Plant Biotechnology 24, 17–25 (2007)

Review

Genomic Databases for Tomato

Kentaro Yano,[a] Koh Aoki, Daisuke Shibata*

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 October 13, 2006; accepted November 2, 2006 (Edited by T. Hashimoto)

Abstract Tomato (*Solanum lycopersicum*) is a model plant of the Solanaceae family. Various biological aspects of tomato have been investigated with molecular biological approaches, and a significant amount of DNA and protein sequencing data on tomato has been accumulated. Recently, the number of tomato genome sequences in the International Nucleotide Sequence Databases has been rapidly increasing due to the progress of the international tomato genome sequencing project. Here, we summarize the current status of tomato genetic and genomic databases currently available in the public domain. The wealth of tomato genetic and genomic information facilitates elucidation of gene functions and metabolic pathways that will lead to the understanding of genetic diversity in the Solanaceae family.

Key words: Expressed sequence tag (EST), genome sequence, metabolome, transcriptome, unigene.

Tomato (*Solanum lycopersicum*, formerly *Lycopersicon esculentum*, $2n=24$) is a vegetable crop consumed worldwide and a model plant of the Solanaceae family. Various biological aspects of tomato, such as carotenoid biosynthesis, hormonal effects, fruit development and pathogenesis, have been investigated with physiological and genetic approaches for a long time (see reviews, e.g., Bramley 2002; Gorguet et al. 2005; Pedley and Martin, 2003). DNA sequencing approaches also have been applied from the 1980s onward, and the number of tomato sequences stored in The International Nucleotide Sequence Databases (INSD) (Brunak et al. 2002) maintained by DDBJ (Okubo et al. 2006), EMBL (Cochrane et al. 2006) and GenBank (Benson et al. 2006), has steadily increased (Table 1). In 2004, the tomato genome sequencing program was launched by the internationally coordinated International Solanaceae Genome Project (SOL) consortium (Mueller et al. 2005a). The current sequencing project aims to completely sequence the entire euchromatic regions (approximately 220 Mb), which contain the majority of genes in the tomato genome, on the distal portions of the arms of each chromosome. In the tomato genome (approximately 950 Mb total) (Arumuganathan and Earle, 1991), the other regions are pericentromeric heterochromatin that is largely devoid of genes (http://www.sgn.cornell.edu/about/tomato_project_overview.pl). Due to recent progress of the tomato sequencing project, a rapidly increasing number of tomato genome survey sequences (GSS) are being deposited in INSD (Table 1). At present, the number of tomato GSS is the largest among plants, followed by *Zea mays*, *Brassica oleracea*, *Sorghum bicolor* and *Arabidopsis* (dbGSS release 090806). Over 200,000 tomato expressed sequence tags (ESTs) are now available from INSD (Table 1); this is the largest number of ESTs among vegetable crops (Table 2).

The rapid accumulation of tomato sequencing data

Table 1. Tomato sequences provided by NCBI.

	Protein	Nucleotide			
		CoreNucleotide[a]	EST[b]	GSS[c]	Total
Jan., 1995	90	149	0	0	149
Jan., 2000	557	685	49869	1251	51805
Jan., 2004	1195	2757	135297	11895	149949
Jan., 2005	1359	4948	139001	11895	155844
Jan., 2006	2101	5622	141429	184832	331883
Sep., 2006	2739	6765	201033	320398	528196

[a] High-quality nucleotide sequences.
[b] Expressed sequence tag.
[c] Genome Survey Sequence.

Abbreviations: BAC, bacterial artificial chromosome; BC, backcross population; CAPS, cleaved amplified polymorphic sequence; CDS, coding sequence; EST, expressed sequence tag; FISH, fluorescence in-situ hybridization; GO, gene ontology; HTC, high throughput cDNA sequence; INSD, International Nucleotide Sequence Databases; NCBI, National Center for Biotechnology Information; ORF, open reading frame; RFLP, restriction fragment length polymorphism; SGN, The Solanaceae Genome Project Network; SOL, International Solanaceae Genome Project; 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/

[a] Present address: Agricultural Bioinformatics Receach Unit, Graduate Scool of Agricultural and Life Sciences, The University of Tokyo, Yoyogi 1–1–1, Bunkyo-ku, Tokyo 113–8657 Japan.

Table 2. Twenty plants rank-ordered on the basis of the number of publicly available ESTs.

Species	EST[a]	GSS[a]	CoreNucleotide[a]	Protein[a]	Genome size (Mb)[b]	*n*[b]
Oryza sativa (rice)	1188992	253150	187329	153190	390	12
Zea mays (maize)	1143830	2017441	93797	8930	2400	10
Triticum aestivum (wheat)	854015	7999	6257	3643	16000	21
Arabidopsis thaliana (thale cress)	622972	440439	165696	124636	120	5
*Hordeum vulgar*e + subsp. *vulgare* (barley)	461471	2033	6707	4249	5000	7
Glycine max (soybean)	359158	100064	4848	2861	1200	20
Pinus taeda (loblolly pine)	329469	1792	2473	1581	ND	ND
Vitis vinifera (wine grape)	316756	109147	1309	931	500	19
Malus x domestica (apple tree)	253992	19	1181	989	750	17
Saccharum officinarum (sugarcane)	246301	0	484	427	ND	ND
Medicago truncatula (barrel medic)	225129	168809	3533	19295	500	8
Solanum tuberosum (potato)	219917	1211	3142	2969	840	12
Sorghum bicolor (sorghum)	204208	591193	6256	826	760	10
Lycopersicon esculentum (tomato)	201033	320398	6765	2739	950	12
Physcomitrella patens subsp. *Patens*	194822	0	1066	1023	510	27
Lotus japonicus	150631	46569	1789	575	470	6
Picea glauca	132624	4	1855	1808	3000	17
Helianthus annuus	94110	0	2337	1320	3000	17
Citrus sinensis	93926	0	242	364	380	9[c]
Populus trichocarpa	89943	297	42366	208	480	19

[a]The number of expressed sequence tags (ESTs), genome survey sequences (GSS), high-quality nucleotide sequences (CoreNucleotide) and proteins were derived from NCBI on September 20, 2006.

[b]Genome sizes and the numbers of haploid chromosomes (*n*) were derived from the Entrez Genome Project database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj).

ND; no data in the Entrez Genome Project database.

enables comparative genomic studies with other plants. Complete genome sequences have been determined for *Arabidopsis thaliana* (120 Mb), rice (*Oryza sativa*; 390 Mb) and *Populus trichocarpa* (480 Mb) (Table 2) (*Arabidopsis* Genome Initiative 2000; International Rice Genome Sequencing Project 2005; Tuskan et al. 2006) and are available online. Sequence comparisons allow us to identify tomato genes that have no orthologue (counterpart) in these plants. Information on such tomato genes facilitates elucidation of gene functions and metabolic pathways that are characteristic of tomato and the Solanaceae family.

Information concerning various biological aspects such as functional annotations of genes and gene products, intron-exon structures, gene expressions and metabolic pathways is available from various tomato-related web sites. These public data advance not only functional genomics but also the emerging field of systems biology. Pioneer studies and developments in tomato functional genomics were well covered in previous reviews (e.g., Fray and Grierson 1993; Mysore et al. 2001; Rick 1991; Shibata 2005). Here, we review the current status of tomato databases that provide information on the tomato genome sequencing project, coding sequences (CDS) with complete sequencing of full-length cDNA clones, ESTs, non-redundant sequence sets derived from ESTs, gene expressions, microarray platforms and metabolic pathways. The web sites mentioned in this article are summarized in Table 3.

The international tomato genome sequencing project

The Solanaceae Genome Project Network (SGN) provides information relevant to the tomato genome sequencing project such as linkage maps containing DNA markers, bacterial artificial chromosome (BAC) clones anchored to these linkage maps (called "seed BAC" clones) and BAC sequences with genomic and functional annotations as described below.

Linkage maps with DNA markers

Information on tomato linkage maps containing DNA markers is available from SGN and the National Center for Biotechnology Information (NCBI). SGN provides four maps that were constructed by using segregation populations and inbred lines derived from crossings between tomato cultivars and closely-related wild species. Restriction fragment length polymorphisms (RFLPs), single nucleotide polymorphisms (SNPs), cleaved amplified polymorphic sequence (CAPS) and simple sequence repeats (SSRs) are included in these maps. Currently, the map "Tomato-EXHIR 1997", derived from the interspecific backcross of cultivar TA209 (E6203) with a backcross population (BC) of *S. habrochaites* (formerly *L. hirsutum*) LA1777, contains 134 RFLP markers (Bernacchi and Tanksley 1997). "Tomato-EXPEN 1992", based on an F_2 population from cultivar VF36-Tm2a and *S. pennellii* (formerly *L. pennellii*) LA716, has two CAPS and 919 RFLP markers

Table 3. Major public databases containing biological information of tomato.

Database	Contents
Sequence Databases	
INSD (http://www.insdc.org/)	The International Nucleotide Sequence Databases, maintained by DDBJ, EMBL and GenBank.
DDBJ (http://www.ddbj.nig.ac.jp/)	A nucleotide sequence database at the National Institute of Genetics (NIG) in Japan.
EMBL (http://www.ebi.ac.uk/embl)	A nucleotide sequence database at the European Bioinformatics Institute (EBI) in UK.
GenBank (http://www.ncbi.nlm.nih.gov/Genbank/index.html)	A nucleotide sequence database maintained by the National Center for Biotechnology Information (NCBI) in USA.
dbEST (http://www.ncbi.nlm.nih.gov/dbEST/index.html)	A EST database for each organism in NCBI.
UniGene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene)	A unigene database for each organism in NCBI.
dbGSS (http://www.ncbi.nlm.nih.gov/dbGSS/index.html)	A genome survey sequences (GSS) database for each organism in NCBI.
Tomato Integrated Databases	
The International Tomato Sequencing Project (http://www.sgn.cornell.edu/about/tomato_sequencing.pl)	A web-site of the International Tomato Sequencing Project.
SOL (http://www.sgn.cornell.edu/solanaceae-project/)	A web-site of the International Solanaceae Genome Project.
SGN (http://www.sgn.cornell.edu/index.pl)	Genomic, genetic and taxonomic information for species in the Solanaceae and related families.
TIGR Tomato Gene Index (http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=tomato)	A database for publicly available ESTs and unigenes for tomato.
MiBASE (http://www.kazusa.or.jp/jsol/microtom/)	A database for tomato unigenes with ESTs from Micro-Tom, gene expressions, metabolic pathways, gene ontologies.
TomDB (http://mips.gsf.de/proj/plant/jsf/tomato/index.jsp)	A database for the tomato genome database.
Tomato Full-Length cDNA	
KaFTom (http://www.pgbio.kazusa.or.jp/kaftom/)	A database for ESTs and full-length sequences from tomato full-length cDNA libraries and their annotations.
Chromosome Maps	
Lycopersicon esculentum (tomato) genome view (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=4081)	A database for tomato linkage maps in NCBI.
Gene Expression	
GEO (http://www.ncbi.nlm.nih.gov/geo/)	Providing data from microarray, serial analysis of gene expression (SAGE), and mass spectrometry proteomics.
ArrayExpress (http://www.ebi.ac.uk/Databases/microarray.html)	A database for gene expression data from microarray experiments in EBI.
SGED (http://www.tigr.org/tdb/potato/SGED_index2.shtml)	A database for Solanaceae expression data using potato cDNA microarrays.
TED (http://ted.bti.cornell.edu/)	A tomato microarray data warehouse and databases for tomato microarray expression data and tomato digital expression data.
CGEP (http://bti.cornell.edu/CGEP/CGEP.html)	A web-site of The Center for Gene Expression Profiling (CGEP) for high quality tomato cDNA microarrays.
Metabolite and Metabolic Pathway	
KEGG (http://www.genome.jp/kegg/)	Databases for metabolic pathways, genes, protein families, ligands, drugs, diseages and so on.
LycoCyc (http://solcyc.sgn.cornell.edu/LYCO/server.html)	A tomato metabolic pathway database.
TOMET (http://tomet.bti.cornell.edu)	Tomato Metabolite Database (TOMET) contains data on metabolites such as ascorbate, carotenoids and sugars.
Metabolome Tomato Database (MoTo DB) (http://appliedbioinformatics.wur.nl/moto/)	A metabolite database dedicated to liquid chromatography-mass spectrometry-based metabolomics of tomato fruit.
Genetics Resources	
TGRC (http://tgrc.ucdavis.edu)	Genebank of wild relatives, monogenic mutants and miscellaneous genetic stocks of tomato at Tomato Genetics Resource Center.

Table 3. continued.

Database	Contents
Others	
RAP-DB (http://rapdb.lab.nig.ac.jp/)	A database of The Rice Annotation Project providing access to the rice annotation data.
TAIR (http://www.arabidopsis.org/)	The Arabidopsis Information Resource (TAIR) which releases a database of genetic and molecular biology data for *Arabidopsis thaliana*.
Populus trichocarpa Genome (http://genome.jgi-psf.org/Poptr1/Poptr1.home.html)	A databases for *Populus trichocarpa* genome.
TIGR (http://www.tigr.org/index.shtml)	A website of The Institute for Genomic Research (TIGR) providing information for analyzing genomes.
The NCBI Entrez Genome Project database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj)	A database for complete and incomplete large-scale sequencing projects for cellular organisms.
Gene Ontology (http://www.geneontology.org/index.shtml?all)	A web site of Gene Ontology Consortium.
Plant Ontology (http://www.plantontology.org/index.html)	A web site of Pant Ontology Consortium.

The hyperlinks for the listed web-sites are available at the web page of database links in MiBASE (http://www.kazusa.or.jp/jsol/microtom/).

(Tanksley et al. 1992). "Tomato-EXPEN 2000", based on an F_2 population of *S. lycopersicum* LA925 and *S. pennellii* LA716, includes 699 CAPS, 1342 RFLP, 19 SNP and 156 SSR markers (Fulton et al. 2002). "Tomato-EXPIMP 2001", derived from the cross of cultivar TA209 with BC and backcross recombinant inbred lines (BCRILs) of *S. pimpinellifolium* (formerly *L. pimpinellifolium*) LA1589, includes one CAPS and 143 RFLP markers (Grandillo and Tanksley 1996; Tanksley et al. 1996; Doganlar et al. 2002).

NCBI provides four linkage maps "Genetic map, 1992" with 239 markers, "*L. esculentum* × *L. pennellii*; 1986" with 112 markers (Bernhatzky and Tanksley 1986), "*L. esculentum* × *L. pennellii*; 1992" with 1054 markers (Tanksley et al. 1992) and "*L. esculentum* × *L. pennellii*; 2000" with 126 markers (Fulton et al. 2000).

SGN and NCBI provide map viewers that show relationships between the different linkage maps.

Bacterial artificial chromosomes and physical maps

SGN provides BAC information, including physical maps. For physical mapping and whole-genome sequencing of tomato, the first BAC library was constructed from the tomato cultivar Heinz 1706 using a *Hind*III partial digestion of megabase-size DNA (Budiman et al. 2000). This BAC library represents a 15.0-fold coverage of the tomato haploid genome and contains 129,024 clones, with an average insert size of 117.5 kb. For genome sequencing, 88,642 BACs from this library were used to generate fingerprints at the Arizona Genome Institute (Mueller et al. 2005b; http://www.sgn.cornell.edu/cview/map.pl?map_id=9&physical=1). The fingerprints are used in selecting subsequent BACs for BAC-by-BAC sequencing. To anchor BACs to the linkage map "Tomato-EXPEN 2000" and construct a tomato physical map , DNA fragments from the fingerprinted BACs were hybridized to 1536 overlapping oligonucleotide (called "overgo") probes (Cai et al. 1998) generated from markers mapped on the current high density linkage map. The overgo probes that matched to one or more BACs are referred to as "anchor points". Currently, more than 650 anchor points are available. The anchored BAC clones (called 'seed BAC clones') have been sequenced as the start point of the tomato genome sequencing project.

SGN also provides information on the use of fluorescence in-situ hybridization (FISH) mapping to verify the linkage and physical maps. A FISH map has been constructed by using the pachytene phase chromosomes with labeled BAC probes (Mueller et al. 2005b; http://www.sgn.cornell.edu/cview/map.pl?map_id=13). Currently, 43 BACs (markers) have been linked to the FISH map.

Genomic sequences and annotations

SGN provides up-to-date information on the progress of the genome sequencing project (Fig. 1). Approximately 1500 seed BAC clones will be anchored to the tomato high density genetic map to help guide the genome sequencing (Mueller et al. 2005b).

Information provided by SGN includes the genomic sequences, exons inferred from computational methods, homologous sequences of tomato ESTs, tomato unigenes (described below), *Arabidopsis* protein sequences and potato ESTs. Genomic sequences, BAC end sequences and homologous sequences are graphically visualized with the multiple alignment viewer "Genome browser (Gbrowse)" (Stein et al. 2002). Putative intron-exon stuructures in the tomato genome are also provided by the database KaFTom as described below.

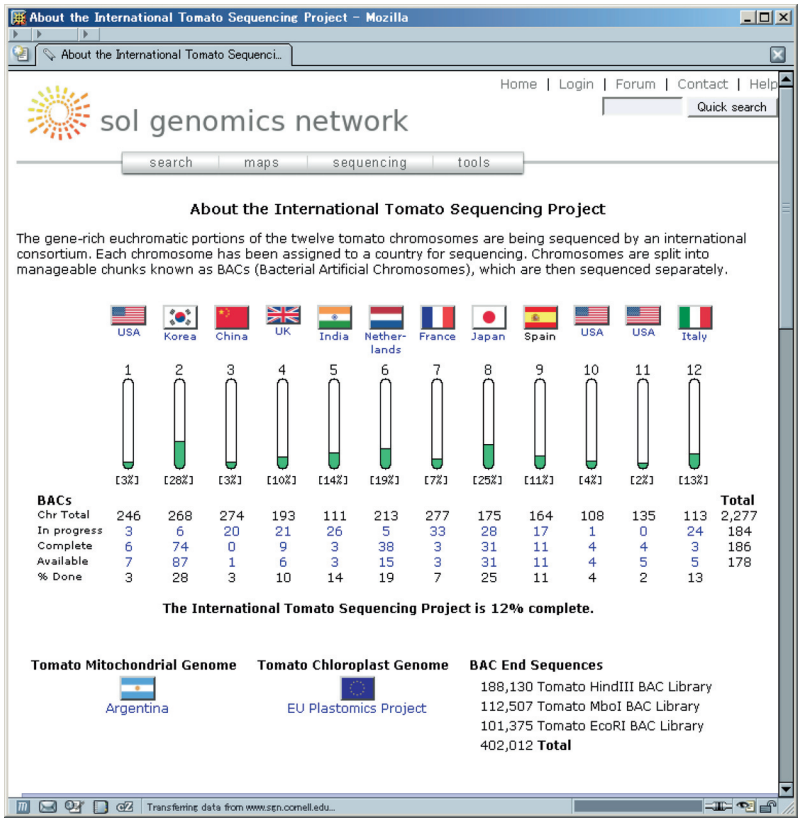

Figure 1. The web page showing the genome sequencing progress of the International Tomato Sequencing Project (September 20, 2006) (http://www.sgn.cornell.edu/about/tomato_sequencing.pl).

Full-length cDNA sequences

Information on tomato full-length cDNA clones is provided by INSD and KaFTom. Full-length cDNA clones are fundamental resources for molecular biology experimental investigations of gene functions, as well as for detection of intron-exon structures in genomes. With the vector-capping protocol developed by Kato et al. (2005), the first tomato full-length cDNA libraries were constructed from the fruit of the miniature tomato cultivar Micro-Tom (Tsugane et al. 2005). Micro-Tom has attracted attention as a model tomato plant for genomic studies 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). Full-length cDNA libraries have been constructed from the fruit at different developmental stages and from pathogen-treated leaves of Micro-Tom.

To date, 57,422 ESTs and 320 draft full-length sequences (high throughput cDNA sequences; HTCs) have been deposited into INSD (September, 2006). Their InterProScan (Zdobnov and Apweiler, 2001) and BLAST (Altschul et al. 1990) annotations are available from KaFTom. KaFTom also provides information on predicted intron-exon structures in the tomato genome. The intron-exon structures are detected from comparisons between tomato HTCs and tomato genomic sequences made by using est2genome (Mott 1997; http://emboss.sourceforge.net/) and BLASTN.

Expressed sequence tags and unigenes

Information on tomato ESTs and a non-redundant sequence set derived from ESTs are provided by INSD and other public databases. ESTs generated from cDNA libraries give us information on transcript sequences and expression patterns in tissues and organs at various developmental stages. Currently, INSD provides information on 201,033 tomato ESTs (Table 1) totalling approximately 100 million bases. By assembling EST sequences, ESTs that appear to come from the same locus produce a consensus sequence, and a non-redundant consensus sequence set is obtained. Availability of non-redundant consensus sequences allows us to use computational approaches such as homology searches and functional annotations to analyze for putative gene functions. Several different notations, such as “unigenes”, “Unigenes”, “UNIGENEs” and “tentative consensus (TC)”, are in use to refer to a non-redundant sequence set.

Information on tomato unigene sequences is provided in the TIGR Tomato Gene Index database (Lee et al. 2005), the Solanaceae Genome Project Network (SGN)

(Mueller et al. 2005b) and the database MiBASE (Yano et al. 2006a).

The current version (Release 11.0 at June 21, 2006) of the TIGR Tomato Gene Index provides information on 41,425 unigene sequences generated by assembling and clustering 215,990 tomato EST sequences, along with corresponding homologous protein sequences, open reading frames (ORFs), gene ontology (GO) terms (Gene Ontology Consortium 2001), SNPs, alternative splicing sequences, cDNA libraries, Enzyme Commision (EC) numbers, names of KEGG metabolic pathways, unique 70-mer oligonucleotide sequences and orthologues in other organisms.

SGN recently updated its unigene information, which includes 34,829 unigene sequences assembled from 239,593 ESTs (version Tomato 200607), along with EST libraries, microarray resources, DNA markers, manual annotations, BLAST annotations based on the non-redundant amino acid sequence database (nr) at NCBI and the *Arabidopsis* protein database at The *Arabidopsis* Information Resource (TAIR) (Rhee et al. 2003), predicted peptide sequences and InterProScan annotations (including domains, GO terms and gene families). DNA markers are searchable with the graphical chromosome maps described above.

MiBASE provides information on 26,363 tomato unigenes (version January, 2005) assembled from 150,581 publicly available ESTs in dbEST (Boguski et al. 1993) and 35,824 Micro-Tom ESTs from fruit and leaves (Yamamoto et al. 2005; Yano et al. 2006a). In MiBASE, BLAST annotations based on the *Arabidopsis* translated protein sequence database in TAIR and the TIGR Gene Indices of rice, soybean (*Glycine max*), maize (*Zea mays*) and tomato have been released, as well as annotations from the NCBI nr database. MiBASE also contains 1935 putative SNPs between Micro-Tom and other inbred tomato lines (cultivars E6203, R11-13, Rio Grande PtoR, R11-12 and a wild relative, *S. pennellii* TA56), obtained by comparing relevant EST sequences. 409 SSRs provided in MiBASE were identified using a Perl script (Temnykh et al. 2001) customized for searching the unigene sequences. The current version of MiBASE contains information on unigenes, DNA markers, EST libraries, BLAST annotations, GO terms, metabolic pathways, and gene expressions obtained from Micro-Tom cDNA arrays (described below).

NCBI has released the database UniGene (Wheeler et al. 2003), which provides accession numbers of ESTs that appear to come from the same locus. In the UniGene database, each unigene (cluster) entry (nucleotide sequences, including ESTs) includes other data such as protein similarities. The current version of UniGene (*Lycopersicon esculentum*: UniGene Build #25) contains 12,847 tomato unigenes constructed from 167,689 EST sequences.

For tomato researchers, it can sometimes be confusing that four different unigene sets of tomato are available in the public domain. The differences between unigene sequences available at the TIGR Tomato Gene Index, SGN and MiBASE are caused by differences in the respective computational methods used for unigene construction. The fact that these methods still need improvement should be kept in mind when using these databases. By contrast, NCBI's UniGene does not provide consensus unigene sequences. Instead, UniGene is updated weekly or monthly, thus providing more recent entry information (accession numbers) for each unigene (cluster). Despite these issues, however, unigene sequences serve as a very useful basis for analyzing ORFs, sequence similarities, and functional domains in protein sequences.

Gene expression data

Gene expression data obtained from tomato cDNA microarray experiments are also available in public databases. The Center for Gene Expression Profiling (CGEP) provides information on the tomato cDNA microarrays TOM1, which were constructed with approximately 12,000 probes (Alba et al. 2004). BLAST annotations and SGN unigene names are provided for each probe. The 5′ and 3′ end sequences of the cDNA clones used to construct TOM1 are searchable in the SGN database. The tomato microarray expression database (TED) (Fei et al. 2006) provides expression data obtained through the use of TOM1. Currently, TED allows users to search expression data from 15 sets of experiments, including annotations, unigene names, sequence similarities and expression patterns. Hierarchical clustering, k-means clustering and self organizing map (SOM) analysis can be also performed against the experimental data in TED. The search function in TED also allows users to obtain information on ascorbate content during fruit development and which probes show similar or inverse expression patterns with reduced ascorbate during development.

MiBASE provides experimental data obtained from cDNA arrays that were constructed from Micro-Tom cDNA clones. From fruit and leaf Micro-Tom cDNA libraries containing 37,972 cDNA clones, Yano et al. (2006b) selected 10,905 cDNA clones as being representative of non-redundant sequences for constructing cDNA arrays. The current version of MiBASE contains BLAST annotations for each probe and the expression data for samples from leaves and fruit. Gene expression data from two sets of experiments can be searched by querying with clone names, annotation keywords, GO terms or expression patterns.

Due to the high level of sequence similarity between potato (*Solanum tuberosum*) and tomato, many samples

from tomato hybridize strongly to potato microarrays. The Solanaceae Gene Expression Database (SGED) provides tomato gene expression data obtained from the potato cDNA microarrays "TIGR Potato cDNA Array". These microarrays contain 15,264 potato cDNA clones. The current version of SGED provides tomato gene expression data from 10 sets of experiments.

The microarray experiment databases GEO (Barrett et al. 2005) at NCBI and ArrayExpress (Parkinson et al. 2005) at European Bioinformatics Institute (EBI) contain tomato gene expression data from six sets of experiments obtained by using different platforms such as "Cornell-CGEP Tomato 13K TOM1" and "Potato 10k cDNA array".

Recently, Affymetrix, Inc. (Affymetrix, Santa Clara, CA) released the GeneChip® Tomato Genome Array, which consists of over 10,000 probes synthesized based on transcripts. *Lycopersicon esculentum* UniGene Build #20 (October 3, 2004) in UniGene was used to design this array.

As described here, several platforms are available for monitoring tomato gene expression. Relational tables for accession numbers in INSD and probes for each platform will facilitate sharing and exchange of experimental data obtained from different platforms.

Metabolite and metabolic pathways

Public databases providing metabolic pathway and metabolite information in tomato are starting to become available. KEGG has released the metabolic pathway database PATHWAY, which includes over 200 metabolic pathways of tomato (Kanehisa et al. 2004). Recently, SGN released a tomato metabolic pathway database LycoCyc, part of the BioCyc collection of databases (Karp et al. 2005). The current version of LycoCyc includes 271 pathways with 833 compounds and 21,546 genes. MiBASE also provides information on tomato putative metabolic pathways, inferred from sequence similarities between tomato and *Arabidopsis* genes, that are related to *Arabidopsis* metabolic reactions listed in the databases AraCyc (Mueller et al. 2003) and KEGG. Tomato metabolite profiles of ascorbate, carotenoids and sugars are collected into the Tomato Metabolite Database (TOMET), which is aimed at helping elucidate correlations between metabolite accumulations and gene expressions. Data from liquid chromatography-mass spectrometry (LC-MS)-based metabolomics of tomato fruit recently became available in the Metabolome Tomato Database (MoTo DB), including retention times and calculated masses of tomato metabolites detected by LC-MS (Moco et al. 2006).

Other databases

Tomato Genetics Resource Center (TGRC) provides information on wild relatives, monogenic mutants and miscellaneous genetic stocks of tomato. Plant ontology (PO) (The Plant Ontology Consortium, 2002) terms, which describe plant structures and growth and developmental stages, are available for *Arabidopsis*, rice and maize. PO annotation efforts in tomato are ongoing (Jaiswal et al. 2005). Munich Information Center for Protein Sequences (MIPS) is currently developing its tomato genome database TomDB. Phenotypes of tomato mutant lines will be provided in a database maintained by the Japan Solanaceae Consortium (Yamazaki Y and Ezura H, personal communication).

References

Alba R, Fei Z, Payton P, Liu Y, Moore SL, Debbie P, Cohn J, D'Ascenzo M, Gordon JS, Rose JKC, Martin G, Tanksley SD, Bouzayen M, Jahn MM, Giovannoni J (2004) ESTs, cDNA microarrays, and gene expression profiling: tools for dissecting plant physiology and development. *Plant-J* 39: 697–714

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

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

Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R (2005) NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic Acids Res* 33: D562–566

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2006) GenBank. *Nucleic Acids Res* 34: D16–20

Bernhatzky R, Tanksley SD (1986) Toward a saturated linkage map in tomato based on isozymes and random cDNA sequences. *Genetics* 112: 887–898

Bernacchi D, Tanksley SD (1997) An interspecific backcross of *Lycopersicon esculentum* × *L. hirsutum*: Linkage analysis and a QTL study of sexual compatibility factors and floral traits. *Genetics* 147: 861–877

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

Bramley PM (2002) Regulation of carotenoid formation during tomato fruit ripening and development. *J Exp Bot* 53: 2107–2113

Brunak S, Danchin A, Hattori M, Nakamura H, Shinozaki K, Matise T, Preuss D (2002) Nucleotide Sequence Database Policies. *Science* 298: 1333

Budiman MA, Mao L, Wood TC, Wing RA (2000) A deep-coverage tomato BAC library and prospects toward development of an STC framework for genome sequencing. *Genome Res* 10: 129–136

Cai WW, Reneker J, Chow CW, Vaishnav M, Bradley A (1998) An anchored framework BAC map of mouse chromosome 11 assembled using multiplex oligonucleotide hybridization. *Genomics* 54: 387–397

Cochrane G, Aldebert P, Althorpe N, Andersson M, Baker W, Baldwin A, Bates K, Bhattacharyya S, Browne P, van den Broek A, Castro M, Duggan K, Eberhardt R, Faruque N, Gamble J, Kanz C, Kulikova T, Lee C, Leinonen R, Lin Q, Lombard V, Lopez R, McHale M, McWilliam H, Mukherjee G, Nardone F, Pastor MP, Sobhany S, Stoehr P, Tzouvara K, Vaughan R, Wu D, Zhu W, Apweiler R (2006) EMBL Nucleotide Sequence Database: developments in 2005. *Nucleic Acids Res* 34: D10–15

Doganlar S, Frary A, Ku HM, Tanksley SD (2002) Mapping Quantitative Trait Loci in Inbred Backcross Lines of *Lycopersicon pimpinellifolium* (LA1589). *Genome* 45: 1189–1202

Fei Z, Tang X, Alba R, Giovannoni J (2006) Tomato Expression Database (TED): a suite of data presentation and analysis tools. *Nucleic Acids Res* 34: D766–770

Fray RG, Grierson D (1993) Molecular genetics of tomato fruit ripening. *Trends Genet* 22: 281–300

Fulton T, van der Hoeven R, Eannetta N, Tanksley S (2002). Identification, Analysis and Utilization of a Conserved Ortholog Set (COS) Markers for Comparative Genomics in Higher Plants. *Plant Cell* 14: 1457–1467

The Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res* 11: 1425–1433

Gorguet B, van Heusden AW, Lindhout P (2005) Parthenocarpic Fruit Development in Tomato. *Plant Biol* (Stuttg) 7: 131–139

Grandillo S, Tanksley SD (1996) QTL analysis of horticultural traits differentiating the cultivated tomato from the closely related species *Lycopersicon pimpinellifolium*. *Theor Appl Genet* 92: 935–951

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

Jaiswal P, Avraham S, Ilic K, Kellogg EA, McCouch S, Pujar A, Reiser L, Rhee SY, Sachs MM, Schaeffer M, Stein L, Stevens P, Vincent L, Ware D, Zapata F (2005) Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages. *Comp-Funct-Genom* 6: 338–397

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

Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahren D, Tsoka S, Darzentas N, Kunin V, Lopez-Bigas N (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res* 33: 6083–6089

Kato S, Ohtoko K, Ohtake H, Kimura T (2005) Vector-capping: a simple method for preparing a high-quality full-length cDNA library. *DNA Res* 12: 53–62

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. *Plant-J* 12: 456–1472

Moco S, Bino RJ, Vorst O, Verhoeven HA, de Groot J, van Beek TA, Vervoort J, de Vos CH (2006) A liquid chromatography-mass spectrometry-based metabolome database for tomato. *Plant Physiol* 141: 1205–1218

Mott R (1997) EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput Appl Biosci* 13: 477–478

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

Mueller LA, Tanksley SD, Giovannoni JJ, Van Eck J, Stack S, Choi D, Kim BD, Chen M, Cheng Z, Li C, Ling H, Xue Y, Seymour G, Bishop G, Bryan G, Sharma R, Khurana J, Tyagi A, Chattopadhyay D, Singh NK, Stiekema W, Lindhout P, Jesse T, Lankhorst RK, Bouzayen M, Shibata D, Tabata S, Granell A, Botella MA, Giuliano G, Frusciante L, Causse M, Zamir D (2005a) The Tomato Sequencing Project, the first cornerstone of the International Solanaceae Project (SOL) *Comp-Funct-Genom* 6: 153–158

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 (2005b) The SOL Genomics Network: a comparative resource for Solanaceae biology and beyond. *Plant Physiol* 138: 1310–1317

Mysore KS, Tuori RP, Martin GB (2001) *Arabidopsis* genome sequence as a tool for functional genomics in tomato. *Genome Biol* 2: REVIEWS1003

Okubo K, Sugawara H, Gojobori T, Tateno Y (2006) DDBJ in preparation for overview of research activities behind data submissions. *Nucleic Acids Res* 34: D6–9

Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, Coulson R, Farne A, Lara GG, Holloway E, Kapushesky M, Lilja P, Mukherjee G, Oezcimen A, Rayner T, Rocca-Serra P, Sharma A, Sansone S, Brazma A (2005) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* 33: D553–D555

Pedley KF, Martin GB (2003) Molecular basis of Pto-mediated resistance to bacterial speck disease in tomato. *Annu Rev Phytopathol* 41: 215–243

The Plant Ontology Consortium (2002) The Plant Ontology Consortium and Plant Ontologies. *Comp Funct Genom* 3: 137–142

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

Rick CM (1991) Tomato paste: a concentrated review of genetic highlights from the beginnings to the advent of molecular genetics. *Genetics* 128: 1–5

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

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S (2002) The generic genome browser: a building block for a model organism system database. *Genome Res* 12: 1599–1610

Tanksley SD, Ganal MW, Prince JP, de Vicente MC, Bonierbale MW, Broun P, Fulton TM, Giovannoni JJ, Grandillo S, Martin GB, Messeguer R, Miller JC, Miller L, Paterson AH, Pineda O, Roder MS, Wing RA, Wu W, Young ND (1992) High density molecular linkage maps of the tomato and potato genomes. *Genetics* 132: 1141–1160

Tanksley SD, Grandillo S, Fulton TM, Zamir D, Eshed Y, Petiard V, Lopez J, Beck-Bunn T (1996) Advanced backcross QTL analysis in a cross between an elite processing line of tomato

and its wild relative *L. pimpinellifolium*. *Theor Appl Genet* 92: 213–224

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 Biotechnol* 22: 161–165

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Dejardin A, Depamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjarvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leple JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouze P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y, Rokhsar D (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313: 1596–1604

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

Yano K, Watanabe M, Yamamoto N, Tsugane T, Aoki K, Sakurai M, Shibata D (2006a) MiBASE: A database of a miniature tomato cultivar Micro-Tom. *Plant Biotechnol* 23: 195–198

Yano K, Tsugane T, Watanabe M, Maeda F, Aoki K, Shibata D (2006b) Non-biased distribution of the tomato genes that have no counterpart in *Arabidopsis thaliana* in expression patterns during fruit maturation. *Plant Biotechnol* 23: 199–202

Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, Wagner L (2003) Database Resources of the National Center for Biotechnology. *Nucleic Acids Res* 31: 28–33

Zdobnov EM, Apweiler R (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17: 847–848