DNA Research 12, 247–256 (2005) doi:10.1093/dnares/dsi011

RARTF: Database and Tools for Complete Sets of Arabidopsis Transcription Factors

Kei Iida,1,2,4,* Motoaki Seki,1,3 Tetsuya Sakurai,1,2,6 Masakazu Satou,1,2,6 Kenji Akiyama,1,2,6 Tetsuro Toyoda,2 Akihiko Konagaya,2 and Kazuo Shinozaki1,3,5,6

Plant Mutation Exploration Team, Plant Functional Genomics Research Group1 and Genomic Knowledge Base Research Team, Bioinformatics Group, RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan2, Laboratory of Plant Molecular Biology, RIKEN Tsukuba Institute, 3-1-1 Koyadai, Tsukuba 305-0074, Japan3, Faculty of Bio-Science, Nagahama Institute of Bio-Science and Technology, 1266 Tamura-cho, Nagahama 526-0829, Japan4 and CREST, Japan Science and Technology Corporation (JST), Japan5

6Present address: RIKEN Plant Science Center (PSC), RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan

(Received 2 March 2005; revised 6 June 2005)

Abstract

More than 5% of all genes in the Arabidopsis thaliana genome have been assumed to code for transcription factors. However, it has been difficult to accurately identify them. To construct proper sets of transcription factors, we used PSI-BLAST and InterProScan, and also checked several families manually. Especially to determine major Arabidopsis transcription factors (MYB, AP2/EREBP, bHLH, NAC, MADS, bZIP, WRKY), we compared the PSI-BLAST search results with those in recent reports. Finally, we identified 1968 proteins as transcription factors (7.4% of all Arabidopsis genes). We established a database named RARTF (RIKEN Arabidopsis Transcription Factor database, http://rarge.gsc.riken.jp/rartf/) based on the identified transcription factors. In RARTF, we provide information on the functional motif of transcription factors, full-length cDNAs, alternative pre-mRNA splicing events and Ac/Ds transposon-tagged mutants. We also provide expression profiles of 400 transcription factor genes in six experiments. We will report expression profiles of all transcription factor genes in various plant tissues under various stress and hormone conditions in the near future.

Key words: Arabidopsis thaliana; transcription factor; database; full-length cDNA; expression profile

1. Introduction

Arabidopsis thaliana is a model plant for studying developmental processes, responses to stress and signal transduction. In 2000, its whole genomic sequence consisting of ~26 000 genes was determined.1 A genome-wide comparative analysis of transcription factors in Arabidopsis, Drosophila melanogaster, Caenorhabditis elegans and Saccharomyces cerevisiae by Riechmann et al.2 revealed that the Arabidopsis genome codes for ~1500 transcription factors. This is ~6% of the total genes, which is a higher percentage than that in either D. melanogaster, C. elegans or S. cerevisiae. These transcription factors often have separate roles for different developmental stages, tissues or stress responses.3–5 Identification of the transcription factors is essential for understanding the life system of Arabidopsis. Owing to the success of full-length cDNA projects,6–8 the annotation of gene structures has been greatly improved, and the information of all transcription factors needs to be updated based on the new gene annotation. Furthermore, some novel transcription factor families have been identified after the report by Riechmann et al.9–18

A new transcription factor database incorporating accumulated information on upstream regulatory sequences of genes, alternative pre-mRNA splicing events,
transposon-tagged mutant lines and so on is needed. In this study, we first searched for the transcription factors with using PSI-BLAST.\textsuperscript{19} Riechmann et al.\textsuperscript{2} classified the transcription factors in Arabidopsis into \textasciitilde30 families. Homology search tools or motif search tools such as BLAST\textsuperscript{19} or InterProScan\textsuperscript{20} are commonly used to search for protein family members by amino acid sequence similarities. Although these algorithms are very useful, none of them are perfect. PSI-BLAST is a powerful algorithm in the search for protein families based on amino acid sequence similarities. With PSI-BLAST, we can search for similar sequences from the database using the position-specific score matrix (PSSM). PSSM places more weight on highly conserved sites. Because transcription factors often have highly conserved DNA-binding domains, we expected to identify all the transcription factors with high sensitivity and specificity from all protein sequences of Arabidopsis with PSI-BLAST. To construct a proper set of Arabidopsis transcription factors, we also used InterProScan and checked for several transcription factor families manually. Finally, we obtained 1968 transcription factors. Based on the transcription factors identified, we constructed a new Arabidopsis transcription factor database named RARTF. We provide the list and the information on all Arabidopsis transcription factors from RARTF.

2. Materials and Methods

2.1. Datasets

We used amino acid sequences of all predicted Arabidopsis proteins released from MAtDB (MIPS A. thaliana Genome Database) of MIPS (Munich Information Centre for Protein Sequences, ftp://ftp.mips.gsf.de/cress/, v110103).\textsuperscript{21} In addition, we used amino acid sequences of a bZIP transcription factor, AREB1 (AB017160), not annotated by the MIPS protein entry code. We used 26 594 protein sequences in total.

2.2. Identification of transcription factors

We used the classification of transcription factor families reported by Riechmann et al.\textsuperscript{2} We added another 12 families as novel transcription factor families of Arabidopsis to this classification.\textsuperscript{9–18} To find novel transcription factor families, we used PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed).

We chose one representative sequence for each transcription factor family or subfamily. We used the sequence in table 3 of Riechmann et al.\textsuperscript{2} when the accession number was found, and we selected a protein sequence which had the InterPro\textsuperscript{22} motif of each transcription factor family when not. We used these representative sequences as query sequences for the PSI-BLAST search. We used the reported sequences for the 12 novel transcription factor families. We used the entire amino acid sequence of each representative transcription factor for PSI-BLAST search except for ‘GARP, ARR-B class’ family (TF10\_2). For this family, we selected the DNA-binding region for PSI-BLAST query. We used PSI-BLAST version 2.2.5 to identify the transcription factors from all Arabidopsis proteins. In the first step, we constructed PSSMs\textsuperscript{19} based on query sequences. The parameters we used were the maximum round of iteration \((j) = 10\), filter \((F) = \text{on}\), \(E\)-value threshold for iteration \((h) = 0.0001\), database \((d) = \text{amino acid sequences of all Arabidopsis proteins}\). In the next step, we searched for transcription factors with the constructed PSSMs. The parameters in the second step were \(E\)-value threshold \((e) = 0.001\), filter \((F) = \text{on}\). After the PSI-BLAST search, we used InterProScan for ‘C2H2-type zinc-finger’ family \((\text{TF}_5\_3)\) to complement PSI-BLAST search results. We performed a manual check for major Arabidopsis transcription factor families (MYB, AP2/EREBP, bHLH, NAC, MADS, bZIP, WRKY) and ‘PcG; Esc class’ transcriptional regulator family \((\text{TF}_6\_0)\). For check processes of major transcription factor families, we referred to the most recent papers.\textsuperscript{23–30}

2.3. Validation of sensitivities and specificities

We validated pre-fixed PSI-BLAST search results on seven major transcription factor families: MYB, AP2/EREBP, bHLH, NAC, MADS, bZIP, WRKY. We calculated the sensitivities and specificities using the new corrected list we made for each transcription factor family based on the most recent papers,\textsuperscript{23–30} which we refer to as the ‘correct list’ hereafter. Sensitivity is calculated as \([\text{number of transcription factors identified in this study in the correct list}] / [\text{number of all transcription factors identified in this study}] = \text{[True positive]} / ([\text{true positive}] + [\text{false negative}])\). Specificity is calculated as \([\text{number of transcription factors identified in this study in the correct list}] / [\text{number of all transcription factors identified in this study}] = \text{[True positive]} / ([\text{true positive}] + [\text{false positive}])\). For comparison, we searched transcription factors by BLAST\textsuperscript{19} and InterProScan.\textsuperscript{20} In the BLAST search, we used a parameter for \(E\)-value threshold \((e) = 0.01\) which revealed the best sensitivities and specificities within \(E\)-value \(= 0.1, 0.01, 0.001, 0.0001, 0.00001\). In InterProScan search, we used IPR001005 motif (PS50090, PF00249, SM0071) for MYB family, IPR001471 motif (PF00847, PR00367, SM0071) for AP2/EREBP family, IPR001092 motif (PS50888, SM00353, PF00010, SSF47459) for bHLH family, IPR003441 motif (PF02365) for NAC family, IPR002100 motif (PS50066, SM00432, PF00319, PR00404, SSF55455, PS00350) for MADS family, IPR004827 motif (PF00170, PS50217, PS00036, SM00338) for bZIP family, IPR003657 motif (PF03106, PS50811) for WRKY family. We used InterProScan version 3.3 and dataset version 7.2 with default
parameters. We listed proteins that had motifs with a true decision or matched sequence pattern.

2.4. Database construction

We classified the transcription factors into families and subfamilies according to Riechmann et al.\textsuperscript{2} In each family or subfamily, we included transcription factors which were identified by PSI-BLAST search with one query sequence. As a result of PSI-BLAST search, we found that some transcription factors belonged to more than two families. To re-organize these transcription factors, we made two lists: the Phase 1 list includes all transcription factors obtained by the search using each query sequence, and the Phase 2 list gives the proteins detected with the smallest E-value. In the Phase 1 list, some transcription factors were annotated to more than two families. In the Phase 2 list, the transcription factor belonged to one family. In the Phase 2 list, there is no redundancy. Based on these two lists, we constructed the transcription factor database: RARTF, in which we related various contents based on AGI gene codes\textsuperscript{21} (Fig. 1).

3. Results and Discussion

3.1. Identification of transcription factors

We analyzed 54 transcription factor families and 7 transcriptional regulator families according to Riechmann et al.\textsuperscript{2} We also analyzed 12 additional transcription factor families\textsuperscript{9–18} reported after 2000 (Table 1). Seventeen families had some subfamilies. We used 121 representative sequences of each family or subfamily. We chose representative sequences from Riechmann et al.\textsuperscript{2} or other reported novel transcription factors\textsuperscript{9–18} (see Materials and Methods). We used the entire amino acid sequence of representative proteins for PSI-BLAST search except for the ‘GARP, ARR-B class’ family (TF10.2). For ‘GARP, ARR-B class’ family, we chose the DNA-binding region for PSI-BLAST search to obtain much superior search results. Using these sequences as a query, we performed a PSI-BLAST\textsuperscript{19} search for all transcription factors in Arabidopsis (Fig. 1). We performed the PSI-BLAST search using as a parameter the maximum round of iteration set 10. With this parameter, converged results were obtained in 116 out of 121 cases. In two cases (TF.32 and TF.46), there was no significant hit on the first search round. Unconverged results were obtained in another three cases (TF.1, TF.17.2 and TF.22.1), and we manually checked them.

In a preliminary analysis, we found it difficult to identify members belonging to the ‘C2H2-type zinc-finger’ family (TF.5) because there were a few conserved sites in this family and there were many gaps within their zinc-finger sequences. By a PSI-BLAST search of the proteins detected by InterProScan,\textsuperscript{20} we listed 171 proteins into TF.5.3. The preliminary analysis revealed another family in which many proteins were clearly not transcription factors, i.e. ‘PCG; Esc class’ transcriptional regulator family (TF.60). We selected three proteins from the 200 listed proteins in this group, by eye.

The Arabidopsis genome has several large transcription factor families that have 100–200 members. These families have been well studied and are known to have important functions. Especially for these major transcription factor families, we compared the PSI-BLAST search results with recent publication. We referred to a review by Stracke et al.\textsuperscript{23} for the MYB family, Gutterson and Reuber\textsuperscript{24} in 2004 and Sakuma et al.\textsuperscript{25} in 2002 for the AP2/EREBP family, Bailey et al. in 2003\textsuperscript{26} for the bHLH family, Ooka et al. in 2003 for the NAC family, Parenicova et al.\textsuperscript{27} in 2003 for the MADS family, Jakoby et al.\textsuperscript{28} for the bZIP family and Ulker and Somssich in 2004\textsuperscript{29} for the WRKY family. We found 15 false negatives and 38 false positives (Table 2). In each case, we checked their transcription factor lists and redescribed them with AGI codes of our dataset. We corrected our PSI-BLAST search results according to this check.

We added AREB1 (accession number: AB017160)\textsuperscript{31} to ‘ABSCISIC ACID-INSUSCEPTIVE’ subfamily of bZIP family. AREB1 is mapped onto chromosome 1 of Arabidopsis, but the locus and entry of AREB1 do not exist in the MIPS dataset. We checked the proteins not identified as transcription factors and added some proteins to the ‘TF_other’ list, if one protein was annotated as transcription factor. We listed 17 proteins as general transcription factors, GTF. In total, we obtained 1968 transcription factors and transcriptional regulators, which account for 7.4% of all Arabidopsis proteins (Table 1).

The number of identified transcription factor genes, 1968, was larger than the 1533 reported by Riechmann et al.\textsuperscript{2} This difference was caused by the added
Table 1. Transcription factor families in *Arabidopsis thaliana*.

| Family                  | Subfamily                        | Query sequence | Genes in family | Phase 1 | Phase 2 |
|-------------------------|----------------------------------|----------------|-----------------|---------|---------|
| MYB superfamily         | CAPRICE                          | BAAA21917      | 238             | 136     | 4       |
|                         | LHY                              | CAA07004       | 229             |         | 8       |
|                         | WEREWOLF                         | AAF18939       | 170             | 12      |         |
|                         | ATR1                             | AAC16897       | 180             | 41      |         |
|                         | GL1                              | AAC97387       | 146             | 95      |         |
|                         | CIRCADIAN CLOCK ASSOCIATED 1     | AAB40525       | 227             | 29      |         |
| AP2/EREBP               | ABSCISIC ACID-INSENSITIVE4       | AACC9489       | 145             | 144     | 2       |
|                         | AINTEGUMENTA                     | AAB17364       | 145             | 25      |         |
|                         | APETALA2                         | AAC13770       | 144             | 18      |         |
|                         | CBF1                             | NP_172721      | 145             | 39      |         |
|                         | DREB2A                           | BAA33794       | 137             | 9       |         |
| bHLH                    | LONG HYPOCOTYL IN FAR-RED1       | AAG40617       | 157             | 138     | 38      |
|                         | TRANSPARENT TESTA8               | Q9FT81         | 141             | 106     |         |
|                         | checked group                    | —              | 13              | 13      |         |
| NAC                     | CUP-SHAPED COTYLEDON2            | BAA19529       | 106             | 106     | 106     |
| C2H2(Zn)                | FIS2                             | AAD09104       | 177             | 7       | 7       |
|                         | SUPERMAN                         | AAC49116       | 65              | 0       |         |
|                         | InterProScan C2H2                | —              | 171             | 170     |         |
| HB                      | ATHB-2                           | Q05466         | 101             | 90      | 19      |
|                         | BELL1                            | A57632         | 24              | 13      |         |
|                         | GLABRA2                          | P66007         | 71              | 6       |         |
|                         | REVOLUTA/INTERFASCICULAR FIBERLESS1 | AAF42938   | 73             | 12      |         |
|                         | SHOOTMERISTEMLESS                | Q88874         | 24              | 9       |         |
|                         | WUSCHEL                         | CAAB9986       | 63              | 10      |         |
|                         | ANTHOCYANINLESS2                 | AAD47139       | 96              | 28      |         |
| MADS                    | PISTILLATA                       | P48007         | 106             | 104     | 8       |
|                         | SHORT VEGETATIVE PHASE           | Q9FVC1         | 105             | 11      |         |
|                         | SOC1                             | O64645         | 103             | 11      |         |
|                         | SHATTERPROOF1                    | AAM64275       | 105             | 7       |         |
|                         | AGAMOUS                          | P17839         | 105             | 11      |         |
|                         | SEPALLATA1                       | CAC83066       | 105             | 13      |         |
|                         | FRUITFULL                        | Q88876         | 105             | 5       |         |
|                         | FLOWERING LOCUS C                | AAN04056       | 105             | 6       |         |
|                         | CAULIFLOWER                      | AAA64789       | 105             | 9       |         |
|                         | APETALA3                         | A12095         | 104             | 9       |         |
|                         | APETALA1                         | P35631         | 105             | 10      |         |
|                         | ANR1                             | CAB09793       | 105             | 6       |         |
| bZIP                    | ABSCISIC ACID-INSENSITIVE5       | AAD21438       | 75              | 73      | 43      |
|                         | PERIANTHIA                       | AAD19660       | 71              | 12      |         |
|                         | checked group                    | —              | 2               | 2       |         |
| WRKY(Zn)                | S2443                            | 72              | 72              |         |         |
| GARP                    | G2-like                          | AAD55941       | 57              | 55      | 38      |
|                         | ARR-B class                      | BAA4528        | 57              | 13      |         |
| C2C2(Zn)                | Dof                              | CAA66900       | 126             | 33      | 33      |
|                         | CO-like                          | A56133         | 51              | 51      |         |
|                         | GATA                              | AAB61058       | 37              | 37      |         |
|                         | YABBY                            | AAD30526       | 5               | 5       |         |
| CCAAT                   | HAP2 type                        | A26771         | 37              | 10      | 10      |
|                         | HAP3 type                        | P13434         | 13              | 9       |         |
|                         | HAP4 type                        | S37936         | 0               | 0       |         |
|                         | HAP5 type                        | Q02516         | 27              | 14      |         |
|                         | Dt1                              | AAB51375       | 13              | 4       |         |
| Family         | Subfamily                     | Query sequence | Genes in family<sup>a</sup> | Phase 1 | Phase 2 |
|---------------|-------------------------------|----------------|-----------------------------|---------|---------|
| GRAS          | SHORT-ROOT                    | AAF75234       | 32                          | 32      | 4       |
| SCARECROW     |                               | AAB06318       | 32                          | 32      | 15      |
|               | Repressor of ga-1             | BAC41902       | 31                          | 31      | 6       |
| PAT1          |                               | AAF73237       | 32                          | 32      | 7       |
| Gibberelin Insensitive |               | CAB51557       | 32                          | 32      | 0       |
| Trihelix      |                               | S39484         | 31                          | 31      | 31      |
| HSF           |                               | CAB63801       | 27                          | 27      | 27      |
| TCP           |                               | AAC26786       | 24                          | 24      | 24      |
| ARF           | MONOPTEROS/ARF5               | P93024         | 119                         | 119     | 37      |
|               | NPH4/ARF7                     | AAF71831       | 119                         | 119     | 37      |
|               | ETTIN/ARF3                    | T03278         | 45                          | 45      | 13      |
| C3H-type 1(Zn)| putative zinc finger protein  | NP_176987      | 38                          | 38      | 37      |
| C3H-type 2(Zn)|                               | CAA65242       | 10                          | 10      | 10      |
| SBP           |                               | CAB56581       | 17                          | 17      | 17      |
| Nin-like      |                               | CAB61243       | 14                          | 14      | 14      |
| AB3/VP1       | FUSCA3                        | AAC5247        | 112                         | 95      | 18      |
|               | ABSCISIC ACID-INSENSITIVE3    | CAA05484       | 112                         | 112     | 33      |
| TUB           |                               | AAC28518       | 11                          | 11      | 11      |
| E2F/DP        |                               | O00716         | 8                           | 8       | 6       |
|               |                               | Q64163         | 2                           | 2       |         |
| CPP(ZN)       |                               | CAA00928       | 11                          | 11      | 8       |
| Alfin-like    |                               | AAA20093       | 60                          | 60      | 47      |
| EIL           |                               | AAC49749       | 6                           | 6       | 6       |
| LFY           |                               | AAA32826       | 3                           | 3       | 3       |
| Other         | HUA2                          | AAD31171       | 27                          | 18      | 15      |
|               | NFR1                          | AAC49611       | 7                           | 7       |         |
|               | STERILE APETALA              | CAA11128       | 1                           | 1       |         |
|               | SPOROCYTELESS/NOZZLE         | AAD45344       | 1                           | 1       |         |
| PAIRED(w/o HB)|                               | AAF47314       | 88                          | 88      | 2       |
| Swi4/Swi6     |                               | Q91YU8         | 1                           | 1       | 1       |
| Aux/IAA       |                               | AAC39440       | 49                          | 49      | 21      |
| HMG-box       |                               | BAA02719       | 15                          | 15      | 11      |
| ARID          |                               | AAC62899       | 6                           | 6       | 6       |
| JUMONJI       |                               | T03254         | 15                          | 15      | 13      |
| PcG; E(z) class | CURLY LEAF | CAA71599       | 33                          | 33      | 20      |
|               | MEDEA                         | AAC39446       | 32                          | 32      | 12      |
| PcG; Esc class | FERTILIZATION-INDEPENDENT ENDOSPERM | AAD23584 | 3 | 3 | 3 |
| CBF5          |                               | O13473         | 2                           | 2       | 2       |
| SW13          |                               | AAO42112       | 11                          | 11      | 6       |
| TGA3          |                               | S46523         | 82                          | 82      | 27      |
| Pti4          |                               | T07686         | 127                         | 127     | 5       |
| Pti5          |                               | T07689         | 144                         | 144     | 5       |
| Pti6          |                               | T07728         | 146                         | 146     | 18      |
| ERF           |                               | BAB62912       | 146                         | 146     | 19      |
| PHD-finger    |                               | CAC69664       | 10                          | 10      | 10      |
| VIP3          |                               | CAB75507       | 1                           | 1       | 1       |
| LIM-domain    |                               | NP_680133      | 6                           | 6       | 6       |
| AT-hook       |                               | NP_565769      | 31                          | 31      | 31      |
| Sir2          |                               | NP_200387      | 2                           | 2       |         |
| Other TF in MIPS |               |                | 40                          | 40      | 40      |
| GTF           |                               | 17             | 17                          | 17      |         |

<sup>a</sup>Number of transcription factors classified into each family. This count is non-redundant within each family, but is redundant when compared with another family.
transcription factor families and additional members of each family identified by PSI-BLAST. Recently, Jiao et al. reported 1864 transcription factors in Arabidopsis, which were similar to those in our list (data not shown).

3.2. Comparison of sensitivity and specificity

We used PSI-BLAST, InterProScan and checked their results manually to obtain a proper transcription factor set. In this process, we referred several papers about their transcription factor families. But this identification strategy is not used in all cases, because not all transcription factor families are well studied and reported in reviews. It is important to recognize the sensitivity and specificity of the PSI-BLAST search results without manual check. We checked the sensitivity and specificity of pre-fixed PSI-BLAST search results on major transcription factor families (MYB, AP2/EREBP, bHLH, NAC, MADS, bZIP, WRKY). We also compared the results with those obtained using BLAST and InterProScan. We selected query sequences according to the list written by Riechmann et al. We thought their sequences were enough to find all proteins in each family with PSI-BLAST, because we got good consistency between PSI-BLAST search results and description of reviews (Table 2). We used the same query sequences for BLAST search for this comparison.

The sensitivity of PSI-BLAST, InterProScan and BLAST was 0.981, 0.980 and 0.924, respectively (Table 2; Fig. 2), there being little difference between PSI-BLAST and InterProScan. These two methods gave a higher sensitivity than BLAST. We evaluated the specificity in six families, because we could not calculate the number of false-positive proteins in the MYB family. The specificity of PSI-BLAST, InterProScan and BLAST was 0.944, 0.981 and 0.962, respectively (Table 2; Fig. 2). Although there was little difference among the three methods, InterProScan showed the highest specificity and PSI-BLAST the lowest. These results showed that the InterProScan is the best method for identifying transcription factors with high sensitivity and specificity. However, PSI-BLAST may be the best tool for detecting all the transcription factors since PSI-BLAST showed the best sensitivity. PSI-BLAST can detect

| TF name | PSI-BLAST Sensitivity | PSI-BLAST Specificity | InterProScan Sensitivity | InterProScan Specificity | BLAST Sensitivity | BLAST Specificity |
|---------|-----------------------|-----------------------|--------------------------|--------------------------|------------------|------------------|
| MYB     | 129/129               | 129/129               | 129/129                  | 129/129                  | 129/129          | 129/129          |
| AP2     | 145/145               | 145/146               | 145/145                  | 145/145                  | 144/145          | 144/146          |
| bHLH    | 144/157               | 144/149               | 153/157                  | 153/155                  | 116/157          | 116/117          |
| NAC     | 106/106               | 106/109               | 100/106                  | 100/100                  | 104/106          | 104/106          |
| MADS    | 106/106               | 106/118               | 101/106                  | 101/105                  | 100/106          | 100/116          |
| bZIP    | 73/75                 | 73/89                 | 74/75                    | 74/79                    | 65/75            | 65/67            |
| WRKY    | 72/72                 | 72/73                 | 72/72                    | 72/73                    | 72/72            | 72/73            |

aThe denominator represents no. of transcription factors in the correct list, the numerator represents no. of true positive.
bThe denominator represents no. of true positive and false positive identified in this study, the numerator represents no. of true positive.

![Figure 2](image-url). Comparison of three methods. (A) PSI-BLAST and InterProScan showed a higher sensitivity than BLAST. (B) PSI-BLAST showed a slightly worse specificity than either InterProScan or BLAST.
sequence similarities of proteins even if the proteins have partial sequences. By this analysis, we found At2g25820 in a PSI-BLAST search result of the AP2/EREBP family. Indeed, At2g25820 has a partial AP2-type DNA-binding domain (Fig. 3), but if we include the upstream genomic region together, it has the entire DNA-binding domain (Fig. 3). No full-length cDNA or EST supported the translation of the upstream region of the annotated first methionine codon of At2g25820, but the amino acid sequence translated from 5′-UTR has a highly conserved AP2-type DNA-binding domain. Thus, there is a high enough probability that At2g25820 with the entire AP2 domain is translated. A motif search based on the Hidden Markov Model algorithm cannot detect this partial DNA-binding domain. This example showed that PSI-BLAST can detect sequence similarities even from partial protein sequences. Genome sequencing projects on Poplar or Lotus japonicus are in progress. In many species, there are issues about correct detection of exons or ORFs or CDSs just after sequencing of their genomes. PSI-BLAST is expected to be useful for detecting transcription factors more correctly from these genomes. When a novel transcription factor family member is found, we can search for other members of the family quickly based on one query sequence. All of these points indicate the usefulness of PSI-BLAST for identifying transcription factors efficiently from accumulating genome information of plants.

3.3. RARTF: RIKEN Arabidopsis transcription factor database

We constructed a transcription factor database named RARTF (RIKEN Arabidopsis transcription factor database, http://rarge.gsc.riken.jp/rartf/) based on detected
transcription factors (Figs 1 and 4). RARTF provides information on transcription factor families and each transcription factor, and search tools such as PSI-BLAST and RPS-BLAST based on the PSSMs. We made two lists. In the Phase 1 list, there is redundancy of listed proteins and the Phase 2 list is non-redundant (see Materials and Methods). Multiple alignments of amino acid sequences and a phylogenic tree are available for each family or subfamily (Fig. 4). For each transcription factor, we provide the PSI-BLAST search results, functional motif information of InterProScan. We also made some links for other Arabidopsis databases such as MAtDB of MIPS, the Arabidopsis Information Resource (TAIR) and the Institute for Genomic Research (TIGR) DB. For some proteins, the amino acid sequences in these databases are different. In these cases, we recommend checking all the databases. RARTF provides a RPS-BLAST (Reverse PSI-BLAST) search tool. By using RPS-BLAST, it is possible to check whether the sequence of the amino acid or DNA or mRNA is one of a transcription factor or not. With the PSI-BLAST search tool on RARTF, we provided PSSMs and datasets of latest Arabidopsis proteins and proteins of O. sativa. The latest lists of transcription factors of A. thaliana and O. sativa are available as results of PSI-BLAST searches.

The contents of RARTF linked with RARGE provide more detailed information about transcription factors (Figs 1 and 4). We collected 1072 RIKEN Arabidopsis full-length (RAFL) cDNAs coding for transcription factors. From the full-length cDNAs, we can identify the exon–intron structures accurately. We can identify the upstream sequences of genes by mapping the full-length cDNAs to the genome. From RARTF, the cis-element database of RARGE is available, and cis-elements located in the upstream region of genes that code for transcription factors can be identified. It is important to study the regulatory region of transcription factor genes to understand the transcription network of Arabidopsis.

### Table 3. Expression profiles of DREB family transcription factors under drought stress.

| AGI code   | Gene name                                      | Time(s) after drought stress treatment* |
|------------|------------------------------------------------|-----------------------------------------|
|            |                                                | 1   | 2   | 5   | 10  | 24  |
| At1g43160  | RAP2.6 (At1g43160)                             | 1.01| 9.37| 26.28| 40.76| 19.62|
| At3g14230  | transcription factor EREBP-like protein        | 0.60| 1.04| 1.16 | 1.14 | 2.10 |
| At5g51190  | unknown protein                                | 0.64| 1.41| 1.48 | 0.63 | 1.52 |
| At3g16770  | AP2 domain containing protein RAP2.3           | 0.70| 0.93| 0.69 | 0.60 | 1.21 |
| At3g16770  | AP2 domain containing protein RAP2.3           | 1.06| 1.71| 1.38 | 0.90 | 2.12 |
| At1g72360  | putative AP2 domain transcription factor       | 0.84| 1.76| 2.40 | 1.38 | 1.61 |
| At1g78080  | putative AP2 domain containing protein (At1g78080) | 1.38| 3.10| 1.21 | 0.81 | 1.19 |
| At2g20880  | AP2 domain transcription factor                | 1.61| 24.89| 4.85 | 3.65 | 3.08 |
| At1g22190  | putative protein                               | 3.62| 5.37| 1.62 | 1.83 | 2.45 |
| At4g28140  | putative DNA binding protein                   | 1.58| 38.55| 10.50| 16.28| 6.99 |
| At1g28370  | putative ethylene responsive element binding factor 4 protein | 2.82| 1.86| 1.21 | 1.27 | 2.44 |
| At4g17500  | ethylene responsive element binding factor 1 (frameshift !) | 1.83| 3.90| 3.05 | 1.96 | 2.49 |
| At4g17500  | ethylene responsive element binding factor 1 (frameshift !) | 1.03| 1.84| 1.23 | 0.64 | 0.96 |
| At1g53910  | unknown protein                                | 0.64| 1.29| 1.74 | 2.33 | 4.18 |
| At5g25190  | Ethylene responsive element-like protein       | 0.58| 0.93| 1.16 | 0.32 | 0.50 |
| At4g17490  | ethylene responsive element binding factor-like protein | 1.77| 2.17| 1.55 | 0.60 | 0.92 |
| At1g74930  | putative AP2 domain transcription factor       | 3.40| 6.80| 3.36 | 3.87 | 2.35 |
| At5g61600  | DNA binding protein like                       | 0.76| 1.42| 1.28 | 0.45 | 1.49 |
| At5g61590  | ethylene responsive element binding factor-like protein | 0.48| 0.68| 0.78 | 0.38 | 3.99 |
| At1g22985  | unknown protein                                | 1.01| 4.79| 3.40 | 4.05 | 3.46 |
| At1g68550  | putative AP2 domain transcription factor       | 0.83| 1.65| 1.92 | 0.91 | 1.77 |
| At1g25560  | putative protein                               | 0.43| 0.79| 0.88 | 0.55 | 1.26 |
| At1g68840  | putative DNA binding protein, protein RAV2     | 0.56| 0.84| 0.94 | 0.48 | 0.77 |
| At5g60120  | APETALA2 protein like                          | 0.80| 1.70| 1.67 | 1.21 | 1.72 |
| At1g16060  | putative transcription factor CKC protein      | 1.02| 1.69| 1.43 | 1.18 | 1.50 |
| At2g28550  | AP2 transcription factor like protein          | 0.71| 1.64| 1.20 | 0.54 | 0.88 |

*Each ratio value is [intensity of each point/intensity at 0 h]. In shaded cells, the ratios in the cells were >3.0. In dotted cells, the ratios in the cells were <0.33.
With RARTF, users can search the database of RARGE for alternative splicing events. Out of 1968 transcription factors, 110 genes which code for transcription factors undergo alternative splicing. Alternative splicing events have been reported in the tissue-specific expression of transcription factors in mice. This suggested that some transcription factors have another isoform and function in each expressed tissue. In Arabidopsis, strong relationships between alternative splicing events and environmental stress conditions and expressed tissues have been reported. It is important to check whether a transcription factor has alternative splicing isoforms or not.

RARGE is the largest database on alternative splicing events in Arabidopsis. It might provide useful information on alternative splicing events of transcription factors.

RARTF provides information on Ac/Ds transposon mutants. There are 836 (~42%) transcription factor genes tagged by Ac/Ds transposons. Seeds of most of these mutant lines are available from RIKEN Bioresource Center (http://www.brc.riken.jp/en/). We made links to genome maps on RARGE with which users can check the locations of the genes and sites tagged by transposons. Expression profiles of transcription factors are also important. Transcription factors often have critical roles for control expression of other genes under various stress conditions or developmental stages.

For example, we showed expression profiles of DREB family transcription factors under drought stress in Table 3. The DREB family is a key transcription factor family that responds to drought stress. Indeed, there are transcription factors with various expression profiles in the DREB family (Table 3). This indicates that transcription factors in one family have a distinct expression manner. We provide the expression profiling results of the transcription factor genes using the 7K RIKEN Arabidopsis full-length (RAF1) cDNA microarray from RARTF. The expression profiling data of 379 transcription factor genes in treatments with cold, drought, high-salinity, absicic acid, high light-stress and rehydration are available from RARTF. In the near future, we will provide the expression profiles of all 1968 transcription factors in Arabidopsis in various tissues at various developmental stages and under various stress conditions from RARTF.

Acknowledgements: We thank Asako Kamiya, Maiko Nakajima, Junko Ishida, Akiko Enju and Mari Narusaka for their excellent technical assistance. We also thank Dr Yoshihide Hayashizaki, Takashi Kuromori, Takuya Ito, Prof. Joseph R. Ecker, Athanasios Theologis and Ronald W. Davis for their collaboration. This work was supported in part by a grant for CREST, Genome Research from RIKEN, BRAIN to K.S. The work was also supported in part by a Grant-in-Aid for Scientific Research on Priority Areas ‘Genome Science’ from MECSST to M.S.

References

1. Arabidopsis Genome Initiative (2000), Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature, 408, 796–815.
2. Riechmann, J. L., Heard, J., Martin, G., et al. 2000, Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes, Science, 290, 2105–2110.
3. Schultz, E. A. and Haughn, G. W. 1991, LEAFY, a Homeotic Gene That Regulates Inflorescence Development in Arabidopsis, Plant Cell, 3, 771–781.
4. Birnbaum, K., Shasha, D. E., Wang, J. Y., et al. 2003, A gene expression map of the Arabidopsis root, Science, 302, 1956–1960.
5. Shinozaki, K. and Yamaguchi-Shinozaki, K. 2000, Molecular responses to dehydration and low temperature: differences and cross-talk between two stress signaling pathways, Curr. Opin. Plant Biol., 3, 217–223.
6. Seki, M., Narusaka, M., Kamiya, A., et al. 2002, Functional annotation of a full-length Arabidopsis cDNA collection, Science, 296, 141–145.
7. Yamada, K., Lim, J., Dale, J. M., et al. 2003, Empirical analysis of transcriptional activity in the Arabidopsis genome, Science, 302, 842–846.
8. Haas, B. J., Volfovsy, N., Town, C. D., et al. 2002, Full-length messenger RNA sequences greatly improve genome annotation, Genome Biol., 3, research0029.1–0029.12.
9. Kendall, A., Hull, M. W., Bertrand, E., Good, P. D., Singer, R. H., and Engelke, D. R. 2000, A CBF5 mutation that disrupts nucleolar localization of early tRNA biosynthesis in yeast also suppresses tRNA gene-mediated transcriptional silencing, Proc. Natl Acad. Sci. USA., 97, 13108–13113.
10. Wagner, D. and Meyerowitz, E. M. 2002, SPLAED, a novel SWI/SNF ATPase homolog, controls reproductive development in Arabidopsis, Curr. Biol., 12, 85–94.
11. Miao, Z. H., Liu, X., and Lam, E. 1994, TGA3 is a distinct member of the TGA family of bZIP transcription factors in Arabidopsis thaliana, Plant Mol. Biol., 25, 1–11.
12. Gu, Y. Q., Wildermuth, M. C., Chakravarthy, S., et al. 2002, Tomato transcription factors pti4, pti5, and pti6 activate defense responses when expressed in Arabidopsis, Plant Cell, 14, 817–831.
13. Fujimoto, S. Y., Ohta, M., Usui, A., Shinshi, H., and Ohme-Takagi, M. 2000, Arabidopsis ethylene-responsive element binding factors act as transcriptional activators or repressors of GCC box-mediated gene expression, Plant Cell, 12, 393–404.
14. Wilson, Z. A., Morroll, S. M., Dawson, J., Swarup, R., and Tighe, P. J. 2001, The Arabidopsis MALE STERILITY1 (MS1) gene is a transcriptional regulator of male gametogenesis, with homology to the PHD-finger family of transcription factors, Plant J., 28, 27–39.
15. Jones, H. D., Kurup, S., Peters, N. C., and Holdsworth, M. J. 2000, Identification and analysis of proteins that interact with the Avena fatua homologue of the maize transcription factor VIVIPAROUS 1, Plant J., 21, 133–142.
16. Mundel, C., Baltz, R., Eliasson, A., et al. 2000, A LIM-domain protein from sunflower is localized to the cytoplasm and/or nucleus in a wide variety of tissues and is associated with...
with the phragmoplast in dividing cells, Plant Mol. Biol., 42, 291–302.

17. Aravind, L. and Landsman, D. 1998, AT-hook motifs identified in a wide variety of DNA-binding proteins, Nucleic Acids Res., 26, 4413–4421.

18. Frye, R. A. 2000, Phylogenetic classification of prokaryotic and eukaryotic Sir2-like proteins, Biochem. Biophys. Res. Commun., 273, 793–798.

19. Altschul, S. F., Madden, T. L., Schaffer, A. A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res., 25, 3389–3402.

20. Zdobnov, E. M. and Apweiler, R. 2001, InterProScan—an integration platform for the signature-recognition methods in InterPro, Bioinformatics, 17, 847–848.

21. Schoof, H., Ernst, R., Nazarov, V., Pfeifer, L., Mewes, H. W., and Mayer, K. F. 2003, MIPS Arabidopsis thaliana Database (MATDB): an integrated biological knowledge resource for plant genomics, Nucleic Acids Res., 32, D373–D376.

22. Mulder, N. J., Apweiler, R., and Attwood, T. K. 2003, The InterPro Database, 2003 brings increased coverage and new features, Nucleic Acids Res., 31, 315–318.

23. Stracke, R., Werber, M., and Weisshaar, B. 2001, The R2R3-MYB gene family in Arabidopsis thaliana, Curr. Opin. Plant Biol., 4, 447–456.

24. Gutierrez, N. and Reuber, T. L. 2004, Regulation of disease resistance pathways by AP2/ERF transcription factors, Curr. Opin. Plant Biol., 7, 465–471.

25. Sakuma, Y., Liu, Q., Dubouzet, J. G., Abe, H., Shinozaki, K., and Yamaguchi-Shinozaki, K. 2002, DNA-binding specificity of the ERF/AP2 domain of Arabidopsis DREBs, transcription factors involved in dehydration- and cold-inducible gene expression, Biochem. Biophys. Res. Commun., 290, 998–1009.

26. Bailey, P. C., Martin, C., Toledo-Ortiz, G., et al. 2003, Update on the basic helix-loop-helix transcription factor gene family in Arabidopsis thaliana, Plant Cell, 15, 2497–2502.

27. Ooka, H., Satoh, K., Doi, K., et al. 2003, Comprehensive analysis of NAC family genes in Oryza sativa and Arabidopsis thaliana, DNA Res., 10, 239–247.

28. Parenicová, L., de Folter, S., Kieffer, M., et al. 2003, Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in Arabidopsis: new openings to the MADS world, Plant Cell, 15, 1538–1551.

29. Jakoby, M., Weisshaar, B., Droge-Laser, W., et al. 2002, bZIP transcription factors in Arabidopsis, Trends Plant Sci., 7, 106–111.

30. Ulker, B. and Somssich, I. E. 2004, WRKY transcription factors: from DNA binding towards biological function, Curr. Opin. Plant Biol., 7, 491–498.

31. Uno, Y., Furuhata, T., Abe, H., Yoshida, R., Shinozaki, K., and Yamaguchi-Shinozaki, K. 2000, Arabidopsis basic leucine zipper transcription factors involved in a abscisic acid-dependent signal transduction pathway under drought and high-salinity conditions, Proc. Natl Acad. Sci. USA., 97, 11632–11637.

32. Jiao, Y., Yang, H., Ma, L., et al. 2003, A genome-wide analysis of blue-light regulation of Arabidopsis transcription factor gene expression during seedling development, Plant Physiol., 133, 1480–1493.

33. Tuskan, G. A., DiFazio, S. P., and Teichmann, T. 2004, Poplar genomics is getting popular: the impact of the poplar genome project on tree research, Plant Biol. (Stuttg.), 6, 2–4.

34. Kato, T., Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., and Tabata, S. 2003, Structural analysis of a Lotus japonicus genome. V. Sequence features and mapping of sixty-four TAC clones which cover the 6.4 Mb regions of the genome, DNA Res., 10, 277–285.

35. Rhee, S. Y., Beavis, W., Berardini, T. Z., et al. 2003, The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community, Nucleic Acids Res., 31, 224–228.

36. Marchler-Bauer, A., Panchenko, A. R., Shoemaker, B. A., Thiessen, P. A., Geer, L. Y., and Bryant, S. H. 2002, CDD: a database of conserved domain alignments with links to domain three-dimensional structure, Nucleic Acids Res., 30, 281–283.

37. Yuan, Q., Ouyang, S., Liu, J., et al. 2003, The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists, Nucleic Acids Res., 31, 229–233.

38. Sakurai, T., Satou, M., Akiyama, K., et al. 2005, RARGE: a large-scale database of RIKEN Arabidopsis resources ranging from transcriptome to phenotype, Nucleic Acids Res., 33, D647–D650.

39. Taneri, B., Snyder, B., Novoradovsky, A., and Gaasterland, T. 2004, Alternative splicing of mouse transcription factors affects their DNA-binding domain architecture and is tissue specific, Genome Biol., 5, R75.

40. Iida, K., Seki, M., Sakurai, T., et al. 2004, Genome-wide analysis of alternative pre-mRNA splicing in Arabidopsis thaliana based on full-length cDNA sequences, Nucleic Acids Res., 32, 5096–5103.

41. Kuromori, T., Hirayama, T., Kiyosue, Y., et al. 2004, A collection of 11 800 single-copy Ds transposon insertion lines in Arabidopsis, Plant J., 37, 897–905.

42. Seki, M., Narusaka, M., Ishida, J., et al. 2002, Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray, Plant J., 31, 279–292.

43. Seki, M., Ishida, J., Narusaka, M., et al. 2002, Monitoring the expression pattern of around 7,000 Arabidopsis genes under ABA treatments using a full-length cDNA microarray, Plant J., 34, 868–887.