Gene Note

Expressed sequence tags of full-length cDNA clones from the miniature tomato (*Lycopersicon esculentum*) cultivar Micro-Tom

Taneaki Tsugane[1], Manabu Watanabe[2], Kentaro Yano[2], Nozomu Sakurai[2], Hideyuki Suzuki[2], Daisuke Shibata[2]*

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

Received March 10, 2005; accepted April 15, 2005 (Edited by K. Yazaki)

Abstract Tomato genome sequencing projects have started to become an internationally coordinated program. To accelerate tomato functional genomics studies in coordination with the complete sequencing of the tomato genome, we prepared a full-length cDNA library from the miniature tomato (*Lycopersicon esculentum*) cultivar Micro-Tom, which has attracted attention as a laboratory-grown model plant. Total RNA from maturing fruits was subjected to a vector-capping protocol for full-length cDNA synthesis. We generated 8,046 expressed sequence tags (ESTs), which comprised 3,484 contigs. We calculated that 80.7% of the cDNA clones in the library met the criteria for full-length clones, and 1,920 non-redundant full-length clones were identified. As a pilot experiment, we chose seven clones, whose encoded proteins shared low homology with *Arabidopsis* proteins, for full sequencing. Of these, three genes had no or very low homology with *Arabidopsis* genes, indicating the usefulness of the library for analyses of "not-found-in-Arabidopsis" genes.

Key words: Full-length cDNA, *Lycopersicon esculentum*, Micro-Tom.

Recently, genome sequencing programs for tomato (*Lycopersicon esculentum*) have started to be the major activity of the internationally coordinated International Solanaceae Genome Project (SOL) consortium (http://www.sgn.cornell.edu/solanaceae-project/). Tomato was chosen as a model of the Solanaceae because of its moderately sized genome of 950 Mb (Arumuganathan et al. 1991), which is estimated to encode ~35,000 genes in gene-rich euchromatin regions (Van der Hoeven et al. 2002). Several genetic and genomic resources of tomato such as inbred lines, DNA markers, mutagenized populations and expressed sequence tags (ESTs) are available (see review, Shibata 2005). Therefore, tomato is a promising model crop for agricultural research. Further development of resources such as a comprehensive set of full-length cDNA clones will expand the potential usefulness of tomato for use in genetic and functional genomics approaches.

A miniature tomato cultivar, Micro-Tom, which was originally bred for home gardening (Scott and Harbaugh 1989), is a suitable host for genetic research (Meissner et al. 1997; Emmanuel and Levy 2002). Its small size of 10–20 cm in height, ability to grow well at high densities and short life cycle of 70–90 days are suitable for cultivation and experimentation in most plant biology laboratories (Meissner et al. 1997; Emmanuel and Levy 2002). Mutant populations of Micro-Tom generated by ethyl methane sulfonate exhibit various phenotypic mutations of leaves, fruits and flower shape and color (Meissner et al. 1997), providing a promising genetic resource. Micro-Tom is also a suitable host for 16 well-known fungal, bacterial, and viral pathogens of tomato (Takahashi et al. 2005). Recently, 35,824 ESTs from leaves and fruits of Micro-Tom became available (BP875611-BP91143, Yamamoto et al. 2005), providing DNA sequence information for the cultivar. A large collection of ESTs from various tomato cultivars or lines and a few wild relatives has been deposited in the NCBI database (Benson et al. 2003) (dbEST, 189,735 ESTs, April 8, 2005). Candidates for SNPs between Micro-Tom and other cultivars were mined from tomato EST data sets, providing DNA markers for map-based cloning from Micro-Tom mutants and for transferring useful traits found in Micro-Tom mutants to commercial cultivars (Yamamoto et al. 2005).

Full-length cDNA clones are a fundamental resource for genomic research, useful not only for functional analysis of proteins but also for prediction of protein coding regions from genome sequences, especially for genes that have no homologous sequences in other

The sequences reported in this article have been deposited in the DDBJ database under accession numbers BW684914-BW692959 for the ESTs and AB211519, AB211521-AB211526 for full-length cDNAs.
Abbreviations: EST, expressed sequence tag; nr, non-redundant protein database.

organisms. Several protocols for preparing full-length cDNA libraries are available (Maruyama and Sugano 1994; Seki et al. 1998). In plants, comprehensive full-length cDNA clone sets of *Arabidopsis thaliana* (Seki et al. 2002) and rice (*Oryza sativa*) (Kikuchi et al. 2003) are available. However, no full-length cDNA library of tomato has been reported.

In this study, we prepared a full-length cDNA library from maturing fruits using a new protocol of Kato et al. (2005) and generated 8,046 ESTs from the library.

Maturing fruits of Micro-Tom, which were grown under natural conditions in a greenhouse, were collected at the mature green stage, the light green stage, the breaker stage, the turning stage, the light red stage and the red ripe stage, as defined according to color changes described in "United States Standards for Grades of Fresh Tomatoes" (United States Department of Agriculture, http://www.ams.usda.gov/standards/). The fruits were immediately frozen in liquid nitrogen, and stored at −80°C until RNA extraction.

RNA was extracted from fruits of the six stages. Briefly, the entire fruit (16 g) was ground to powder in liquid nitrogen by a mortar and pestle and mixed with RNA extraction buffer (4.2 M guanidine thiocyanate, 17 mM Sarkosyl, 25 mM trisodium citrate, and 0.1% Antifoam). Phenol/chloroform extraction was performed three times and RNA was precipitated with isopropyl alcohol as described in Sambrook et al. (1989). Total RNA fractions of six stages (6 μg each) were mixed and further purified using sugar precipitation in the presence of 0.1 M sodium acetate. Sugar precipitation was repeated four times.

We ordered construction of a full-length cDNA library from the total RNA using a vector-capping protocol (Kato et al. 2005) from Hitachi Instruments Service Co., Ltd. (Tokyo, Japan). The vector used for library construction is shown in Figure 1. The cDNA fragments were directionally inserted into the cloning site, which carries rare restriction sites for *Sfi I* at both ends, and introduced them into *Escherichia coli* strain DH12S.

From the full-length cDNA library, 9,792 cDNA clones were randomly selected and single-pass sequenced from the 5′ ends. Plasmid inserts of selected clones were amplified by PCR using forward (5′-CCCAGTCACGACGTTGTAAAACG-3′) and reverse (5′-AGCGGATAACAATTTCACACAGG-3′) primers. The 5′ ends of amplified cDNA fragments were sequenced using the forward primer and BigDye terminator cycle sequencing kit (PE Applied Biosystems, USA), and run on an automated DNA sequencer (ABI PRISM 3730xl DNA sequencer, PE Applied Biosystems, USA). Vector-derived and ambiguous sequences (PHRED quality value<20) were eliminated using a combination of the Phred program (Ewing et al. 1998) and CROSS_MATCH software (http://www.phrap.org). Poly(A) tail sequences in the ESTs processed by our Perl script were at most 10 bases. Subsequently, ESTs whose sequences were <50 bp were omitted from the final data set. A total of 8,046 ESTs were generated and submitted to the DDBJ database (accession numbers BW684914-

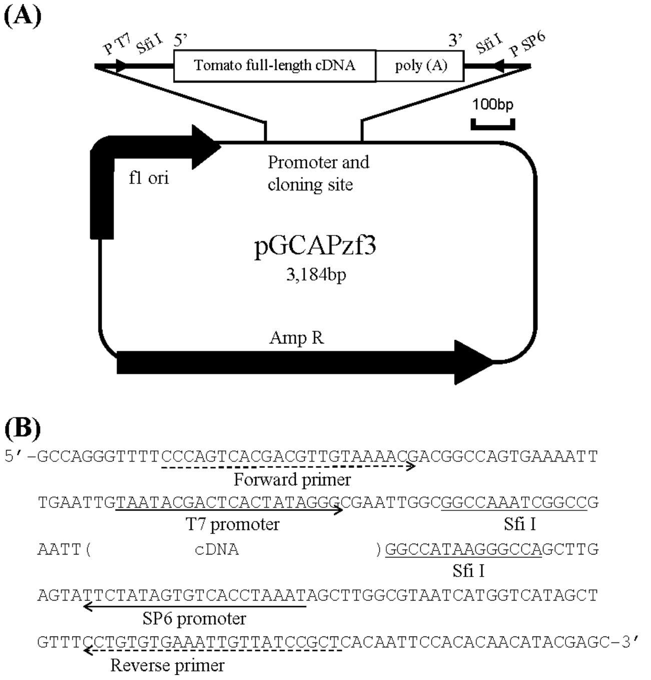
(A)
P T7
Sfi I
5'
Tomato full-length cDNA
poly (A)
3'
Sfi I
P SP6
100bp
f1 ori
Promoter and cloning site
pGCAPzf3
3,184bp
Amp R
(B)
5′-GCCAGGGTTTTCCCAGTCACGACGTTGTAAAACGACGGCCAGTGAAAATT
Forward primer
TGAATTGTAATACGACTCACTATAGGGCGAATTGGCGGCCAAATCGGCCG
T7 promoter
Sfi I
AATT(cDNA)GGCCATAAGGGCCAGCTTG
Sfi I
AGTATTCTATAGTGTCACCTAAATAGCTTGGCGTAATCATGGTCATAGCT
SP6 promoter
GTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAACATACGAGC-3′
Reverse primer

Figure 1. The plasmid vector used for the construction of the full-length cDNA library of Micro-Tom. (A) The structure of the vector pGCAPzf3 and the cloning site. The cDNA synthesized was inserted directionally into the cloning site. (B) The sequence of the cloning site. Locations of forward and reverse primers used for PCR amplification and sequencing and the SP6 and T7 promoters are shown. The full-length cDNA fragment can be excised from the vector by digestion with the 8-base restriction enzyme *Sfi I*.

Table 1. Tomato ESTs in Genbank dbEST.

Inbred line*	Origins of EST	No. of EST
E6203	root, shoot, flower, flower bud, fruit, seed, callus, suspension culture, carpel, crown gall	123,772
Micro-Tom**	leaf, fruit	35,824
Micro-Tom (This work)	fruit	8,046
Rio Grande Pto R	leaf	10,014
R11-13	leaf	5,966
R11-12	leaf	5,402
TA56 (*Lycopersicon pennellii*)	pollen	5,427

* The tomato cultivars of which more than 5,000 ESTs have produced.

** Yamamoto et al. (2005)

Table 2. Classification of molecular functional annotation of 3,808 Micro-Tom contigs found in the 8,046 ESTs.

GO slim term	Number of contigs
molecular function unknown	705
other enzyme activity	415
hydrolase activity	301
structural molecule activity	259
transporter activity	257
transferase activity	243
other molecular functions	233
other binding	222
DNA or RNA binding	179
protein binding	137
transcription factor activity	102
kinase activity	98
nucleic acid binding	87
nucleotide binding	83
receptor binding or activity	23

Classification of functional annotation for biological process and cellular components is available at the Micro-Tom database MiBASE (http://www.kazusa.or.jp/microtom/).

BW692959). The ESTs of Micro-Tom and other tomato cultivars available are listed in Table 1.

The 8,046 ESTs were assembled into 3,808 contigs using the PHRAP program (http://www.phrap.org). Similarity searches of the EST sequences were carried out using the BLASTN program (Altschul et al. 1990, 1997) against the Micro-Tom ESTs (35,824 sequences) that were previously generated from leaves and fruits (Yamamoto et al. 2005), and the ESTs of other cultivars (150,581 sequences from dbEST, Table 1). To the 8,046 ESTs, 82% (6,573 ESTs) and 89% (7,178 ESTs) of the previously identified Micro-Tom ESTs and the other tomato ESTs from dbEST, respectively, had matches with an E-value of <E-30. The clones with no matches could either be derived from novel genes or be too short to match against the previously identified sequences of short clones.

The 3,484 contigs were classified into functional categories based on Arabidopsis gene ontology (GO Slim) (Berardini et al. 2004) (Table 2). Information about ontology for each contig is available at the Micro-Tom database MiBASE (http://www.kazusa.or.jp/microtom/).

We estimated the abundance of cDNA clones encoding full-length proteins in the 8,046 ESTs. BLASTX (1E-5) searches against the non-redundant protein database (nr) in NCBI (ftp://ftp.ncbi.nlm.nih.gov/blast/db/blastdb.txt) were performed. They showed that 6,913 of the 8,046 ESTs have significant homology to existing entries and 5,579 ESTs of the 6,913 ESTs (80.7%) extend further upstream than homologous entries, indicating that they contain full-length cDNA inserts. The 5,579 ESTs of the candidate full-length cDNA clones were assembled into 1,920 non-redundant contigs (including 1,038 singletons) using the Phrap program. The average and standard deviation of the contig sequence length were 708.8 and 134.0 bases, respectively. As the abundance of full-length cDNA clones in the previously identified EST population was 37% of 35,824 ESTs (Yamamoto et al. 2005), the cDNA library constructed in this study is satisfactory as a source of full-length cDNA clones. We have listed 1,920 non-redundant full-length cDNA clones in the Micro-Tom database MiBASE.

It was reported that about 30% of tomato genes have no significant correspondence to *Arabidopsis* genes, and the function of the majority of these genes remains unknown (Van der Hoeven et al. 2002). Thus, we searched the nr database with the sequences of the non-redundant 1,920 contigs and found clones whose sequences shared no or low homology with Arabidopsis proteins. As a pilot experiment, we chose seven clones and subjected them to full sequencing. The sequences obtained were submitted to the DDBJ database (accession numbers AB211519, and AB211521 to AB211526). The full sequences obtained were searched against the nr database (Table 3). The gene product of AB211523 shared sequence homology with a *Mus musculus* protein at a low level, but no sequence homology with any *Arabidopsis* protein. The gene product of AB211521 shared sequence homology with *Plasmodium falciparum* 3D7 and *Arabidopsis* proteins at low levels. The gene product of AB211525 was highly homologous with a potato (*Solanum tuberosum*) protein, but shared low-level sequence homology with an *Arabidopsis* major latex-like protein. The other four genes shared homology with *Arabidopsis* proteins at various levels. The AB211526 gene had high homology

Table 3. Full-length cDNA clones isolated from maturing tomato fruits. The sequences were searched against the nr database with the BLASTX program. The top hit genes and E-values for corresponding *Arabidopsis thaliana* genes are listed.

Tomato gene ID	Organism	Gene product	Gene ID	E-value
AB211523	*Mus musculus*	Unknown (protein for MGC:12025)	AAH05782	3.7
	Arabidopsis thaliana	No hit	—	—
AB211521	*Plasmodium falciparum* 3D7	Hypothetical protein PFL1430c	NP_701648	0.16
	Arabidopsis thaliana	Putative splicing factor	At4g36690	3.8
AB211525	*Solanum tuberosum*	pSTH-2 protein	AAA03019	2e-71
	Arabidopsis thaliana	Major latex-like protein	At1g24020	9e-08
AB211522	*Euphorbia esula*	60S ribosomal protein L35	AAF34800	2e-56
	Arabidopsis thaliana	60S ribosomal protein L35	At5g02610	3e-54
AB211519	*Oryza sativa*	Putative u1 small nuclear ribonucleoprotein C	BAD27897	9e-27
	Arabidopsis thaliana	Putative C-type U1 snRNP	At4g03120	5e-26
AB211524	*Arabidopsis thaliana*	Unknown protein	At5g02160	3e-23
AB211526*	*Arabidopsis thaliana*	Eukaryotic pantothenate kinase family protein	At1g60440	e-125

* The N-terminal half of the gene product has high homology with that of the *Arabidopsis* protein.

at a conserved domain with a eukaryotic pantothenate kinase family protein of *Arabidopsis*, but lacked the C-terminal half of the gene product. These results imply that full-length tomato cDNA clones will be useful for understanding the function of "not-found-in-Arabidopsis" genes.

Full-length cDNA sequences are also crucial components for genome annotation. Genefinder programs such as GlimmerM, Exonomy and Unveil need to be trained with known sequences such as cDNA of the same organism to predict better the gene structures from genome sequences (Majoros et al. 2003). Full sequencing of the 1,920 non-redundant full-length clones identified in this study and subsequent application of the resulting sequences to genefinder program training would facilitate genome annotation when the wealth of the upcoming tomato genome sequence becomes available. We are currently working on full sequencing of these clones.

Acknowledgements

This work was supported by a grant from the Kazusa DNA Research Institute Foundation and a grant from the Japanese Ministry of Agriculture, Forestry and Fisheries for research projects utilizing advanced technologies in agriculture, forestry and fisheries.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 5: 403–410

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res* 25: 3389–3402

Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. *Plant Mol Biol Rep* 9: 208–218

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2003) GenBank. *Nucl Acids Res* 31: 23–27

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

Boguski MS, Lowe TMJ, Tolstoshev CM (1993) dbEST—database for "expressed sequence tags". *Nature Genet* 4: 332–333

Emmanuel E, Levy AA (2002) Tomato mutants as tools for functional genomics. *Curr Opin Plant Biol* 5: 112–117

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using *Phred*. I. Accuracy assessment. *Genome Res* 8: 175–185

Kato S, Ohtoko K, Ohtake H, Kimura T (2005) Vector-capping: a simple method for preparing a high-quality full-length cDNA library. *DNA Research* (in press)

Kikuchi S, Satoh K, Nagata T, Kawagashira N, Doi K, Kishimoto N, Yazaki J, Ishikawa M, Yamada, Ooka H, Hota I, Kojima K, Namiki T, Ohneda E, Yahagi W, Suzuki K, Li CJ, Ohtsuki K, Shishiki T, Otomo Y, Murakami K, Iida Y, Sugano S, Fujimura T, Suzuki Y, Tsunoda Y, Kurosaki T, Kodama T, Masuda H, Kobayashi M, Xie Q, Lu M, Narikawa R, Sugiyama A, Mizuno K, Yokomizo S, Niikura J, Ikeda R, Ishibiki J, Kawamata M, Yoshimura A, Miura J, Kusumegi T, Oka M, Ryu R, Umeda M, Matsubara K, Kawai J, Carninci P, Adachi J, Aizawa K, Arakawa T, Fukuda S, Hara A, Hashizume W, Hayatsu N, Imotani K, Ishii Y, Itoh M, Kagawa I, Kondo S, Konno H, Miyazaki A, Osato N, Ota Y, Saito R, Sasaki D, Sato K, Shibata K, Shinagawa A, Shiraki T, Yoshino M, Hayashizaki Y, Yasunishi A (2004) Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science* 301: 376–379

Majoros WH, Pertea M, Antonescu C, Salzberg SL (2003) GlimmerM, Exonomy and Unveil: three *ab initio* eukaryotic genefinders. *Nucl Acid Res* 31: 3601–3604

Maruyama K, Sugano S (1994) Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene* 138: 171–174

Meissner R, Jacobson Y, Melamed S, Levyatuv S, Shalev G, Ashri A, Elkind Y, Levy A (1997) A new model system for tomato genetics. *Plant J* 12: 1465–1472

Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular cloning: A laboratory manual, 2nd edition*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York

Scott JW, Harbaugh BK (1989) MicroTom—a miniature dwarf tomato. *Florida Agr Exp Sta Circ* 370: 1–6

Seki M, Carninci P, Nishiyama Y, Hayashizaki Y, Shinozaki K (1998) High-efficiency cloning of *Arabidopsis* full-length cDNA by biotinylated CAP trapper. *Plant J* 15: 707–720

Seki M, Narusaka M, Kamiya A, Ishida J, Satou M, Sakurai T, Nakajima M, Enjy A, Akiyama K, Oono Y, Muramatsu M, Hayashizaki Y, Kawai J, Carninci P, Itoh M, Ishii Y, Arakawa T, Shibata K, Shinagawa A, Shinozaki K (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science* 296: 141–145

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

Takahashi H, Shimizu A, Arie T, Rosmalawati S, Fukushima S, Kikuchi M, Hikichi Y, Kanda A, Takahashi A, Kiba A, Ohnishi K, Ichinose Y, Taguchi F, Yasuda C, Kodama M, Egusa M, Masuta C, Sawada H, Shibata D, Hori K, Watanabe Y (2005) Catalog of Micro-Tom tomato responses to common fungal, bacterial, and viral pathogens. *J Gen Plant Pathol* 71: 8–22

Van der Hoeven R, Ronning C, Giovannoni J, Martin G, Tanksley S, (2002) Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. *Plant Cell* 14: 1441–1456

Yamamoto N, Tsugane T, Watanabe M, Yano K, Maeda F, Kuwata C, Torki M, Ban Y, Nishimura S, Shibata D (2005) *Gene* (in press)