16
AthaMap, a Database for the Identification and Analysis of Transcription Factor Binding Sites in the *Arabidopsis thaliana* Genome

Reinhard Hehl

Abstract

The genome-wide Identification of Transcription Factor Binding Sites (TFBS) was created when the genome sequence of the first higher plant, *Arabidopsis thaliana*, was reported. Since then, the recent completion of the sequencing of many plant genomes permits the mapping of TFBS on a genome-wide scale. There are several approaches to mapping TFBS within genomic sequences. If a regulatory sequence has been identified experimentally, a bioinformatic approach to obtain positional information for these sequences in the genome may involve pattern recognition programs such as MatInspector, Match, Patser, or PatMatch [1–4]. This positional information can be stored in databases. For example AGRIS, AthaMap, Athena, and ATTED-II are all database resources that contain pre-calculated TFBS within genomic sequences of *A. thaliana* [5–8]. While Athena, ATTED-II, and AGRIS focus on upstream regions and use consensus sequences for the identification of putative regulatory sequences, AthaMap is the first database that generates a genome-wide map of putative TFBS mainly based on alignment matrices. AthaMap is freely available at http://www.athamap.de/. In this chapter, we will present the AthaMap database and its applications.

16.1
Introduction

AthaMap was first generated by matrix based sequence searches using alignment matrices derived from many binding sites of single transcription factors (TFs) [6]. The early version of AthaMap contained a simple search function that requires a chromosomal position or a locus identifier that would result in a sequence display window with indicated binding sites. In this early version of AthaMap the genes were simply underlined, beginning with either the transcription or the translation start site. The next version of AthaMap increased its functionality by incorporating a co-localization function [9]. This function permits the identification of chromosomal

The Handbook of Plant Functional Genomics: Concepts and Protocols.
Edited by Günter Kahl and Khalid Meksem

ISBN: 978-3-527-31885-8

positions of putative combinatorial elements. Such combinatorial elements were also pre-calculated and annotated to AthaMap based on TFs that are known to interact and on TFs that contain two DNA binding sites. Furthermore, a new function permits the restriction of displayed TFBS to those which are highly conserved. In a next step, AthaMap was extended with functionally verified single TFBS and TFBS that where predicted on the basis of these functionally verified sites [10]. The most recent update of AthaMap contains a gene analysis function that permits the identification of common or missing TFBS in a set of genes [11]. This function may be useful for the analysis of co-regulated genes and for genes that are members of the same metabolic pathway. Furthermore, the complete gene structure consisting of upstream and downstream untranslated regions, introns and exons is now displayed in the genomic sequence.

For the annotation of gene structure and the determination of TFBS in AthaMap, XML flatfiles containing sequence and gene structure information (release 5.0) were downloaded from the TIGR web site [12]. These flatfiles were parsed using a Perl script. Positional information for 5′ and 3′ UTRs, exons and introns were annotated to AthaMap. These regions are displayed in AthaMap with a color code similar to that used by TAIR, the Arabidopsis Information database [11,13].

Currently, AthaMap contains TFBS that were determined using two different methods. First, TFBS were detected with alignment matrices and second, TFBS were identified with single experimentally verified sites [6,10]. Using the pattern search program Patser [1] more than 9×10^6 putative TFBS for 49 TFs from 22 different families were detected within the *Arabidopsis thaliana* genome. With two exceptions, all detected TFBS were annotated to the database. Only in case of the CAT- and TATA-box binding factors CBF and TBP was positional information used to restrict the number of annotated TFBS to those that occur within a defined region upstream of the transcription or translation start site [9]. In some cases an alignment matrix was only available for a TF from a species other than *A. thaliana*. These alignment matrices were also used for the detection of putative *A. thaliana* TFBS because TFs and their binding site specificities are not plant species specific.

In a second approach single published TFBS were annotated to AthaMap. For many TFs no alignment matrix was available but a single binding site had been determined in a gene. This binding site may also occur at other genomic positions. Therefore, novel putative binding sites were determined within the genome that were identical to the sequence of the experimentally verified site adjacent to the core sequence of the TFBS [10] To detect TFBS based on single transcription factor binding sites, a Perl script was written for pattern-based screenings of the *Arabidopsis thaliana* genome. Both strands of the annotated genome were screened resulting in records harboring absolute positional information and orientation. In this case only sites determined with *A. thaliana* TFs were included. In total 94 191 TFBS for 55 factors from 15 TF families were identified using this method.

The third class of TFBS that was pre-calculated and annotated to AthaMap is combinatorial elements [9,11]. For this, TFBS determined with alignment matrices were used for a co-localization analysis. Combinatorial elements were identified for TFs that are either known to interact or that are known to harbor two binding

domains. In total 359 867 sites for six combinatorial elements were annotated to AthaMap.

16.2
Methods and Applications

16.2.1
Using the Web Interface at http://www.athamap.de/

16.2.1.1 The Search Function

To display TFBS at any chromosomal position, the user can choose one of two options on the search page of AthaMap as shown in Figure 16.1. Either a position on a selected chromosome or a gene identifier (*Arabidopsis* genome identification number: AGI) can be submitted. Furthermore, it is possible to restrict the display to highly conserved TFBS.

A typical result screen is shown in Figure 16.2. The result window displays the nucleotide sequence 500 bp upstream and 500 bp downstream of the gene start or the chromosomal position defined by the search mode. In this case a known transcription start site (TSS) has been annotated. The gene structure is shown with a color code. The three types of TFBS are indicated on the result page with three different symbols. Matrix based TFBS (->), combinatorial element (==), and TFBS based on single sites (>>). The names of the factors or combinatorial elements for which TFBS were detected are linked to pop-up windows that show the underlying data for that particular site (Figure 16.3A–C). Furthermore, a tool tip box will open with specific positional information on the TFBS and with further parameters for matrix based TFBS (Figure 16.3D). At the bottom of the result page, two arrows allow forward and backward scrolling of the sequence window by 500 bp (not shown). Furthermore, a short description of the gene is given and links to external databases for further

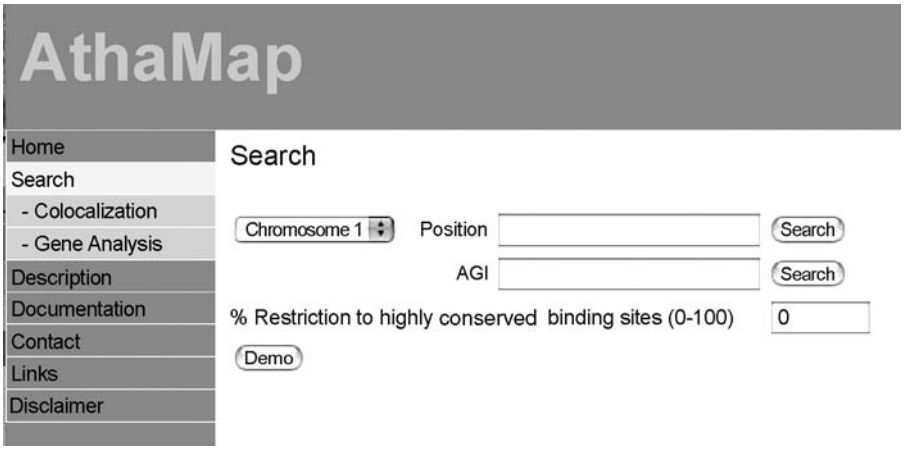


Figure 16.1 The search function of AthaMap.

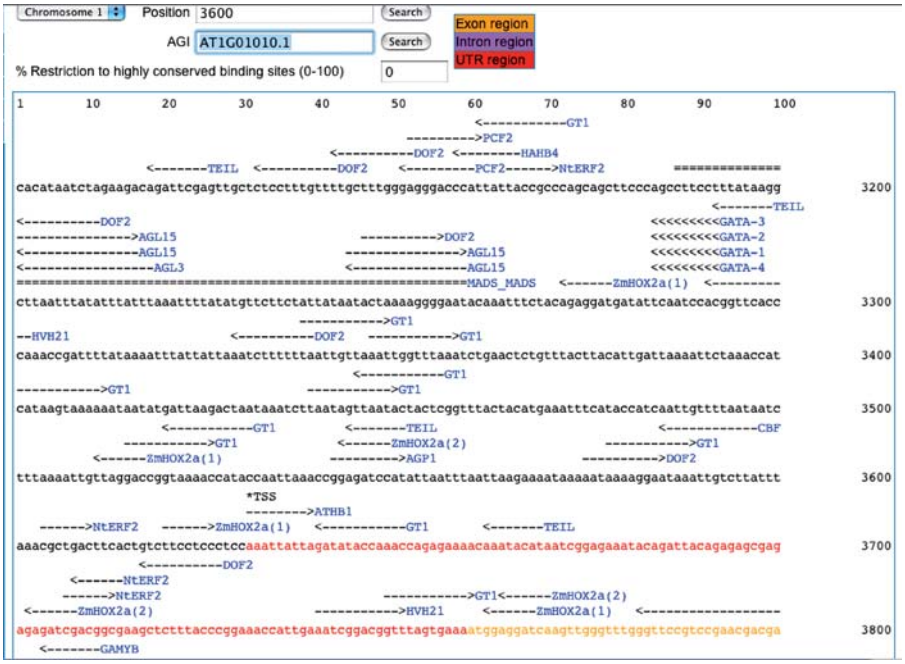


Figure 16.2 Partial screen of the search result for gene AT1G01010.1.

information on the displayed gene are implemented below the sequence display window (not shown).

Figure 16.3A and B show additional information in pop-up windows linked to every TF for which a binding site was detected in the genome. This information contains the name of the TF, the family of the TF, and the plant species. Either the matrix (Figure 16.3A) or the single sequences used for TFBS determination (Figure 16.3B) are shown. The reference from which these sequences were extracted is directly linked to the PubMed database. If the factor is annotated to the TRANSFAC database, the TRANSFAC accession number links to the factor description in the TRANSFAC professional database for licensed TRANSFAC users [14]. If the TF is from *A. thaliana*, the AGI links to the gene locus in TAIR [13]. Further specific information derived from matrix based searches is the maximum score and the threshold determined for a matrix by the search program Patser (Figure 16.3A). Each matrix-based TFBS has an individual score between threshold and maximum score which is an indication of the conservation of the binding site. A high score close to the maximum score means that this particular binding site contains nucleotides that are more frequently encountered at the corresponding position in the matrix. To obtain this information each binding site is linked to a tool tip box that opens when the cursor is moved over the site determined with a matrix. Figure 16.3D shows a tool tip box for a specific binding site. Here, in addition to the positional information, maximum score and threshold score of the matrix, the individual score of the binding site is also shown. Each matrix-based binding site has a specific score.

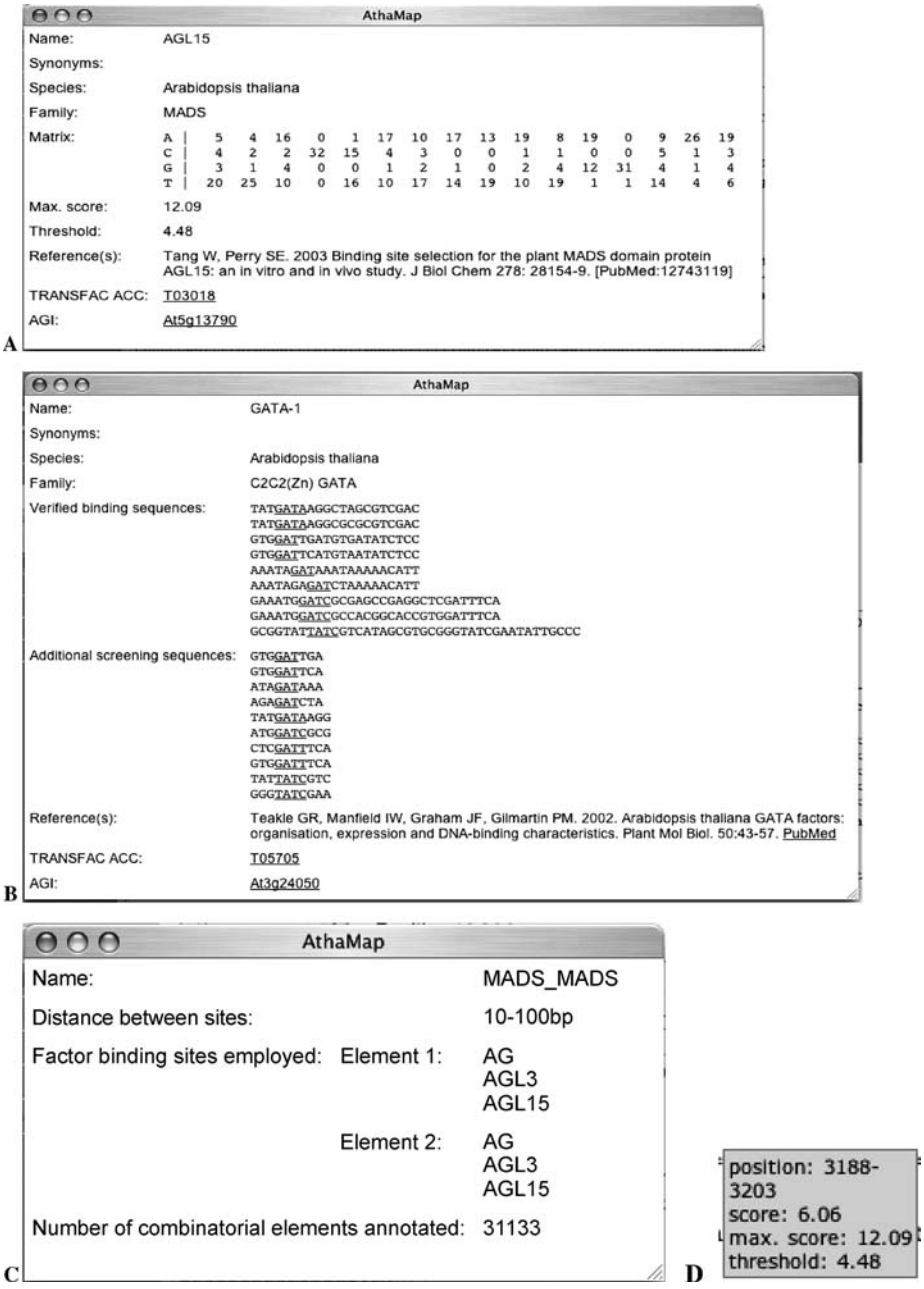


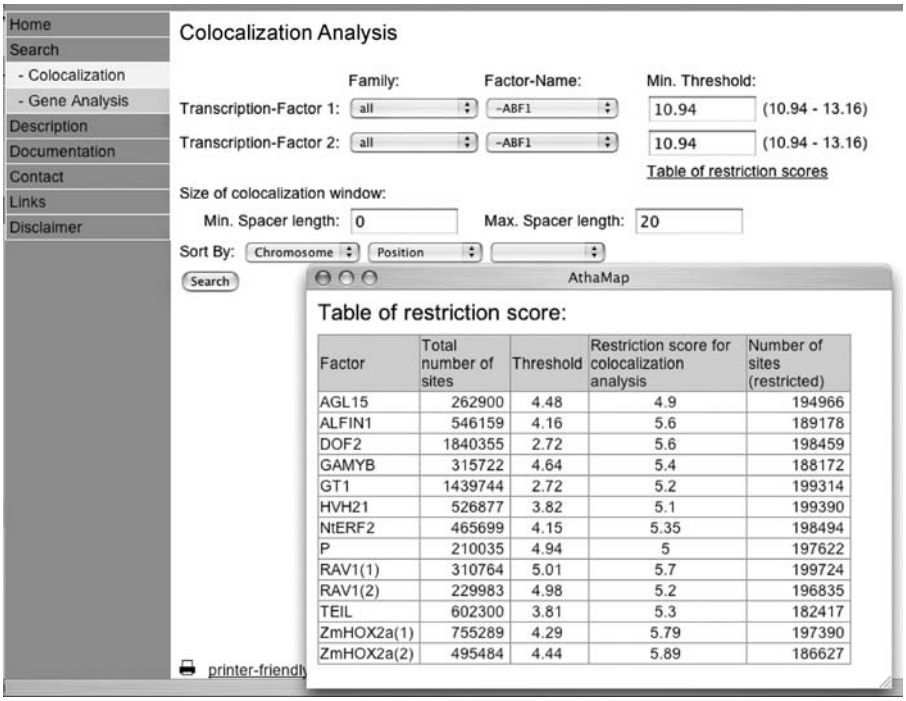
Figure 16.3 Pop-up windows linked to TFs that have binding sites determined with a matrix (A), with single experimentally verified sites (B), or that were used for identification of a combinatorial element (C). (D) A tool tip box linked to a TFBS identified with the matrix shown in A.

Figure 16.3C shows the information that is provided in a pop-up window for a combinatorial element. The distance between the first nucleotides of the two sites is indicated as well as all TFs that were used to determine the combinatorial element. Also, the total number of these combinatorial elements is shown.

It is also possible to restrict the number of displayed TFBS based on their sequence conservation. To restrict a search to those TFBS that are highly conserved a restriction value between 1 and 99 can be entered in the search window (Figure 16.1). A value of 50 means that only those TFBS are displayed that have a score that results when 50% of the difference between maximum score and threshold is added to the threshold score. If, for example, the maximum score is 6 and the threshold is 2, entering a 50 would result in the display of TFBS that have a score of at least 4. This restriction is useful, for example when multiple TFs of the same family are proposed to bind a specific TFBS. Applying a restriction will uncover those TFs that may have a higher binding affinity to this TFBS.

16.2.1.2 **Co-Localization Analysis**

Another option in AthaMap is the detection of co-localizing TFBS. This is useful for detecting TFBS of interacting TFs that occur in close proximity to their target genes. Figure 16.4 shows a composite screen of the Co-localization Analysis tool [9,11].



Factor	Total number of sites	Threshold	Restriction score for colocalization analysis	Number of sites (restricted)
AGL15	262900	4.48	4.9	194966
ALFIN1	546159	4.16	5.6	189178
DOF2	1840355	2.72	5.6	198459
GAMYB	315722	4.64	5.4	188172
GT1	1439744	2.72	5.2	199314
HVH21	526877	3.82	5.1	199390
NtERF2	465699	4.15	5.35	198494
P	210035	4.94	5	197622
RAV1(1)	310764	5.01	5.7	199724
RAV1(2)	229983	4.98	5.2	196835
TEIL	602300	3.81	5.3	182417
ZmHOX2a(1)	755289	4.29	5.79	197390
ZmHOX2a(2)	495484	4.44	5.89	186627

Figure 16.4 The Co-localization Analysis web tool and the 'table of restriction scores' pop-up window.

Although an unrestriced co-localization analysis with all TFBS in AthaMap is desirable, certain restrictions due to the time-performance of the web server apply. First, the difference between maximal and minimal spacer must not exceed 50 (size of co-localization window, Figure 16.4). In addition to 0 and 50 as the minimal and maximal spacer, respectively, other values can also be entered, for example, 100 and 150. Second, for 13 TFs the number of TFBS that can be entered into an online co-localization analysis should be restricted to about 200 000. This applies to all matrix-based TFBS for which more than 200 000 have been annotated. The link 'Table of restriction scores' will show a table in a pop-up window (Figure 16.4) that displays all TFs for which a restriction has been implemented. This can be achieved by selecting a restriction score higher than the threshold score required to obtain less than 200 000 TFBS.

For a co-localization analysis, the user can select two TFs from the list of all TFs for which TFBS are annotated in AthaMap. Also combinatorial elements can be selected. The three different types of TFBS that can be selected are indicated in the selection list (Factor-Name, Figure 16.4). Matrix-based TFBS are preceded by '–' as shown in Figure 16.4 for ABF1. Combinatorial elements are preceded by '=' and TFBS based on single sites by '>' in front of the TF name. Because the list of displayed TFBS is extensive, it is also possible to restrict this list to those that belong to a specific factor family (Family, Figure 16.4). For matrix-based TFBS it is also possible to increase the threshold score to restrict the search to higher conserved TFBS. For this, the threshold score and the maximum score determined by the program Patser is displayed next to the factor name (10.94–13.16 in case of ABF1, Figure 16.4). In the case shown, a user-defined threshold score has to be higher than 10.94.

The result page shows information on the TFs selected, the number of TFBS that were used for the co-localization analysis, the spacer length, and the minimum threshold. Below this information a list of all combinatorial elements is shown. This list displays the position of the two TFBS, their orientation, the spacer between both, and the nearest gene with the distance to the start codon. A minus means that the element occurs upstream of the closest translation start site. Links are implemented from the table to permit the display of the gene or combinatorial element and to show the sequence and TFBS context. To analyze the detected genes further, a link permits the export of the gene IDs to the Gene Analysis web tool of AthaMap (see below). Another link that exports the gene IDs to the PathoPlant database was also implemented [15,16]. This allows the analysis of the identified genes for co-regulation during plant pathogen interactions. The 'show overview' link on the result page is useful if a very extensive list of co-localizations is obtained. This results in a table that summarizes the number of co-localizations with the same spacer length.

16.2.1.3 **Gene Analysis**

The gene analysis web tool serves to determine common or missing TFBS in a set of genes [11]. This can be used, for example, to analyze a set of co-regulated genes. Figure 16.5 shows a screen of the Gene Analysis tool after activating the 'Demo' button. In this example a list of three gene IDs is submitted. The default area of these genes inspected for TFBS is -500 to $+50$ relative to the start codon. This region can

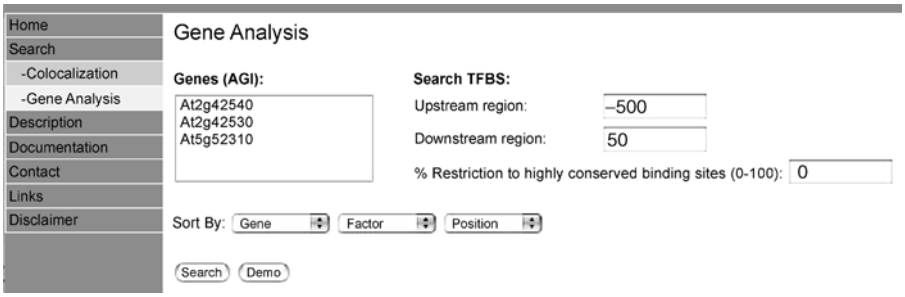


Figure 16.5 The Gene Analysis web tool (Demo).

be changed but the area inspected must not exceed 2000 bp upstream and downstream. Also, the list of gene IDs entered must not be longer than 100. It is also possible to select how the result is sorted. It is possible to sort by submitted gene, TF family, TFBS position, orientation, and distance from the identified TFBS to the start codon. When the list of genes is submitted, the result will be shown in the same window in a table. This lists the genes submitted for analysis and all factors in the corresponding factor family for which positions were detected in the selected region of the submitted genes. The positions are linked to the sequence display window. Further information on matrix-based TFBS such as maximum score and threshold score and the score of the identified TFBS is also shown. The relative distance from the start codon also indicates whether the TFBS is identified upstream or downstream of the translation start and whether the orientation of the TFBS is the same (+) or opposite (−) to the direction of the transcription of the gene. If Internet Explorer is used as a web browser, this table can be directly exported into a Microsoft Excel table for further analyses. Because these result tables are usually very long, further display options are provided. 'Show overview' will summarize the total number of TFBS detected for a specific TF. 'Show factors that are common in genes' will show all TFs for which TFBS were found in all of the genes. In this table, TFBS that occur in the submitted genes are displayed hierarchically, starting with those at the top that occur in most or all of the submitted genes and those at the bottom of the list that do not occur in the genes. This list also shows the total number of respective TFBS detected in the selected region of the submitted genes and compares this number with the theoretical number of TFBS that would be expected. These values can be subjected to a statistical analysis to obtain an indication of the significance of the observation [11].

16.2.1.4 **External Links**

AthaMap has been linked with other databases for further information on TFs and on all *Arabidopsis* genes. If the TF for which TFBS were determined is annotated to the TRANSFAC database, the TRANSFAC accession number (Figure 16.3A) directly leads to the factor table in the TRANSFAC database [14]. This information is only displayed for licensed users of the TRANSFAC professional database. If the factor is from *A. thaliana*, the gene ID (Figure 16.3A, AGI) links to the

gene locus in TAIR [13]. Other links show up below the sequence display window on a search result (not shown). If a gene is shown in the sequence display window it is linked to the TAIR, MIPS and TIGR databases [12,13,17]. Additional external links are listed on the Links page of the website. In addition to TRANSFAC, TAIR, MIPS and TIGR, a link to the database of *Arabidopsis* transcription factors DATF is also implemented [18].

References

1 Hertz, G.Z. and Stormo, G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

2 Kel, A.E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V. and Wingender, E. (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Research*, **31**, 3576–3579.

3 Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids*, **23**, 4878–4884.

4 Yan, T., Yoo, D., Berardini, T.Z., Mueller, L.A., Weems, D.C., Weng, S., Cherry, J.M. and Rhee, S.Y. (2005) PatMatch: a program for finding patterns in peptide and nucleotide sequences. *Nucleic Acids*, **33**, W262–266.

5 Davuluri, R.V., Sun, H., Palaniswamy, S.K., Matthews, N., Molina, C., Kurtz, M. and Grotewold, E. (2003) AGRIS: arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. *BMC Bioinformatics*, **4**, 25.

6 Steffens, N.O., Galuschka, C., Schindler, M., Bülow, L. and Hehl, R. (2004) AthaMap: an online resource for *in silico* transcription factor binding sites in the *Arabidopsis thaliana* genome. *Nucleic Acids*, **32**, D368–372.

7 O'Connor, T.R., Dyreson, C. and Wyrick, J.J. (2005) Athena: a resource for rapid visualization and systematic analysis of Arabidopsis promoter sequences. *Bioinformatics*, **21**, 4411–4413.

8 Obayashi, T., Kinoshita, K., Nakai, K., Shibaoka, M., Hayashi, S., Saeki, M., Shibata, D., Saito, K. and Ohta, H. (2007) ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. *Nucleic Acids*, **35**, D863–D869.

9 Steffens, N.O., Galuschka, C., Schindler, M., Bülow, L. and Hehl, R. (2005) AthaMap web tools for database-assisted identification of combinatorial *cis*-regulatory elements and the display of highly conserved transcription factor binding sites in *Arabidopsis thaliana*. *Nucleic Acids*, **33**, W397–402.

10 Bülow, L., Steffens, N.O., Galuschka, C., Schindler, M. and Hehl, R. (2006) AthaMap: from *in silico* data to real transcription factor binding sites. *In Silico Biology*, **6**, 0023.

11 Galuschka, C., Schindler, M., Bülow, L. and Hehl, R. (2007) AthaMap web-tools for the analysis and identification of co-regulated genes. *Nucleic Acids*, **35**, D857–D862.

12 Haas, B.J., Wortman, J.R., Ronning, C.M., Hannick, L.I., Smith, R.K., Jr., Maiti, R., Chan, A.P., Yu, C., Farzad, M., Wu, D. *et al.* (2005) Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release. *BMC Biology*, **3**, 7.

13 Rhee, S.Y., Beavis, W., Berardini, T.Z., Chen, G., Dixon, D., Doyle, A., Garcia-

Hernandez, M., Huala, E., Lander, G., Montoya, M. *et al*. (2003) The arabidopsis information resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids*, **31**, 224–228.

14 Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids*, **31**, 374–378.

15 Bülow, L., Schindler, M., Choi, C. and Hehl, R. (2004) PathoPlant®: a database on plant–pathogen interactions. *In Silico Biology*, **4**, 529–536.

16 Bülow, L., Schindler, M. and Hehl, R. (2007) PathoPlant®: a platform for microarray expression data to analyze co-regulated genes involved in plant defense responses. *Nucleic Acids*, **35**, D841–D845.

17 Schoof, H., Ernst, R., Nazarov, V., Pfeifer, L., Mewes, H.W. and Mayer, K.F. (2004) MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource for plant genomics. *Nucleic Acids*, **32**, D373–D376.

18 Guo, A., He, K., Liu, D., Bai, S., Gu, X., Wei, L. and Luo, J. (2005) DATF: a database of Arabidopsis transcription factors. *Bioinformatics*, **21**, 2568–2569.