ACS SyntheticBiology

Technical Note

pubs.acs.org/synthbio

Database Construction for PromoterCAD: Synthetic Promoter Design for Mammals and Plants

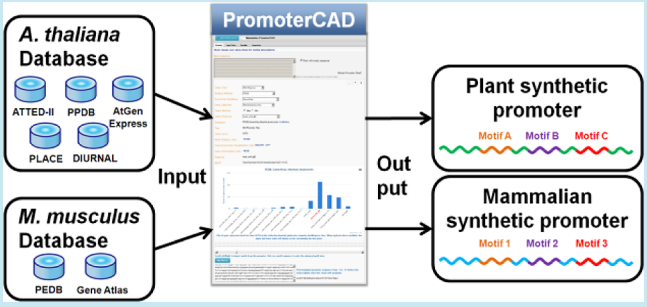
Koro Nishikata,[†] Robert Sidney Cox, III,[‡] Sayoko Shimoyama,[†] Yuko Yoshida,[†] Minami Matsui,[‡] Yuko Makita,[†] and Tetsuro Toyoda*[,†]

[†]Integrated Database Unit, Advanced Center for Computing and Communication (ACCC), RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

[‡]Synthetic Genomics Research Team, Biomass Engineering Program Cooperation Division, Center for Sustainable Resource Science (CSRS), RIKEN, 1-7-22 Suehiro, Tsurumi, Yokohama, Kanagawa 230-0045, Japan

Supporting Information

ABSTRACT: Synthetic promoters can control a gene's timing, location, and expression level. The PromoterCAD web server (http://promotercad.org) allows the design of synthetic promoters to control plant gene expression, by novel arrangement of *cis*-regulatory elements. Recently, we have expanded PromoterCAD's scope with additional plant and animal data: (1) PLACE (Plant *Cis*-acting Regulatory DNA Elements), including various sized sequence motifs; (2) PEDB (Mammalian Promoter/Enhancer Database), including gene expression data for mammalian tissues. The plant PromoterCAD data now contains 22 000 *Arabidopsis thaliana* genes, 2 200 000 microarray measurements in 20 growth conditions and 79 tissue organs and developmental stages, while the new mammalian PromoterCAD data contains 679 *Mus musculus* genes and 65 000 microarray measurements in 96 tissue organs and cell types (http://promotercad.org/mammal/). This work presents step-by-step instructions for adding both regulatory motif and gene expression data to PromoterCAD, to illustrate how users can expand PromoterCAD functionality for their own applications and organisms.



KEYWORDS: *promoter design, bioinformatics, synthetic biology, web application, database construction*

Promoter sequences are collections of *cis*-regulatory motifs controlling associations between transcription factors and basal transcriptional apparatus. Using motif recognition methods[1−3], functional predictions can be made of new combinations of *cis*-regulatory motifs into natural or user-specific basal sequences, resulting in synthetic promoters. The synthetic promoter designs can then be used as functional modular sequences ("CAD bricks") in larger genomic design.[4] PromoterCAD web server is an open set of tools for mining gene expression and *cis*-regulatory motif data, along with a GUI for selecting and arranging the retrieved motifs.[5]

PromoterCAD Application Architecture (Figure 1A). PromoterCAD is constructed on the LinkDataApp backend.[6] PromoterCAD provides motif operation algorithms MotifExpress and MotifCircadian by default, which are used along with the sequence design workflow to create an output sequence. The main application loads input table data uploaded on the LinkData registry.[6] Previously described data sources include regulatory sequence motif data (ATTED-II,[3] PPDB[2]) and microarray expression data (AtGenExpress,[7] DIURNAL[8]).

Overview of PromoterCAD Data Architecture (Figure 1B,C). The PromoterCAD data architecture is composed of (i) a promoter *cis*-regulatory motif−microarray expression mash-up table, and (ii) a gene expression rank list. Figure 1B and C illustrate data structures prepared for PLACE[9] and PEDB[10] motif analysis in this study (for combinations of ATTED-II, PPDB motifs + AtGenExpress, DIURNAL expressions, see Figure S1 of ref 5). This alignment of motif analysis and gene expression data allow for rapid correspondences between them. Rank lists are precomputation for each algorithm and allow the application to quickly return genes ranked by properties such as gene expression level (MotifExpress).

Motif Operation Algorithm. On the basis of this data architecture, motif operation tools find genes with desired expression properties in a given experimental condition. MotifExpress finds genes associated with maximum or minimum expression, while MotifCircadian searches for genes with largest circadian amplitudes. PromoterCAD finds appropriate motifs by selecting columns and rows of a mash-up table (for details of motif finding algorithms, see Figure S4 and Methods of ref 5). Users can prepare new data tables to expand PromoterCAD input data's scope. Below we present two examples of extending PromoterCAD functionality: (i) showing how to add new data

Special Issue: SB6.0

Received: November 1, 2013
Published: December 23, 2013

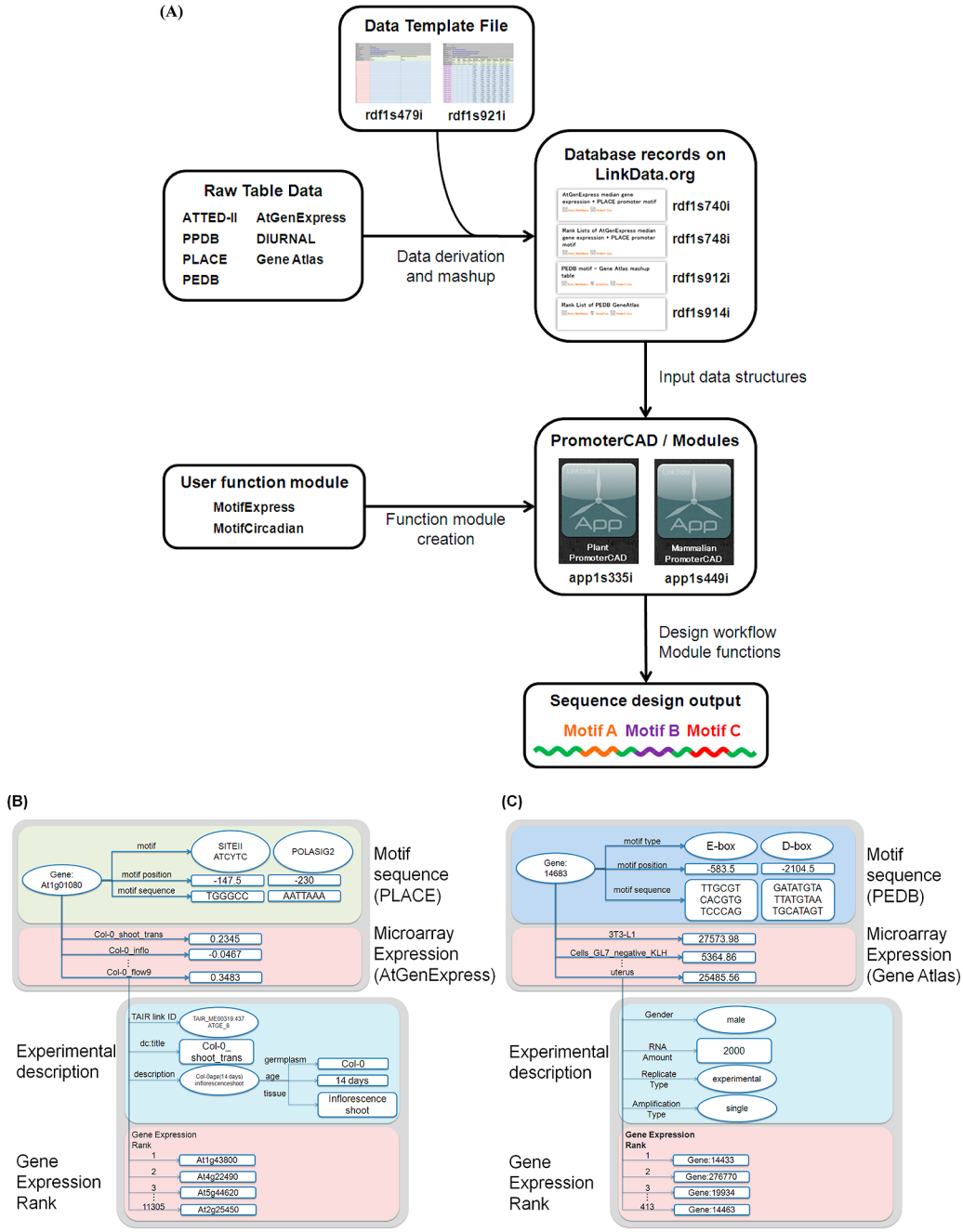


Figure 1. (A) Overall architecture of PromoterCAD. Data URLs are provided in Table S1, Supporting Information. (B,C) Data structure mashup table. Data structures were prepared for PLACE (B) and PEDB (C) motif analysis. Each data structure includes both motif analysis and gene expression information. For PLACE + AtGenexpress example (B), the *Arabidopsis* gene locus At1g01080 has hexamer element SITEIIATCYC and heptamer element POLASIG2 as its promoter motifs, with positions from TSS −147.5 and −230, with sequences TGGGCC and AATTAAA ("Motif sequence" box). The gene locus has expression values 0.2345 under the condition Col-0_shoot_trans, −0.0467 under Col-0_inflo, ..., 0.3483 under Col-0_flow9 ("Microarray Expression" box). For PEDB + GeneAtlas example (C), mouse Entrez Gene 14683 has 18mer and 24mer elements as its promoter motifs E-box and D-box, with positions from the TSS −583.5 and −2104.5, with sequences TTGCGTCACGTGTCCCAG and GATATGTATTATGTAATGCATAGT ("Motif sequence" box). The gene locus has expression values 27573.98 under the cell type 3T3-L1, 5364.86 under Cells_GL7_negative_KLH, ..., 25485.56 under uterus ("Microarray Expression" box). Mouse Gene Loci are ranked by expression values under each condition such as 3T3-L1 ("Gene Expression Rank" box). *Arabidopsis* Gene Loci are ranked by expression values under each condition such as Col-0_shoot_trans ("Gene Expression Rank" box). Additionally, Gene Expression Property Rank Lists allow the application to quickly return genes using expression mining tools MotifExpress and MotifCircadian.(A).

sets such as PLACE,[9] and (ii) how to expand PromoterCAD to a new organism using mammalian data from the PEDB.[10]

METHODS AND RESULTS

Example 1. PLACE as New Database Including Sequence Motifs with Multiple Sizes. We show how to add new motif data set PLACE[9] to PromoterCAD and mash it up with gene expression databases AtGenExpress or DIURNAL (Figure 1B).

Find PLACE Motifs in Upstream Sequences of *Arabidopsis* Gene Loci. We searched for presence of PLACE motifs in 4 regions (−2000 to −1501 bp, −1500 to −1001 bp,

dx.doi.org/10.1021/sb400178c | *ACS Synth. Biol.* 2014, 3, 192−196

−1000 to −501 bp, −500 to −1 bp from the transcription start site (TSS, +1 bp)) and selected motifs that occurred more significantly in −500 to −1 bp than in other regions, on the basis of frequency analysis (Figure S2, Supporting Information). We searched for motifs based on IUPAC nucleotide code. For motifs describing multiple possible sequences, we recorded actual sequence strings present in the natural promoter.

Format and Load LinkData Table (Figure 1A, Table S1, Supporting Information). We provide template files to make new PromoterCAD data tables (*e.g.*, PromoterCAD_AtGenExpress_mean_Flowering_Template.txt in http://linkdata.org/work/rdf1s921i, Table S1, Supporting Information). For plant PromoterCAD, these template files contain AtGenExpress or DIURNAL experiment measurements about each *Arabidopsis* gene locus. When a template is filled in with the PLACE motif information prepared above, a Motif-Expression mash-up table is generated (*e.g.*, AtGenExpress_PLACE_Flowering_median.txt in http://linkdata.org/work/rdf1s740i, Table S1, Supporting Information). This combines Developmental Microarray Expression Data (AtGenExpress) of plant developmental tissues with PLACE motif database, for Flower related developmental tissues. Gene loci without expression data or motif data were removed from this database. Prepare mash-up tables for other tissues (Fruit and Seeds, Leaf, Root, Seedling, Stem, and Whole plant) in similar ways. Next upload these tables as new works in LinkData (http://linkdata.org/upload).

Load Your Data to Your Application. Return to your own workspace page and click button "Add LinkData work (LinkData)". Select your new LinkData table (*e.g.*, rdf1s740i, Figure 1A, Table S1, Supporting Information) and click "Add" button. This adds the data table created in previous step to "Input data". Select the new data table from your own application menu. You must select PLACE motifs with multiple sizes and obtain designed synthetic promoters including PLACE motifs. Just as in the previous paper,[5] design procedure starts off by default as a 500 bp blank sequence with 50 bp CaMV 35S minimal promoter (Figure S3, Supporting Information). Additionally, selecting PLACE analysis method returns PLACE motifs. For example, gene locus AT5G47600 is displayed and its 16 promoter motifs UP2ATMSD, ..., SITEIIATCYTC, with positions 56.5, ..., 533.5 from TSS, with sequences AAACCCTA, ..., TGGGCC. As a result, we added to PromoterCAD: 22 000 genes from *Arabidopsis thaliana* (TAIR), 97 literature curated motifs (PLACE), 2 200 000 microarray measurements under 20 growth conditions (DIURNAL) and under 79 tissue organs and developmental stages (AtGenExpress). You can get the designed promoter sequences where these PLACE motifs are introduced (Figure S3, Supporting Information).

Example 2. PEDB as a New Motif Database and Gene Atlas as a New Expression Database from Mouse Biological Resources. We show how to combine a motif database PEDB[10] with expression data sets Gene Atlas[12] (Figure 1C) and how to add them to PromoterCAD.

PEDB includes three types of regulatory elements, which constitute the mammalian circadian clock: Clock/Bmal1-binding elements (E-box) (CACGTG), DBP/E4BP4 binding

Figure 2. continued

dx.doi.org/10.1021/sb400178c | *ACS Synth. Biol.* 2014, 3, 192−196

(B)

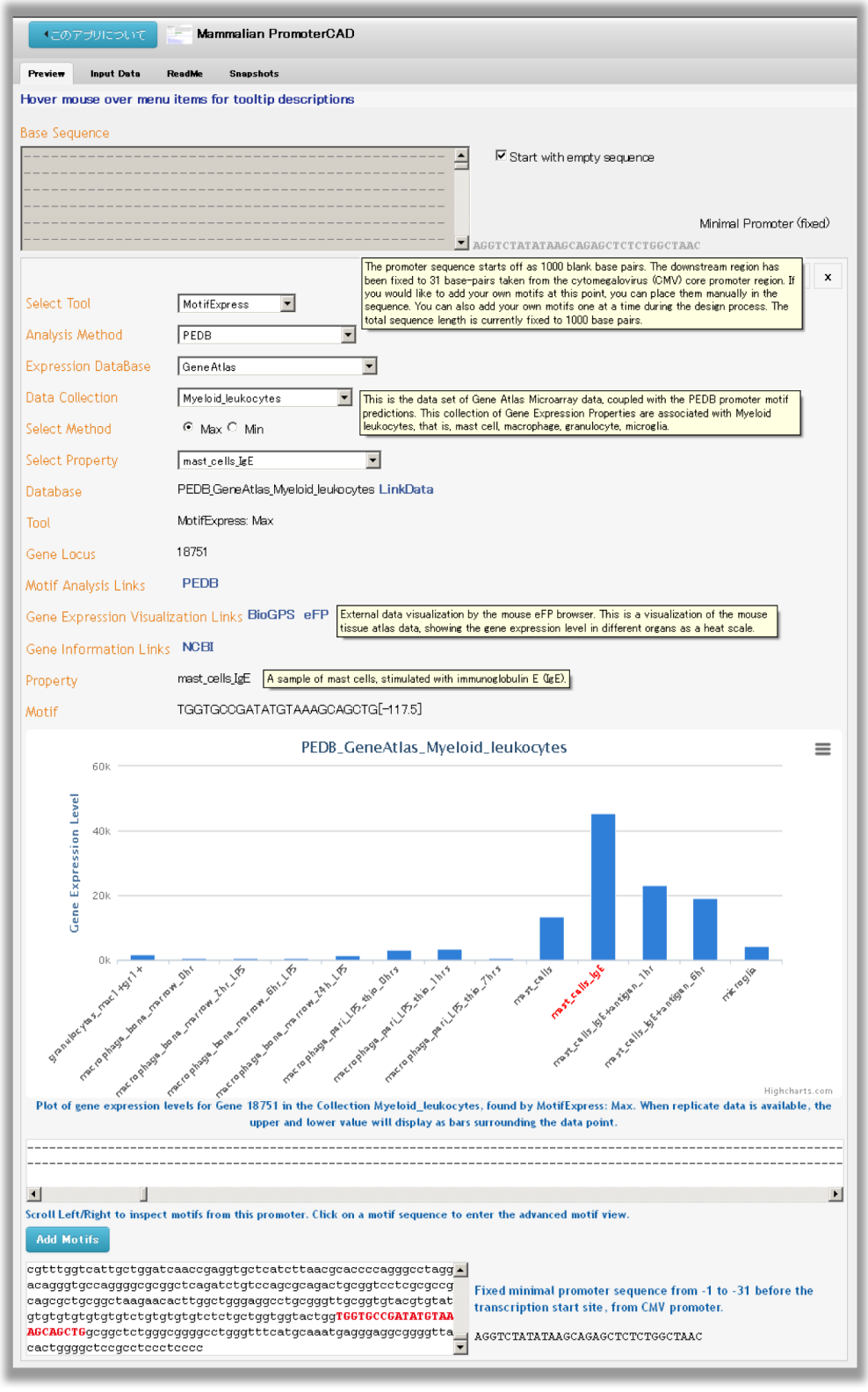


Figure 2. (A) Classification of the 96 tissues, organs, and cell types of gene expression properties in Gene Atlas. For simplicity, we show details only for tissues and cell types referred in the manuscript. Full map is described in Figure S4, Supporting Information. Annotations we made manually using Cell Type Ontology are in parentheses. Detail explanations of each sample are described in http://linkdata.org/work/rdf1s915i/tooltip_GeneAtlas.html. (B) Result of adding PEDB databases to PromoterCAD application.

elements (D-box) (TTATG[T/C]AA), and RevErbA/ROR binding elements (RRE) ([A/T]A[A/T]NT[A/G]GGTCA). Positions of these motifs range around from −100 kb to +100 kb relative to TSS. We focused on motifs from −10 000 to −1, which contain 11% of the whole motif annotation records of PEDB.

dx.doi.org/10.1021/sb400178c | *ACS Synth. Biol.* 2014, 3, 192−196

Multiple probe sets are from the Gene Atlas microarray data set[12] (GSE10246, GPL1261). We selected the probe set with highest mean intensity across all samples in the data set as the representative gene. We mashed up PEDB motifs and Gene Atlas expression profiles using the template file. This mash-up table includes information on 96 sample conditions of mouse cell resources. We added annotation to each sample using Cell Type Ontology.[13] Then we classified the mash up table into 19 different tables based on specific tissues and cell types (*e.g.*, Lymphocytes, Myeloid leukocytes, Dendritic cells, ..., Trunk and Abdomen organ, Figure 2A).

We used NCBI genome build 33, in line with PEDB, to prepare mouse baseline promoter sequences. You can select mouse promoter motifs and design synthetic mammalian promoters. The design procedure starts off as a 10 000 bp blank sequence with 31 bp taken from the cytomegalovirus (CMV) core promoter region (−31 to +93), which includes TATA box, initiator (Inr), and downstream sequence[14] by default (Figure 2B). The fixed sequence region (31 bp) and promoter size limit (10 000 bp) can be modified by forking the application. At first, select MotifExpress tool, PEDB motif analysis method, and Gene Atlas expression database. Next select "Myeloid leukocyte" data collection and the gene expression property "mast_cells_IgE". This indicates a sample of mast cells, stimulated with immunoglobulin E (IgE).[12] Then the application returns Gene 18751 named Prkcb, and its promoter motif D-box with position −117.5 from TSS, with sequence TGGTGCCGATAT-GTAAAGCAGCTG, and visualizes the expression profile of the mouse Entrez Gene 18751. External links and visualizations inform you how to decide which motifs to incorporate into the design. These links include the Entrez Gene page for gene locus,[15] PromoterCAD data files on LinkData.org and PEDB motif analysis[10] and BioGPS database[11] pages. Links to the mouse eFP browser[16] provide images of different mouse organs with gene expression level shown as a color scale (Figure 2B). After clicking "Add Motifs" and "Finish Now", you get the designed sequence including motifs selected above. As a result, we added to PromoterCAD: 679 *Mus musculus* genes (Entrez), 753 circadian clock-related motifs (PEDB), 65 000 microarray measurements in 96 tissue organs and cell types (Gene Atlas). Thus the application can be a "Mammalian PromoterCAD" (http://promotercad.org/mammal/) as well as a "Plant PromoterCAD" (Figure 2B).

Conclusions. We demonstrated promoter design procedures and result sequences for each input database, suggesting we can expand PromoterCAD scope with additional plant and animal resource data (*e.g.*, tomato, mouse). We added PLACE sequence motifs with multiple sizes and PEDB mouse data to PromoterCAD. Thus users can design not only plant promoters as they like, but also mammalian promoters, by adding new data for a new organism, experimental measurements, or motif identification methods. This will realize synthetic biology for medical applications as well as biomass engineering.

■ ASSOCIATED CONTENT

Supporting Information

Supporting tables and figures. This material is available free of charge *via* the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: tetsuro.toyoda@riken.jp.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We would like to thank Chanaka Perera, Gayan Hewathanthri, Kazuro Fukuhara (Axiohelix) for web application and LinkData development, David Gifford for comments and critical reading of the manuscript. This work was supported by RIKEN and the National Bioscience Database Center (NBDC) of Japan Science and Technology Agency (JST).

■ REFERENCES

(1) Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res. 34*, W369−W373.

(2) Yamamoto, Y. Y., and Obokata, J. (2008) ppdb: a plant promoter database. *Nucleic Acids Res. 36*, D977−D981.

(3) Obayashi, T., and Kinoshita, K. (2010) Coexpression landscape in ATTED-II: usage of gene list and gene network for various types of pathways. *J. Plant Res. 123*, 311−319.

(4) Toyoda, T. (2011) Methods for open innovation on a genome design platform associating scientific, commercial, and educational communities in synthetic biology. *Methods Enzymol. 498*, 189−203.

(5) Cox, R. S., Nishikata, K., Shimoyama, S., Yoshida, Y., Matsui, M., Makita, Y., and Toyoda, T. (2013) PromoterCAD: Data-driven design of plant regulatory DNA. *Nucleic Acids Res. 41*, W569−W574.

(6) Shimoyama, S., Cox, R. S., Gifford, D., and Toyoda, T. (2013) Organizing Scientific Competitions on the Semantic Web. *Lect. Notes Comput. Sci. 8055*, 311−318.

(7) Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., Scholkopf, B., Weigel, D., and Lohmann, J. U. (2005) A gene expression map of *Arabidopsis thaliana* development. *Nat. Genet. 37*, 501−506.

(8) Mockler, T. C., Michael, T. P., Priest, H. D., Shen, R., Sullivan, C. M., Givan, S. A., McEntee, C., Kay, S. A., and Chory, J. (2007) The diurnal project: diurnal and circadian expression profiling, model-based pattern matching, and promoter analysis. *Cold Spring Harbor Symp. Quant. Biol. 72*, 353−363.

(9) 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.

(10) Kumaki, Y., Ukai-Tadenuma, M., Uno, K. D., Nishio, J., Masumoto, K. H., Nagano, M., Komori, T., Shigeyoshi, Y., Hogenesch, J. B., and Ueda, H. R. (2008) Analysis and synthesis of high-amplitude Cis-elements in the mammalian circadian clock. *Proc. Natl. Acad. Sci. U. S. A. 105*, 14946−14951.

(11) Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., Hodge, C. L., Haase, J., Janes, J., Huss, J. W., 3rd, and Su, A. I. (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol. 10*, R130.

(12) Lattin, J. E., Schroder, K., Su, A. I., Walker, J. R., Zhang, J., Wiltshire, T., Saijo, K., Glass, C. K., Hume, D. A., Kellie, S., and Sweet, M. J. (2008) Expression analysis of G Protein-Coupled Receptors in mouse macrophages. *Immunome Res. 4*, 5.

(13) Meehan, T. F., Masci, A. M., Abdulla, A., Cowell, L. G., Blake, J. A., Mungall, C. J., and Diehl, A. D. (2011) Logical development of the cell ontology. *BMC Bioinf. 12*, 6.

(14) Gendra, E., Colgan, D. F., Meany, B., and Konarska, M. M. (2007) A sequence motif in the simian virus 40 (SV40) early core promoter affects alternative splicing of transcribed mRNA. *J. Biol. Chem. 282*, 11648−11657.

(15) Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res. 39*, D52−D57.

(16) Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G. V., and Provart, N. J. (2007) An 'Electronic Fluorescent Pictograph' browser for exploring and analyzing large-scale biological data sets. *PLoS One 2*, e718.