Published online 16 November 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D901–D905***
doi:10.1093/nar/gkl921

TomatEST database: *in silico* exploitation of EST data to explore expression patterns in tomato species

Nunzio D'Agostino, Mario Aversano, Luigi Frusciante[1] and Maria Luisa Chiusano*

Department of Structural and Functional Biology, University 'Federico II', 80126 Naples, Italy
and [1]Department of Soil, Plant and Environmental Sciences, University 'Federico II', 80055 Portici, Naples, Italy

Received August 1, 2006; Revised and Accepted October 17, 2006

ABSTRACT

TomatEST is a secondary database integrating expressed sequence tag (EST)/cDNA sequence information from different libraries of multiple tomato species. Redundant EST collections from each species are organized into clusters (gene indices). A cluster consists of one or multiple contigs. Multiple contigs in a cluster represent alternatively transcribed forms of a gene. The set of stand-alone EST sequences (singletons) and contigs, representing all the computationally defined 'Transcript Indices', are annotated according to similarity versus protein and RNA family databases. Sequence function description is integrated with the Gene Ontologies and the Enzyme Commission identifiers for a standard classification of gene products and for the mapping of the expressed sequences onto metabolic pathways. Information on the origin of the ESTs, on their structural features, on clusters and contigs, as well as on functional annotations are accessible via a user-friendly web interface. Specific facilities in the database allow Transcript Indices from a query be automatically classified in Enzyme classes and in metabolic pathways. The 'on the fly' mapping onto the metabolic maps is integrated in the analytical tools. The TomatEST database website is freely available at http://biosrv.cab.unina.it/tomatestdb.

INTRODUCTION

Solanum lycopersicum is a tomato species with a modest-sized diploid genome, and it is tolerant to inbreeding. This is why it has been selected as a model organism to study several topics of plant biology, including fruit development (1), response to biotic/abiotic stress and plant diversification and adaptation.

The International Tomato Genome Sequencing Project is ongoing (2) and it is paralleled by the intensive production of expressed sequence tags (ESTs) from different tomato and other Solanaceae species (3–5) to support the study of Solanaceae biology and to provide a consistent resource for expression studies (4), for gene discovery, for genome annotation (6) and for comparative genomics.

The 'tag' nature and the vast quantity of ESTs require suitable approaches to harvesting the full potential from this data source. Hence, several efforts, based on bioinformatics methodologies, are focused on the construction of information frameworks, where the fragmented and error-prone EST data are organized into tentative consensus (TC) sequences, representing possible alternative transcripts of a gene, for investigations on functional roles and expression mechanisms (7,8).

Several specific EST repositories from *S.lycopersicum* are available worldwide. The TIGR Tomato Gene Index (LeGI) is a collection of high-fidelity virtual TC sequences constructed by clustering and assembling ~163 000 ESTs (release 10.1) generated in the laboratories of the TIGR Institute, of the Cornell University and of the Boyce Thompson Institute. The SOL Genomics Network (SGN) (9), a website dedicated to the biology of Solanaceae family, organizes and distributes ESTs (~176 000), sequenced from 35 different cDNA libraries from *S.lycopersicum*, *Solanum pennellii*, *Solanum habrochaites* and the corresponding 'combined' consensus sequences. Other EST resources are (i) the Tomato Stress EST Database (TSED), which contains ESTs from more than 10 stress-treated substractive cDNA libraries from *S.lycopersicum*; (ii) the Micro-Tom Database (MiBASE) (10), which distributes ~8000 ESTs from a full-length cDNA library from the fruit of Micro-Tom (a miniature and dwarf tomato cultivar); (iii) the PlantGDB (11), which collects PlantGDB-assembled Unique Transcripts (PUT) from *S.lycopersicum* generated from EST sequences available at the NCBI dbEST database (12).

We present here TomatEST, a secondary database of EST/cDNA sequences from 105 libraries from all the tomato species available at dbEST. TomatEST has been designed

*To whom correspondence should be addressed. Tel/Fax: +39 081679186; Email: chiusano@unina.it

to provide a workbench for mining the complexity of EST sequence information content from multiple tomato species (i) for expression pattern analysis and (ii) for gene discovery in the framework of the *S.lycopersicum* genome project.

IMPLEMENTATION AND ARCHITECTURE

TomatEST architecture consists of a relational database, a web interface created using HTML and PHP scripts which dynamically execute MySQL queries. It operates under an Apache web server on a Fedora Linux system. We have developed an entity relationship data model for TomatEST raw and processed data as shown in Figure 1.

DATABASE CONTENT

TomatEST is designed to support investigations on expressed sequence data from multiple tomato species. The current release includes 200 438 ESTs from *S.lycopersicum*, 8346 from *S.pennellii*, 8000 from *S.habrochaites* and 1008 from *S.lycopersicum* × *Solanum pimpinellifolium*. All EST sequence information were collected from 105 libraries covering different tissues, developmental stages and treatments, downloaded from the NCBI dbEST database (release 020106).

EST/cDNA collections were processed by ParPEST, a pipeline for comprehensive EST data analyses (13). The database contains: (i) raw data; (ii) higher-quality sequences obtained by the EST pre-processing; (iii) the TC sequences (contigs) obtained from the assembling phase; (iv) clusters with single or multiple contigs; (v) the functional annotations based on BLAST similarity searches.

After quality checking and vector trimming, EST sequences sharing >85% identity over a region longer than 60 nts are grouped into clusters. Sequences in a cluster are assumed to represent the same gene, this is why each cluster is defined as a gene index. The EST set in a cluster can be assembled in one or multiple contigs. Indeed, since the clustering process is a simple 'tentative closure' procedure, the clustering program will find the overlaps among EST sequences, not considering if they make sense all together. When sequences in a cluster cannot be all reconciled into a consistent multiple alignment during the much more rigorous assembly phase they are split accordingly into multiple assemblies/contigs. Possible interpretations of multiple contigs from a cluster are: (i) alternative transcription, (ii) paralogy or (iii) protein domain sharing.

A summary of the information collected in the database is shown in Table 1, where the Transcript Indices represent the number of singletons plus contigs we assume to indicate the number of classified transcripts per species.

Functional annotation is performed both on EST sequences and on contigs, to allow checking on annotation consistencies when ESTs sequences are assembled. The functional annotation is based on the detection of similarities (E-value ≤ 0.001) with both proteins and non-protein coding

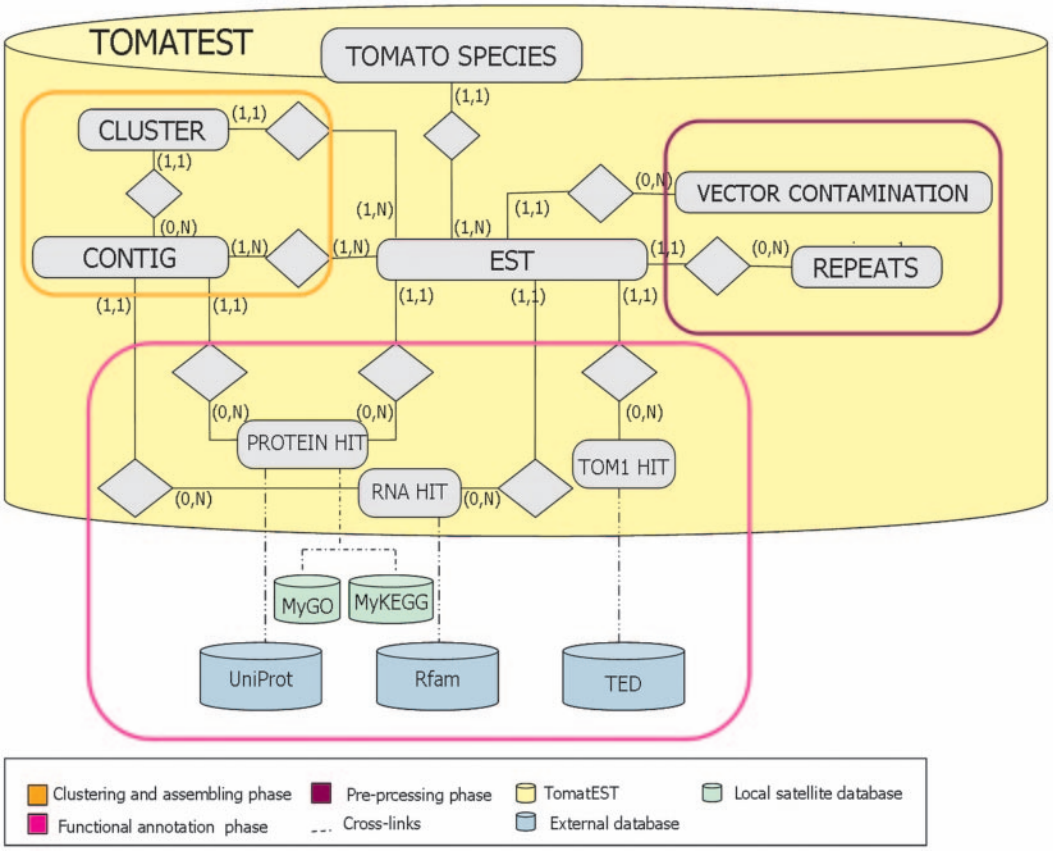

Figure 1. Database organization.

Table 1. Current status of the TomatEST database

Source	ESTs	Libraries	Gene indices	Transcript indices	Contigs	Singletons
Solanum lycopersicum	200 438	95	42 261	43 370	16 888	26 482
Solanum pennelli	8346	2	4349	4355	741	3614
Solanum habrochaites	8000	2	4263	4309	1088	3221
S.lycopersicum × *Solanum pimpinellifolium*	1008	6	746	746	96	650
All	217 792	105	51 619	52 780	18 813	33 967

RNAs, by BLAST searches versus the UniProt (14) and the Rfam databases (15), respectively. Protein and RNA identifiers are used to build cross-references to the corresponding external databases. TomatEST is integrated with two local satellite databases: myGO, a mirror of the Gene Ontology database (16), and myKEGG, built from KEGG (17) XML formatted files and the related maps in GIFF format. When the UniProt identifier is recorded in myGO, Gene Ontologies are associated to the transcript, to integrate the UniProt annotation with an international standard. If the Enzyme Commission (EC) number is present in the BLAST hit description lines, a cross-reference to the ENZYME database (18) is provided in order to include information such as the enzyme name and its synonyms, reaction(s), substrate(s) and product(s). Proteins that are associated to EC number(s) are also hyperlinked to myKEGG allowing the mapping of the expressed sequences onto known metabolic pathways.

EST reads similar (E-value $\leqslant 10^{-5}$) to the $\sim$12 000 sequences spotted on the TOM1 cDNA microarray, are cross-linked to the 'microarray expression data' section of the Tomato Expression Database (TED) (19). The genome coordinates indicating the start/end positions of the ESTs/contigs when mapped onto BAC sequences available from the ongoing International Tomato Genome Sequencing Project are also included and cross-references to the genome sequence annotation pages are provided.

QUERYING THE DATABASE

TomatEST web application supports data retrieval through a pre-defined query system. Data can be inspected via three different HTML forms to allow distinctive queries on (i) EST sequences, (ii) clusters and (iii) Transcript Indices.

The first HTML form produces an '*ESTs report page*', displaying each EST as a green bar, with vector contaminations, or low complexity sub-sequences and repeats as highlighted regions, when present. The EST bar is linked to the nucleotide sequence. Protein as well as non-protein coding RNA matching regions are drawn as grey bars on the length of the query sequence; each bar is linked to details on the local alignments.

The second HTML form results in a '*Clusters report page*', where data are presented in a summary table and cluster ids are listed and linked to the cluster structure. The cluster structure is represented by contigs as orange bars; EST bars are drawn along the contig bar length to represent the assembly. The protein and non-protein-coding RNA matching regions of the contig sequence are also reported. Each contig bar is linked to the EST multiple alignment which contig was generated from.

The third HTML form results in a '*Transcript Indices report page*', where data corresponding to user-selected criteria are listed in a table summarizing their structure and function annotation. Data can be also analysed considering two different classes of objects: the enzymes and the metabolic pathways. Enzymes are classified into classes, subclasses and sub-subclasses according to the guidelines of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB). They are listed as HTML-based tree menus (Figure 2A). For each enzyme in the list, we reported all the transcripts which, in the functional annotation phase, have been associated to the same enzyme. Redundancy may occur because (i) more UniProt proteins referenced in the ENZYME repository with the same EC identifier; (ii) different transcripts encoding for different subunits of the same enzyme; (iii) different transcripts representing different segments of the same mRNA not assembled because of the 'tag' nature of the ESTs. Because one enzyme can contribute to more than one metabolic pathway, all the pathways which the enzyme belongs to are also listed in the tree menu. The description of the metabolic pathways is based on the KEGG collection of metabolic maps. The class '*metabolic pathway*' is useful to investigate on a specific map and on its 'coverage'; indeed, the enzymes associated to Transcript Indices resulting from a query are mapped 'on the fly' onto the pathways in which they occur. All the metabolic maps are listed as HTML-based tree menus; for each map we report the number of the enzymes mapped and which of them are specific in that map (Figure 2B). Metabolic pathways can be always accessed as GIFF images which are modelled as graphs where a node represents an enzyme and an edge represents an interaction. For each map we report the 'activated' nodes (enzymes) highlighted in red (Figure 2C). A BLAST service is also provided.

CONCLUDING REMARKS

TomatEST has been designed to manage and to explore the vast amount of ESTs from collections of tomato species providing a reference for expression pattern analysis, for gene discovery and for genome sequencing in the frame of the Tomato Genome Project. The EST sequences are from 105 cDNA libraries from various tissues at defined developmental stages and treatments. Table 1 (EST and Libraries) reflects the current status of the worldwide tomato EST sequencing projects, where libraries from *S.lycopersicum* represent the majority (95/105). This wealth of EST information can be properly investigated for tissue specific expression pattern analysis. The complete set of raw EST sequences are maintained in the database allowing users investigate on library quality and on single EST structural features (vector contamination, repeat regions, function annotation).

The removal of contaminating vector sequences (13) before the clustering procedure, slightly reduced the dataset

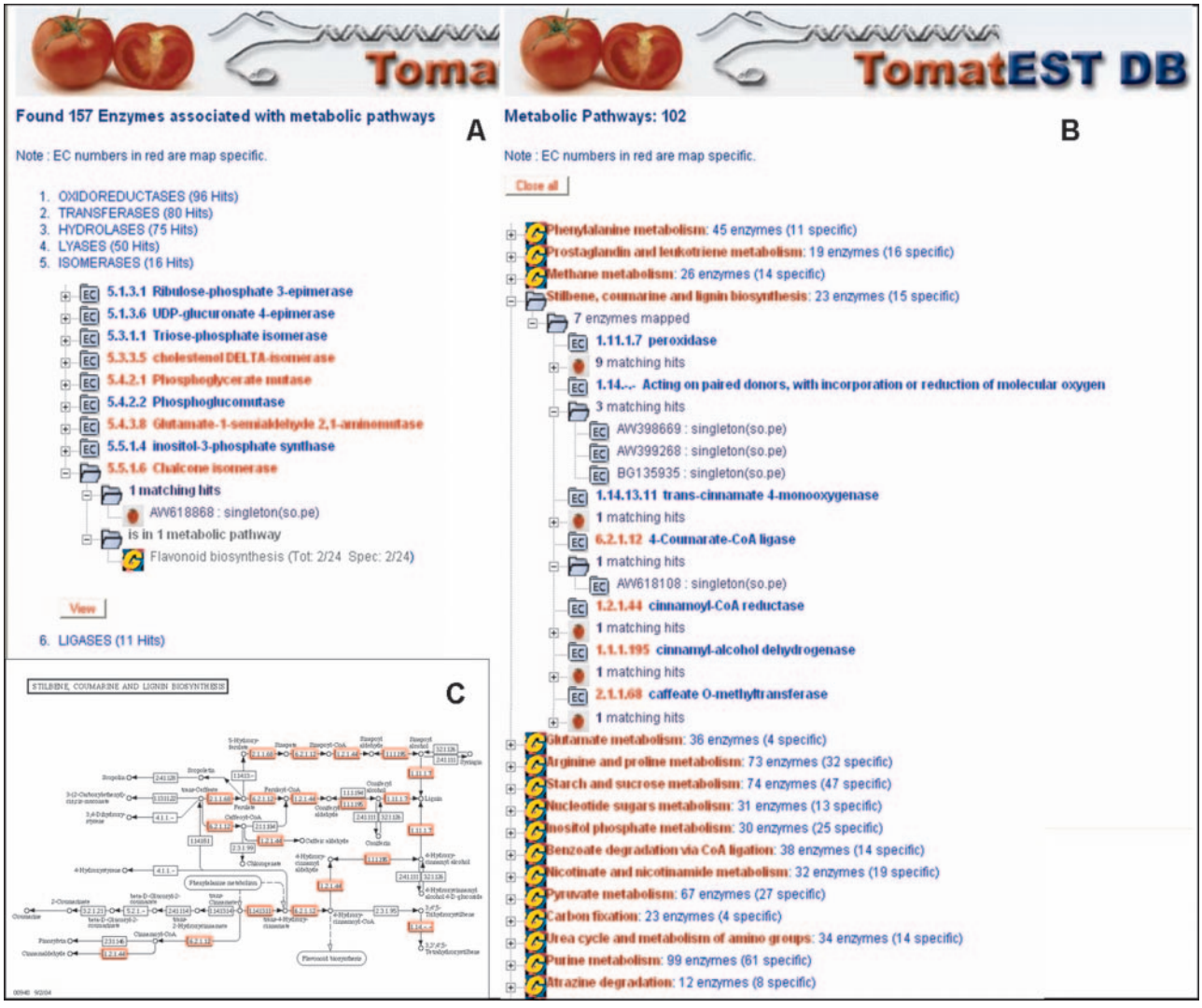

Figure 2. Snapshots of the TomatEST web interface. The panels show examples resulting from a general query on *S.pennellii*. (**A**) Shows an example of the tree menu listing all the metabolic enzymes annotated for the species. The node corresponding to chalcone isomerase (EC 5.1.1.6) is expanded showing the transcript associated with the enzyme and the metabolic pathway which include the enzyme. (**B**) Shows an example of the tree menu listing all the metabolic pathways associated with *S.pennellii*. The node corresponding to 'Stilbene, coumarine and lignine biosynthesis' is expanded. (**C**) Shows the pathway schema where seven enzymes highlighted in red are mapped.

of *S.lycopersicum* (from 200 438 to 200 280) and of *S.pennellii* (from 8346 to 7815); while less than 10 sequences were removed from *S.habrochaites* and *S.lycopersicum* × *S.pimpinellifolium* datasets. The report on regions matching non-protein coding RNAs within EST sequences also allows to investigate on possible contaminations still present, such as yet unprocessed RNAs and intron retaining. As an example, of 1329 ESTs matching sequences from the Rfam database, 1092 share similarity also with proteins; among these sequences 834 ribosomal RNAs are still present.

TomatEST is the first EST database of tomato species that includes the possibility to check on clusters organization. In Table 2, the number of contigs per cluster for each species is shown. The investigation on EST clusters may provide a useful tool for the detection of possible alternative, paralog or domain sharing transcripts. The computational engine, implemented to automatically classify single EST as well as Transcript Indices according to the functional annotation, aims to support the analysis and the mining on genome functionalities. The organization of Transcript Indices associated with enzymes in classes and in metabolic pathways represents a novelty in EST database organization. Moreover, the 'on the fly' mapping of the transcripts to the corresponding maps (Figure 2C), allows friendly investigations on the 'coverage' of the pathways.

TomatEST is a species specific workbench for EST data management and analysis, designed to offer the possibility to investigate on different libraries, from different tissues, at different developmental stages. The database has been built to permit the study of species specific expression patterns and their time course, in normal or pathological conditions and/or under specific biotic or abiotic stimula, exploiting the large amount of libraries today available for Tomato species. Moreover, such a collection also represents a reference to support the ongoing Tomato Genome Sequencing Project. We may well hope that this database will

Table 2. Contig distribution per cluster in the tomato species

S.lycopersicum × *S.pimpinellifolium*		*S.pennellii*		*S.habrochaites*		*S.lycopersicum*	
Cluster	Contigs per cluster	Cluster	Contigs per cluster	Cluster	Contigs per cluster	Cluster	Contigs per cluster
96	1	730	1	1025	1	15 133	1
—	—	4	2	10	2	515	2
—	—	1	3	2	3	66	3
—	—	—	—	1	4	19	4
—	—	—	—	1	6	9	5
—	—	—	—	1	9	5	6
—	—	—	—	1	18	2	7
—	—	—	—	—	—	1	8
—	—	—	—	—	—	2	9
—	—	—	—	—	—	1	10
—	—	—	—	—	—	2	11
—	—	—	—	—	—	1	13
—	—	—	—	—	—	1	19
—	—	—	—	—	—	1	23
—	—	—	—	—	—	1	249

contribute to the comprehension of the structure and the functionality of the tomato genome.

AVAILABILITY

The TomatEST database is freely available at http://biosrv.cab.unina.it/tomatestdb. All questions, comments and requests should be sent by email to chiusano@unina.it.

ACKNOWLEDGEMENTS

We thank Dr Alessandra Traini and Dr Enrico Raimondo for useful feedback on the database usage. We thank Prof. Gerardo Toraldo for technical support and useful discussions. We thank Dr James Giovannoni for kindly providing data and feedback on TOM1 array. This work is supported by the AGRONANOTECH project (Italian Ministry of Agriculture) and by the EU-SOL project (VI frame programme of the European Community). Funding to pay the Open Access publication charges for this article was provided by the AGRONANOTECH project.

Conflict of interest statement. None declared.

REFERENCES

1. Giovannoni,J.J. (2004) Genetic regulation of fruit development and ripening. *Plant Cell*, **16**, S170–S180.
2. Mueller,L.A., Tanksley,S.D., Giovannoni,J.J., van Eck,J., Stack,S., Choi,D., Kim,B.D., Chen,M., Cheng,Z., Li,C. *et al*. (2005) The Tomato Sequencing Project, the first cornerstone of the International Solanaceae Project (SOL). *Comp. Funct. Genom*., **6**, 153–158.
3. Fei,Z., Tang,X., Alba,R.M., White,J.A., Ronning,C.M., Martin,G.B., Tanksley,S.D. and Giovannoni,J.J. (2004) Comprehensive EST analysis of tomato and comparative genomics of fruit ripening. *Plant J*., **40**, 47–59.
4. Rensink,W.A., Lee,Y., Liu,J., Iobst,S., Ouyang,S. and Buell,C.R. (2005) Comparative analyses of six solanaceous transcriptomes reveal a high degree of sequence conservation and species-specific transcripts. *BMC Genomics*, **6**, 124.
5. Ronning,C.M., Stegalkina,S.S., Ascenzi,R.A., Bougri,O., Hart,A.L., Utterbach,T.R., Vanaken,S.E., Riedmuller,S.B., White,J.A., Cho,J. *et al*. (2003) Comparative analyses of potato expressed sequence tag libraries. *Plant Physiol*., **131**, 419–429.
6. Brendel,V., Xing,L. and Zhu,W. (2004) Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same locus. *Bioinformatics*, **20**, 1157–1169.
7. Lee,Y., Tsai,J., Sunkara,S., Karamycheva,S., Pertea,G., Sultana,R., Antonescu,V., Chan,A., Cheung,A. and Quackenbush,J. (2005) The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Res*., **33**, D71–D74.
8. Pontius,J.U., Wagner,L. and Schuler,G.D. (2003) UniGene: a unified view of the transcriptome. In *The NCBI Handbook*. National Center for Biotechnology Information, Bethesda, MD, Chapter 21, pp. 1–12.
9. Mueller,L.A., Solow,T.H., Taylor,N., Skwarecki,B., Buels,R., Binns,J., Lin,C., Wright,M.H., Ahrens,R., Wang,Y. *et al*. (2005) The SOL Genomics Network: a comparative resource for *Solanaceae* biology and beyond. *Plant Physiol*., **138**, 1310–1317.
10. Yamamoto,N., Tsugane,T., Watanabe,M., Yano,K., Maeda,F., Kuwata,C., Torki,M., Ban,Y., Nishimura,S. and 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.
11. Dong,Q., Lawrence,C.J., Schlueter,S.D., Wilkerson,M.D., Kurtz,S., Lushbough,C. and Brendel,V. (2005) Comparative plant genomics resources at PlantGDB. *Plant Physiol*., **139**, 610–618.
12. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST—database for 'expressed sequence tags'. *Nature Genet*., **4**, 332–333.
13. D'Agostino,N., Aversano,M. and Chiusano,M.L. (2005) ParPEST: a pipeline for EST data analysis based on parallel computing. *BMC Bioinformatics*, **6** (Suppl. 4), S9.
14. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res*., **33**, D154–D159.
15. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*., **33**, D121–D124.
16. The Gene Ontology Consortium (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res*., **34**, D322–D326.
17. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*., **34**, D354–D357.
18. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res*., **28**, 304–305.
19. Fei,Z., Tang,X., Alba,R. and Giovannoni,J. (2006) Tomato Expression Database (TED): a suite of data presentation and analysis tools. *Nucleic Acids Res*., **34**, D766–D770.