KAGIANA: An Excel-Based Tool for Retrieving Summary Information on Arabidopsis Genes

Yoshiyuki Ogata¹, Nozomu Sakurai², Koh Aoki¹, Hideyuki Suzuki¹, Koei Okazaki¹, Kazuki Saito² and Daisuke Shibata¹,*

¹Kazusa DNA Research Institute, Kazusa-Kamatari 2-6-7, Kisarazu, Chiba, 292-0818 Japan
²Graduate School of Pharmaceutical Science, Chiba University, Yayoi-cho 1-33, Inage-ku, Chiba, 263-8522 Japan

Various public databases provide Arabidopsis gene information via the internet. It is useful to abstract information obtained from such databases. We have developed the KAGIANA tool, which allows a user to retrieve summary information obtained from selective databases and to access pages for a gene of interest in those databases. The tool is based on Microsoft Excel and provides several macro programs for gene expression analyses. It can assist plant biologists in accessing omics information for plant biology. The KAGIANA tool is freely available at http://pmnedo.kazusa.or.jp/kagiana/.

Keywords: Annotation • Arabidopsis • Database • Gene expression • Omics.

Abbreviations: AGI, Arabidopsis Genome Initiative; GO, Gene Ontology; NCBI, The National Center for Biotechnology Information; TAIR, The Arabidopsis Information Resource.

Since the completion of the genome sequence of the model plant Arabidopsis thaliana (Arabidopsis Genome Initiative 2000), advances in genome and gene expression analysis have resulted in a vast number of data sets generated for Arabidopsis. Data sets of the Arabidopsis genome sequence are available at GenBank (Benson et al. 2008; http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide) and DDBJ (Sugawara et al. 2008; http://www.ddbj.nig.ac.jp/). In Arabidopsis, the sequence is separated into five chromosomes, which have 33,282 gene loci according to TAIR8 (Swarbreck et al. 2008; http://www.arabidopsis.org/). Amino acid sequences of proteins encoded by genes have been analyzed for various purposes. On the basis of localization signals included in such sequences, subcellular localization of proteins has been predicted using tools such as WolFP SORT (Horton et al. 2007; http://wolfsort.org/) and TargetP (Emanuelsson et al. 2007; http://www.cbs.dtu.dk/services/TargetP/). Domain structures, which show functional properties in proteins, have been predicted on the basis of amino acid sequences using tools such as SCOP (Andreeva et al. 2008; http://scop.berkeley.edu/), and can be found in databases such as InterPro (Mulder et al. 2007; http://www.ebi.ac.uk/interpro/). Analysis of transmembrane domains has been performed using tools such as TMHMM (Moller et al. 2001; http://www.cbs.dtu.dk/services/TMHMM/) and that of hydropathy can be found in databases such as SUBA (Heazlewood et al. 2007; http://www.plantenergy.uwa.edu.au/suba2/). Data sets of gene expression are available from databases such as the Gene Expression Omnibus (GEO) database (Barrett et al. 2006; http://www.ncbi.nlm.nih.gov/geo/). Several databases such as ATTED-II (Obayashi et al. 2007; http://atted.jp/) and Genevestigator (Zimmermann et al. 2004; https://www.genevestigator.ethz.ch/) provide a function to perform gene expression analysis. Using a vast number of gene expression data sets, approaches for detecting co-expressed genes, such as the ARACNE tool (Margolin et al. 2006; http://amdec-bioinfo.cu-genome.org/html/ARACNE.htm), the average clustering coefficient index (Gupta et al. 2006) and the Confeito algorithm (http://pmnedo.kazusa.or.jp/kagiana/coexprocess/) have been developed. On the basis of these analyses, molecular function, subcellular localization and biological processes of genes have finally been consistently assigned to ‘molecular function’, ‘cellular component’ and ‘biological process’, respectively, the three aspects of the Gene Ontology (GO) terminology (Gene Ontology Consortium 2008; http://geneontology.org/).

To obtain genomic and transcriptomic information on genes of interest, a user can visit these databases and access these tools via the internet or download them for personal
use. However, to retrieve the information, users generally require knowledge of the omics information published in the databases; for example, how to select an adequate website and how to set an adequate threshold value such as the gene-to-gene correlation coefficient for acquiring data of interest in the website. For biological users, unfamiliar with omics analyses such as genomics and transcriptomics, it is useful to have access to abstracted gene information from such databases and analyses and to use quick links to these databases.

We have developed the KAGIANA (Kazusa Arabidopsis Gene Information And Network Analysis) tool to summarize various Arabidopsis omics analyses from the above-mentioned databases and tools, and to provide links to pages for genes of interest in the databases. The tool is based on Microsoft Excel (version 2003 or higher) and thus requires only enough skill for basic Excel operation. The implementation of this tool is verified using Windows XP or higher for PC, and OS X or higher for Macintosh. The macro programs of the tools are available only for Windows users as of November 2008. Our goal is to assist plant biologists in accessing information from omics analyses so that they can incorporate it into their plant biology research.

The KAGIANA tool is downloadable as a ZIP-format file at http://pmnede.kazusa.or.jp/kagiana/. The KAGIANA tool is formatted as a Microsoft Excel workbook file, composed of five worksheets [one database sheet (‘Data20080524’), two readme sheets (‘ReadMe_1st’ and ‘ReadMe_Tools’) and two retrieval sheets (‘Selected_Link’ and ‘Selected_GO’)] and one macro program (‘Tools’) comprising four analysis tools (‘Confeito’, ‘GX bar chart’, ‘GO pie chart’ and ‘ATTED chart’). In KAGIANA, AGI codes (e.g. At1g01010) are used for the retrieval and performance of the tool.

The database sheet is composed of the following information for 33,362 loci (Fig. 1A), which was obtained from the TAIR database. First, the A to D columns represent AGI codes, a short description, description, and identifiers for NCBI, respectively. Secondly, the E to J columns display representative GO terms, which certainly accompany the evidence codes, and their Evidence Code categories, which are abbreviated as ‘X’ (experimental) for EXP, IDA, IPI, IMP, IGI and IEP; ‘S’ (statement) for TAS and IC; ‘C’ (computational) for ISS, ISO, ISA, ISM, IGC and RCA; ‘L’ (electronic) for IEA; and ‘N’ (not available) for NAS and ND, in the three aspects of GO terminology, i.e. cellular component (the E and F columns), molecular function (the G and H columns) and biological process (the I and J columns).
process (the I and J columns), respectively. A GO term was selected as the representative term for each aspect for a gene, according to the order of Evidence Code categories, i.e. X, S, C, L and N. The following columns represent information from the analytical tools. The K and L columns represent data from WoLF PSORT, which predicts the subcellular localization of proteins, and the reliability index, whose best score is 14, respectively. The M column represents information from TargetP, which also predicts subcellular localization, and the reliability index, ranging from 0 to 9 at the maximum. The N and O columns represent that from SCOP, which predicts domains of proteins, and the reliability index, which is a negative logarithm of the actual value, respectively. The P column represents TMHMM, which predicts the number of transmembrane domains of proteins.

The ‘Selected_Link’ sheet provides hyperlinks to 19 selected public databases for information retrieval of genes of interest and their Short Description and Description (Fig. 1B). These hyperlinks lead a user to the pages for individual genes in the individual databases by the following steps: (i) input AGI code(s) in the A column from the A4 cell to the lower cells (e.g. input ‘At1g01010’ in the A4 cell and ‘At1g01020’ in the A5 cell); (ii) select the range of the B4 to the W4 cells; and (iii) double-click the right lower corner (a black square) to copy the equations in the fourth row into the lower rows in the same columns (e.g. copy the B4–W4 into the B5–W5).

Then, a user can access a database of interest from among the C to U columns (e.g. click the T4 cell for access to the page for the query gene in the KaPPA-View tool). The tool provides access to the databases shown in Table 1. The way to use this sheet is also described in the ‘ReadMe_1st’ sheet.

Table 1 Abstract of databases linkable from the KAGIANA tool

| Abbreviation | Databases | URL |
|--------------|-----------|-----|
| TA           | TAIR (Swarbreck et al. 2008) | http://www.arabidopsis.org/ |
| EG           | Entrez Gene (Maglott et al. 2005) | http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene |
| MP           | MPSS (Brenner et al. 2000) | http://mpss.udel.edu/at/ |
| IP           | InParanoid (Berglund et al. 2008) | http://inparanoid.sbc.su.se/cgi-bin/index.cgi |
| TD           | SiGNAL (Alonso et al. 2003) | http://signal.salk.edu/cgi-bin/tdnaexpress |
| SB           | SUBA (Heazlewood et al. 2007) | http://www.plantenergy.uwa.edu.au/applications/suba2/ |
| KG           | KEGG (Kanehisa et al. 2008) | http://www.genome.jp/kegg/ |
| GR           | GRAMENE (Liang et al. 2008) | http://www.gramene.org/ |
| DN           | NASCArrays Digital Northern (Craigen et al. 2004) | http://affymetrix.arabidopsis.info/narrays/digitalnorthern.pl |
| SH           | NASCArrays Spot History (Craigen et al. 2004) | http://affymetrix.arabidopsis.info/narrays/spothistory.pl |
| GA           | Genevestigator Gene Atlas (Zimmermann et al. 2004) | https://www.genevestigator.ethz.ch/gv/index.jsp |
| GC           | Genevestigator Gene Chronologer (Zimmermann et al. 2004) | https://www.genevestigator.ethz.ch/gv/index.jsp |
| RV           | Genevestigator Response Viewer (Zimmermann et al. 2004) | https://www.genevestigator.ethz.ch/gv/index.jsp |
| eB           | eFP Browser (Winter et al. 2007) | http://bbc.botany.utoronto.ca/efp/cgi-bin/efpWeb.cgi |
| Av           | AtGenExpress Visualization Tool | http://jsp.weigelworld.org/expviz/expviz.jsp |
| AT           | ATTED-II (Obayashi et al. 2007) | http://atted.jp/ |
| AC           | AraCyc (Zhang et al. 2005) | http://www.arabidopsis.org/biocyc/index.jsp |
| KV           | KaPPA-View 3 (Sakurai and Shibata 2006) | http://kpv.kazusa.or.jp/kappa-view3/ |
| AP           | AtProteome (Baerenfaller et al. 2008) | http://fgcz-atproteome.unizh.ch/ |
Fig. 2 Composition of the KAGIANA tools. (A) The window of the ‘Tools’ macro program, including analyses of ‘Confeito’, ‘GX bar chart’, ‘GO pie chart’ and ‘ATTED chart’. (B) The result from the ‘GX bar chart’ analysis. (C) The result from the ‘GO pie chart’ analysis, showing the aspect of molecular function.
Detailed steps for using these tools are described in the 'ReadMe_Tools' sheet in the KAGIANA workbook. Briefly, the steps are (i) click 'Tools' in the menu bar; (ii) select 'Macro' and click 'Macros'; (iii) select 'Tools' in the macro box and click 'Execute' (open the 'Tools' window); (iv) select a tool in the 'Analysis' frame in the window; (v) input AGI codes into different lines in the textbox left of the frame; (vi) select the option frame when selecting 'GX bar chart' and 'GO pie chart' tools; and then (vii) click the 'OK' button if the character color on the button is black (otherwise, there is insufficient information for retrieval).

**Funding**

The New Energy and Industrial Technology Development (NEDO) program, part of the 'Development of Fundamental Technologies for Controlling the Material Production Process of Plants' project.

**References**

Alonso, J.M., Stepanova, A.N., Leisse, T.J., Kim, C.J., Chen, H., Shinn, P., et al. (2003) Genome-wide insertional mutagenesis of Arabidopsis thaliana. *Science* 301: 653–657.

Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E., Hubbard, T.J., Chothia, C., et al. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.* 36: D419–D425.

**Arabidopsis Genome Initiative** (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796–815.

Baerenfaller, K., Grossmann, J., Grobei, M.A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., et al. (2008) Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. *Science* 320: 938–941.

Barrett, T., Troup, D.B., Willhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., et al. (2006) NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res.* 35: D760–D766.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2008) GenBank. *Nucleic Acids Res.* 36: D25–D30.

Berglund, A.C., Sjölund, E., Östlund, G. and Sonnhammer, E.L. (2008) InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res.* 36: D263–D266.

Brenner, S., Williams, S.R., Vermaas, E.H, Storck, T., Moon, K., McCollum, C., et al. (2000) In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. *Proc. Natl. Acad. Sci. USA* 18: 630–634.

Craigon, D.J., James, N., Okyere, J., Higgins, J., Jotham, J. and May, S. (2004) NASCArrays: a repository for microarray data generated by NASC’s transcriptomics service. *Nucleic Acids Res.* 32: D575–D577.

Emanuelsson, O., Brunak, S., von Heijne, G. and Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* 2: 953–971.

Gene Ontology Consortium (2008) The Gene Ontology project in 2008. *Nucleic Acids Res.* 36: D440–D444.

Gupta, A., Maranas, C.D. and Albert, R. (2006) Elucidation of functional activity for co-expressed genes: predicting intra-operon termination sites. *Bioinformatics* 22: 209–214.

Heazlewood, J.L., Verboom, R.E., Tonti-Filippini, J., Small, I. and Millar, A.H. (2007) SUBA: the Arabidopsis Subcellular Database. *Nucleic Acids Res.* 35: D213–D218.

Horton, P., Park, KJ, Obayashi, T., Fujita, N., Harada, H., Adams-Collier, CJ, et al. (2007) WolF PSORT: protein localization predictor. *Nucleic Acids Res.* 35: W585–W587.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., et al. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* 36: D480–D484.

Liang, C., Jaiswal, P., Hebbard, C., Avraham, S., Buckler, E.S., Casstevens, T., et al. (2008) Gramene: a growing plant comparative genomics resource. *Nucleic Acids Res.* 36: D947–D953.

Maglott, D., Ostell, J., Pruitt, K.D. and Tatusova, T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* 33: D54–D58.

Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla-Favera, R., et al. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *Bioinformatics* 7: S7.

Möller, S., Croning, M.D.R. and Apweiler, R. (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* 17: 646–653.

Mulder, NJ, Apweiler, R., Attwood, T., Bairoch, A., Bateman, A., Binns, D., et al. (2007) New developments in the InterPro database. *Nucleic Acids Res.* 35: D224–D228.

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

Sakurai, N. and Shibata, D. (2006) KaPPA-View for integrating qualitative and quantitative data sets on plant metabolic pathway maps. *J. Pestic. Sci.* 31: 293–295.

Sugawara, H., Ogasawara, O., Okubo, K., Gojobori, T. and Tateno, Y. (2008) DDBJ with new system and face. *Nucleic Acids Res.* 36: D22–D24.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T.Z., Garcia-Hernandez, M., Foerster, H., et al. (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* 36: D1009–D1014.

Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G.V. and Provart, NJ. (2007) An ‘electronic fluorescent pictograph’ browser for exploring and analyzing large-scale biological data sets. *PLoS ONE* 2: e718.

Zhang, P., Foerster, H., Tissier, C.P., Mueller, L., Paley, S., Karp, P.D., et al. (2005) MetaCyc and AraCyc: Metabolic pathway databases for plant research. *Plant Physiol.* 138: 27–37.

Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L. and Gruissem, W. (2004) GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. *Plant Physiol.* 136: 2621–2632.

(Received October 22, 2008; Accepted November 17, 2008)