Plant Biotechnology 23, 195–198 (2006)

Bioinformatics Note

MiBASE: A database of a miniature tomato cultivar Micro-Tom

Kentaro Yano[1], Manabu Watanabe[2], Naoki Yamamoto[1], Taneaki Tsugane[2], Koh Aoki[1], Nozomu Sakurai[1], Daisuke Shibata[1]*

[1] Kazusa DNA Research Institute, Kazusa-Kamatari 2-6-7, Kisarazu, Chiba, 292-0818 Japan; [2] Chiba Prefectural Agriculture Research Center, Daizenno-Cho 808, Midori-ku, Chiba, Chiba 266-0006, Japan
* E-mail: shibata@kazusa.or.jp Tel: +81-438-52-3947 Fax: +81-438-52-3948

Received October 14, 2005; accepted October 24, 2005 (Edited by T. Hashimoto)

Abstract The miniature cultivar Micro-Tom has attracted much attention as a model of tomato (*Solanum lycopersicum*) because it has small size (10–20 cm in height), a short life cycle (70–90 days), and grows well in ordinary laboratory spaces. Recently, expressed sequence tag data and full-length cDNA sequences have been accumulated for Micro-Tom. To provide genomic information resource for Micro-Tom, we constructed the MiBASE database, which can be access *via* the Internet at http://www.kazusa.or.jp/jsol/microtom/. In addition to sequence information, this database contains information on simple sequence repeats, single nucleotide polymorphisms between other tomato inbred lines, nonredundant sequence sets, gene ontology terms, metabolic pathway names, and gene expression data.

Key words: Database, expressed sequence tags (ESTs), full-length cDNA, gene ontology, tomato (*Solanum lycopersicum*).

Due to rapid increases in the amount of information for expressed sequence tags (ESTs) and genomic nucleotide sequences from various plant species in public databases such as DDBJ/GenBank/EMBL, many databases have been constructed from the primary data sets. Various sets of genomic information for *Arabidopsis thaliana* and rice (*Oryza sativa*), whose full genome sequences have been determined (*Arabidopsis* Genome Initiative 2000; International Rice Genome Sequencing Project 2005), are available online. These include The *Arabidopsis* Information Resource (TAIR) (Rhee et al. 2003) and the Rice Genome Research Program (http://rgp.dna.affrc.go.jp/). The Institute for Genome Research (TIGR) Gene Indices database provides information on expressed transcripts with functional annotations for 88 species including 33 plants (October 2005) and is based on publicly available EST information (Lee et al. 2005).

Genomic information for tomato (*Solanum lycopersicum*, formerly *Lycopersicon esculentum*, $2n=24$; genome size=950 Mb) has accumulated in public databases. The Solanaceae Genome Project Network (Mueller et al. 2005) provides up-to-date information on the status of the tomato genome sequence projects, which is one of the major activities of the internationally coordinated International Solanaceae Consortium (see review, Shibata 2005). To date, 199,278 tomato ESTs have been deposited in the dbEST database of GenBank (release 100705; http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). Information on ESTs and nonredundant sequence sets is available from the TIGR Tomato Gene Index (http://www.tigr.org/tdb/tgi/index.shtml) and the Solanaceae Genome Project Network (http://www.sgn.cornell.edu/index.pl).

The miniature tomato cultivar Micro-Tom has attracted much attention as a model tomato plant because it is small (10–20 cm in height), has a short life cycle (70–90 days), and grows well at densities as high as 1357 plants/m^2 (Meissner et al. 1997) and even in ordinary laboratory spaces (Shibata 2005). Yamamoto et al. (2005) generated 35,824 ESTs from leaf and fruit cDNA libraries of Micro-Tom and constructed 26,363 nonredundant sequence sets (UNIGENEs) by combining Micro-Tom ESTs with the publicly available tomato ESTs. They also identified 2160 candidate single nucleotide polymorphisms (SNPs) between Micro-Tom and other inbred tomato lines (Yamamoto et al. 2005). A full-length cDNA library was constructed from Micro-Tom fruit and used to generate 8046 ESTs from the clones as a genomic resource (Tsugane et al. 2005).

Here, we constructed the database MiBASE, which contains genetic information on Micro-Tom (accessible at http://www.kazusa.or.jp/jsol/microtom/). The current version of MiBASE provides information of EST sequences, EST annotations, full-length cDNA clones,

Abbreviations: EST, expressed sequence tag; GO, gene ontology; SNP, single nucleotide polymorphism; SSR, simple sequence repeat; TAIR, The *Arabidopsis* Information Resource; TIGR, The Institute for Genomic Research.
This article can be found at http://www.jspcmb.jp/

UNIGENEs, SNPs between other tomato inbred lines, simple sequence repeats (SSRs), gene ontology (GO) terms, metabolic pathway names, gene expressions, and sequence similarites with other plant genes (Figure 1).

Database construction

The MiBASE database was constructed using MySQL (http://www.mysql.com/) and Hypertext Preprocessor (http://www.php.net/) in the Sun FireV880 web server (Solaris 8). The sets of information that are provided from the database were obtained by computationally processing the EST sequences of Micro-Tom as described below.

Using the BLAST program (Altschul et al. 1990), the EST sequences were searched against nonredundant amino acid sequences in the National Center for Biotechnology Information database with a search threshold of 1e-1, *Arabidopsis* Coding Sequences in TAIR with a search threshold of 1e-5, and the TIGR Tomato Gene Index (Release 10.1) with a search threshold of 1e-20. Each EST was assigned the annotation of the gene that gave the highest hit score for the EST sequence in the selected database.

Full-length cDNA sequences were obtained from two sources. First, possible full-length clones were identified from the Micro-Tom sequences as described by Yamamoto et al. (2005). The current version of the database contains information on 11,093 possible candidate full-length cDNA clones. The second source was the 5′-end sequences of cDNA clones from a full-length cDNA library that was prepared from maturing fruit by the vector-capping protocol for full-length cDNA synthesis (Tsugane et al. 2005). The current version of the database contains data on 8046 full-length clones. This information is useful for identifying the coding sequences in the tomato genome as they are generated by the genome sequencing projects as well as for finding full-length cDNA clones for functional studies of tomato genes. The result page from an EST search is shown in Figure 1A.

To ensure the correct annotation of ESTs, we prepared a set of nonredundant sequences by assembling publicly available tomato ESTs including the Micro-Tom sequences. Using the Phrap program (http://www.phrap.org), we assembled 35,824 Micro-Tom ESTs and 150,581 ESTs from other tomato lines (http://www.ncbi.nlm.nih.gov/dbEST/index.html) into a set of 26,363 nonredundant sequence groups, which comprise 18,436 contigs and 7927 singletons. We define these groups as the UNIGENEs. The UNIGENEs were named Contig1 to Contig18436 and Singlet1 to Singlet7927 in an arbitrary order.

Using the BLAST programs, the UNIGENE sequences were searched against nonredundant amino acid sequences in the National Center for Biotechnology

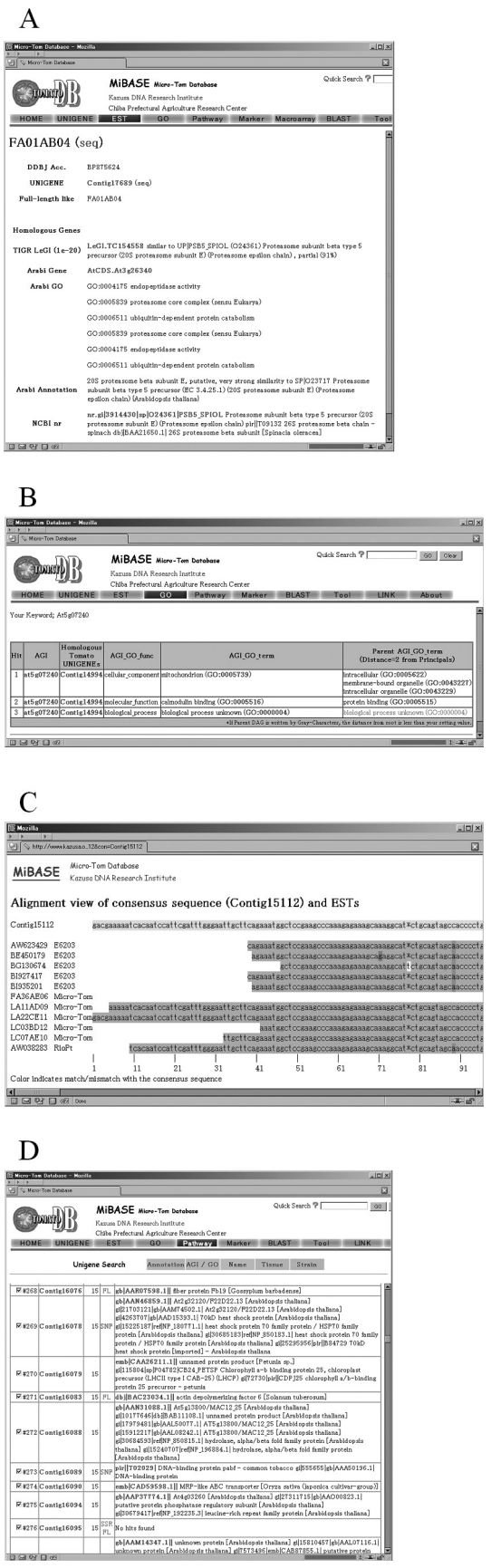

Figure 1. Main pages of the Micro-Tom genomic information database, MiBASE. (A) Result page of an EST search. (B) Result page of a GO search. (C) Multiple sequence alignment viewer for UNIGENE and ESTs. (D) Result page of a UNIGENE search.

Information database with a search threshold of 1e-5, the *Arabidopsis* translated protein sequences at TAIR with a search threshold of 1e-10, and the TIGR Gene Indices of rice (Release 10.1), soybean (Release 12.0), maize (Release 15.0) and tomato (Release 10.1) with search thresholds of 1e-10, 1e-10, 1e-10 and 1e-50, respectively. Annotations for each UNIGENE were determined as described above for ESTs.

GO terms for each EST or UNIGENE were assigned according to those for the *Arabidopsis* genes with sequence similarities (Berardini et al. 2004). We developed a tool for searching GO terms assigned to tomato sequences in the database. This tool provides information on not only the tomato GO terms matching the user's queries but also the parent GO terms, which are found in the structured ontology vocabularies (Figure 1B). All child GO terms also can be retrieved from the vocabularies.

When the annotations of an EST or UNIGENE sequence assigned from the sequence similarity search with *Arabidopsis* genes were related to a metabolic reaction, the sequence was given the name of the metabolic pathway in which the reaction is involved. The names of the metabolic pathway were obtained from the Kyoto Encyclopedia of Genes and Genomes (Kanehisa et al. 2004) or from the *Arabidopsis thaliana* Biochemical Pathways (Mueller et al. 2003).

SNPs between Micro-Tom and other inbred tomato lines (cultivars E6203, R11-13, Rio Grande PtoR, and R11-12 and a wild relative, *S. pennellii* TA56) were identified by comparing EST sequences (Yamamoto et al. 2005). Currently, 2160 candidate SNPs are available in the database. SSRs were identified using a Perl script (Temnykh et al. 2001) for searching the UNIGENE sequences (see below). Currently, 409 candidate SSRs are available in the database. We created a viewer to show the alignments of the UNIGENE and EST sequences in which the SNP and SSR nucleotides are highlighted (Figure 1C).

To obtain information on gene expression, ESTs in each UNIGENE were classified into 27 groups according to the origins of the cDNA libraries used for EST sequencing, and the number of the ESTs in each group was counted. By selecting groups, the UNIGENEs that match the conditions are presented on the screen (Figure 1D). The search function serves as a rough measure of the expression of tomato genes.

To compare the sequence similarity between genes of other plant species or organisms, a data set was produced by the BLAST programs against the National Center for Biotechnology Information nonredundant amino acid sequences, *Arabidopsis* Coding Sequences, and TIGR Gene Indices of rice, soybean, and maize. The search function for the data set provides a list of UNIGENEs that match the conditions.

Search functions of MiBASE

The information in MiBASE can be searched *via* the Internet by querying with keywords or sequences. The results of the searches are viewed in the result pages, wherein detailed information can be accessed *via* hyperlinks to the DDBJ, TIGR, TAIR, and GO consortium (http://www.geneontology.org/index.shtml) databases.

Acknowledgements

This work was supported by the Research Project for Utilizing Advanced Technologies in Agriculture Forestry and Fisheries of the Ministry of Agriculture, Forestry, and Fisheries in Japan and by a grant from the Kazusa DNA Research Institute.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–410

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

Berardini TZ, Mundodi S, Reiser L, Huala E, Garcia-Hernandez M, Zhang P, Mueller LA, Yoon J, Doyle A, Lander G., Moseyko N, Yoo D, Xu I, Zoeckler B, Montoya M, Miller N, Weems D, Rhee SY (2004) Functional annotation of the *Arabidopsis* genome using controlled vocabularies. *Plant Physiol* 135: 745–755

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* 436: 793–800

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res* 32: D277–280

Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J (2005) The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Res* 33: D71–74

Meissner R, Jacobson Y, Melamed S, Levyatuv S, Shalev G, Ashri A, Elkind Y, Levy A (1997) A new model system for tomato genetics. *The Plant Journal* 12: 1456–1472

Mueller LA, Zhang P, Rhee SY (2003) AraCyc: A Biochemical Pathway Database for *Arabidopsis*. *Plant Physiol* 132: 453–460

Mueller LA, Solow TH, Taylor N, Skwarecki B, Buels R, Binns J, Lin C, Wright MH, Ahrens R, Wang Y, Herbst EV, Keyder ER, Menda N, Zamir D, Tanksley SD (2005) The SOL Genomics Network: a comparative resource for Solanaceae biology and beyond. *Plant Physiol* 138: 1310–1317

Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, Miller N, Mueller LA, Mundodi S, Reiser L, Tacklind J, Weems DC, Wu Y, Xu I, Yoo D, Yoon J, Zhang P (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

Shibata D (2005) Genome sequencing and functional genomics approaches in tomato. *J Gen Plant Pathol* 71: 1–7

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res* 11: 1441–1452

Tsugane T, Watanabe M, Yano K, Sakurai N, Suzuki H, Shibata D (2005) Expressed sequence tags of full-length cDNA clones from the miniature tomato (*Lycopersicon esculentum*) cultivar Micro-Tom. *Plant Biotechnology* 22: 161–165

Yamamoto N, Tsugane T, Watanabe M, Yano K, Maeda F, Kuwata C, Torki M, Ban Y, Nishimura S, Shibata D (2005) Expressed sequence tags from the laboratory-grown miniature tomato (*Lycopersicon esculentum*) cultivar Micro-Tom and mining for single nucleotide polymorphisms and insertions/deletions in tomato cultivars. *Gene* 356: 127–134