

This dataset was retrieved from the [Tomato SBM DataBase](#), which had been released by the Kazusa DNA Research Institute.

Tomato SBM DataBase

Kazusa DNA Research Institute

Tomato SBM (selected BAC clone mixture) contigs

To enrich gene-rich regions within the tomato genome, two sets of BAC clone pools were generated using the BAC end sequences available at the SGN website, and denoted as SBM (selected BAC clone mixture).

In the first set (SBM-I), a total of 20,000 BAC clones, in which the sequence at neither end showed any homology to any of the tomato repetitive sequences registered in the SGN tomato genome repeat data, were selected from the BAC libraries (10,000 clones from the *HindIII* library, and 5,000 clones each from the *MboI* and *EcoRI* libraries).

The second set (SBM-II) selected included clones that contained repetitive sequences at one end but unique sequences at the other. A total of 10,800 clones (5,400 clones from the *HindIII* library, and 2,700 clones each from the *MboI* and *EcoRI* libraries) were pooled.

Shotgun libraries were generated from the SBM-I and SBM-II sets, and sequencing was performed by the Sanger method. The total number of accumulated reads was 4,248,000: 2,916,000 from SBM-I and 1,332,000 from SBM-II. The sequences were then subjected to assembly, along with the BAC and fosmid end sequences that had been made publicly available by SGN. As a result, non-redundant sequences of 540,588,968 bp that consisted of 100,783 contigs have been generated. The statistics of the assembly are summarized here.

Scaffolds were formed by ordering and orienting the contig sequences, by taking the relative positions of the BAC and fosmid end sequences into account. The number of scaffolds thus generated was 3,996 containing 15,119 contigs.

The sequence information of the SBM contigs is available through the public DNA databases (DDBJ/Genbank/EMBL) under the accession numbers BABP01000001-BABP01100783.

This dataset was retrieved from the [Tomato SBM DataBase](#), which had been released by the Kazusa DNA Research Institute.

Contig name

Assembled contigs, excluding those shorter than 1,000 bp in size, were designated as SISBM (*Solanum lycopersicum* Selected BAC clone Mixture) followed by the assembly consensus names containing an indication of the assembly status. Note that these contig names should be regarded as temporal. The SBM sequence data are currently being integrated into the whole genome assembly data by the International Tomato Sequencing Consortium, and new standardized contig names will be provided in the future.

■ SISBM_S*****_**

The contigs in the scaffold groups are designated by the alphabetical code "S" and a five-digit group number followed by a two-digit sequential number for each contig.

■ SISBM_L*****_01

The singlet contigs containing the BAC and fosmid-end sequences as assembly components are indicated by the alphabetical code "L" and a five-digit contig number followed by the consensus code "01".

■ SISBM_*****_01

The contigs exclusively composed of SISBM shotgun sequences are designated with a six-digit contig number followed by the consensus code 01.

Statistics of SBM assembly

SBM assembly statistics	
Total number of contigs	100,783
Total length of contigs	540,588,968 bp
G+C%	34.52 %
Average length of contigs	5,364 bp
Max length of contig	202,166 bp
N50 length	9,735 bp

This dataset was retrieved from the [Tomato SBM DataBase](#), which had been released by the Kazusa DNA Research Institute.

Total number of scaffolds	3,996
Number of contigs in scaffolds	15,199
Total length of scaffolds	230,039,600 bp
Average length of scaffolds	57,567 bp
Max length of scaffold	782,429 bp

Annotation

Gene prediction and modeling were performed using an automatic gene assignment software that employs *ab initio* gene-finding and similarity search. The *ab initio* gene-finding software used in this pipeline included GeneMark.hmm and SNAP with the *Arabidopsis thaliana* and *Lotus japonicus* trained matrices. Similarity searches to detect potential protein-coding exons were performed using the BLASTX function of BLAST against the TrEMBL library. A total of 77,779 potential protein-coding genes have been assigned, of which 37,738 were identified as being related to transposable elements.

IDs for the predicted genes

The protein-encoding genes thus assigned have been denoted with an ID consisting of the contig name followed by a sequential number in an increment of 10 from one end to another. Note that these gene IDs should be regarded as temporal. The SBM sequence data are currently being integrated into the whole genome assembly data by the International Tomato Sequencing Consortium, and new standardized IDs will be provided in the future.

Codes for the predicted genes

Codes for the gene prediction categories

/A	genes that were predicted by two gene finding software, GeneMark.hmm and SNAP, and were similar to LGI or full length cDNA sequences from microTom.
/B	genes that were predicted by both of the software mentioned above but were not similar to LGI or full length cDNA sequences.

This dataset was retrieved from the Tomato SBM DataBase, which had been released by the Kazusa DNA Research Institute.

/C	genes that were predicted by either one of the software mentioned above and were similar to LGI or full length cDNA sequences.
/D	genes predicted by either one of the software mentioned above but were not similar to any of the known genes.

Codes for the degree of gene coverage

/nf	sequences with significant similarities to the registered genes in the TrEMBL database that cover 70% or more of the coding regions.
/p	sequences with significant similarities to the registered genes in the TrEMBL database that cover less than 70% of the coding regions.
no indication	sequences with no significant similarities to the registered genes in the TrEMBL database.

Other codes for the predicted genes

/TE	genes related to transposable elements.
/partial	genes lacking either the initiation or the termination codon as they are located at an contig end.
/tRNA	transfer RNA genes

Data Release Policy

The tomato SBM assembly may be freely downloaded and used in analyses under the following conditions;

1. Users are free to use the data in scientific papers analyzing specific genes and genomic regions, provided that the Kazusa DNA Research Institute (KDRI) is properly acknowledged.
2. The tomato SBM sequence data will be integrated into the whole-genome assembly data, and the initial large-scale analysis of the assembled whole-genome data will be published under the name of "The International Tomato Sequencing Consortium SOL" (ITSC-SOL). Therefore, KDRI, as a member of the ITSC-SOL, reserves the right to perform and publish the initial large-scale analysis of the tomato SBM assembly dataset. For this reason, the downloaded data should not be subjected to large-scale analysis of any sort before the intended ITSC-SOL publication.

This dataset was retrieved from the [Tomato SBM DataBase](#), which had been released by the Kazusa DNA Research Institute.

Acknowledgements

We thank the Solanaceae Genome Network at Cornell University supported by the US National Science Foundation Plant Genome Program for providing all the tomato BAC and fosmid resources. We also thank the EU-SOL project for providing the tomato fosmid end sequences, and the JSOL project for providing the full-length cDNA sequences from micro-Tom.