Chapter 3
Databases for Solanaceae and Cucurbitaceae Research

Masaaki Kobayashi, Hajime Ohyanagi, and Kentaro Yano

3.1 Outline of the Omics Databases in Plants

Currently, the vast majority of plant science information is available via the Internet. It covers publications, experimental resources, protocols and results, analysis tools, and so on. In this section, major databases for general plant sciences will be briefly introduced.

The primary sequence data are provided by the International Nucleotide Sequence Databases (INSD) (Nakamura et al. 2013) and Universal Protein Resource (UniProt) (The UniProt Consortium 2014). The INSD, maintained by the DNA Data Bank of Japan (DDBJ) (Kosuge et al. 2014), the European Nucleotide Archive (ENA) (Pakseresht et al. 2014), and GenBank (Benson et al. 2014), stores nucleotide sequence data. UniProt provides high-quality information on protein sequences and their biological functions. Through the INSD and UniProt, sequence data for Solanaceae and Cucurbitaceae crops are also accessible.

M. Kobayashi • K. Yano (✉)
Bioinformatics Laboratory, School of Agriculture, Meiji University, Kawasaki, Japan
e-mail: kyano@meiji.ac.jp

H. Ohyanagi
Bioinformatics Laboratory, School of Agriculture, Meiji University, Kawasaki, Japan

Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia

H. Ezura et al. (eds.), *Functional Genomics and Biotechnology in Solanaceae and Cucurbitaceae Crops*, Biotechnology in Agriculture and Forestry 70,
DOI 10.1007/978-3-662-48535-4_3

3.1.1 Information and Tools from NCBI

On the website of the National Center for Biotechnology Information (NCBI) (NCBI Resource Coordinators 2014), sequence and annotation data stored in the INSD are easily accessible. From the sequence databases nucleotide, EST, and protein in NCBI, which can be specified with a pull-down selector on the top menu, each sequence entity (record, entry) can be retrieved by a text-based keyword search. For a keyword search, each phrase should be in quotes (e.g., "Solanum lycopersicum"); a field-specific search function is also available. For example, sequences of tomato can be retrieved by the query "Solanum lycopersicum" [ORGN]. NCBI provides useful field search functions, such as [ORGN] for the source of the sequence records, [ECNC] for the Enzyme Commission (EC) number (Webb 1992), and [GENE] for the gene name (see http://www.ncbi.nlm.nih.gov/books/NBK49540/).

NCBI's database of expressed sequence tags (dbESTs) (Boguski et al. 1993) and UniGene database (UniGene) (Wheeler et al. 2003) provide information on expressed sequence tags (ESTs) from a broad variety of organisms. The UniGene database contains information on protein similarities, gene expression, cDNA clones, and genomic locations with a list of accession numbers of ESTs. The UniGene database is regularly updated on a weekly or a monthly basis.

Besides sequence data, NCBI also provides information on genomes and genes. The genome database provides information on genome sequencing and annotation projects. For example, a genome map viewer and information on genome sequencing projects for tomato can be browsed by the query "Solanum lycopersicum" [ORGN]. With the gene database, information on genomic structure, positions on the genome, and literature is provided.

The Basic Local Alignment Search Tool (BLAST) algorithm (Boratyn et al. 2013) assists in retrieving sequences homologous to a query sequence. By taking advantage of the web interfaces for a BLAST search, sequences homologous to a nucleotide or protein query sequence in the database can be comprehensively identified. In the case of the NCBI web tool for BLAST, a single database should be specified from the preset databases. In the case of a stand-alone BLAST search, by installing the program distributed from NCBI's FTP site, a custom sequence database, for example, a sequence database for transcription factors in tomato, can be constructed with a set of arbitrary sequences on each personal computer or server.

3.1.2 The Gene Index Databases

The Dana-Farber Cancer Institute (DFCI) maintains a gene index database (http://compbio.dfci.harvard.edu/tgi/) to provide comprehensive information on expressed genes in animals, plants, protists, and fungi. It contains a nonredundant consensus sequence set, called a tentative consensus (TC), generated by assembling and clustering methods (Lee et al. 2005). In the database, information on variants and functional and structural annotations is also available. The database stores

information on homologous protein sequences, open reading frames (ORFs), gene ontology (GO) terms, single nucleotide polymorphisms (SNPs), alternative splicing sequences, cDNA libraries, EC numbers of the International Union of Biochemistry and Molecular Biology, Kyoto Encyclopedia of Genes and Genomes (KEGGs) metabolic pathways (Kanehisa et al. 2014), unique 70-mer oligonucleotide sequences, and orthologs in other organisms. The current version of gene indices provides omics information on plants such as tomato, *Nicotiana benthamiana*, tobacco, eggplant, potato, pepper, petunia, and coffee.

3.1.3 Sequence Data with Next-Generation Sequencing Technologies

Genomic DNA and cDNA sequencing data obtained from next-generation sequencing (NGS) technologies have been rapidly accumulated in the public databases (Wheeler et al. 2008), i.e., the NCBI Sequence Read Archive (SRA) (NCBI Resource Coordinators 2014), the ENA of The European Bioinformatics Institute—Part of the European Molecular Biology Laboratory (EMBL-EBI) (Pakseresht et al. 2014), and the DDBJ Sequence Read Archive (DRA) (Kosuge et al. 2014). Genomic DNA sequence data can be employed to determine genome sequences by de novo assembly or to identify DNA polymorphisms including SNPs and simple sequence repeats (SSRs) by reference mapping methodologies. With cDNA or mRNA sequencing data (RNA-Seq data) generated under multiple experimental conditions, gene expression profiles across these conditions can be obtained. To obtain the gene expression profiles, genome sequences or unigene sequences (a nonredundant sequence set of transcripts derived from ESTs and draft full-length sequences) are required as reference sequences.

3.2 Genome and Transcriptome Data

The families Solanaceae (nightshade-related species) and Cucurbitaceae include numbers of agronomically and economically significant flowering plants. Among the Solanaceae, tomato (*Solanum lycopersicum*) is one of the most important agricultural crops in human culture and history. Together with the Brassicaceae and Fabaceae families, the Solanaceae family has been widely used for evolutionary analysis. The family Cucurbitaceae, sometimes called the gourd family, consists of over a hundred genera. It includes cucumber, watermelon, melon, and pumpkin, as well as other food plants.

To date, Solanaceae and Cucurbitaceae plants have been employed as model plants. Major databases are being maintained to provide omics information on these model plants. The databases maintained by genome sequencing projects generally provide information on genome and transcriptome data, since both types of data are often required for structural and functional annotation of the genome. In this

Table 3.1 Web databases for Solanaceae and Cucurbitaceae

Name of database	URL
International Nucleotide Sequence Databases Collaboration (INSDC)	http://insdc.org/
Universal Protein Resource (UniProt)	http://www.uniprot.org/
DNA Data Bank of Japan (DDBJ)	http://www.ddbj.nig.ac.jp/
European Nucleotide Archive (ENA)	http://www.ebi.ac.uk/ena/
GenBank	http://www.ncbi.nlm.nih.gov/genbank/
National Center for Biotechnology Information (NCBI)	http://www.ncbi.nlm.nih.gov/
NCBI's database of expressed sequence tags (dbESTs)	http://www.ncbi.nlm.nih.gov/dbEST/
UniGene database (UniGene)	http://www.ncbi.nlm.nih.gov/unigene
Basic Local Alignment Search Tool (BLAST)	http://blast.ncbi.nlm.nih.gov/Blast.cgi
Dana-Farber Cancer Institute (DFCI) Gene Index Project	http://compbio.dfci.harvard.edu/tgi/
Kyoto Encyclopedia of Genes and Genomes (KEGGs)	http://www.genome.jp/kegg/pathway.html
NCBI Sequence Read Archive (SRA)	http://www.ncbi.nlm.nih.gov/sra
DDBJ Sequence Read Archive (DRA)	http://trace.ddbj.nig.ac.jp/dra/index.html
Sol Genomics Network (SGN)	http://solgenomics.net/
TOMATOMICS	http://bioinf.mind.meiji.ac.jp/tomatomics/
MiBASE	http://www.pgb.kazusa.or.jp/mibase/
KaFTom	http://www.pgb.kazusa.or.jp/kaftom/
Solanum lycopersicum project in PGSB	http://pgsb.helmholtz-muenchen.de/plant/tomato/
National BioResource Project (NBRP) TOMATO	http://tomato.nbrp.jp/indexEn.html
Potato Genome Sequence Consortium (PGSC)	http://www.potatogenome.net/index.php/Main_Page
CuGenDB	http://www.icugi.org/cgi-bin/ICuGI/index.cgi
MELONOMICS	https://www.melonomics.net/
MeloGene	http://www.melogene.net/
CucurbiGene	http://www.cucurbigene.net/
Plant Metabolic Network (PMN)	http://www.plantcyc.org/
SolCyc	http://solcyc.solgenomics.net/
KaPPA-View4	http://kpv.kazusa.or.jp/
Platform for RIKEN Metabolomics (PRIMe)	http://prime.psc.riken.jp/
KNApSAcK	http://kanaya.naist.jp/KNApSAcK/
KNApSAcK core system	http://kanaya.naist.jp/knapsack_jsp/top.html
SHared Information of GENetic resources (SHIGEN) project	http://www.shigen.nig.ac.jp/indexja.htm
Tomato Genetic Resource Center (TGRC)	http://tgrc.ucdavis.edu/index.aspx
National BioResource Project (NBRP)	http://www.nbrp.jp/
PhylomeDB	http://phylomedb.org/

(continued)

Table 3.1 (continued)

Name of database	URL
The Arabidopsis Information Resource (TAIR)	http://www.arabidopsis.org/
The Rice Annotation Project Database (RAP-DB)	http://rapdb.dna.affrc.go.jp/
MSU Rice Genome Annotation Project Database	http://rice.plantbiology.msu.edu/
InParanoid	http://inparanoid.sbc.su.se/cgi-bin/index.cgi
OrthoMCL	http://orthomcl.org/orthomcl/
InterProScan	http://www.ebi.ac.uk/interpro/search/sequence-search
UniProt Knowledgebase (UniProtKB)	http://www.uniprot.org/uniprot/
Plant Omics Data Center (PODC)	http://bioinf.mind.meiji.ac.jp/podc/

section, leading databases providing large-scale genome and transcriptome data in tomato, potato, tobacco, cucumber, melon, and other species are briefly introduced (Table 3.1).

3.2.1 Tomato

The Sol Genomics Network (SGN) (http://solgenomics.net), funded by the NSF and USDA CSREES and hosted at the Boyce Thompson Institute of Cornell University, NY, USA (Bombarely et al. 2011), represents one of the major versatile Solanaceae databases. It stores and serves complete genomic DNA sequences of potato (Potato Genome Sequencing Consortium 2011) and cultivated tomato (Tomato Genome Consortium 2012). It also archives the draft genome DNA sequences of wild tomato, *Nicotiana benthamiana*, and other species. Information on these genomes and annotations is available via the GBrowse genome browser and FTP server for the database. The SGN also covers resources for the SOL-100 project (a comprehensive genome sequencing project) and other Solanaceae genome and annotation information. Transcriptome data such as microarray data, ESTs, cDNA clones, and unigenes for tomato, potato, pepper, *Nicotiana* species, petunia, and coffee are accessible from the SGN website. Each record (entry in the database) has been integrated and assigned internal SGN identifiers.

The TOMATOMICS database (http://bioinf.mind.meiji.ac.jp/tomatomics/), now open to the public, is an omics database designed especially for tomato. Aside from genome sequences and annotations, it integrates various types of biological information on tomato such as ESTs, nucleotide variant information (SNPs and insertion/deletions or InDels) among inbred lines, DNA markers, microarray data, gene expression networks, and metabolic pathways. It has been constructed and maintained by Meiji University, Japan. TOMATOMICS contains information on ESTs, draft full-length sequences (high-throughput cDNA sequences or HTCs), and unigenes (a nonredundant sequence set derived from ESTs and HTCs) provided from the MiBASE (Yano et al. 2006c) and KaFTom databases (Aoki et al. 2010). In this database, unigenes are called Kazusa tomato unigenes (KTUs). The current

version of the database provides information on KTU version 4 (KTU4). The ESTs and HTCs were generated from the "Micro-Tom" model plant (Aoki et al. 2010; Tsugane et al. 2005; Yamamoto et al. 2005). SNPs/InDels stored in TOMATOMICS were derived from analyses of ESTs and genomes. With EST analysis, SNPs in transcripts were detected among inbred lines (Yano et al. 2006c). By using NGS technology, over 1.2 million SNP sites between the Heinz 1706 and Micro-Tom genomes were detected (Kobayashi et al. 2014). In TOMATOMICS, users can browse comprehensive SNP information with the GBrowse. For the genome sequence of Micro-Tom, TOMATOMICS also provides BAC-end sequence data via GBrowse (Asamizu et al. 2012). The sequences and functional annotation data for Micro-Tom assist in designing effective strategies for elucidating of molecular mechanisms involved in each trait and biological process by using experimental resources (mutant lines and cDNA clones) from Micro-Tom. The Micro-Tom experimental resources are distributed by the National BioResource Project (NBRP) of Japan (http://tomato.nbrp.jp/indexEn.html) (Yamazaki et al. 2010) (see Sect. 3.3).

In addition, the Solanum lycopersicum project of PGSB (http://pgsb.helmholtz-muenchen.de/plant/tomato/) maintains its own integrated database for tomato.

3.2.2 Potato

The potato (*Solanum tuberosum*) genome was sequenced by the Potato Genome Sequencing Consortium (PGSC), an international group of scientists from 14 countries, in 2011 (Potato Genome Sequencing Consortium 2011). Information on the genome sequence, including annotations, has been served primarily by the consortium's website (http://www.potatogenome.net/index.php/Main_Page). The SGN mirrors and integrates the potato information and provides it on their website.

3.2.3 Tobacco

Regarding tobacco (genus *Nicotiana*), a couple of draft genome sequence datasets have been provided. Genome sequence data on the experimental model plant *Nicotiana benthamiana* are available from SGN (Bombarely et al. 2011).

3.2.4 Other Species

The Cucurbit Genomics Database (CuGenDB) (http://www.icugi.org/cgi-bin/ICuGI/index.cgi) is an integrative database of the Cucurbitaceae. From the CuGenDB, information on the genomes of cucumber (Huang et al. 2009) and watermelon (Guo et al. 2013) is available. Beijing Genomics Institute (BGI) also

examined and published genome resequencing data on a broad variety of cucumber lines in late 2013 (Qi et al. 2013). The CuGenDB contains a massive amount of cucumber genome information. The genomic sequences of newly sequenced cultivars are also available in CuGenDB or its associated website (http://cmb.bnu.edu.cn/Cucumis_sativus_v20/resequence/). The draft genome sequences of an elite Chinese watermelon inbred line were published in 2013 (Guo et al. 2013). Genomic data and genome resequencing information on 20 watermelon accessions are available in the CuGenDB.

The CuGenDB also provides information such as annotations, ESTs, pathways, nucleotide variants, SSRs, and genetic maps. ESTs stored in CuGenDB were generated from melon, cucumber, watermelon, and *Cucurbita pepo* (Ando and Grumet 2010; Blanca et al. 2011a, b; Clepet et al. 2011; Guo et al. 2010, 2011; Levi et al. 2006). Data on these ESTs and unigenes are downloadable from the database. BLAST searches against the ESTs and CDSs can be performed. In the current version of CuGenDB, information on SSR markers for melon and pathway data for melon, cucumber, watermelon, and *Cucurbita pepo* are accessible.

Moreover, the CucurbiGene database serves transcriptome data from *Cucurbita pepo* derived from de novo assembly of next-generation sequences (Blanca et al. 2011a).

The genome sequence, physical map, annotations, and transcriptome sequences of melon are available from the MELONOMICS (https://www.melonomics.net/) (Blanca et al. 2011b; Garcia-Mas et al. 2012; González et al. 2010). Additionally, the MeloGene (http://www.melogene.net/) offers GBrowse interface including unigenes derived from EST assembly, SSRs, and SNPs among several botanical varieties obtained by mapping their next-generation sequences against genome (Blanca et al. 2011b, 2012; Esteras et al. 2013; Gonzalez-Ibeas et al. 2007).

3.3 Data on Metabolic Pathways and Compounds

Many metabolic pathway databases are provided from the Plant Metabolic Network (PMN) (http://www.plantcyc.org). These databases contain information on genes, enzymes, compounds, reactions, and pathways involved in primary and secondary metabolism. SolCyc (http://solcyc.solgenomics.net), one of the databases maintained by the PMN, provides metabolic data in the Solanaceae, including tomato (LycoCyc), pepper (CapCyc), petunia (PetCyc), potato (PotatoCyc), and tobacco (TobaccoCyc). The information on biochemical pathways for melon (MelonCyc), cucumber (CucCyc), watermelon (WmnCyc), and *Cucurbita pepo* is available from CuGenDB.

KEGG stores information on the molecular interaction and reaction networks involved in metabolic pathways (Ogata et al. 1998). In KEGG, metabolites and ligands can be searched by keyword. The web page containing the retrieved data shows graphical pathway maps.

Information on tomato metabolic pathways is also available from the KaPPA-View4 database (Sakurai et al. 2011). In KaPPA-View4, not only a browser for metabolic pathways is available but also an analysis function for custom gene expression data. In addition, in the viewer for the metabolic pathways in KaPPA-View4, gene expression data from the MiBASE and TOMATOMICS databases are shown. Using the microarray search function in TOMATOMICS, microarray expression data can be retrieved with hyperlinks to KaPPA-View4. With these hyperlinks, upregulated and downregulated genes and correlations in expression are graphically shown in the KaPPA-View4 viewer.

Metabolic data for tomato and cucumber are freely available from public databases. Spectral data of metabolites measured by nuclear magnetic resonance (NMR), gas chromatography/mass spectrometry (GC/MS), liquid chromatography/mass spectrometry (LC/MS), and capillary electrophoresis/mass spectrometry (CE/MS) are accessible from the Platform for RIKEN Metabolomics (PRIMe) (http://prime.psc.riken.jp/). In the PRIMe, spectral data can be searched for by keywords for compound name, PubChem ID, KEGG ID, and chemical formula. The KNApSAcK database (http://kanaya.naist.jp/KNApSAcK/) stores multiple species-wide metabolite data (Nakamura et al. 2014). Metabolite data are available through the keyword search function by names of organisms including tomato, potato, pepper, petunia, tobacco, melon, cucumber, and *Cucurbita* in the KNApSAcK core system (http://kanaya.naist.jp/knapsack_jsp/top.html).

3.4 Information on Experimental Resources

An infrastructure of experimental resources such as seeds and DNA clones facilitates more efficient research strategies. The SHared Information of GENetic resources (SHIGEN) project (http://www.shigen.nig.ac.jp) provides information on experimental materials and databases for various organisms including tomato. The experimental materials (i.e., frozen embryos, plant seeds, cultured cells, DNA clones, live animal stocks, etc.) in SHIGEN are available upon request. The Tomato Genetic Resource Center (TGRC) (http://tgrc.ucdavis.edu) also provides information on tomato genetic resources, including wild relatives.

Information on DNA resources in tomato is accessible through web databases. The information pages for BAC and EST clones in the SGN (Bombarely et al. 2011) contain hyperlinks to order the clones when they are freely available. The NBRP (Yamazaki et al. 2010) in Japan has also established a bioresource infrastructure for many model organisms, including tomato. The tomato NBRP has enhanced the research infrastructure with mutant collections and DNA resources generated from Micro-Tom. Information on the mutant lines is accessible from the TOMATOMA database (Saito et al. 2011). Information on cDNA clones containing full-length cDNA is available from the TOMATOMICS database (Kobayashi et al. 2014).

3.5 Omics Information and Biological Knowledge on Specific Characters in the Solanaceae and Cucurbitaceae

A wealth of omics information allows comprehensive analysis of the genome, transcriptome, metabolome, and other omes. A large-scale analysis, especially comparison of omics information among inbred lines and/or species, provides clues to understand the molecular mechanisms behind specific traits in the Solanaceae and Cucurbitaceae. Comparisons with model plants such as *Arabidopsis* and rice have been widely performed to find species-specific characters.

For example, a set of expressed genes in tomato that have no counterpart in *Arabidopsis* has been identified (Yano et al. 2006b). Omics information in *Arabidopsis* and rice is available from databases such as the Arabidopsis Information Resource (TAIR) (Lamesch et al. 2012), the Rice Annotation Project Database (RAP-DB) (Sakai et al. 2013), and the MSU Rice Genome Annotation Project Database (Ouyang et al. 2007). Ortholog information is useful to compare genes and proteins among species. Information on orthologs is provided in the InParanoid (Ostlund et al. 2010) and OrthoMCL databases (Chen et al. 2006). In terms of orthology determination, the protein phylogenic trees provided by the database PhylomeDB (http://phylomedb.org/) are significant (Huerta-Cepas et al. 2014). With the interface for the phylogenic trees, protein domains are graphically shown along with their lineage relationship. The current version of the PhylomeDB offers phylogenetic trees for melon, cucumber, cacao, and other 20 model plants.

Gene expression networks (GENs) permit genome-wide views of the similarities of gene expression profiles. In GENs, nodes represent genes, and two nodes with significantly similar expression profiles are connected by an edge. Therefore, a gene set with similar expression profiles (a gene module) can be simultaneously identified from a graphical viewer of GENs. The genes with similar expression profiles can be used as candidates for being involved in the same biological process. With accumulating data on genome sequences and transcriptomes (gene expression) in plants, GEN analyses have been widely used to discover genes and elucidate their biological functions.

The construction of GENs from large-scale expression data becomes difficult while GENs demonstrate substantial usefulness. This is mainly due to the current method for assessment of similarities in expression profiles by Pearson correlation coefficients (PCCs). With the increments in the numbers of genes and samples (experimental conditions) for the analysis with PCC, CPU resources (CPUs, memory, and calculation time) required for the analysis are drastically increased to compute PCCs for gene and/or sample pairs. While the accumulation of data promises to improve the accuracy of the computational analysis, the calculation becomes difficult with large-scale datasets. To handle large-scale omics data, such as NGS and microarray data, with the personal computers or workstations generally

used in laboratories, new bioinformatics and statistical approaches should be developed.

To date, one approach to quickly mine genes with similar expression profiles from large-scale transcriptome data has been introduced (Yano et al. 2006a; Hamada et al. 2011; Manickavelu et al. 2012). The method allows detection of genes with similar expression profiles and construction of GENs even with the use of personal computers. The method can be performed by GUI software distributed by Meiji University (http://bioinf.mind.meiji.ac.jp/lab/index.php?catid=15&blogid=1).

Comparisons of GENs among plant species uncover new biological knowledge. Comparative analysis provides information on species-specific genes and gene modules controlling specific traits in each species. Besides identifying similarities of gene expression profiles, accurate functional annotations also facilitate discovery of genes and elucidation of their biological functions. The current annotations provided from most databases have been assigned based on sequence similarity analysis. Sequence similarity searches by programs such as BLAST and InterProScan (Quevillon et al. 2005) permit genome-wide analysis to quickly predict biological functions of genes. On the other hand, accurate annotations are difficult to be assigned, based on such high-throughput homology search methods. To resolve this issue and provide accurate functional annotations for proteins, the UniProt Knowledgebase (UniProtKB) (The UniProt Consortium 2014) has collected high-quality information based on manual annotations with the literature and curator-evaluated computational analysis. Manual annotation is a time-consuming and labor-intensive procedure requiring expert knowledge and training, but these more sophisticated descriptions of the functions and interactions of genes, gene products, and biological conditions will undoubtedly more effectively reveal complicated biological mechanisms and behaviors. Combined with an analysis tool for comparison of GENs among species including tomato, the Plant Omics Data Center (PODC) database (http://bioinf.mind.meiji.ac.jp/podc/) is being maintained to provide information on both automated and manual annotations of genes (Ohyanagi et al. 2015). Such web services with graphical and intuitive interfaces enable both understanding and making use of species-specific characters in the Solanaceae and Cucurbitaceae.

References

Ando K, Grumet R (2010) Transcriptional profiling of rapidly growing cucumber fruit by 454-pyrosequencing analysis. J Am Soc Hortic Sci 135:291–302

Aoki K, Yano K, Suzuki A et al (2010) Large-scale analysis of full-length cDNAs from the tomato (*Solanum lycopersicum*) cultivar Micro-Tom, a reference system for the Solanaceae genomics. BMC Genomics 11:210

Asamizu E, Shirasawa K, Hirakawa H et al (2012) Mapping of Micro-Tom BAC-end sequences to the reference tomato genome reveals possible genome rearrangements and polymorphisms. Int J Plant Genomics 2012:437026

Benson DA, Clark K, Karsch-Mizrachi I et al (2014) GenBank. Nucleic Acids Res 42:D32–D37

Blanca J, Cañizares J, Roig C et al (2011a) Transcriptome characterization and high throughput SSRs and SNPs discovery in *Cucurbita pepo* (Cucurbitaceae). BMC Genomics 12:104

Blanca J, Cañizares J, Ziarsolo P et al (2011b) Melon transcriptome characterization: simple sequence repeats and single nucleotide polymorphisms discovery for high throughput genotyping across the species. Plant Genome 4:118–131

Blanca J, Esteras C, Ziarsolo P et al (2012) Transcriptome sequencing for SNP discovery across Cucumis melo. BMC Genomics 13:280

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

Bombarely A, Menda N, Tecle IY et al (2011) The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl. Nucleic Acids Res 39:D1149–D1155

Boratyn GM, Camacho C, Cooper PS et al (2013) BLAST: a more efficient report with usability improvements. Nucleic Acids Res 41:W29–W33

Chen F, Mackey AJ, Stoeckert CJ Jr et al (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res 34:D363–D368

Clepet C, Joobeur T, Zheng Y et al (2011) Analysis of expressed sequence tags generated from full-length enriched cDNA libraries of melon. BMC Genomics 12:252

Esteras C, Formisano G, Roig C et al (2013) SNP genotyping in melons: genetic variation, population structure, and linkage disequilibrium. Theor Appl Genet 126(5):1285–1303

Garcia-Mas J, Benjak A, Sanseverino W et al (2012) The genome of melon (Cucumis melo L.). Proc Natl Acad Sci U S A 109(29):11872–11877

González VM, Garcia-Mas J, Arús P et al (2010) Generation of a BAC-based physical map of the melon genome. BMC Genomics 11:339

Gonzalez-Ibeas D, Blanca J, Roig C et al (2007) MELOGEN: an EST database for melon functional genomics. BMC Genomics 8:306

Guo S, Zheng Y, Joung JG et al (2010) Transcriptome sequencing and comparative analysis of cucumber flowers with different sex types. BMC Genomics 11:384

Guo S, Liu J, Zheng Y et al (2011) Characterization of transcriptome dynamics during watermelon fruit development: sequencing, assembly, annotation and gene expression profiles. BMC Genomics 12:454

Guo S, Zhang J, Sun H et al (2013) The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. Nat Genet 45:51–58

Hamada K, Hongo K, Suwabe K et al (2011) OryzaExpress: an integrated database of gene expression networks and omics annotations in rice. Plant Cell Physiol 52:220–229

Huang S, Li R, Zhang Z et al (2009) The genome of the cucumber, *Cucumis sativus* L. Nat Genet 41:1275–1281

Huerta-Cepas J, Capella-Gutiérrez S, Pryszcz LP et al (2014) PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res 42:D897–D902

Kanehisa M, Goto S, Sato Y et al (2014) Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 42:D199–D205

Kobayashi M, Nagasaki H, Garcia V et al (2014) Genome-wide analysis of intraspecific DNA polymorphism in 'Micro-Tom', a model cultivar of tomato (*Solanum lycopersicum*). Plant Cell Physiol 55:445–454

Kosuge T, Mashima J, Kodama Y et al (2014) DDBJ progress report: a new submission system for leading to a correct annotation. Nucleic Acids Res 42:D44–D49

Lamesch P, Berardini TZ, Li D et al (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res 40:D1202–D1210

Lee Y, Tsai J, Sunkara S et al (2005) The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. Nucleic Acids Res 33:D71–D74

Levi A, Davis A, Hernandez A et al (2006) Genes expressed during the development and ripening of watermelon fruit. Plant Cell Rep 25:1233–1245

Manickavelu A, Kawaura K, Oishi K et al (2012) Comprehensive functional analyses of expressed sequence tags in common wheat (*Triticum aestivum*). DNA Res 19:165–177

Nakamura Y, Cochrane G, Karsch-Mizrachi I (2013) The international nucleotide sequence database collaboration. Nucleic Acids Res 41:D21–D24

Nakamura Y, Afendi FM, Parvin AK et al (2014) KNApSAcK metabolite activity database for retrieving the relationships between metabolites and biological activities. Plant Cell Physiol 55:e7

NCBI Resource Coordinators (2014) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 42:D7–D17

Ogata H, Goto S, Fujibuchi W et al (1998) Computation with the KEGG pathway database. Biosystems 47:119–128

Ohyanagi H, Takano T, Terashima S et al (2015) Plant Omics Data Center: an integrated web repository for interspecies gene expression networks with NLP-based curation. Plant Cell Physiol 56:e9

Ostlund G, Schmitt T, Forslund K et al (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res 38:D196–D203

Ouyang S, Zhu W, Hamilton J et al (2007) The TIGR rice genome annotation resource: improvements and new features. Nucleic Acids Res 35:D883–D887

Pakseresht N, Alako B, Amid C et al (2014) Assembly information services in the European Nucleotide Archive. Nucleic Acids Res 42:D38–D43

Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. Nature 475:189–195

Qi J, Liu X, Shen D et al (2013) A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. Nat Genet 45:1510–1515

Quevillon E, Silventoinen V, Pillai S et al (2005) InterProScan: protein domains identifier. Nucleic Acids Res 33:W116–W120

Saito T, Ariizumi T, Okabe Y et al (2011) TOMATOMA: a novel tomato mutant database distributing Micro-Tom mutant collections. Plant Cell Physiol 52:283–296

Sakai H, Lee SS, Tanaka T et al (2013) Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. Plant Cell Physiol 54:e6

Sakurai N, Ara T, Ogata Y et al (2011) KaPPA-View4: a metabolic pathway database for representation and analysis of correlation networks of gene co-expression and metabolite co-accumulation and omics data. Nucleic Acids Res 39:D677–D684

The UniProt Consortium (2014) Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res 42:D191–D198

Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–641

Tsugane T, Watanabe M, Yano K et al (2005) Expressed sequence tags of full-length cDNA clones from the miniature tomato (*Lycopersicon esculentum*) cultivar Micro-Tom. Plant Biotechnol 22:161–165

Webb EC (1992) Enzyme nomenclature. Academic Press, San Diego, CA

Wheeler DL, Church DM, Federhen S et al (2003) Database resources of the National Center for Biotechnology. Nucleic Acids Res 31:28–33

Wheeler DL, Barrett T, Benson DA et al (2008) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 36:D13–D21

Yamamoto N, Tsugane T, Watanabe M et al (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

Yamazaki Y, Akashi R, Banno Y et al (2010) NBRP databases: databases of biological resources in Japan. Nucleic Acids Res 38:D26–D32

Yano K, Imai K, Shimizu A et al (2006a) A new method for gene discovery in large-scale microarray data. Nucleic Acids Res 34:1532–1539

Yano K, Tsugane T, Watanabe M et al (2006b) Non-biased distribution of tomato genes with no counterparts in *Arabidopsis thaliana* in expression patterns during fruit maturation. Plant Biotechnol 23:199–202

Yano K, Watanabe M, Yamamoto N et al (2006c) MiBASE: a database of a miniature tomato cultivar Micro-Tom. Plant Biotechnol 23:195–198