated gene models and dynamic structures. In particular, we refer to the combination of ARTADE2 and ORA as 'ARTADE2-ORA'.

We expect our ORA results to contribute to the community's efforts to describe gene functions in Arabidopsis. Therefore, the reliability of the ORA results should be thoroughly examined. To determine their reliability, we compared them with the GO/PO/annotation terms found in the original annotation of the gene itself. As described in the Materials and Methods section, we did not use any GO/PO/annotation terms found in the original gene annotations for ORA. Instead, we tested the enrichment of GO/PO/annotation terms in genes that are highly co-expressed with the tested genes, excluding the gene itself. In this test, when a gene has at least one GO/PO/annotation term in common with the original annotation, we treated the gene as a positive result. To summarize the results of this test, we classified genes into three categories: (A) well-annotated genes; (B) genes annotated on the basis of a

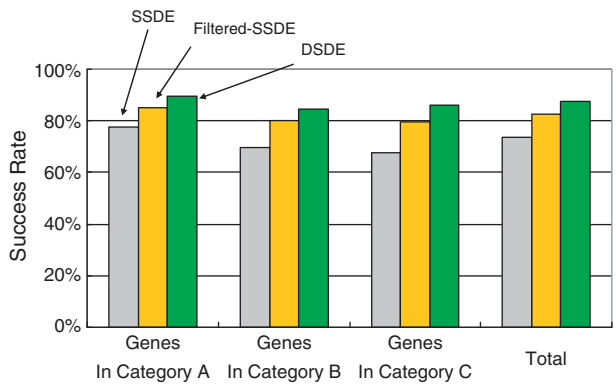

Fig. 1 Success rates of functional annotation terms of genes with ORA. In the graph, only GO terms were considered. Genes are categorized into four groups: genes with annotation (category A), genes annotated on the basis of similarities (category B), unknown genes (category C) and pseudogenes/transposable elements (data not shown). Summaries of gene categories A, B and C are also shown. Gray, yellow and green bars represent the results of annotated genes (SSDE), annotated genes with corresponding ARTADE2 gene models (filtered SSDE) and ARTADE2 gene models (DSDE), respectively. A similar graph drawn with the ORA results considering all of GO, PO and other annotation terms can be found in **Supplementary Fig. S1**.

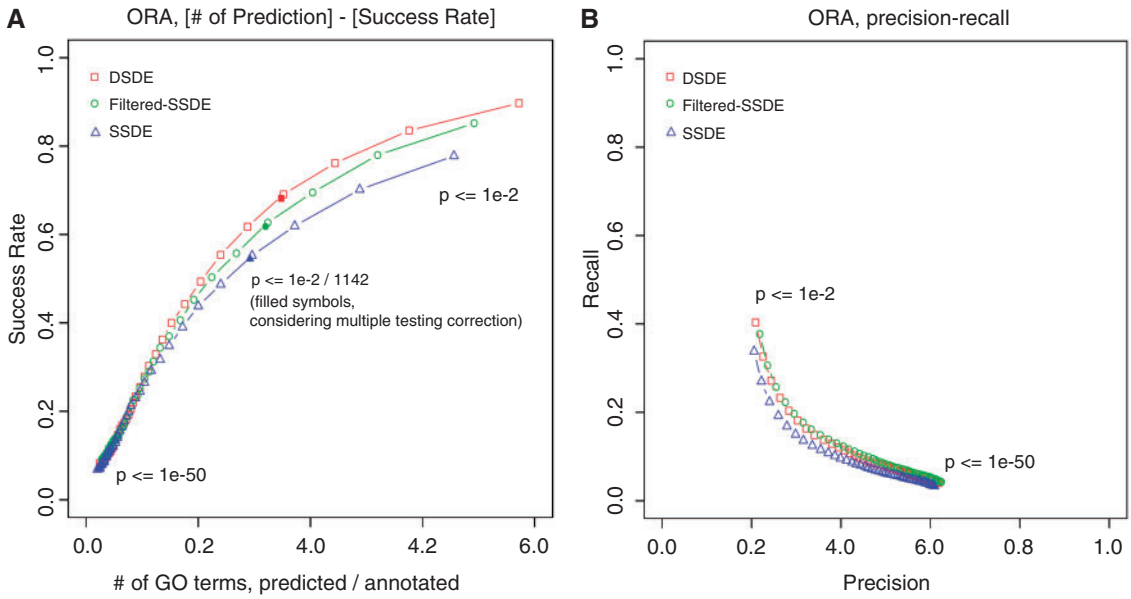

Fig. 2 (A) Success rate vs. number of predicted GO terms. The number of GO terms is shown with respect to the number of GO terms appearing in the annotation. The *P*-value thresholds for ORA range from 1e-2 (upper right) to 1e-50 (lower left). For DSDE, about twice as many GO terms are described as with ORA, and the result showed a success rate of >90%, with the most relaxed threshold. The success rate at the threshold is about 85 and 78% in the filtered SSDE and SSDE, respectively. In this study, 1,142 GO/PO/annotation terms were tested with ORA. We plotted the filled symbols to show the results with a *P*-value threshold; $P < 8.76e\text{-}06$ which corresponded to $P < 1e\text{-}2$ when considering multiple testing correction. (B) Precision Recall graph of DSDE, filtered SSDE and SSDE. Although all results were similar, DSDE and filtered SSDE showed slightly better performances. More detailed results used for the graphs are given in **Supplementary Table S1**.

similarity; and (C) unknown or hypothetical genes. First, we tested the ORA results for GO terms only. For category A, we obtained positive results for 8,446 annotated genes out of 10,863 genes tested (success rate = 77.8%, **Fig. 1**, **Supplementary Table S2**). This was the result of SSDE analysis because it was based on static gene structures. For comparison, we performed two other types of analysis: filtered SSDE that uses only static structures with corresponding dynamic structures, and DSDE. We found improved success rates for filtered SSDE (85.2%) and DSDE (89.7%) (**Fig. 1**, **Supplementary Table S2**). For the genes in category B, we obtained similar results to those for category A. The success rates were 69.4, 79.9 and 84.6% for SSDE, filtered SSDE and DSDE, respectively (**Fig. 1**, **Supplementary Table S2**). More than half of the genes in category C have no GO annotations (**Supplementary Table S2**). However, some may have GO annotations based on expression profiles or information about subcellular localization. ORA also showed a high success rate for the genes in category C. Overall, when we tested the genes in all categories, the success rates were 73.7, 82.8 and 87.6% for SSDE, filtered SSDE and DSDE, respectively (**Fig. 1**, **Supplementary Table S2**). Moreover, when we considered PO terms and other annotation terms found in TAIR annotations, the success rate became higher than when considering only GO terms (**Supplementary Fig. S1, Supplementary Table S3**). The results showed that ORA is a powerful way to reconstruct annotations based on GO/PO/annotation terms, and our ARTADE2 gene models can improve the predictions. Although our ARTADE2DB supports annotated gene models, we recommend referring information based on ARTADE2 gene models when it is available because of the high reliability of the ORA results. The details of assessing ORA results are described at **Supplementary Tables S4** (DSDE) and S5 (SSDE).

We selected ORA results that met the *P*-value threshold $P \leq 1e\text{-}2$. With this parameter, we obtained about twice as many annotation terms as there were GO terms in the original annotations. At the same time, the ORA results showed a success rate of about 90%. When we slid the thresholds, the number of predicted GO terms and the success rate decreased (**Fig. 2A**). The selected *P*-value ($P < 1e\text{-}2$) is a parameter with which we can expect to yield at least one true positive term within the predicted terms. In this study, we tested 1,142 GO/PO/annotation terms in total (see Materials and Methods for details). When considering an effect of multiple testing, a threshold for giving results with a significance level of $P < 1e\text{-}2$ becomes $P < 8.76e\text{-}06$ under the multiple testing correction. The results using the corrected threshold can be seen in **Fig. 2A**.

Even thoght DSDE showed the best results, the relationship between the numbers of predicted terms and success rates were similar in SSDE and filtered SSDE. **Fig. 2B** shows the relationship between precision ([# of truly predicted GO terms]/[# of GO terms predicted]) and recall ([# of truly predicted GO terms]/[# of GO terms found in annotations]) when we used sliding thresholds. This graph showed that DSDE and filtered SSDE had almost the same performance concerning precision

recall, and were slightly better than SSDE. This graph suggested that the improvement in ORA results in DSDE analysis was caused by filtering out the genes which were not expressed in the studied conditions. The small improvement found in DSDE analysis compared with filtered SSDE analysis may be caused by differences in exon–intron structures between these analyses. Although DSDE and filtered SSDE showed similar performances, we note that only the DSDE analysis can show novel gene candidates and assign function predictions to them (**Table 1**).

Our SSDE and DSDE analyses made several predictions about annotation terms for genes with unknown function. We can assign some function predictions to about 98% of the functionally unknown genes (**Supplementary Tables S4, S5**). Note that only DSDE described novel gene candidates and predicted their functions. We expect that the database can contribute to functional analysis of unknown genes.

Publishing ARTADE2DB

The ARTADE2DB has a gene information page for each of the 17,591 ARTADE2 gene models. The database also provides web pages for annotated gene models consisting of 39,361 TAIR9 gene models including alternative splicing variants and genes annotated as pseudogenes, 188 miRNA genes described in miRBase (Release 12) and 189 snoRNAs described in the plant snoRNA database (**Fig. 3**). A user can choose ARTADE2 gene models (DSDE) or annotated gene models (SSDE) and browse the list of all genes. Moreover, tables for annotated gene models with overlapping ARTADE2 gene models (filtered SSDE) and ARTADE2 gene models without corresponding annotated gene models were prepared for browsing a specified set of gene models.

A better way to browse the database is to use the search function. ARTADE2DB is published on the RIKEN database publication infrastructure, Scientist Networking Systems (SciNetS), and the SciNetS search engine developed with GRASE (General and Rapid Association Study Engine; Kobayashi and Toyoda 2008) also provides a search function for ARTADE2DB. Several types of keyword can be used to find gene models (**Fig. 4**), e.g. gene IDs, gene definitions and other information and terms explaining the gene models, which are connected by semantic links (described in detail below).

Users can examine details by following the links from each gene name. The web page for each gene model provides information including gene positions (**Supplementary Fig. S2A**), related information about the annotated gene model (**Supplementary Fig. S2B**), expression profile data (**Supplementary Figs. S2C, S2D**) and other computational analysis results including function prediction.

Contents of the database (1): correlation plots

As described above, ARTADE2 constructs gene models on the basis of correlations among tiling array probes. ARTADE2DB provides images of probe-wise correlation plots or correlation plots for each gene model (**Fig. 5**). The correlation plot shows the Pearson's correlation coefficient (PCC) between each tiling array probe located on the locus. A series of expression values from 55 tiling array experiments was used to calculate the PCC (see the Materials and Methods for details). ARTADE2 constructed gene models based on these PCC values. The PCC tends to be high between two probes located on exon regions of ARTADE2 genes. On the other hand, for probe pairs with a

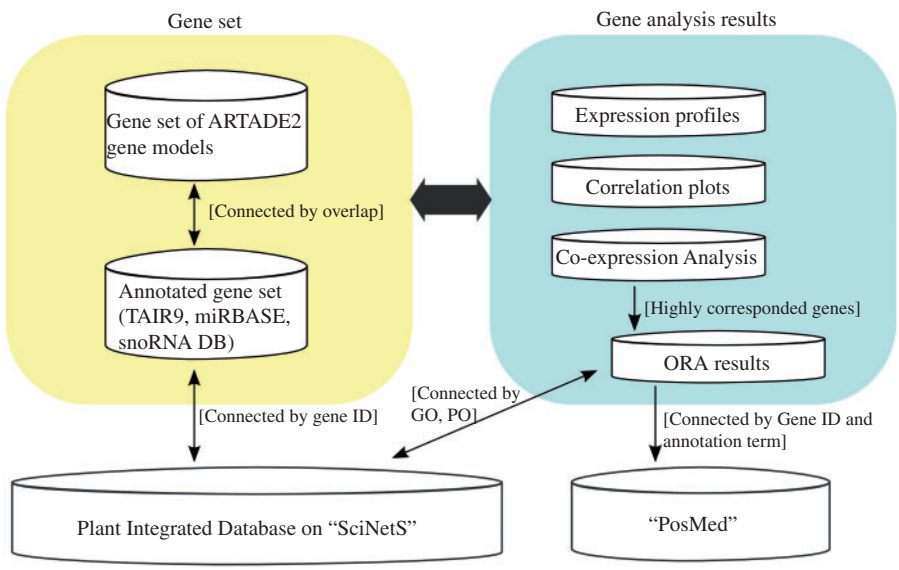

Fig. 3 Schema of information stored in ARTADE2DB. We have two sets of gene models: a set of ARTADE2 gene models and a set of annotated gene models. The two sets are connected on the basis of their overlap. Each gene model in either set contains information about the expression profile, the correlation plot, a list of co-expressed genes and the ORA result. Annotated gene models and GO or PO terms act as gates to information stored in SciNetS.

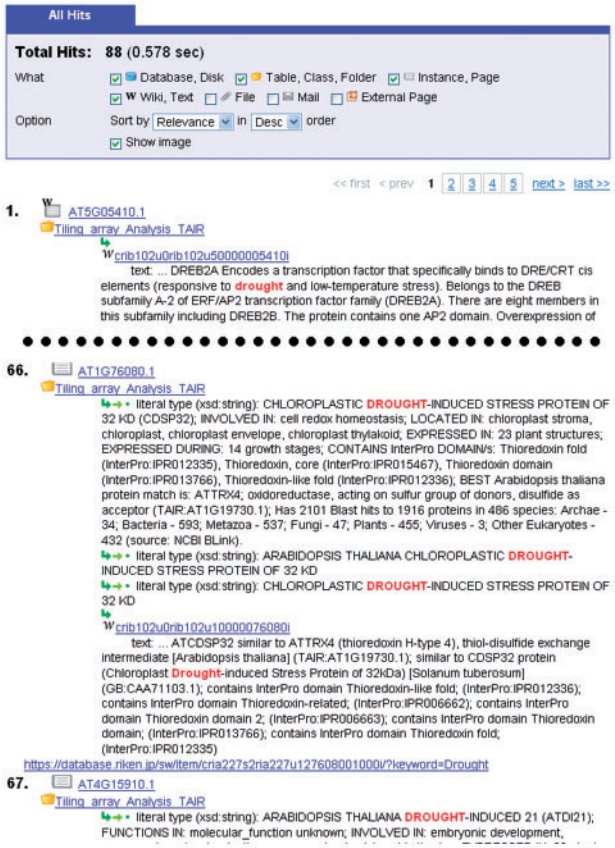

Fig. 4 Example of search results with the query term 'drought'. The database search engine locates gene models containing the query term. Terms appearing in entry descriptions that have semantic links against the gene models can be query words.

low PCC, one or both of the probes are not located on exons or they are located on different transcription units. Even if they are located on exons of the same gene, a low PCC is possible if the gene is not expressed in the examined conditions.

A correlation plot of tiling array experiments is a powerful way to find transcription units. Here we describe what the correlation plots show. One example can be found on the locus for putative primary miRNA (Kurihara and Watanabe 2004). The locus of MIR156A, which has a pre-miRNA length of about 120 bases, includes a region covered by high PCC plots (**Supplementary Fig. S3A**). This region corresponds to the ARTADE2 gene model named OMAT2P104540 with a length of about 1,700; this model fully covers the region of pre-miRNA for MIR156A. The gene model seems to be a primary RNA for this miRNA. A similar situation is found on the gene locus for OMAT5P014680 (**Supplementary Fig. S3B**). A region with high PCC plots found at this locus may be the primary RNA for the snoRNA cluster from which AtsnoR29-1, AtsnoR30 and AtsnoR31 are processed. As shown here, the correlation plot reveals transcribed regions with high sensitivities. This method also makes it possible to survey genes that were not described previously.

The correlation plots also give some suggestions about alternative splicing. One example can be found on the locus for AT3G53500, which encodes a Ser/Arg-rich (SR) protein member, RSZ32, which is reported to have an alternative splicing event (Iida and Go 2006). The alternatively spliced region is located on the 5′ part of the third exon of AT3G53500.1 (shown with red lines in **Supplementary Fig. S4**), whose tiling array probes have a high PCC. Other constitutive exons also showed high PCC values among the probes within the region. However, the PCC values between probes located on the alternatively spliced region and the constitutive exons are relatively low. Thus, these two regions have different expression profiles, i.e. this correlation plot suggests an alternative splicing event here.

Contents of the database (2): ORA

In this study, GO/PO/annotation terms with *P*-values for ORA of $<$1e-2 were stored as function prediction results. Of these, the top 20 terms are described in the middle of the web page (**Fig. 6**). They have semantic links to other content stored on RIKEN SciNetS via GO/PO terms. All ORA results, including the top 20 results, are described at the bottom of the web page (**Supplementary Fig. S2F**). A table includes additional features such as links to Positional Medline (PosMed) *P*-values (Makita et al. 2009, Yoshida et al. 2009) and flags showing whether the GO/PO/annotation terms are described in the original gene annotations. Prediction results which are supported by both ARTADE2-ORA and PosMed may be very reliable, because PosMed evaluates the relationship between the genes and annotation terms in a different way from the ARTADE2-ORA. In ARTADE2DB, when a paired gene and GO/PO/annotation term have a PosMed *P*-value of $<$1e-4, the PosMed *P*-value and a link to the PosMed system are displayed. Details can be found on the PosMed web page.

A gene model's web page describes the top 10 genes with the highest PCC relative to the selected gene. These are some of the genes used in ORA. Users can trace the co-expressed genes via these links. The web page also shows the bottom 10 genes with the highest negative PCCs, which helps clarify the genes' relationships with respect to negative transcription factors and miRNA.

Examples of positive results of ARTADE2-ORA

We showed through detailed examples that the ORA analysis works well. We classified Arabidopsis genes based on TAIR9 annotation. However, gene functions are cleared up every day. Therefore, some genes that are described as an unknown gene have not yet been characterized well; the gene AT2G30280 is one of the best examples. This gene is currently named RNA-directed DNA methylation4 (RDM4) (He et al. 2009), but it was annotated as an unknown protein in TAIR9. This gene is reported to be involved in epigenetic silencing of transposons via cytosine methylation. The authors showed that small RNAs requiring RNA-directed DNA methylation were

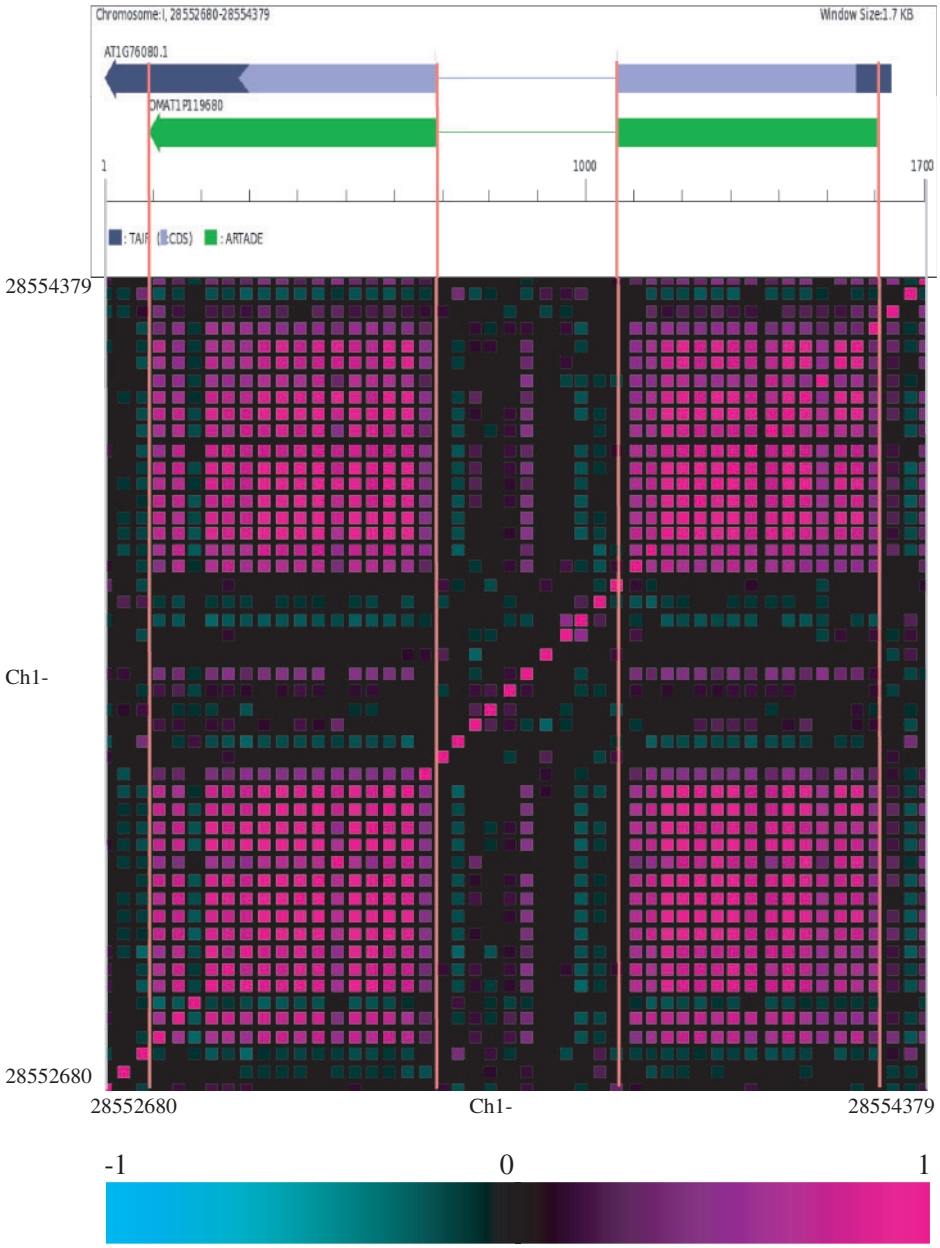

Fig. 5 An example of a dynamic gene structure and supporting correlation plot. In the upper figure, exon–intron structures of a dynamic gene model (OMAT1P119680) and the overlapping TAIR gene model (AT1G76080.1) are drawn. Boxes show exons and lines show introns. The light blue region on the TAIR gene model shows the CDS region. Arrows indicate the directions of the gene models. The lower figure is a correlation plot on the locus. The upper and lower figures share the *x*-axis. In the correlation plot, positive probe-wise correlation values are shown in red and negative values in blue. We found high positive correlation values within the first and second exons of the ARTADE2 gene model. In addition, correlation between probes located on the first and second exon is also high. Similar figures are available at the gene information page of ARTADE2DB.

decreased in *rdm4* mutant plants. Our ARTADE2-ORA results for the ARTADE2 gene model OMAT2P106260, which corresponds to this gene, includes the following GO terms: GO:0006396 (RNA processing, $P = 9.78e\text{-}10$), GO:0016070 (RNA metabolic process, $P = 4.69e\text{-}5$), GO:0010467 (gene expression, $P = 1.18e\text{-}4$) and many other GO terms suggesting that this gene is involved in RNA processing (**Supplementary Table S6A**). We think our database might be the most powerful tool when combined with a genetics approach, as in this example. Other good examples can be seen at the web page for OMAT1P108150 (AT1G26110), which is reported to be the mRNA decapping-related gene DCP5 (Xu and Chua 2009) and the web page for OMAT2P009840 (AT2G37940), which is reported to involve programmed cell death associated with defense (Wang et al. 2008). Both gene models have ARTADE2-ORA results that are very consistent with their

Relative_ontology_term

ontology_analysis_result

	ontology_analysis_result (Tiling_array_Analysis_TAIR)	in_quantity_of	in_organ_named	has_ontology	other_ontology	ORA_P_value	Posmed_P_value
chloroplast part (Relative_ontology_term)	AT1G76080.1			chloroplast part (Gene Ontology)		4.11E-93	
plastid part (Relative_ontology_term)	AT1G76080.1			plastid part (Gene Ontology)		5.55E-92	
thylakoid (Relative_ontology_term)	AT1G76080.1		thylakoid (Gene Ontology)			5.56E-91	6.75E-18
plastid (Relative_ontology_term)	AT1G76080.1			plastid (Gene Ontology)		2.14E-85	3.64E-21
thylakoid part (Relative_ontology_term)	AT1G76080.1			thylakoid part (Gene Ontology)		4.12E-83	
plastid thylakoid (Relative_ontology_term)	AT1G76080.1			plastid thylakoid (Gene Ontology)		1.13E-80	5.82E-19
organelle subcompartment (Relative_ontology_term)	AT1G76080.1			organelle subcompartment (Gene Ontology)		1.74E-80	
thylakoid membrane (Relative_ontology_term)	AT1G76080.1		thylakoid membrane (Gene Ontology)			2.1E-76	
photosynthetic membrane (Relative_ontology_term)	AT1G76080.1			photosynthetic membrane (Gene Ontology)		5.07E-76	5.46E-18
plastid thylakoid membrane (Relative_ontology_term)	AT1G76080.1			plastid thylakoid membrane (Gene Ontology)		6.84E-76	
cytoplasmic part (Relative_ontology_term)	AT1G76080.1			cytoplasmic part (Gene Ontology)		1.19E-57	
intracellular organelle part (Relative_ontology_term)	AT1G76080.1			intracellular organelle part (Gene Ontology)		1.07E-55	
organelle part (Relative_ontology_term)	AT1G76080.1			organelle part (Gene Ontology)		1.16E-55	
cytoplasm (Relative_ontology_term)	AT1G76080.1			cytoplasm (Gene Ontology)		1.45E-53	
				intracellular			

Fig. 6 An example of ORA results (OMAT1P119680). GO and PO are classified as belonging to the rule of the RIKEN SciNetS. Several database constructed on the SciNetS share the data format which may help data integration in the future. Added to ORA *P*, PosMed *P* are described. The table is available at the gene information page of the ARTADE2DB. Other examples of the ARTADE2DB contents can found in **Supplementary Fig. S2**.

reported functions (**Supplementary Tables S6B, C**). Especially for the latter example, the PosMed system contributes in highlighting the presumed GO terms in several ORA results.

Contents of the database (3): cooperation with phenome data

One of the advantages of our ARTADE2DB is that the database was constructed as part of the semantic web (Berners-Lee and Hendler 2001). Users can find Resource Description Framework (RDF) files via links at the top of the web pages. These RDF files (**Supplementary Fig. S1G**) are suitable for computational analysis. Researchers who use computational methods to survey the database can use these RDF files. The semantic web provides not only the RDF files but also semantic links to information stored in SciNetS that is helpful for understanding the genes. Semantic links indicate the relationship between linked objects and also connect them. As described above, semantic links via GO/PO terms are an example. Recently, we also developed an integrated plant database on SciNetS. SciNetS stores information about Arabidopsis genes such as *Ac*/*Ds* transposon mutant lines and phenome study results (Sakurai et al. 2005, Kuromori et al. 2009), full-length Arabidopsis cDNA (Seki et al. 2002) and other data. These related data resources can help to enrich genetic knowledge. Such semantic links are centralized into TAIR gene models. ARTADE2 gene models were connected with this information via semantics links to the TAIR gene models.

We now show an example where OMAT1P110110, one of the ARTADE2 gene models, corresponds to AT1G32080, which is annotated as 'membrane protein, putative'. Users can find semantic links to entries in the RIKEN Arabidopsis Phenome Information Database (**Supplementary Fig. S5A**). *Ds* transposon mutant lines 15-5500-1 and 16-1202-1, which break this gene, showed desaturated green leaves or multicolored leaves (**Supplementary Fig. S5B, C**; a wild-type plant is shown in S5D as a control). ORA results described in ARTADE2DB (**Supplementary Figs. S5E**) suggest that this gene is related to photosynthesis. Integration of information in the SciNetS is useful for surveying information associated with a gene.

Summary

We showed that our DSDE approach, in which we sequentially constructed gene models on the basis of multiple transcriptome measurements and predicted gene functions, can greatly improve gene function predictions. In addition to reconstructing previously described information, our studies provide function predictions for most of the functionally unknown genes and genes annotated only on the basis of similarities to *A. thaliana*. Furthermore, we identified more than a thousand novel gene candidates and predicted their functions. When studying gene functions, obtaining the initial trigger to narrow down the possible gene functions is important and often difficult. ARTADE2DB is designed to provide such triggers for functional genomics. We expect that ARTADE2 gene

models, correlation plots and function predictions made by using ORA should provide useful information for researchers.

Materials and Methods

Whole-genome tiling array experiments

In this study, we used expression profiles obtained with the GeneChip Arabidopsis tiling array set (1.0F Array and 1.0R Array, Affymetrix). The data set consists of 55 pairs of 1.0F and 1.0R arrays which were used to observe transcriptomes under 18 different conditions, comprising nine stress-related conditions and nine organs (**Supplementary Table S7**). The data sets for drought stress, high-salinity stress, cold stress, ABA treatment, dry seed and 24 h imbibed seed used were as previously reported in Matsui et al. (2008) and Okamoto et al. (2010). To obtain approximately 3-week-old leaf and root, plants were grown on a 0.5% gelangum (Wako) containing 0.5% sucrose and 1/2 MS at 22°C under continuous light. In the case of sampling for other tissues (stem, flower, early silique, middle silique and late silique), plants were grown on soil in pots at 22°C under a 16 h light/8 h dark cycle. Total RNA was isolated using ISOGEN (Nippon gene) or an RNAqueous Kit (Qiagen). These whole-genome tiling array results are available at GEO (http://www.ncbi.nlm.nih.gov/geo/info/linking.html) under the accession numbera GSE9646, GSE15700 and GSE26074.

Data set of annotated gene models

For the annotated gene set, we used TAIR9 genes (39,361 gene models on 33,239 loci), 188 miRNA genes described in miRBase (Release 12) (Griffiths-Jones et al. 2008) and 189 snoRNAs described in the plant snoRNA database (Brown et al. 2003). In the database, each genomic position is transformed into those counted on the TAIR8 genome because that genome is more suitable for a genome tiling array design than the TAIR9 genome. We treated a total of 39,738 annotated gene models in the database.

Preprocessing of tiling array data

We compared nucleotide sequences of perfect match (pm) probes of genome tiling arrays with the Arabidopsis genome sequences (from TAIR, version 8). For each probe, when we found a unique genomic locus whose sequence completely matched with the pm probe sequence, we defined the genomic position as the position of the pm probe and the probe was used for following analysis. After removing the spatial effect within a single array with the NMPP program (Wang and He 2006), the differential intensity between pm and mm (mismatch) was calculated in each position with the MAS5 (Affymetrix Microarray Analysis Suite v5.0) algorithm (Li and Wong 2001) in the R software environment (http://www.r-project.org). Tukey biweight values required as background correction in this procedure were calculated in each chromosome with the Bioconductor (http://www.bioconductor.org) affy package. The normalized quantiles function in the R software was used for between-array normalization.

Calculating probe-wise correlation values

As a result of the pre-processing of genome tiling array data, expression values were located on 25 base regions on the chromosomes where pm probes were perfectly mapped. Then we made vectors for each mapped pm probe which consisted of expression values of 55 genome tiling array experiments. PCC was then calculated for pairs of pm probe positions, which were called 'probe-wise correlation values'. The values are used for the following manipulations, and were used for drawing the correlation plots (**Fig. 5**, **Supplementary Fig. S3, S4**).

Constructing dynamic gene models

We constructed dynamic gene models (or the ARTADE2 gene models) using ARTADE2 which integrate the probe-wise correlation values and the genome sequence score from the Markov model constructed with a set of full-length cDNAs. Details of score functions based on the Markov model have been described previously (Toyoda and Shinozaki 2005). A total of 2,813 full-length cDNAs (Seki et al. 2004) which were mapped on the plus strand of chromosome 1 were used for estimation of parameters on the ARTADE2 model. With the method, we surveyed genomic segments where probe-wise correlation values were considerably high. Then, genomic regions around the segment were evaluated with the probe-wise correlation values and the Markov model to generate an exon–intron structure. Details of the method will be published elsewhere (S. Kawaguchi et al. in preparation). A list describing the locations of every ARTADE2 gene model is found at the URL: https://database.riken.jp/sw/links/en/cria227s904-ria227s7i/.

Dynamic gene models are classified into three groups based on their relationship with the annotated gene models. When a dynamic gene model has at least one overlapping annotated gene model, the dynamic gene model is classified as a 'known gene'. When an overlapping gene is found on the opposite strand of an annotated gene model, the dynamic gene model is classified as an 'antisense gene'. When a dynamic gene model has no annotated gene models overlapping, the gene model is classified as a 'novel gene'. To check the overlap between gene models, we compared the start and end positions of two gene models. When the overlapping region accounted for >30% of the length of either gene model, we treated the two gene models as overlapping.

Calculating expression values of the gene models

For both static and dynamic gene model sets, we calculated expression values per gene model, per experiment. We collected pm probes located within exon regions of each gene

model, and then we calculated a median of the expression signals of pm probes as a representative expression value of the gene model. These expression values of the gene models were used for calculating PCC between gene models.

Over-representation analysis (ORA)

Let N be the number of all genes and M be the frequency of genes associated with a certain GO, PO or TAIR9 annotation term. Then, we select the 200 highest PCC gene subsets for each gene and count the frequency m of genes associated with the term in the subset. Consequently, the P-values of genes associated with the term are given by $P = p(m)$, where p follows the hypergeometric distribution H_G (N,M,200) and m is the frequency. When a P-value was below the threshold (1e-2), the term was described as a result. GO and PO have hierarchical structures; if genes have GO/PO terms at different levels of the hierarchy, the prediction power is decreased. To avoid this problem, we transfer the GO/PO terms of every term into the upper hierarchy. For example, when a gene has a GO of GO:0016070; 'RNA metabolic process', this gene is treated as having an upper GO of GO:0090304; 'nucleic acid metabolic process' and the next upper GO:0006139; 'nucleobase, nucleoside, nucleotide and nucleic acid metabolic process'. We counted GO/PO hierarchy trees from GO:0008150(biological process), GO:0005575(cellular component), GO:0003674 (molecular function), PO:0009012(plant growth and development stages) or PO:0009011(plant structure). They are counted as the first rank of GO/PO. We used from the third to the fifth GO/PO terms on ORA. In addition, we defined that GO/PO terms used in ORA must be described at at least one hundred TAIR genes. Finally, we chose 284, 72, 94, 65 and 18 terms for the GO/PO classes as the order described above, respectively.

For SSDE analysis and filtered SSDE analysis, we listed 200 representative TAIR gene models which have the highest gene correlation values against a selected gene. In the case of DSDE, we listed 200 ARTADE2 genes with the highest gene correlation values, then they were replaced with overlapped TAIR genes. ARTADE2 genes without overlapping TAIR genes were ignored. When a single ARTADE2 gene model has multiple overlapping TAIR9 genes, we chose the longest overlapping one.

For ORA with annotation terms, we counted each term which appeared on the gene descriptions by TAIR. Similar to the ORA using GO/PO, we defined that the term must be used at at least one hundred TAIR genes. With this rule, we selected 609 annotation terms. In total, 1,142 GO/PO/annotation terms were used at ORA. With consideration of an effect of multiple testing, we employed Bonferonni correction. Because we tested 1,142 GO/PO/annotation terms, a threshold for giving results with a significance level $P < 1e\text{-}2$ was $P < 8.76e\text{-}06$ under the correction.

Assessing the ORA results

ORA was performed for every static and dynamic gene structure. To assess the results, we prepared three sets of analyses groups: SSDE, filtered SSDE and DSDE. The first set is ORA results for 27,641 representative TAIR gene models, called the SSDE results. Genes annotated as transposable elements are excluded from this set. The second is filtered SSDE, which is a subset of SSDE. TAIR gene models with overlapping ARTADE2 gene models are put into this group. Definitions of 'overlapping' are described above. Some ARTADE2 gene models had sufficient overlap with multiple TAIR gene models, and thus the numbers of gene models manipulated in the filtered SSDE set (15,889) are larger than those in the DSDE set (15,887). The last set is the DSDE set containing all ARTADE2 gene models. To approach details of ORA results, we classified gene models into three categories based on gene description annotated by TAIR. Category A has gene models whose functions are well characterized and have clear gene names. Category B has gene models whose function are described based on sequence similarities. For example -family proteins, -protein like or -domain-containing proteins were classified here. Genes for unknown proteins or hypothetical proteins, or any other genes without clear definitions were classified into category C. This categorization is defined based on TAIR9 gene models. For ARTADE2 gene models, we transferred categories of overlapping TAIR9 genes into ARTADE2 gene models. When a single ARTADE2 gene model has multiple overlapping TAIR9 genes, we chose the longest overlapping one.

Supplementary data

Supplementary data are available at PCP online.

Funding

This work was supported by the the Ministry of Education, Culture, Sports, Science and Technology, Japan [Research Program of Innovative Cell Biology by Innovative Technology (T.T.)].

Acknowledgments

We thank Ms. Yukiko Kanda and Dr. Yuko Makita for the artwork used in ARTADE2DB.

References

Alexa, A., Rahnenfuhrer, J. and Lengauer, T. (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 2213: 1600–1607.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389–3402.

Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408: 796–815.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25: 25–29.

Avraham, S., Tung, C.W., Ilic, K., Jaiswal, P., Kellogg, E.A., McCouch, S. et al. (2008) The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. *Nucleic Acids Res.* 36: D449–D454.

Berners-Lee, T. and Hendler, J. (2001) Publishing on the semantic web. *Nature* 410: 1023–1024.

Brown, J.W., Echeverria, M., Qu, L.H., Lowe, T.M., Bachellerie, J.P., Hüttenhofer, A. et al. (2003) Plant snoRNA database. *Nucleic Acids Res.* 31: 432–435.

Filichkin, S.A., Priest, H.D., Givan, S.A., Shen, R., Bryant, D.W., Fox, S.E. et al. (2010) Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Res.* 20: 45–58.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E. et al. (2010) The Pfam protein families database. *Nucleic Acids Res.* 38: D211–D222.

Griffiths-Jones, S., Saini, H.K., van Dongen, S. and Enright, A.J. (2008) miRBase: tools for micro-RNA genomics. *Nucleic Acids Res.* 36: D154–D158.

Grossmann, S., Bauer, S., Robinson, P.N. and Vingron, M. (2007) Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics* 23: 3024–3031.

He, X.J., Hsu, Y.F., Zhu, S., Liu, H.L., Pontes, O., Zhu, J. et al. (2009) A conserved transcriptional regulator is required for RNA-directed DNA methylation and plant development. *Genes Dev.* 23: 2717–2722.

Iida, K. and Go, M. (2006) Survey of conserved alternative splicing events of mRNAs encoding SR proteins in land plants. *Mol. Biol. Evol.* 23: 1085–1094.

Kobayashi, N. and Toyoda, T. (2008) Statistical search on the Semantic Web. *Bioinformatics* 24: 1002–1010.

Kurihara, Y. and Watanabe, Y. (2004) Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. *Proc. Natl Acad. Sci. USA* 101: 12753–12758.

Kuromori, T., Takahashi, S., Kondou, Y., Shinozaki, K. and Matsui, M. (2009) Phenome analysis in plant species using loss-of-function and gain-of-function mutants. *Plant Cell Physiol.* 50: 1215–1231.

Li, C. and Wong, W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA* 98: 1–6.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H. et al. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* 133: 523–536.

Makita, Y., Kobayashi, N., Mochizuki, Y., Yoshida, Y., Asano, S., Heida, N. et al. (2009) PosMed-plus: an intelligent search engine that inferentially integrates cross-species information resources for molecular breeding of plants. *Plant Cell Physiol.* 50: 1249–1259.

Matsui, A., Ishida, J., Morosawa, T., Mochizuki, Y., Kaminuma, E., Endo, T.A. et al. (2008) Arabidopsis transcriptome analysis under drought, cold, high-salinity and ABA treatment conditions using a tiling array. *Plant Cell Physiol.* 49: 1135–1149.

Obayashi, T., Hayashi, S., Saeki, M., Ohta, H. and Kinoshita, K. (2009) ATTED-II provides coexpressed gene networks for Arabidopsis. *Nucleic Acids Res.* 37: D987–D991.

Okamoto, M., Tatematsu, K., Matsui, A., Morosawa, T., Ishida, J., Tanaka, M. et al. (2010) Genome-wide analysis of endogenous abscisic acid-mediated transcription in dry and imbibed seeds of Arabidopsis using tiling arrays. *Plant J.* 62: 39–51.

Sakurai, T., Satou, M., Akiyama, K., Iida, K., Seki, M. and Kuromori, T. (2005) RARGE: a large-scale database of RIKEN Arabidopsis resources ranging from transcriptome to phenome. *Nucleic Acids Res.* 33: D647–D650.

Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T. et al. (2002) Functional annotation of a full-length Arabidopsis cDNA collection. *Science* 296: 141–145.

Srinivasasainagendra, V., Page, G.P., Mehta, T., Coulibaly, I. and Loraine, A.E. (2008) CressExpress: a tool for large-scale mining of expression data from Arabidopsis. *Plant Physiol.* 147: 1004–1016.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T.Z., Garcia-Hernandez, M., Foerster, H. et al. (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* 36: D1009–D1014.

Toyoda, T. and Shinozaki, K. (2005) Tiling array-driven elucidation of transcriptional structures based on maximum-likelihood and Markov models. *Plant J.* 43: 611–621.

van Noort, V., Snel, B. and Huynen, M.A. (2003) Predicting gene function by conserved co-expression. *Trends Genet.* 19: 238–242.

Wang, X.F. and He, H. (2006) NMPP: a user-customized NimbleGen Microarray Data Processing Pipeline. *Bioinformatics* 22: 2955–2957.

Wang, W., Yang, X., Tangchaiburana, S., Ndeh, R., Markham, J.E., Tsegaye, Y. et al. (2008) An inositolphosphorylceramide synthase is involved in regulation of plant programmed cell death associated with defense in Arabidopsis. *Plant Cell* 20: 3163–3179.

Xu, J. and Chua, N.H. (2009) Arabidopsis decapping 5 is required for mRNA decapping, P-body formation, and translational repression during postembryonic development. *Plant Cell* 21: 3270–3279.

Yamada, K., Lim, J., Dale, J.M., Chen, H., Shinn, P., Palm, C.J. et al. (2003) Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science* 302: 842–846.

Yoshida, Y., Makita, Y., Heida, N., Asano, S., Matsushima, A., Ishii, M. et al. (2009) PosMed (Positional Medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning. *Nucleic Acids Res.* 37: W147–W152.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W., Chen, H. et al. (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell* 126: 1189–1201.