Learning sequence patterns of AGO-sRNA affinity from high-throughput sequencing libraries to improve *in silico* functional small RNA detection and classification in plants

Lionel Morgado[1,*], Ritsert C. Jansen[1] and Frank Johannes[2,3]

[1] Groningen Bioinformatics Centre, University of Groningen, Groningen, 9747 AG Groningen, The Netherlands
[2] Department of Plant Sciences, Technical University of Munich, Freising, 85354 Freising, Germany
[3] Institute for Advanced Study, Technical University of Munich, Garching, 85748 Garching, Germany
* To whom correspondence should be addressed. Email: lionelmorgado@gmail.com

ABSTRACT

The loading of small RNA (sRNA) into Argonaute (AGO) complexes is a crucial step in all regulatory pathways identified so far in plants that depend on such non-coding sequences. Important transcriptional and post-transcriptional silencing mechanisms can be activated depending on the specific AGO protein to which sRNA bind. It is known that sRNA-AGO associations are at least partly encoded in the sRNA primary structure, but the sequence features that drive this association have not been fully explored. Here we train support vector machines (SVM) on sRNA sequencing data obtained from AGO-immunoprecipitation experiments to identify features that determine sRNA affinity to specific AGOs. Our SVM reveal that AGO affinity is strongly determined by complex k-mers in the 5' and 3' ends of sRNA, in addition to well-known features such as sRNA length and the base composition of the first nucleotide. Moreover, we find that these k-mers tend to overlap known transcription factor (TF) binding motifs, thus highlighting a close interplay between TF and sRNA-mediated transcriptional regulation. We embedded the learned SVM in a computational pipeline that can be used for de novo functional classification of sRNA sequences. This tool, called SAILS, is provided as a web portal accessible at http://sails.eu.nu.

INTRODUCTION

The small RNA (sRNA) is a class of non-coding RNA with significant roles in developmental biology, physiology, pathogen interactions, and more recently in genome stability and transposable element control (1). Plants have two major classes of sRNA: the micro-RNA (miRNA), which is processed from imperfectly self-folded hairpin precursors derived from miRNA genes (2); and the small-interfering RNA (siRNA) that is produced from double-stranded RNA duplexes (3) (Fig. 1). siRNAs can be further divided into three major groups: secondary siRNA such as trans-acting (ta)-siRNAs, which are promoted by miRNA cleavage of messenger RNA; natural antisense transcript (nat)-siRNAs derived from the overlapping regions of antisense transcript pairs naturally present in the genome; and heterochromatin-associated (hc)-siRNAs, mostly generated from transposons, heterochromatic and repetitive genomic regions and involved in DNA methylation and heterochromatin formation.

Apart from their biogenesis, the mode of action of a given sRNA is tightly related with the Argonaute protein to which it can bind (4). Argonautes form the core of all sRNA-guided silencing complexes

identified so far. Once loaded into Argonaute, sRNA guide the silencing machinery to targets through base pairing principles. Argonautes are highly conserved proteins with family members in most eukaryotes (4-6). Although there are two main subfamilies of Argonautes in eukaryotes: AGO and PIWI; only AGO proteins can be found in plants. Also in plants, AGOs can be grouped into three phylogenetic clades with a highly variable number of elements from species to species (5) (Fig. 2). Arabidopsis has ten members with specialized or redundant functions among them: AGO1, AGO5 and AGO10 in the first clade; AGO2, AGO3 and AGO7 form the second clade; and the third clade is composed by AGO4, AGO6, AGO8 and AGO9 (4). Members of the first and second clade are involved in post-transcriptional silencing (PTS) by inhibiting translation or by promoting messenger RNA cleavage, and AGOs in the third clade are chromatin modifiers that induce transcriptional silencing (TS) via mechanisms such as DNA methylation (7-9). Understanding the mechanisms that determine the loading of sRNA to specific AGOs is essential for predicting their biological function, and for identifying their putative silencing targets.

High throughput sequencing in combination with immunoprecipitation (IP) techniques have made possible to determine the sequences of sRNA that are bound to different AGO families. AGO-IP experiments have been performed for AGO1, AGO2, AGO4, AGO5, AGO6, AGO7, AGO9 and AGO10. The low expression level of AGOs 3 and 8 suggests that they may not be functionally relevant. Previous analyses have shown that AGO-sRNA associations are partly determined by the 5' terminus and the length of a sRNA sequence (10, 11). Sequences of approximately 21 nucleotides (nt) tend to be involved in PTS, while 24 nt sRNAs are characteristic of TS. Although sequence length has been widely used as a way to infer the silencing pathway a given sRNA is most likely implicated in, by itself it is an inaccurate predictor since many sequencing products can lack other structural features known to enable AGO loading (10-13). Contrary to animals and flies, in plants the 5' nucleotide is also recognized as a strong indicator of AGO sorting. Enrichment for sequences starting with pyrimidines are frequent in AGOs from the first clade (AGO1/10: uridine and AGO5: cytosine), adenosine dominates the third clade (AGO4/6/9) as well as AGO2 from the second clade. Furthermore, direct experimental evidence show that mutating the 5' nucleotide of a sRNA can redirect its AGO destination (14); however, other relevant features, such as sequence motifs encoded by the primary structure appear to play a role (7, 15-18).

While sRNA that participate in PTS have been intensively studied, leading to the discovery of many structural features that influence activation, much less is known about AGO-associated hc-sRNA in transcriptional silencing. Indeed, most studies that use hc-sRNAs give a strong emphasis to sequence length ignoring other important aspects of the sRNA sequence that promote an active role in genomic regulation. Studying AGO-bound sRNA is a starting point to fill this gap and to improve our understanding on the relationship between the structure and function of sRNA in plants. Here we train support vector machines (SVM) on sRNA sequencing data obtained from AGO-IP experiments to identify features that determine sRNA affinity to specific AGOs. Our SVM reveal that AGO affinity is strongly determined by complex k-mers in the 5' and 3' ends of sRNA, in addition to well-known features such as sRNA length and the base composition of the first nucleotide. Moreover, we find that these k-mers tend to overlap known transcription factor (TF) binding motifs, thus highlighting a

close interplay between TF and sRNA-mediated transcriptional regulation. We incorporated the learned SVM in an online computational pipeline that can be used for *de novo* sRNA functional classification. The classification pipeline is suitable for individual sRNA but also for high-throughput sRNA-seq datasets.

MATERIAL AND METHODS

Data sets

Table 1 summarizes the *A. thaliana* deep-sequencing sRNA libraries used in this study. For SVM model training and testing we used one Col-0 wild-type sRNA library as well as eight AGO-IP datasets. SVM model validation was afterwards performed on additional AGO-IP data from different tissues and from pathogen infected plants, in addition to a set of putative ta-siRNAs from several plant species. All sRNA-seq datasets were pre-processed and mapped to the Col-0 *A. thaliana* reference genome. Reads with at least one perfect match were collapsed into unique sRNA sequences and single copy sRNAs were removed. sRNA sequences in the genome-wide library that were not present in any of the AGO-IP sets, were isolated as a new group named “noAGO”. The “noAGO” set was used in the SVM training in combination with AGO-IP data to learn discriminative rules to identify sequences with low potential to load to an AGO and therefore with small chance of becoming functional sRNAs.

Learning procedure

We developed a supervised machine learning approach rooted in the Support Vector Machine (SVM) algorithm to learn classifiers capable of determining AGO-sRNA affinity from the sRNA sequences alone (Supplementary Notes). Briefly, the complete inference system comprises 3 layers (Figure 3A): layer 1 includes a binary SVM model that filters out sequences that do not show strong evidence for binding to any of the known plant AGOs, and that therefore are expected to be inactive; layer 2 is composed by an ensemble of binary one-vs-one classifiers, each trained to explore the dissimilarities in AGO-bound sRNA sequences in a pairwise fashion; finally, layer 3 consists of a voting system that assigns a single score to each AGO using the decision values produced in the previous layer. Layers 2 and 3 are interconnected, since the outputs of the classifiers from the 2nd layer serve as inputs to the 3rd layer, and the 3rd layer combines them to provide a more informative result. Layer 1 is independent of all other classifiers and can therefore be decoupled if desired. All sRNA sequences used in training, testing or validation were transformed into sets of features comprising:

i. Position specific base composition (PSBC)

   One way to convert a string into a numerical representation is by using flags for the presence of a given nucleotide in a determined sequence position. This is equivalent to mapping each sequence position to a four-dimensional feature space that represents each of the four possible bases in the DNA alphabet as follows:

$$A=\langle 1,0,0,0\rangle,\ C=\langle 0,1,0,0\rangle,\ G=\langle 0,0,1,0\rangle,\ T=\langle 0,0,0,1\rangle$$

By using such a format, a sequence can be mapped to a feature space of dimensionality *F=A.L*, where *A* is the number of possibilities in the alphabet of nucleic acids and *L* expresses the dependence on the sequence length. Since the length of the sRNA sequences here analysed is not constant but can vary in an interval with an upper right limit here defined as being 27 bases, all sRNA must be projected into a space of size 27x4=108. In the case of sequences with a length shorter than 27, the extra positions in the feature vector can simply remain empty for all nucleotides. Although this approach is a reasonable solution to cope with the variation observed in length, it has the disadvantage of introducing noise in the representation that increases as the 3' end of the model sequence is approached. This happens because the right most nucleotides in the real sRNA sequences are projected into more central positions of the model sequence, blurring 3' side positional patterns when looking across instances with variable size. To compensate for that effect, the same kind of projection but starting at the 3' position of the sRNA instead of the 5' was additionally considered in a feature set here mentioned as PSBC2.

ii. k-mer composition

Approaches based on k-mers map the presence or absence of subwords with a given length in the sRNA sequence into a feature space that represents all possible k-mers of that length. Taking as example k-mers of size 2, there are 4^2=16 possibilities in the 4 letters universe of DNA. The 16 length 2-gram vector for the DNA 'ACGT' alphabet would then be:

⟨AA,AC,AG,AT,CA,CC,CG,CT,GA,GC,GG,GT,TA,TC,TG,TT⟩

As an example, mapping the sequence "ATGCATG" onto this vector space, considering the presence or absence of each of the possible k-mers yields:

⟨0,0,0,2,1,0,0,0,0,1,0,0,0,0,2,0⟩

It is important to note that this method focuses on the frequency of patterns rather than their position in the sequence. Here k-mers of length 1 to 5 were explored.

iii. Shannon entropy scores

Entropy as a measure of information content gives an indication about the degree of repetitiveness in a sequence. Among several flavours, Shannon entropy is one of the most popular and consists in a score given by:

$$H(X)=-\sum_i p(x_i) \log_2 ((x_i))$$

, with *X* a sequence with length *l* and $p(x_i)$ the frequency of the character at position *i* .

iv. Sequence length

This feature entails the number of nucleotides that compose each sRNA. Although simple, the enrichment for certain sizes in specific pathways is a consistent observation.

The learning methodology applied to layers 1 and 2 was similar. Prior to model training, highly correlated features (|Pearson score|>0.75) were removed keeping a single representative randomly selected from each correlated set. The remaining ones were normalized in the range between 0 and 1, to avoid dominance effects and numerical difficulties in the downstream calculations (19). SVM learning with recursive feature elimination (SVM-RFE) (20) was then applied to find models for layers 1 and 2 with a reduced and more informative feature set. To circumvent computational problems that typically arise when standard SVM algorithms are applied to large datasets, a linear kernel was employed with a specialized linear solver (21). A 5-fold cross-validation procedure was implemented to modulate data variation in each feature selection round and the mean ROC-AUC was calculated to assess the quality of the classifiers. Each round, 1/3 of the features with the lowest contribution for the discriminative model were eliminated, until 10 features were left. From here on, features were excluded one by one until no more features were available. The optimal feature subset for each classifier from layers 1 and 2 was determined with an elbow method applied to the curve formed by the mean cross-validation ROC-AUC values recorded during the feature selection process. The best features were subsequently used to train classifiers applying LIBSVM (22) with radial basis function (RBF) kernels to explore non-linear relationships in the data. A cascade scheme was implemented in this task to tackle computational problems that otherwise would not allow learning with such large datasets (Supplementary Notes, Figure 3B). To avoid biases in learning, training was performed with balanced datasets by under-sampling the largest class involved in each binary problem. Sequences with the highest read abundance were prioritized, and the remaining spots were occupied by instances randomly selected from the remaining pool(s).

In layer 3, a balanced dataset composed of sRNAs from all 8 AGO groups available was created. Decision values were computed for each of the sequences using the classifiers obtained for layer 2, and served as input for a 5-fold cross-validation procedure used to train and test the inference scheme applied to the 3rd layer. Three strategies were explored to combine the outputs from layer 2: a voting system, where the winner is the AGO protein with the largest number of decisions in its favour; a weighted rule learned with a linear SVM algorithm; and a weighted rule learned with a non-linear SVM using a RBF kernel.

All SVM hyperparameters in layers 1, 2 and 3, were tuned by means of a grid search. A more detailed description of the SVM approach is provided in Supplementary Notes.

RESULTS

Detection of high confidence functional sRNAs

We explored a series of SVM classifiers to discriminate putatively functional from non-functional sRNA based on various sRNA sequence features. As outlined above (see section Material and Methods),

SVM were trained on sequenced *A. thaliana* Columbia (Col-0) AGO-IP sRNA libraries in comparison with libraries of Col-0 total sRNA. Specific sRNA sequences contained in the AGO-IP sRNA libraries were labelled “AGO”, whereas sequences contained in the total library (but not in the AGO-sRNA libraries) were labelled “noAGO”. To minimize sequencing noise, single copy sRNA were removed

5’ and 3’ k-mers are important in distinguishing functional from non-functional sRNA

We further assessed whether the k-mers correspond to known sequence motifs. To that end, we performed a motif analysis using the “AGO” and the “noAGO” sets with MEME (24), a computational framework for *de novo* and known motif identification. Then, we focused on k-mers with a size of 5 nt and mapped them to the motifs retrieved in the previous step. We found that the majority of k-mers (38 of 46) corresponded to segments of known or predicted motifs (Table 2), and noted a significant enrichment for k-mer matches in “AGO” when comparing with “noAGO” (100 and 32, respectively). Interestingly, the known motifs mapping k-mers were consistently related to stress response and development, which are processes known to be highly regulated by sRNAs (Supplementary Table S6). The TF enrichment was not surprising for 21 nt sRNA which are known to act on genic sequences that frequently contain TF binding motifs. However, similar TF patterns were also found for 24 nt sRNA, suggesting a role in gene regulation beyond the well-documented function of 24 nt in heterochromatin silencing. We identified the 5-mer “AGAAG” as the k-mer that showed stronger enrichment in 24 nt sequences compared with 21 nt sRNA from the “AGO” set and noted that this subsequence was associated with 3 motifs: At1g68670, a G2-like protein involved in phosphate homeostasis; SVP, a MADS protein that acts as a floral repressor and functions within the thermosensory pathway; and MYB77, which is expressed in response to potassium deprivation and auxin. All these motifs have been shown to have higher DNA-binding capacity when the targets are unmethylated (24), revealing a possible bridge between 24 nt heterochromatic sRNA, DNA methylation and TF occupancy.

Classification and validation of specific AGO-sRNA associations

In the previous section our goal was to build a classifier that can distinguish functional from non-functional sRNA. We achieved this by training on “AGO” versus “noAGO” sRNA. A related problem is to find properties of sRNA that allow to infer their differential loading to specific AGO proteins, as this will determine their particular mode of action. To achieve this we trained binary one-vs-one classifiers to find sequence features that discriminate between the eight different Col-0 AGO-IP libraries (layer 2, Figure 3). The ensemble of one-vs-one classifiers was then subjected to a voting system that assigns a single score to each AGO (layer 3, Figure 3).

We used a 5-fold cross-validation scheme to determine the accuracy of the inference system at three levels: 1. AGO: fraction of sequences for which the AGO-bound protein was correctly predicted; 2. Clade: fraction of sequences for which the AGO prediction falls within the correct clade; and 3. Function: fraction of sequences for which the functional group can be correctly assigned based on the AGO prediction, translating predictions for AGO4/6/9 into a potential for involvement in TS, and assignments to other AGOs as suggestion of PTS activity. In addition to the 5-fold cross-validation

scheme we also validated the classification system using 37 additional *A. thaliana* AGO-IP datasets that were never seen during the training phase by the classifier. These validation datasets were collected from different tissues and experimental conditions, which allowed us to evaluate the robustness of the classifier.

Classification accuracy of sRNA at AGO, clade and function level

Results from our 5-fold cross-validation analysis showed that our classifiers achieved very high accuracy (Figure 6), indicating that differences between sRNA bound to specific AGOs can indeed be detected. Classification accuracy at the level of specific-AGO proteins was around 60% on average, ranging from as low as 40% (AGO7) to as high as 85% (AGO5), see Figure 6B. Since AGO proteins within a clade are highly homologous and similar in function, it is likely that sRNA-AGO binding is promiscuous in nature, which would render the search for discriminative features more challenging. Indeed, classification accuracy at the level of the clade and function were considerably higher than at the level of specific AGOs (clade accuracy: 85%, function accuracy 93%, see Figure 6A). Moreover, the final intra-clade classifiers were generally more complex than inter-clade classifiers, containing on average 122 features compared with 75 features, respectively (Figure 6C). Looking to the features retained in the final classifiers, intra-clade classifiers contained proportionally more k-mers of length larger than 1 nt compared with inter-clade classifiers (86% and 79% of the features in intra-clade and inter-clade models, respectively). Hence, factors that govern AGO-specific affinities within a clade appear to involve more complex sequence determinants. Similar to the AGO versus noAGO analysis presented above (section Detection of high confidence functional sRNAs), we found that the k-mers in the final classifiers showed strong positional bias toward the 5' and 3' ends of sRNA sequences (Supplementary Figure S3). This finding indicates that the same sRNA regions that differentiate functional from non-functional sequences also contribute to the binding affinity of sRNA to specific AGO proteins.

To study the nature of the information contained in the k-mers selected by SVM-RFE in more detail, we compared them with a motif analysis performed for each AGO library using the MEME suite (23). For a matter of simplicity, we focused on 5-mers (Supplementary Notes). We found that nearly 40% of the 5-mers were identified as motifs or derived segments (Table 2), often related to known transcription factors with roles in development and response to stress. For instance, in models involving AGO2 (an AGO linked to ta-siRNAs) we identified 5-mers matching the motif ETT, an experimentally validated target of an evolutionarily conserved ta-siRNA denominated tasiR-ARF (25). Targeting of ETT by ta-siRNA has been extensively studied and is known to have pleiotropic effects on Arabidopsis flower development by interference with the auxin pathway (26). These results thus support the conclusion that SVM learning captured sequence information with relevant biological meaning. A full overview of k-mers overlapping known or predicted motifs is provided in Table 2.

Validation of the classification system

The AGO-sRNA classification framework was validated by testing on 35 AGO-IP libraries never seen during the training phase, including material acquired from different tissues and from plants under

different treatment conditions (Table 1). These diverse datasets allowed us to evaluate the robustness of our classification framework. AGO, clade and function-based sensitivity were determined for each AGO library, in a procedure similar to the one applied to layer 3. Additionally, validation was also performed using two datasets of experimentally confirmed ta-siRNAs. One of these two datasets contained ta-siRNA only from *A. thaliana* while the other datasets comprised ta-siRNA from a large collection of different plant species. Since the ta-siRNA databases did not contain a record for the specific AGO to which the sequences load, function-based inference was the only adequate assessment in this case. Because no dataset of validated hc-siRNA is currently available, the quality of function inference for this kind of sRNA could not be measured.

Validation analysis revealed that our classifiers are relatively sensitive, even across biologically very heterogeneous datasets (Figure 7). Sensitivity was particularly high at the level of clade and function (clade: 70% and function: 84.3%), and moderate at the level of specific AGOs (AGO: 42%). Interestingly, when looking to the performance obtained for the ta-siRNA dataset, we observed high sensitivity values both for the set isolated for Arabidopsis (~80%) and also for the complete plant set (~95%). The sensitivity is considerably higher for the second set compared with the first which supports the idea that the inference system can identify PTS sRNAs not only from the species used in the learning phase but is also extensible to other plant species.

In conclusion, the inference system demonstrates robustness for tissue and treatment variation and can recognize sRNA membership independently of the cellular origin showing good generalization as desired for the discovery of new functional units.

DISCUSSION

To our knowledge, this is the first time that classifiers were built to infer AGO-sRNA affinity from the sRNA sequence alone. Using adequate solvers and learning architectures, SVMs could be applied to large genomic datasets and discovered highly discriminative rules, both to distinguish sequences that bind to AGOs from other sequencing products, as well to infer AGO-sRNA kinship. In addition to the known 5' nucleotide composition and the sequence length, feature selection revealed the contribution of other features, mostly complex k-mers, in defining sRNA preference for certain AGO proteins. The question how robust specific sRNA-AGO affinities are to changes in certain sequence properties is an interesting topic for future research and can shed light on evolutionary constraints on sRNA-mediated transcriptional and post-transcriptional silencing pathways. To answer such questions, additional experiments are necessary that can manipulate sRNA sequence in a precise and targeted fashion, for example by use of the CRISPR-Cas9 system. Although our inference method appear to be highly accurate in predicting the putative function of sRNA sequences, it is important to keep in mind that the actual biological activity of a given sRNA is dependent on other factors beyond the AGO loading step, such as the degree of complementarity between a sRNA and the target sequence, as well as the presence of specific chromatin states at the target locus.

AGO-sorting information has the potential to decrease the very high number of false positives reported by currently available PTS target prediction tools, but the true impact needs to be further evaluated. In any case, the computational framework here developed displays a discriminative power

that makes it suitable for early screens in genome-wide sRNA libraries when looking for candidates with the highest chance to have certain functional roles, and when sorting sRNAs by TS and PTS classes is needed. The tool is ultimately a more affordable alternative to expensive and laborious AGO-IP experiments, since it can get AGO-sRNA profiles from a single genome-wide sequencing library. Another interesting application of the framework is to explore if specific sRNA from exogenous sources, such as artificially designed sequences or those derived from pathogens, correspond to functional plant sRNAs and which silencing pathways they are likely to target in a plant.

More sophisticated models can be developed using for example the expression patterns of the AGO proteins at the moment that the sRNA sequencing experiments take place. This extra information can eventually improve the capacity to correctly predict AGO affinity under particular situations, but on the other hand demands additional and more complex inputs, thus limiting the range of applications of the sRNA classifiers

REFERENCES

1. Borges,F., Martienssen,R.A. (2015) The expanding world of small RNAs in plants. *Nat. Rev. Mol. Cell Biol.*, 16, 727–741, doi:10.1038/nrm4085.

2. Kurihara,Y., Takashi,Y. and Watanabe,Y. (2006) The interaction between DCL1 and HYL1 is important for efficient and precise processing of pri-miRNA in plant microRNA biogenesis. *RNA*, 12, 206-212.

3. Shen,B. and Goodman,H.M. (2004) Uridine addition after microRNA-directed cleavage. *Science*, 306:997.

4. Vaucheret,H. (2008) Plant ARGONAUTES. *Trends Plant Sci.*, 13, 350-358.

5. Mirzaei,K., Bahramnejad,B., Shamsifard,M.H. and Zamani,W. (2014) *In silico* identification, phylogenetic and bioinformatic analysis of Argonaute genes in plants. *Int. J. of Genomics*, ID 967461

6. Rodríguez-Leal,D., Castillo-Cobián,A., Rodríguez-Arévalo,I., Vielle-Calzada,J.-P. (2016) A primary sequence analysis of the ARGONAUTE protein family in plants. *Front. Plant Sci.*, 7, 1347, doi: 10.3389/fpls.2016.01347

7. Montgomery,T.A., Howell,M.D., Cuperus,J.T., Li,D., Hansen,J.E., Alexander,A.L., Chapman,E.J., Fahlgren,N., Allen,E. and Carrington,J.C. (2008) Specificity of ARGONAUTE7–miR390 interaction and dual functionality in TAS3 trans-acting siRNA formation. *Cell*, 133, 128-141.

8. Havecker,E.R., Wallbridge,L.M., Hardcastle,T.J., Bush,M.S., Kelly,K.A., Dunn,R.M., Schwach,F., Doonan,J.H. and Baulcombe,D.C. (2010) The Arabidopsis RNA-directed DNA methylation argonautes functionally diverge based on their expression and interaction with target loci. *Plant Cell*, 22, 321-34.

9. Ji,L., Liu,X., Yan,J., Wang,W., Yumul,R.E., Kim,Y.J., Dinh,T.T., Liu,J., Cui,X., Zheng,B. *et al*. (2011). ARGONAUTE10 and ARGONAUTE1 regulate the termination of floral stem cells through two microRNAs in Arabidopsis. *PLoS Genet*., 7, e1001358 10.1371/journal.pgen.1001358.

10. Eamens,A.L., Smith,N.A., Curtin,S.J., Wang,M.B. and Waterhouse,P.M. (2009) The Arabidopsis thaliana double stranded RNA binding protein DRB1 directs guide strand selection from microRNA duplexes. *RNA*, 15, 2219-2235.

11. Khvorova,A., Reynolds,A. and Jayasena,S.D. (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell*, 115, 209-216.

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

13. Allen,E. and Howell,M.D. (2010) miRNAs in the biogenesis of trans-acting siRNAs in higher plants. Semin. *Cell Dev. Biol*., 21, 798-804, doi: 10.1016/j.semcdb.2010.03.008.

14. Mi,S., Cai,T., Hu,Y., Chen,Y., Hodges,E., Ni,F., Wu,L., Li,S. *et al*. (2008) Sorting of small RNAs into Arabidopsis Argonaute complexes is directed by the 5' terminal nucleotide. *Cell*, 133, 116-127.

15. Ma,J.B., Ye,K. and Patel,D.J. (2004) Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain. *Nature*, 429, 318-322.

16. Parker,J.S., Roe,S.M. and Barford,D. (2005) Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex. *Nature*, 434, 663-666.

17. Zhu,H., Hu,F., Wang,R., Zhou,X., Sze,S.H., Liou,L.W., Barefoot,A., Dickman,M. and Zhang,X. (2011) Arabidopsis Argonaute10 specifically sequesters miR166/165 to regulate shoot apical meristem development. *Cell*, 145, 242-56.

18. Kim,V.N. (2008) Sorting Out Small RNAs. *Cell*, 133, 25-26. doi: 10.1016/j.cell.2008.03.015.

19. Hsu,C.-W., Chang,C.-C. and Lin,C.-J. (2003) A practical guide to Support Vector Classification. Technical report, Department of Computer Science, National Taiwan University.

20. Guyon,I., Weston,J., Barnhill,S. and Vapnik,V. (2002) Gene selection for cancer classification using support vector machines. *Mach. Learn.*, 46, 389-422.

21. Fan,R.-E., Chang,K.-W., Hsieh,C.-J., Wang,X.-R. and Lin,C.-J. (2008) LIBLINEAR: A library for large linear classification. *The J. of Mach. Learn. Res.*, 9, 1871-1874.

22. Chang,C.-C. and Lin,C.-J. LIBSVM: a library for Support Vector Machines. 2005. Technical Report. Department of Computer Science, National Taiwan University.

23. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*., 37, W202-W208.

24. O'Malley,R.C., Huang,S.C., Song,L., Lewsey,M.G., Bartlett,A., Nery,J.R., Galli,M., Gallavotti,A. and Ecker,J.R. (2016) Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell*, 19, 1280-92, doi: 10.1016/j.cell.2016.04.038

25. Williams,L., Carles,C.C., Osmont,K.S. and Fletcher,J.C. (2005) A database analysis method identifies an endogenous trans-acting short-interfering RNA that targets the arabidopsis ARF2, ARF3, and ARF4 genes. *Proc. Natl. Acad. Sci. U.S.A.*, 102, 9703-8, doi:10.1073/pnas.0504029102.

26. Yan,J., Cai,X., Luo,J., Sato,S., Jiang,Q., Yang,J., Cao,X., Hu,X., Tabata,S., Gresshoff,P.M. *et al*. (2010) The REDUCED LEAFLET genes encode key components of the trans-acting small interfering RNA pathway and regulate compound leaf and flower development in Lotus japonicas. *Plant Physiol.*, 152, 797–807.

Figure 1

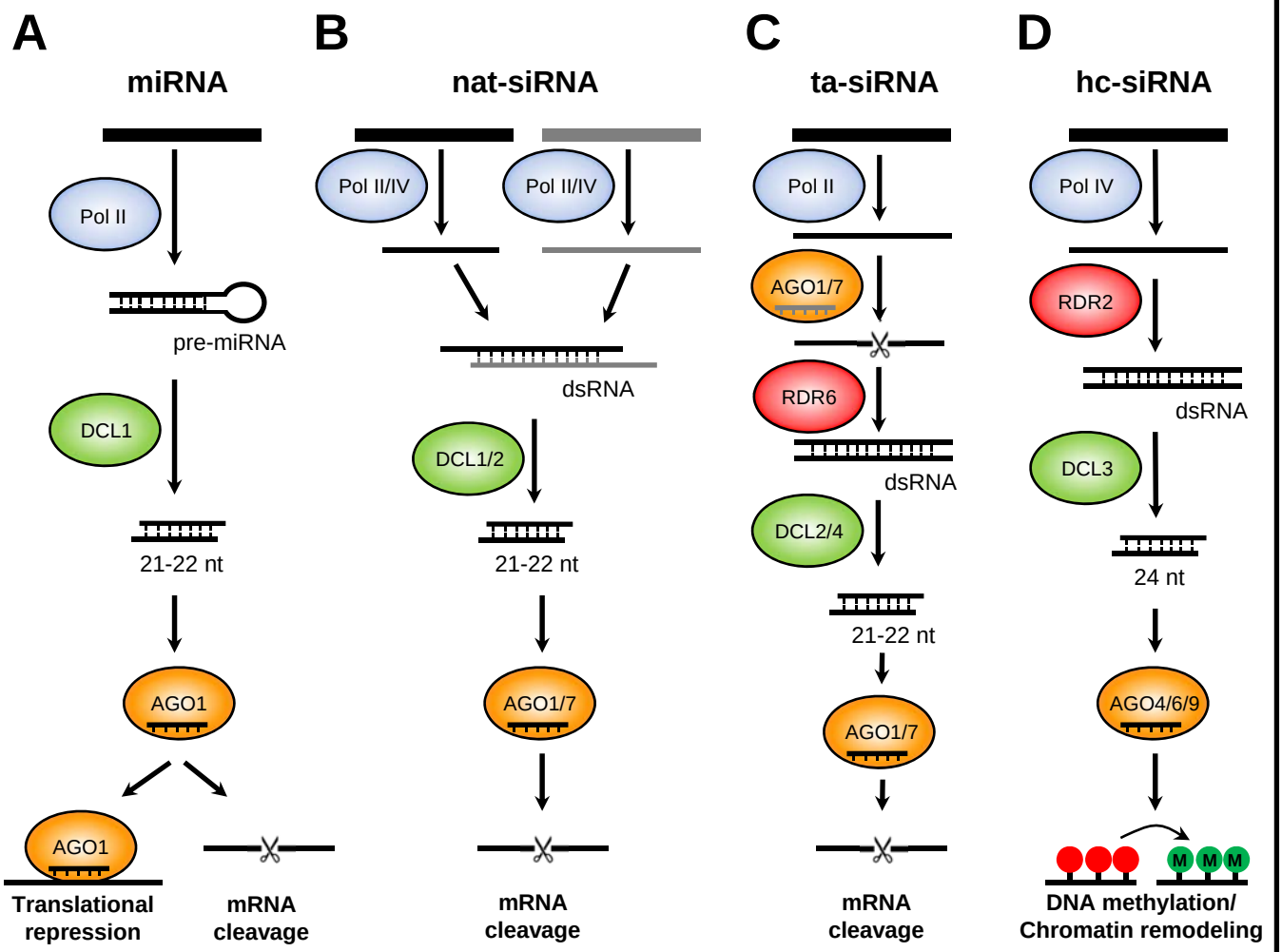

Figure 1. The main endogenous sRNA pathways in plants: miRNA (A), nat-siRNA (B), ta-siRNA (C) and hc-siRNA (D). DCL: dicer; dsRNA: double-stranded RNA; Pol: RNA polymerase; RDR: RNA-dependent RNA polymerase.

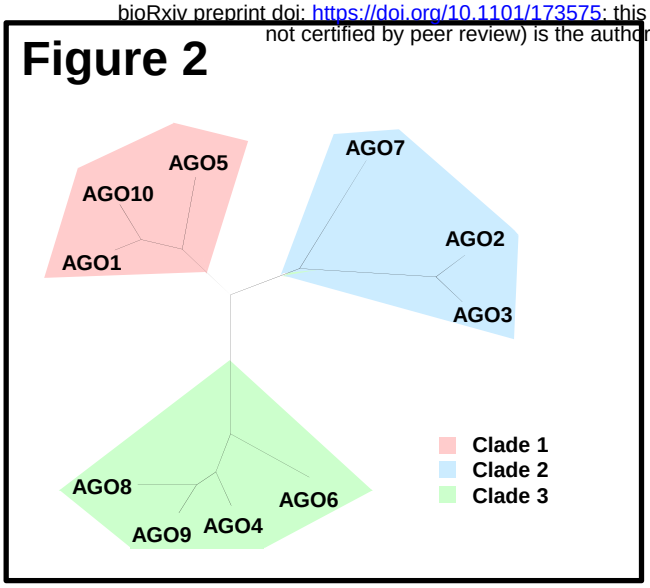

Figure 2. Phylogenetic tree for AGO in *A. thaliana*.

Figure 3

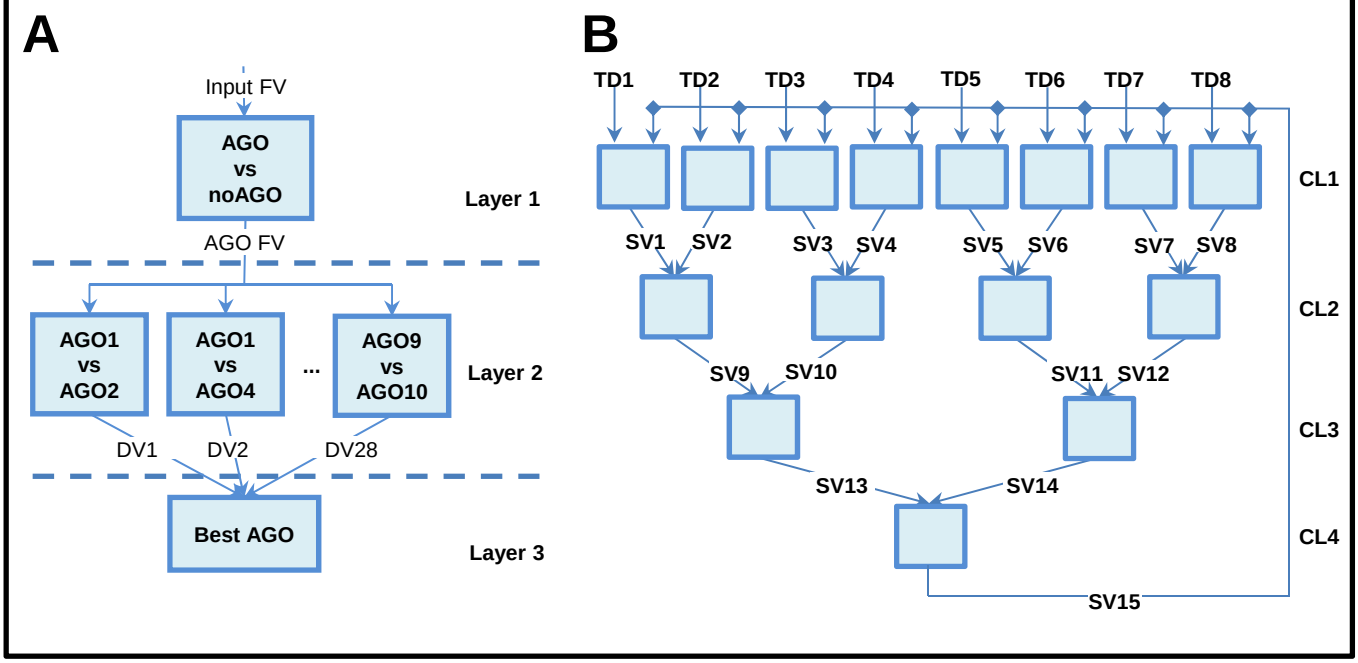

Figure 3. Machine learning architectures used in the current work: (A) architecture of the inference system; (B) cascade SVM implemented. In the cascade scheme, the data are split into subsets and each one is evaluated individually for SVs in the first layer. The results are combined two-by-two and entered as training sets for the next layer. The resulting SVs are tested for global convergence by feeding the result of the last layer into the first layer, together with the non-SVs. FV: feature vectors; DV_i: decision values from classifier i; TD_i: Training data partition i; SV_j: SVs produced by optimization j; CL_k: cascade layer k.

Figure 4

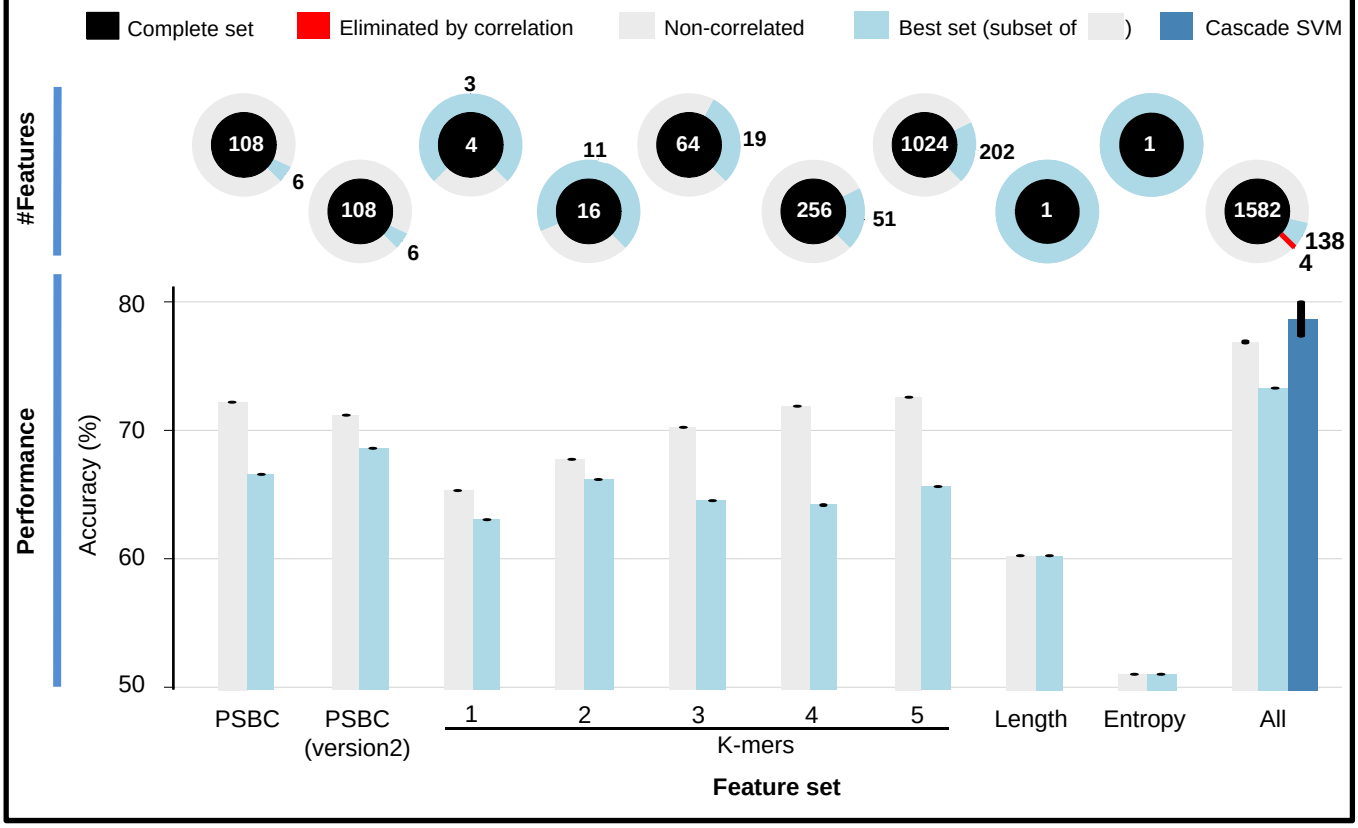

Figure 4. Results for the classifiers trained to detect sRNA from the sets AGO vs noAGO: number of features and performance. Pie charts represent the fraction of features in the complete set (the total number is in the central black section of the pie) that was eliminated during the correlation analysis (red section in the external ring) or kept (grey section in the external ring). The remaining features were subjected to selection with SVM-RFE, of which a subset comprises the optimal features (light blue section). For each initial conglomerate of features, accuracy was measured: using all features in a set (grey bars) and after feature selection (light blue bars). The optimized features determined when all feature sets were analysed together was then subjected to non-linear learning using a cascade SVM scheme (dark blue bar). Each bar plot has on top the standard deviation for the accuracy calculated from the 5-fold cross-validation procedure.

Figure 5

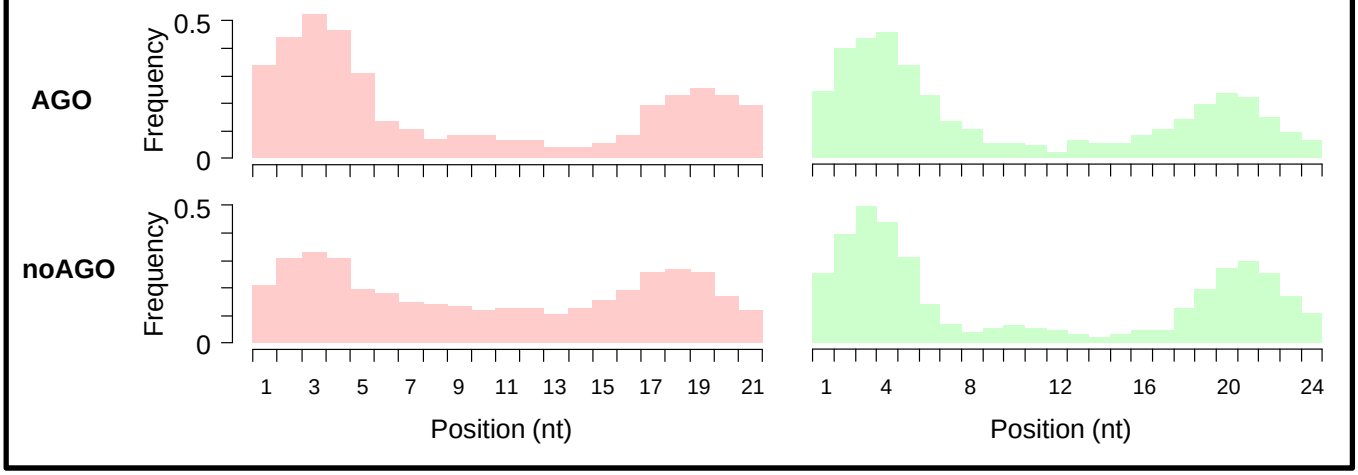

Figure 5. Density distribution of k-mers retained in the final classifiers discriminating between functional (AGO) versus non-functional (noAGO) sRNA (layer 1). The k-mers show a clear positional bias toward the 5' and 3' ends of 21 nt (salmon) and 24 nt (green) sRNA sequences.

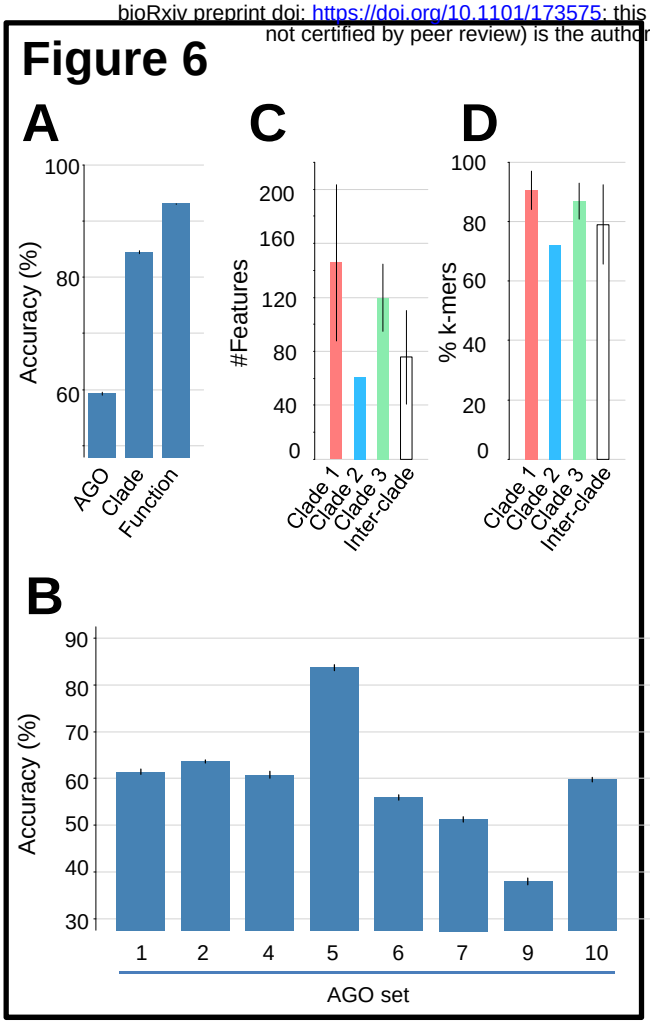

Figure 6: Details of the AGO-sorting inference (layers 2+3). Mean accuracy across the 5-fold cross-validation scheme for the voting system used in layer 3, measured by: (A) AGO, clade and function level; (B) separated by individual AGO sets. The average number of features composing the intra- and inter-clade classifiers from layer 2 (C), and the percentage of those features which are k-mers (D). Error bars: standard deviation.

Figure 7

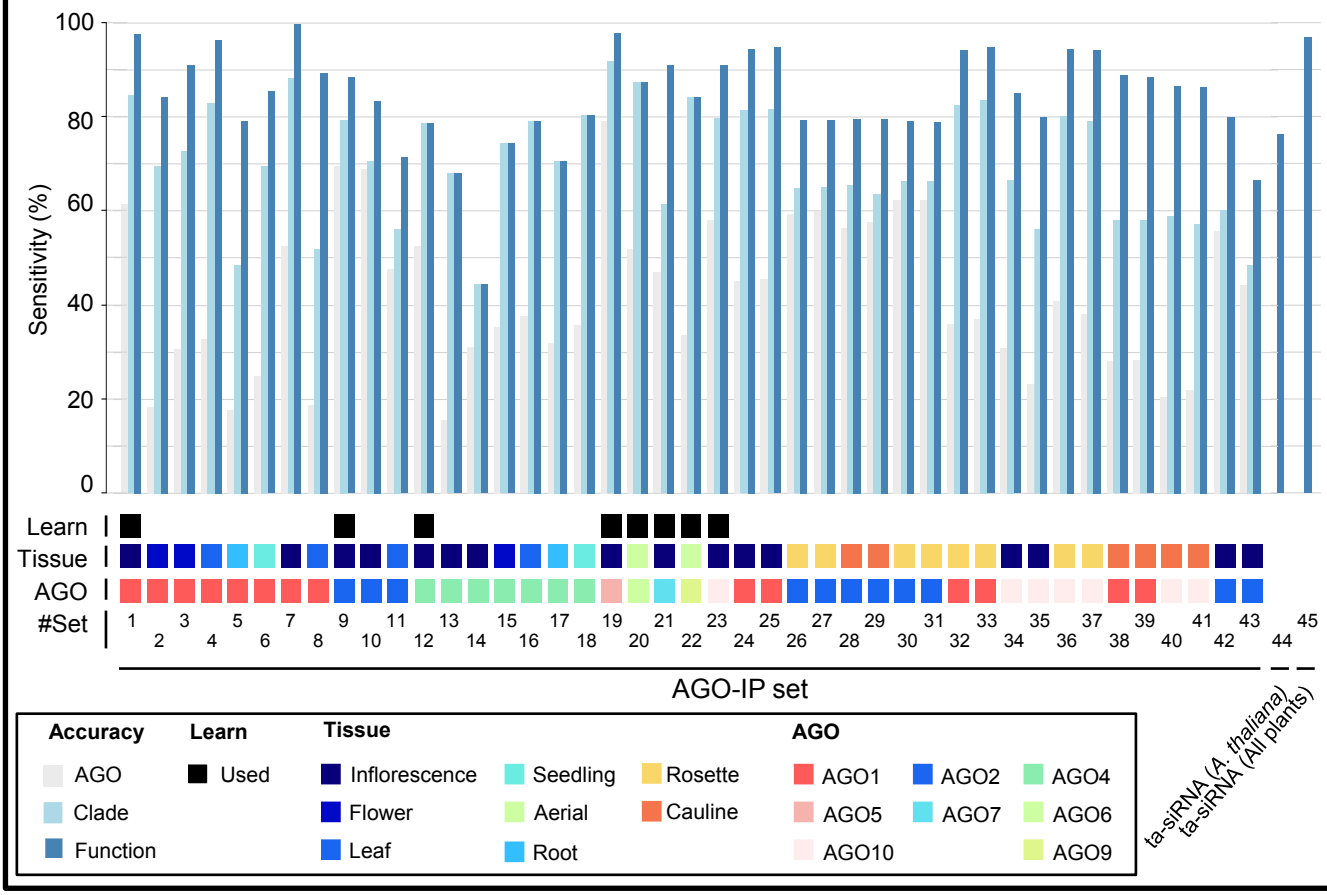

Figure 7. Sensitivity for the validation sets. Bar plots show results for AGO (light grey), clade (light blue) and function (dark blue) inference. Indication about the use of the data sets during learning, the sampled tissue and the AGO with which they were immunoprecipitated is shown by coloured squares under the bars. After these, the data sets reference numbers are indicated, following the enumeration in Table 1.

Table 1. Libraries of sRNA used in the current work. TOT: total sRNA; AGO-IP: AGO immunoprecipitated; NA: Not Applicable. T: used for training; V: used for validation.

#Set	Type	Notes	Species	Tissue	Reference	T	V
1	AGO1-IP	-	*A. thaliana*	Inflorescence	Mi, 2008	Y	Y
2	AGO1-IP	-	*A. thaliana*	Flower	Zhu, 2011	N	Y
3	AGO1-IP	-	*A. thaliana*	Flower	Wang, 2011	N	Y
4	AGO1-IP	-	*A. thaliana*	Leaf	Wang, 2011	N	Y
5	AGO1-IP	-	*A. thaliana*	Root	Wang, 2011	N	Y
6	AGO1-IP	-	*A. thaliana*	Seedling	Wang, 2011	N	Y
7	AGO1-IP	-	*A. thaliana*	Inflorescence	Qi, 2006	N	Y
8	AGO1-IP	Pseudomonas syringae infection set	*A. thaliana*	Leaf	Zhang, 2011	N	Y
9	AGO2-IP	-	*A. thaliana*	Inflorescence	Montgomery, 2008	Y	Y
10	AGO2-IP	-	*A. thaliana*	Inflorescence	Mi, 2008	N	Y
11	AGO2-IP	Pseudomonas syringae infection set	*A. thaliana*	Leaf	Zhang, 2011	N	Y
12	AGO4-IP	-	*A. thaliana*	Inflorescence	Mi, 2008	Y	Y
13	AGO4-IP	-	*A. thaliana*	Inflorescence	Havecker, 2010	N	Y
14	AGO4-IP	-	*A. thaliana*	Inflorescence	Qi, 2006	N	Y
15	AGO4-IP	-	*A. thaliana*	Flower	Wang, 2011	N	Y
16	AGO4-IP	-	*A. thaliana*	Leaf	Wang, 2011	N	Y
17	AGO4-IP	-	*A. thaliana*	Root	Wang, 2011	N	Y
18	AGO4-IP	-	*A. thaliana*	Seedling	Wang, 2011	N	Y
19	AGO5-IP	-	*A. thaliana*	Inflorescence	Mi, 2008	Y	Y
20	AGO6-IP	-	*A. thaliana*	Aerial	Havecker, 2010	Y	Y
21	AGO7-IP	-	*A. thaliana*	Inflorescence	Montgomery, 2008	Y	Y
22	AGO9-IP	-	*A. thaliana*	Aerial	Havecker, 2010	Y	Y
23	AGO10-IP		*A. thaliana*	Inflorescence	Zhu, 2011	Y	Y
24	AGO1-IP	TuMV set: HA-AGO1-DAH(Col-0)_Mock_IP	*A. thaliana*	Inflorescence	Garcia-Ruiz, 2015	N	Y
25	AGO1-IP	TuMV set: HA-AGO1-DAH(Col-0)_TuMV_IP	*A. thaliana*	Inflorescence	Garcia-Ruiz, 2015	N	Y
26	AGO2-IP	TuMV set: HA-AGO2-DAD(ago2-1)_Mock_IP	*A. thaliana*	Rosette	Garcia-Ruiz, 2015	N	Y
27	AGO2-IP	TuMV set: HA-AGO2-DAD(ago2-1)_TuMV-AS9_IP	*A. thaliana*	Rosette	Garcia-Ruiz, 2015	N	Y
28	AGO2-IP	TuMV set: HA-AGO2-DAD(ago2-1)_Mock_IP	*A. thaliana*	Cauline	Garcia-Ruiz, 2015	N	Y
29	AGO2-IP	TuMV set: HA-AGO2-DAD(ago2-1)_TuMV-AS9_IP	*A. thaliana*	Cauline	Garcia-Ruiz, 2015	N	Y
30	AGO2-IP	TuMV set: HA-AGO2-DAD(ago2-1)_Mock_IP	*A. thaliana*	Rosette	Garcia-Ruiz, 2015	N	Y
31	AGO2-IP	TuMV set: HA-AGO2-DAD(ago2-1)_TuMV_IP	*A. thaliana*	Rosette	Garcia-Ruiz, 2015	N	Y
32	AGO1-IP	TuMV set: HA-AGO1-DAH(Col)_Mock_IP	*A. thaliana*	Rosette	Garcia-Ruiz, 2015	N	Y
33	AGO1-IP	TuMV set: HA-AGO1-DAH(Col)_TuMV_IP	*A. thaliana*	Rosette	Garcia-Ruiz, 2015	N	Y
34	AGO10-IP	TuMV set: HA-AGO10-DDH(Col)_Mock_IP	*A. thaliana*	Inflorescence	Garcia-Ruiz, 2015	N	Y
35	AGO10-IP	TuMV set: HA-AGO10-DDH(Col)_TuMV_IP	*A. thaliana*	Inflorescence	Garcia-Ruiz, 2015	N	Y
36	AGO10-IP	TuMV set: HA-AGO10-DDH(Col)_Mock_IP	*A. thaliana*	Rosette	Garcia-Ruiz, 2015	N	Y
37	AGO10-IP	TuMV set: HA-AGO10-DDH(Col)_TuMV_IP	*A. thaliana*	Rosette	Garcia-Ruiz, 2015	N	Y
38	AGO1-IP	TuMV set: HA-AGO1-DAH(ago2-1)_Mock_IP	*A. thaliana*	Cauline	Garcia-Ruiz, 2015	N	Y
39	AGO1-IP	TuMV set: HA-AGO1-DAH(ago2-1)_TuMV-AS9_IP	*A. thaliana*	Cauline	Garcia-Ruiz, 2015	N	Y
40	AGO10-IP	TuMV set: HA-AGO10-DAH(ago2-1)_Mock_IP	*A. thaliana*	Cauline	Garcia-Ruiz, 2015	N	Y
41	AGO10-IP	TuMV set: HA-AGO10-DAH(ago2-1)_TuMV-AS9_IP	*A. thaliana*	Cauline	Garcia-Ruiz, 2015	N	Y
42	AGO2-IP	TuMV set: HA-AGO2-DAD(ago2-1)_Mock_IP	*A. thaliana*	Inflorescence	Garcia-Ruiz, 2015	N	Y
43	AGO2-IP	TuMV set: HA-AGO2-DAD(ago2-1)_TuMV_IP	*A. thaliana*	Inflorescence	Garcia-Ruiz, 2015	N	Y
44	ta-siRNA	All known tasiRNA from Arabidopsis	*A. thaliana*	NA	Zhang, 2014	N	Y
45	ta-siRNA	All known tasiRNA in plants	All plants	NA	Zhang, 2014	N	Y
46	TOT	Total sRNA from WT background	*A. thaliana*	Inflorescence	Slotkin, 2009	Y	N

Table 2. Number of 5-mers in the final classifiers from layer 2 matching motifs found with MEME in each AGO-IP library. The 5-mers were separated by the signal of their weight *w* in the final classifier.

#	Set a	Set b	# features in the final classifier	# 5-mers in the the final classifier				% 5-mers mapping to a motif
				Favouring		Mapping to a motif		
				AGOa (w>0)	AGOb (w<0)	AGOa	AGOb	
1	AGO1	AGO2	92	22	30	2	19	40,4
2	AGO1	AGO4	27	9	3	0	2	16,7
3	AGO1	AGO5	92	20	18	8	8	42,1
4	AGO1	AGO6	41	14	6	7	1	40,0
5	AGO1	AGO7	92	35	16	19	10	56,9
6	AGO1	AGO9	92	27	17	11	0	25,0
7	AGO1	AGO10	138	53	24	14	9	29,9
8	AGO2	AGO4	61	8	6	7	0	50,0
9	AGO2	AGO5	61	16	14	15	6	70,0
10	AGO2	AGO6	92	25	16	18	3	51,2
11	AGO2	AGO7	61	10	5	3	5	53,3
12	AGO2	AGO9	92	19	18	12	3	40,5
13	AGO2	AGO10	61	11	8	8	4	63,2
14	AGO4	AGO5	18	2	1	1	0	33,3
15	AGO4	AGO6	137	48	17	9	1	15,4
16	AGO4	AGO7	92	35	16	7	11	35,3
17	AGO4	AGO9	131	36	17	8	3	20,8
18	AGO4	AGO10	61	2	6	2	5	87,5
19	AGO5	AGO6	41	4	5	3	1	44,4
20	AGO5	AGO7	61	19	16	10	8	51,4
21	AGO5	AGO9	41	5	5	1	1	20,0
22	AGO5	AGO10	207	65	42	31	21	48,6
23	AGO6	AGO7	138	46	41	7	20	31,0
24	AGO6	AGO9	91	8	18	0	0	0,0
25	AGO6	AGO10	92	9	16	1	11	48,0
26	AGO7	AGO9	138	36	43	22	2	30,4
27	AGO7	AGO10	138	38	49	22	23	51,7
28	AGO9	AGO10	61	6	5	0	3	27,3