

ppdb: plant promoter database version 3.0

Ayaka Hieno¹, Hushna Ara Naznin¹, Mitsuro Hyakumachi^{1,2}, Tetsuya Sakurai³,
Mututomo Tokizawa¹, Hiroyuki Koyama^{1,2}, Naoki Sato⁴, Tomoaki Nishiyama⁵,
Mitsuyasu Hasebe^{6,7}, Andreas D. Zimmer⁸, Daniel Lang⁸, Ralf Reski⁸, Stefan
A. Rensing⁹, Junichi Obokata¹⁰ and Yoshiharu Y. Yamamoto^{1,2,3,*}

¹The United Graduate School of Agricultural Science, Gifu University, 1-1 Yanagido, Gifu City, Gifu 501-1193 Japan, ²Faculty of Applied Biological Sciences, Gifu University, 1-1 Yanagido, Gifu City, Gifu 501-1193 Japan, ³Center for Sustainable Resource Science, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045 Japan, ⁴Graduate School of Arts and Sciences, The University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902 Japan, ⁵Advanced Science Research Center, Kanazawa University, Takaramachi 13-1, Kanazawa City, Ishikawa 920-0934, Japan, ⁶National Institute for Basic Biology, Nishigonaka 38, Myodaiji, Okazaki City, Aichi 444-8585 Japan, ⁷Department of Basic Biology, School of Life Science, Graduate University for Advanced Studies, Okazaki 444-8585, Japan, ⁸Faculty of Biology, University of Freiburg, Schänzlestrasse 1, D79104 Freiburg, Germany, ⁹Plant Cell Biology, Faculty of Biology, University of Marburg, Karl-von-Frisch-Strasse 9, 35043 Marburg, Germany and ¹⁰Graduate School of Environmental Life Science, Kyoto Prefectural University, 1-5 Hangi-cho, Shimogamo, Sakyo-ku, Kyoto 606-8522 Japan

Received September 9, 2013; Revised and Accepted October 8, 2013

ABSTRACT

ppdb (<http://ppdb.agr.gifu-u.ac.jp>) is a plant promoter database that provides information on transcription start sites (TSSs), core promoter structure (TATA boxes, Initiators, Y Patches, GA and CA elements) and regulatory element groups (REGs) as putative and comprehensive transcriptional regulatory elements. Since the last report in this journal, the database has been updated in three areas to version 3.0. First, new genomes have been included in the database, and now ppdb provides information on *Arabidopsis thaliana*, rice, *Physcomitrella patens* and poplar. Second, new TSS tag data (34 million) from *A. thaliana*, determined by a high throughput sequencer, has been added to give a ~200-fold increase in TSS data compared with version 1.0. This results in a much higher coverage of ~27 000 *A. thaliana* genes and finer positioning of promoters even for genes with low expression levels. Third, microarray data-based predictions have been appended as REG annotations which inform their putative physiological roles.

INTRODUCTION

Gene regulation is a central part of morphogenesis and environmental adaptation of higher plants, and it is controlled by the promoter of each gene. Therefore,

understanding of promoter structure is crucial to understand these fundamental processes of plants.

There are three aspects to promoter structure: (i) the position, direction and strength of the transcription start sites (TSSs) that indicate actual promoter position; (ii) the type and position of the core promoter elements such as TATA boxes and Initiators (Inrs) that are thought to be the major determinants of the direction and position of promoters and (iii) the type and position of transcriptional regulatory elements that are involved in gene regulation.

In our last report (1), we introduced the plant promoter database (ppdb), which provided promoter information about TSS clusters, core promoter elements [TATA boxes, Inrs, Y Patches, GA and CA elements (2,3)] and regulatory element groups [REGs, putative position-sensitive transcriptional regulatory elements that are extracted by local distribution of short sequences (LDSS) analysis (2)] as putative and comprehensive sets of transcriptional regulatory elements. The database of the original version 1.0 contained information of two plant species, *Arabidopsis thaliana* and rice.

MAJOR EXTENSIONS FROM VERSION 1.0

The major amendment in version 3.0 is the addition of the *Physcomitrella patens* and poplar genomes to the database. The sources used for the information of the four genomes, including *A. thaliana* and rice, are shown in Table 1. The promoter elements of the moss genome have been extracted by the LDSS method (2). During

*To whom correspondence should be addressed. Tel: +81 58 293 2848; Email: yyy@gifu-u.ac.jp

extraction, we noticed that considerable numbers of moss genes are driven by a similar type of promoter that is located within long terminal repeats. These promoters affect the extraction process due to tight sequence conservation that is not related to promoter function and for this reason they were excluded from the LDSS analysis. *A. thaliana* promoter elements have been applied to the poplar genome because the Brassicaceae and Malpighiales are phylogenetically close.

A new function called 'Homologue Gene Search' has been added to facilitate the comparison of promoter structures of orthologous genes within a species or between different species. Orthologue groups have been determined by Gclust, a system that classifies orthologues according to the presence or absence of protein motifs (16).

New *A. thaliana* TSS data of 34 million tags, which corresponds to a ~200-fold increase in the previous data, have been added (Figure 1). REG annotations have also been appended and show functional predictions based on microarray data of responses to plant hormones (AUX: auxin, BR: brassinosteroid, CK: cytokinin, ABA: abscisic acid, ET: ethylene, JA: jasmonic acid, SA: salicylic acid), responses to a hormone-like chemical (H₂O₂) and

some environmental stress-related responses (drought, DREB1A overexpression) (7). Functional annotation of 53 of 308 REGs is now available in version 3.0 (Figure 2).

BROWSING PROMOTER STRUCTURE

The major function of ppdb is to give an indication of a possible promoter structure for each gene in a genome based on the established lists of LDSS-positive elements. The information can be directly called by the gene ID (e.g. AT1G67090 or Os01g0791600), or selected from a list of 'Keyword Search' or 'Homologue Gene Search'. Pages for individual genes show the following information: (i) DNA sequence, (ii) TSS distribution (direction and strength at a 1-bp resolution), (iii) core promoter structure and (iv) REG data.

At the sequence window, promoter elements including REGs and core elements are highlighted in a position-dependent manner as the default setting. Care should be taken that promoters without any TSS information do not show any elements as default. For an indication of the promoter elements of these genes, the 'Reliable' button should be clicked which changes the state to 'All'

Table 1. Source of ppdb version 3.0

	Specification	Source	Size
<i>A. thaliana</i>			
Genome sequence and gene annotation	TAIR9	http://www.arabidopsis.org/ , (4)	
TSS information	Selected RAFL cDNA Cap signature CT-MPSS tags Oligo-Cap Illumina data	http://rarge.gsc.riken.jp/ , (5) Tokizawa M, Yamanaka H, Koyama H, Sakurai T, Kurotani A, Shinozaki K, Suzuki Y, Sugano S, Obokata J, Yamamoto YY (unpublished data)	62 108 (clones ^a) 158 237 (tags ^b) 34 206 936 (tags ^b)
Promoter elements	<i>A. thaliana</i> LDSS-positive octamers Annotation for LDSS elements: PLACE	(2,3) http://www.dna.affrc.go.jp/PLACE/ , (6)	659 (octamers ^c) 21 (only matched motifs ^d)
	Annotation for LDSS elements: stress and hormonal responses	(7)	53 (only matched motifs ^c)
Rice (<i>Oryza sativa</i>)			
Genome sequence and gene annotation	RGSP build 4.0	http://rapdb.lab.nig.ac.jp/ , (8)	—
TSS information	Carefully selected fl-cDNA (from KOME)	http://cdna01.dna.affrc.go.jp/cDNA/ , (9)	17 286 (clones ^a)
Promoter elements	Rice LDSS-positive octamers Annotation for LDSS elements: PLACE	(2,10) http://www.dna.affrc.go.jp/PLACE/ , (6)	660 (octamers ^c) 4 (only matched motifs ^d)
Moss (<i>P. patens</i>)			
Genome sequence and gene annotation	JGI version 1.1, COSMOSS V1.6	http://www.cosmos.org/ , (11,12)	—
TSS information	5' CAGE	(13)	1 122 382 (tags ^b)
Promoter elements	<i>P. patens</i> LDSS-positive octamers	This work	198 (octamers ^c)
Poplar (<i>Populus trichocarpa</i>)			
Genome sequence and gene annotation	Phytozome6	http://www.phytozome.net/poplar/ , (14)	—
TSS information	FL-cDNA info from GenBank	(15)	15 256 (clones ^a , BP921855–937111) 36 103 (clones ^a , DB874873–910976)
Promoter elements	<i>A. thaliana</i> LDSS-positive octamers	(2,3)	659 (octamers ^c)
Orthologue gene			
Orthologue group	Gclust	(16)	336 689 (families ^e)

^aclone number, ^btag number, ^cnumber of octamer sequences, ^dnumber of motifs and ^enumber of orthologue families.

PlantPromoterDB promoter information of AT1G02780.1

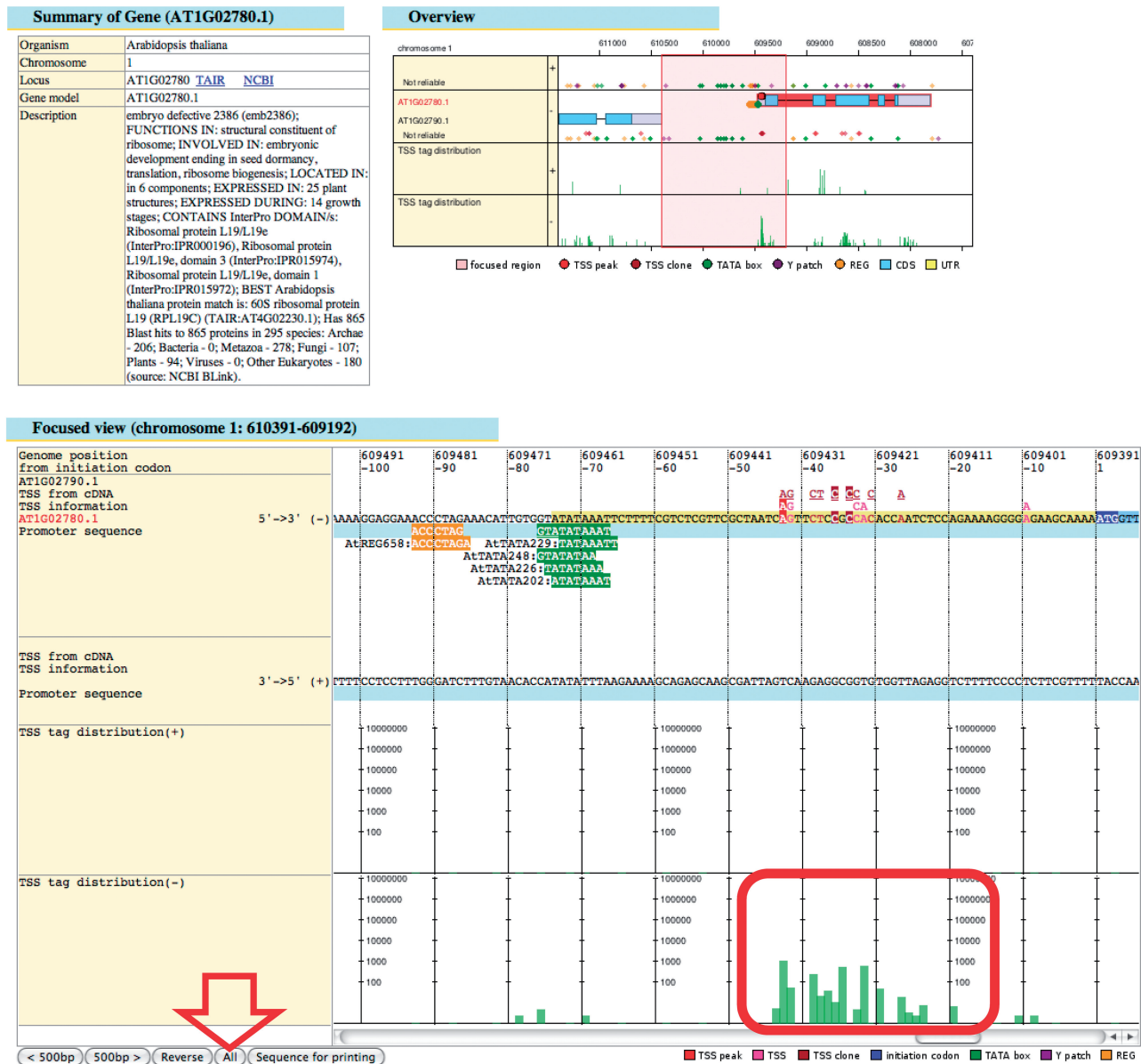


Figure 1. Indication of individual promoters. An *Arabidopsis* gene, AT1G02780.1, is shown. The information is composed of five panels: ‘Summary of Gene’, ‘Overview’, ‘Focused view’ and also ‘Promoter Summary’ (not shown) and ‘Other Reliable Promoter Summary’ (not shown). The top TSS (TSS Peak) is shown in the second column of the ‘Focused view’ as white letters on a red background. New TSS tag data (34 million) are shown at the bottom of ‘Focused view’, highlighted in a red rectangle with rounded corners.

(Figure 1, red arrow). This button is a toggle switch between ‘Reliable’ and ‘All’. ‘Reliable’ is a default setting where only elements at appropriate positions relative to the peak TSS are detected. The setting ‘All’ removes the positional restriction as an indication of promoter elements, allowing global detection. The sensitive area in the ‘Reliable’ mode for each element group is described on the front page of the database.

The ‘TSS tag distribution’ columns in the ‘Focused view’ provide the expressional strength of each TSS. The expression is the sum of six TSS tag libraries that are prepared from leaves, roots, inflorescences, etiolated seedlings and shoots from low light-grown and high light-grown seedlings.

The ‘Core promoter information’ table shows the presence or absence of core promoter elements (TATA boxes, Inrs, Y Patches, GA and CA elements).

The ‘REG information’ table shows a REG list together with the corresponding PPDB motifs (2,3) and PLACE motifs (6). REG sequences, as well as PPDB and PLACE motifs, are linked to other pages containing biological information. New REG annotations for *A. thaliana* obtained from predicted *cis*-regulatory elements based on microarray data (7) have been included (Figure 2).

Selection of the ‘All’ button (Figure 1) adds another category, ‘Not Reliable Promoter Summary’ below ‘Other Reliable Promoter Summary’. This category can be used when searching for regulatory elements (REGs)

REG information

Type	Sequence	Annotation	Genome position		
			strand	Start	End
REG	AAACGGCA			21240505	21240512
AtREG500	AAACGGCA		PPDB Motif	AAACG(C/G)	PLACE Motif
REG	ATTGGCCCATCA			21240556	21240567
AtREG446	ATTGGCCC	CK	PPDB Motif	GCCCA	PLACE Motif GGGCC
AtREG484	TTGGCCCA		PPDB Motif	GCCCA	PLACE Motif GGGCC , TGGGCT
AtREG490	TGGCCCAT		PPDB Motif	GCCCA	PLACE Motif GGGCC , TGGGCT
AtREG403	GGCCCATC		PPDB Motif	GCCCA	PLACE Motif GGGCC , TGGGCT
AtREG635	GCCCATCA		PPDB Motif	GCCCA	PLACE Motif
REG	AGTCGGTC			21240575	21240582
AtREG638	AGTCGGTC	DREB1Aox, ABA	PPDB Motif	CCGAC	PLACE Motif CCGAC , RCCGAC
REG	ACGCGTGT			21240730	21240737
AtREG536	ACGCGTGT	ABA, Drought	PPDB Motif		PLACE Motif ACACNNG
REG	GGACACGTA			21240785	21240793
AtREG472	GGACACGT	ABA, DREB1Aox, Drought	PPDB Motif	ACGT	PLACE Motif ACGT , ACGTG , ACGTGKC , ACGTGTC
AtREG557	GACACGTA	ABA, DREB1Aox	PPDB Motif	ACGT	PLACE Motif ACGT , ACGTG , ACGTGKC , ACGTGTC
REG	TGCCGTTT			21240505	21240512
AtREG500	TGCCGTTT		PPDB Motif	AAACG(C/G)	PLACE Motif
REG	TGATGGGCCAAT			21240556	21240567
AtREG635	TGATGGGC		PPDB Motif	GCCCA	PLACE Motif
AtREG403	GATGGGCC		PPDB Motif	GCCCA	PLACE Motif GGGCC , TGGGCT
AtREG490	ATGGGCCA		PPDB Motif	GCCCA	PLACE Motif GGGCC , TGGGCT
AtREG484	TGGGCCAA		PPDB Motif	GCCCA	PLACE Motif GGGCC , TGGGCT
AtREG446	GGGCCAAT	CK	PPDB Motif	GCCCA	PLACE Motif GGGCC
REG	GACCGACT			21240575	21240582
AtREG638	GACCGACT	DREB1Aox, ABA	PPDB Motif	CCGAC	PLACE Motif CCGAC , RCCGAC
REG	ACACGCGT			21240730	21240737
AtREG536	ACACGCGT	ABA, Drought	PPDB Motif		PLACE Motif ACACNNG
REG	TACGTGTCC			21240785	21240793
AtREG557	TACGTGTC	ABA, DREB1Aox	PPDB Motif	ACGT	PLACE Motif ACGT , ACGTG , ACGTGKC , ACGTGTC
AtREG472	ACGTGTCC	ABA, DREB1Aox, Drought	PPDB Motif	ACGT	PLACE Motif ACGT , ACGTG , ACGTGKC , ACGTGTC

Figure 2. REG information. REG information of the AT5G52310.1 (*RD29A*) promoter is shown. REG annotations, added in version 3.0, are highlighted in a red rectangle with rounded corners.

from wider regions or when there is no TSS information on the promoter of interest.

ADDITIONAL PAGES

A whole list of REGs for each of the genomes can be viewed by selecting a cell in the table of 'Index of Genes' at the top of the page. The lists present the relationships between REG ID, sequence, PPDB motifs, PLACE motifs and also functional annotations. Selection of a specific REG entry leads to 'Summary of the REG' and 'Entry Sequences' that show the whole gene lists containing the corresponding REG, together with gene annotations.

ACKNOWLEDGEMENTS

We would like to acknowledge Mr Takeaki Taniguchi of Mitsubishi Research Institute for excellent technical assistance. We also appreciate Dr Dong Nguyen Tien and Mr Hirokazu Murakami for maintenance of the database.

FUNDING

Grant-in-Aid for Scientific Research on Priority Areas 'Comparative Genomics' (in part) (to Y.Y.Y. and J.O.); Scientific Research on Priority Areas 'Perceptive Plants' (in part) (to Y.Y.Y.); Grant-in-Aid for Publication of Scientific Research Results 'Databases' (in part) (to J.O. and Y.Y.Y.) from the Ministry of Education, Culture, Sports, Science and Technology of Japan; Advanced Low Carbon Technology Research and Development Program from Japan Science and Technology Agency (JST ALCA) (in part) (to Y.Y.Y.). Funding for open access charge: Ministry of Education, Culture, Sports, Science and Technology, Japan.

Conflict of interest statement. None declared.

REFERENCES

1. Yamamoto, Y.Y. and Obokata, J. (2008) ppdb, a plant promoter database. *Nucleic Acids Res.*, **36**, D977–D981.
2. Yamamoto, Y.Y., Ichida, H., Matsui, M., Obokata, J., Sakurai, T., Satou, M., Seki, M., Shinozaki, K. and Abe, T. (2007) Identification of plant promoter constituents by analysis of local distribution of short sequences. *BMC Genomics*, **8**, 67.
3. Yamamoto, Y.Y., Yoshitsugu, T., Sakurai, T., Seki, M., Shinozaki, K. and Obokata, J. (2009) Heterogeneity of *Arabidopsis* core promoters revealed by high density TSS analysis. *Plant J.*, **60**, 350–362.
4. Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M. *et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, D1202–D1210.
5. Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M., Enju, A., Akiyama, K., Oono, Y. *et al.* (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science*, **296**, 141–145.
6. Higo, K., Ugawa, Y., Iwamoto, M. and Korenaga, T. (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.*, **27**, 297–300.
7. Yamamoto, Y.Y., Yoshioka, Y., Hyakumachi, M., Maruyama, K., Yamaguchi-Shinozaki, K., Tokizawa, M. and Koyama, H. (2011) Prediction of transcriptional regulatory elements for plant hormone responses based on microarray data. *BMC Plant Biol.*, **11**, 39.
8. International_Rice_Genome_Sequencing_Project. (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.
9. Kikuchi, S., Satoh, K., Nagata, T., Kawagashira, N., Doi, K., Kishimoto, N., Yazaki, J., Ishikawa, M., Yamada, H., Ooka, H. *et al.* (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science*, **301**, 376–379.
10. Yamamoto, Y.Y., Ichida, H., Abe, T., Suzuki, Y., Sugano, S. and Obokata, J. (2007) Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis. *Nucleic Acids Res.*, **35**, 6219–6226.
11. Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.F., Lindquist, E.A., Kamisugi, Y. *et al.* (2008) The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science*, **319**, 64–69.
12. Zimmer, A.D., Lang, D., Buchta, K., Rombauts, S., Nishiyama, T., Hasebe, M., Van de Peer, Y., Rensing, S.A. and Reski, R. (2013) Reannotation and extended community resources for the genome of the non-seed plant *Physcomitrella patens* provide insights into the evolution of plant gene structures and functions. *BMC Genomics*, **14**, 498.
13. Nishiyama, T., Miyawaki, K., Ohshima, M., Thompson, K., Nagashima, A., Hasebe, M. and Kurata, T. (2012) Digital gene expression profiling by 5'-end sequencing of cDNAs during reprogramming in the moss *Physcomitrella patens*. *PLoS ONE*, **7**, e36471.
14. Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A. *et al.* (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, **313**, 1596–1604.
15. Nanjo, T., Sakurai, T., Totoki, Y., Toyoda, A., Nishiguchi, M., Kado, T., Igasaki, T., Futamura, N., Seki, M., Sakaki, Y. *et al.* (2007) Functional annotation of 19,841 *Populus nigra* full-length enriched cDNA clones. *BMC Genomics*, **8**, 448.
16. Sato, N. (2009) Gclust: trans-kingdom classification of proteins using automatic individual threshold setting. *Bioinformatics*, **25**, 599–605.