Large-scale analysis of full-length cDNAs from the tomato (*Solanum lycopersicum*) cultivar Micro-Tom, a reference system for the Solanaceae genomics

Koh Aoki1*, Kentaro Yano2†, Ayako Suzuki2, Shingo Kawamura2, Nozomu Sakurai1, Kunihiro Suda1, Atsushi Kurabayashi1, Tatsuya Suzuki3, Taneaki Tsugane3, Manabu Watanabe3, Kazuhide Ooga1, Maiko Torii1, Takanori Narita4, Tadasu Shin-4, Yuji Kohara4, Naoki Yamamoto2, Hideki Takahashi5, Yuichiro Watanabe6, Mayumi Egusa7, Motochiro Kodama7, Yuki Ichinose8, Mari Kikuchi9, Sumire Fukushima9, Akiko Okabe9, Tsutomu Arië9, Yoko Sato10, Katsumi Yazawa10, Shinobu Satoh10, Toshikazu Omura11†, Hiroshi Ezura11†, Daisuke Shibata1

Abstract

**Background:** The Solanaceae family includes several economically important vegetable crops. The tomato (*Solanum lycopersicum*) is regarded as a model plant of the Solanaceae family. Recently, a number of tomato resources have been developed in parallel with the ongoing tomato genome sequencing project. In particular, a miniature cultivar, Micro-Tom, is regarded as a model system in tomato genomics, and a number of genomics resources in the Micro-Tom-background, such as ESTs and mutagenized lines, have been established by an international alliance.

**Results:** To accelerate the progress in tomato genomics, we developed a collection of fully-sequenced 13,227 Micro-Tom full-length cDNAs. By checking redundant sequences, coding sequences, and chimeric sequences, a set of 11,502 non-redundant full-length cDNAs (nrFLcDNAs) was generated. Analysis of untranslated regions demonstrated that tomato has longer 5’- and 3’-untranslated regions than most other plants but rice. Classification of functions of proteins predicted from the coding sequences demonstrated that nrFLcDNAs covered a broad range of functions. A comparison of nrFLcDNAs with genes of sixteen plants facilitated the identification of tomato genes that are not found in other plants, most of which did not have known protein domains. Mapping of the nrFLcDNAs onto currently available tomato genome sequences facilitated prediction of exon-intron structure. Introns of tomato genes were longer than those of Arabidopsis and rice. According to a comparison of exon sequences between the nrFLcDNAs and the tomato genome sequences, the frequency of nucleotide mismatch in exons between Micro-Tom and the genome-sequencing cultivar (Heinz 1706) was estimated to be 0.061%.

**Conclusion:** The collection of Micro-Tom nrFLcDNAs generated in this study will serve as a valuable genomic tool for plant biologists to bridge the gap between basic and applied studies. The nrFLcDNA sequences will help annotation of the tomato whole-genome sequence and aid in tomato functional genomics and molecular breeding. Full-length cDNA sequences and their annotations are provided in the database KaFTom http://www.pgb.kazusa.or.jp/kaftom/ via the website of the National Bioresource Project Tomato http://tomato.nbrp.jp.

* Correspondence: kaoki@kazusa.or.jp
† Contributed equally

1Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, 292-0818, Japan

© 2010 Aoki et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Background
The Solanaceae family comprises 1000-2000 species that show wide morphological variability and ecological adaptability [1]. This taxon includes a number of vegetable crops including fruit-bearing vegetables, tuber-bearing vegetables, and ornamental plants, many of which have economic importance. Tomato (Solanum lycopersicum) is a member of the Solanaceae, and it has served as a model system for fruit development [2] and plant defense [3,4]. A number of studies have accumulated substantial information regarding the genetics and physiology of tomato.

To understand the physiological processes of tomato at the molecular level, it is necessary to link traits to DNA sequence. Whole-genome sequencing is one of the approaches to achieve this goal. A project for sequencing the tomato genome was launched by The International Solanaceae Project (SOL) in the year 2003, and the project is currently in the final stage of sequence polishing (March, 2010) [4]. Accumulation of expressed sequence tags (ESTs) is an alternative approach to collecting DNA sequence information. A large number of ESTs have been produced to provide a resource for gene finding and expression studies [5-9]. For tomato ESTs, several repositories are available worldwide, such as the Dana-Farber Cancer Institute (DFCI) Tomato Gene Index, The SOL Genomics Network (SGN) [5], MiBASE [10], PlantGDB [11], and TomatoEST [12], and the number of tomato ESTs accumulated in the databases is 296,957 (January 8, 2010). To elucidate the functions of individual genes, attempts have been made to organize ESTs into consensus sequences, which are referred to as unigenes or tentative consensus (TC) sequences. However, a more direct approach to collecting DNA information of individual genes is to analyze full-length cDNA sequences. Major advantages of sequencing full-length cDNA over the EST assembly include that the data is derived from a single clone rather than the assembly of multiple ESTs which can generate contigs containing sequences from multiple transcripts. Another major advantage of this approach is that full-length cDNA clones have complete sequences of transcripts including coding regions (CDSs) and untranslated regions (UTRs). This facilitates the subsequent annotation and prediction of genomic structures. The complete CDSs in full-length cDNAs allow the prediction of protein sequences, conserved domains, and conserved motifs. Furthermore, full-length cDNAs are easy to use with gene transfer systems, permitting the functional analysis of individual genes through reverse-genetics approaches. Thus, full-length cDNAs are a powerful genomic research tool.

In tomato genomics, a miniature tomato cultivar of S. lycopersicum, Micro-Tom [13], is regarded as an excellent model system. Micro-Tom has characteristics that make it suitable for experimental study, such as small size, short generation time, and ease of transformation [14,15]. Recently, resources such as various mutagenized lines have been developed in the Micro-Tom background [16]. Thus, a collection of Micro-Tom full-length cDNAs would boost the molecular bioengineering of tomato by synergistically combining with other resources. However, no full-length cDNA collection of Micro-Tom, or of tomato, has yet been established.

In this study, we developed a large-scale collection of Micro-Tom full-length cDNAs. We constructed full-length-enriched cDNA libraries from Micro-Tom leaf, fruit, and root tissues, and we obtained full-length sequences of 13,227 cDNAs. We then checked for redundancy between sequences, for the presence of CDSs, and for the contamination of chimeric sequences, thus generating a set of 11,502 non-redundant full-length cDNA sequences (nrFLcDNAs) for sequence analysis. We report the results of UTR analyses, including length distribution, base composition; and the results of CDS analyses, including classification of functions of deduced proteins, metabolic pathway annotations, and comparisons with genes of other plants. We also report the results of mapping nrFLcDNAs onto the draft of tomato whole-genome sequence. Mapping onto genome sequence suggested that introns of tomato genes are relatively long compared to Arabidopsis and rice. We also found that the frequency of nucleotide mismatch between exon regions of Micro-Tom and those of the genome-sequencing cultivar, Heinz 1706, is 0.061%. Finally, the value of Micro-Tom full-length cDNA information will be discussed in terms of integration with the tomato genome sequence and other tomato resources.

Results and discussion

Construction of full-length-enriched cDNA libraries
We constructed four full-length-enriched cDNA libraries from Micro-Tom tissues including leaves, fruits, and roots. RNA samples were prepared from plants grown under 63 different conditions to increase the variation of transcripts in the libraries (see Additional file 1: Micro-Tom tissues used for RNA preparation). Leaves were treated with 9 pathogens, 5 chemical and biochemical elicitors, salicylic acid, and methyl jasmonic acid (Table 1). Leaf RNA samples were prepared from 43 tissue types under different treatment conditions (see Additional file 1: Micro-Tom tissues used for RNA preparation). Fruit RNA samples were prepared from the pericarp at four ripening stages (mature green, breaker, turning, and red ripe) harvested in the year 2003 and 2004 (Table 1, see Additional file 1: Micro-Tom tissues used for RNA preparation). Root RNA samples were prepared from plants both with and without flowers, expecting that tissues of
different growth phases have different transcripts. We also prepared RNA samples from roots treated with *Fusarium oxysporum* race 2 (Table 1, see Additional file 1: Micro-Tom tissues used for RNA preparation). RNA samples prepared from leaf, fruit tissues harvested in the year 2003, fruit tissues harvested in the year 2004, and root-tissues were mixed, and the respective mixtures were used as templates for four separate cDNA libraries, designated LEFL1, FC, LEFL2, and LEFL3. The FC library was constructed by the vector-capping method [17], and the LEFL 1, LEFL2, and LEFL3 libraries were constructed by the CAP-trapper method [18]. cDNAs derived from pathogens were excluded by hybridization to mRNAs prepared from the pathogens. The subtraction process was effective, since 5’-end sequences of only 37 clones out of 109,459 randomly selected clones matched pathogen sequences in the NCBI GenBank, and these were excluded from further data processing.

**Table 1** Micro-Tom tissues used for RNA preparation

| Organ          | Treatment                                                                 | Tissue          | Condition                  | Library |
|----------------|---------------------------------------------------------------------------|-----------------|----------------------------|---------|
| Leaf           | CMV (strain TN) and satellite RNA                                         | Inoculated leaves, systemic leaves | 16 dpi (21-day old)       | LEFL1   |
| ToMV (strain Lta1, L11 A, L2a, Ltb1, L8) | Inoculated leaves, systemic leaves                                      | 16 dpi (21-day old)       | LEFL1   |
| Non-pathogenic *Alternaria alternata* | Inoculated leaves                                          | 6, 24 hpi (5-week old)     | LEFL1   |
| *Alternaria alternate* f. sp. *lycopersici* (As-27) | Inoculated leaves                                       | 6, 24 hpi (5-week old)     | LEFL1   |
| *Corynespora cassicola* (isolated from tomato and cucumber) | Inoculated leaves | 6, 24 hpi (3-week old)     | LEFL1   |
| *Cladosporium fulvum* (strain 210, 211, 217) | Inoculated leaves | 6, 24 hpi (3-week old)     | LEFL1   |
| *Fusarium oxysporum* f. sp. *lycopersici* race 1 and 2 | Leaves                                             | Inoculated to root, 7 dpi (4-week old) | LEFL1   |
| Non-pathogenic *Fusarium oxysporum* | Leaves | Inoculated to root, 7 dpi (4-week old) | LEFL1   |
| *Pseudomonas syringae* pv. *tomato* (DC3000 wild type), pv *tabaci* isolate 6605 (wild type, ΔfliC, ΔfliD) | Spray-inoculated leaves | 24 hpi (6-week old)       | LEFL1   |
| Flagellin from *P. syringae* pv. *tabaci* isolate 6605, fgl22 | Sprayed leaves                                        | 3 hps (6-week old)         | LEFL1   |
| Probenazole (100 μg/ml) | Sprayed leaves                                         | 4 dps (4-week old)         | LEFL1   |
| Valdiamycin A (100 μg/ml) | Sprayed leaves                                         | 4 dps (4-week old)         | LEFL1   |
| Acibenzolar-S-methyl (100 μg/ml) | Sprayed leaves                                         | 4 dps (4-week old)         | LEFL1   |
| Validoxylamine A (100 μg/ml) | Sprayed leaves                                         | 2 dps (4-week old)         | LEFL1   |
| Salicylic acid (0.5 mM) | Sprayed leaves                                         | 2 dps (4-week old)         | LEFL1   |
| Methyl jasmonic acid (50 μM) | Sprayed leaves                                         | 2 dps (4-week old)         | LEFL1   |
| Fruit           | Mature green                                                                 | Pericarp         | 40 daa, year 2003, 2004     | FC/LEFL2 |
|                 | Breaker                                                                                  | Pericarp         | 40 daa, year 2003, 2004     | FC/LEFL2 |
|                 | Turning                                                                                  | Pericarp         | 40 daa, year 2003, 2004     | FC/LEFL2 |
|                 | Red ripe                                                                                 | Pericarp         | 50 daa, year 2003, 2004     | FC/LEFL2 |
| Root            | Roots from plants without flower                                                      | Roots            | 6, 10-week old               | LEFL3   |
|                 | Roots from flowered plants                                                              | Roots            | 10, 12-week old              | LEFL3   |
|                 | *Fusarium oxysporum* race 2                                                             | Roots            | 1, 7, 14, 21 dpi, 4- to 11-week old | LEFL3   |

Abbreviations: dpi, days post inoculation; hpi, hours post inoculation; hps, hour post spray; dps, days post spray; daa, days after anthesis. More detailed description of the samples is available as Additional file 1.

**Full-length sequencing of cDNAs**

A schematic flow of the sequencing process is presented in Figure 1A. We randomly selected 109,422 independent clones from leaf, fruit, and root libraries, and sequenced them from the 5’-end. After trimming vector-derived sequences and low-quality sequences, 89,872 5’-end sequences (30,679, 8046, 18,697 and 27,216 sequences from the LEFL1-, FC-, LEFL2- and LEFL3 libraries, respectively) were combined with Micro-Tom ESTs reported previously [9] and tomato ESTs registered in the SGN database. A total of 322,813 ESTs were grouped into 76,276 clusters. From 76,276 clusters, 22,900 clusters containing FC and LEFL clones were selected. The number of FC or LEFL clones in each cluster ranged from 1 to 137 (Figure 1B). The FC or LEFL sequence having the longest 5’-end extension in each cluster was selected as the representative of that cluster.
Full-length sequencing was performed using the representative clones by the clone-by-clone primer walking method. Out of 22,900 clones, full-length inserts of 13,227 cDNA clones were sequenced (see Additional file 2: Clone number list). Mean of phred quality value of each base was 65 (i.e., one error in $3 \times 10^6$ bases), and more than 70% of bases exceeded quality value 68. This collection was registered to the high-throughput cDNA sequence (HTC) division of the DDBJ.

**Generation of a non-redundant set of the full-length cDNAs**

We checked for redundancy within the set of 13,227 HTCs, and then a set of 12,106 HTCs representing non-redundant transcripts was sent for further processing (see Additional file 2: Clone number list). To check whether each of the 12,106 HTCs has a full-length CDS or not, we first searched for cDNAs derived from non-coding RNAs using the NONCODE version 2.0 dataset [19]. Three HTCs (FC25DB10, LEFL3003O13, and LEFL3054J22) were identified as cDNAs derived from non-coding RNA (see Additional file 3: cDNA derived from ncRNA) and excluded from further analysis. Next, we confirmed that all of the HTCs did not match DNA sequences from pathogens used to attack leaf and root tissues, suggesting that the probability of contamination with pathogen-derived transcripts was negligible (see Additional file 2: Clone number list, for the accession numbers of pathogen-derived sequences). We then checked for chimeric clones putatively generated during cDNA cloning process. According to the criteria described in Methods, 36 HTCs were regarded as chimeric clones (see Additional file 2: Clone number list). We then identified sets of HTCs generated by the alternative splicing event, and excluded 470 HTCs containing retained introns (see Additional file 2: Clone number list). From the remaining 11,597 HTCs, CDSs were predicted by using FrameDP [20] and GeneMark.hmm-E [21] programs, and by selecting open reading frame encoding the longest amino acid sequence. Essentially, a CDS encoding amino acid sequence that had the highest similarity to proteins registered in either the nr or the tomato SBM protein databases was selected as representative CDS of the HTC (Additional file 4: Predicted CDS of 11,597 non-redundant HTCs). HTCs with CDSs shorter than nine amino acid (aa) long were excluded from further analysis, since the shortest length tomato protein registered in UniProt was 10 aa (UniProt accession number, Q6TS30; description, ENOD40). Out of 11,597 non-redundant HTCs, CDS encoding full-length protein was predicted for 11,502 HTCs. Full-length CDS were not predicted for 73 HTCs and no coding regions were predicted for 22 HTCs. Finally, a set of 11,502 HTCs (see Additional file 2: Clone number list) were used for the subsequent sequence analyses.

**Figure 1 Data processing scheme of tomato full-length cDNA sequences**

(A) Scheme for data processing of tomato full-length cDNA sequences. Four separate full-length-enriched libraries, LEFL1, FC, LEFL2, and LEFL3, were constructed. From randomly chosen clones, we obtained high-quality 5'-end sequences, 30,679, 8046, 18,697, and 27,216 sequences from the LEFL1, FC, LEFL2, and LEFL3 libraries, respectively. These high-quality 5'-end sequences were registered in the EST division of the DDBJ. These were combined with 238,157 public tomato ESTs and then clustered into 76,276 groups. Clusters containing FC or LEFL sequences as a member were selected. The FC or LEFL sequence with the longest 5'-extension was chosen as the representative of each cluster and sent for full-length sequencing. Full-length sequencing was finished for 13,227 cDNAs, which were registered in the high throughput cDNA (HTC) division of the DDBJ. From 13,227 HTCs, 12,106 non-redundant full-length cDNAs were chosen. The 12,106 full-length cDNA set was tested for whether it contained non-coding RNA-derived cDNAs, pathogen transcript-derived cDNAs, chimeric clones, and cDNAs containing retained introns. After excluding these sequences, a set of 11,597 non-redundant HTCs was checked for CDS. Finally, a set of 11,502 non-redundant full-length cDNAs (nFLcDNAs) was generated for subsequent sequence analyses. (B) Distribution of the number of 5'-end sequences derived from FC and LEFL cDNA libraries in each cluster.
list) was generated, hereafter referred to as nrFLcDNA. This nrFLcDNA set was subjected to subsequent sequence analyses.

Nucleotide sequences of 11,502 nrFLcDNAs were compared with DFCI Tomato Gene Index release 12.0 tentative consensus (TC), SGN tomato unigenes (SGN_tomato_20090805), and the prerelease of tomato genome shotgun sequence (S_lycopersicum_scaffolds_20091201) (Figure 2A). Percentage of nrFLcDNAs that did not match (E-value greater than 1e-10) was 1.7%, 2.0%, and 0.5% against DFCI tomato TCs, SGN tomato unigenes, and tomato genome sequences, respectively. Predicted amino acid sequences of 11,502 HTCs (Additional file 4: Predicted CDS of 11,597 non-redundant HTCs) were compared with proteins registered in NCBI nr dataset and Tomato SBM protein dataset (protein_sequence, ftp://ftp.kazusa.or.jp/pub/tomato/) (Figure 2B). Distribution of E-value was similar against both datasets. Approximately 22% of the nrFLcDNAs had very high similarity (E-value < 1e-180) to nr and SBM proteins, and more than 75% of the nrFLcDNAs have E-value smaller than 1e-50.

Length distributions of UTRs and CDSs

Figures 3A and 3B show the distributions of cDNA insert length and the CDS length, respectively. The

![Figure 2 Similarity of nrFLcDNAs with public tomato sequences](image)

**Figure 2** Similarity of nrFLcDNAs with public tomato sequences. (A) Similarity of the 11,502 nrFLcDNAs to DFCI tomato tentative consensus (TC), SGN tomato unigene, and the prerelease of tomato genome shotgun sequence (Tomato GSS). Black bars: nrFLcDNAs that showed very high similarity (E-value < 1e-180), white bars; nrFLcDNAs that showed very low similarity (E-value > 1e-10). (B) Similarity of amino acid sequences predicted from the 11,502 nrFLcDNAs to proteins in nr (white rectangle) and the Tomato SBM database (black rectangle).

![Figure 3 Distribution of characteristic lengths of nrFLcDNAs](image)

**Figure 3** Distribution of characteristic lengths of nrFLcDNAs. (A) cDNA insert length, (B) CDS length, (C) 5'-UTR length, and (D) 3'-UTR length. 5'- and 3'-UTR lengths were longer than those of Arabidopsis, soybean, poplar, and maize, and slightly shorter than those of rice.
average insert length was 1418 bp, which is shorter than
those of Arabidopsis (1445 bp [22]) and soybean (1539
bp [23]) (Table 2). The median value for the insert
length was 1324 bp, which is shorter than Arabidopsis
(1459 bp) and rice (1548 bp) [24], but longer than
poplar (990 bp [25]) (Table 2). The average length of
the CDS was 938 bp, corresponding to an average poly-
peptide length of 313 aa, which was also shorter than
those of Arabidopsis, rice, and soybean, and longer than
poplar and maize [26,27] (Table 2). The nrFLcDNA set
contains 904 sequences derived from the FC clones that
were cloned by the vector-capping method, and these
harbor shorter cDNA inserts (average insert length 773
bp, average CDS length 482 bp) than those derived from
the LEFL clones that were cloned by the CAP-trapper
method (average insert length 1474 bp, average CDS
length 977 bp). This probably explains why the average
lengths are shorter than those reported for full-length
cDNAs of other plants produced by the CAP-trapper
method.

Figures 3C and 3D show the distributions of the 5'-
and 3'-UTR lengths, respectively. In contrast to CDS
length, tomato nrFLcDNAs had longer UTRs compared
to those of other plants. The average 5'- and 3'-UTR
lengths were 175 and 257 bp, respectively. The median
5'- and 3'-UTR lengths were 106 and 220 bp, re-
spectively. On both average and median bases, the UTRs
of tomato are longer than those of Arabidopsis, soybean,
poplar, and maize, and slightly shorter than those of rice
(Table 2). The relatively long UTRs of tomato likely
contribute to the regulation of mRNA transcription,
translation, and stability [28].

### Nucleotide composition

The nucleotide composition of nrFLcDNAs differed
among the 5'-UTRs, CDSs, and 3'-UTRs (Table 3). The
AT-content of the full-length cDNAs was 59.7%, which is
slightly higher than that of Arabidopsis full-length
cDNAs (57%) [24]. AT richness was more pronounced in
the 5'- and 3'-UTRs (63.5% and 66.3%, respectively), and
less pronounced in the CDSs (57.2%). The 5'-UTRs had
more Cs than the 3'-UTRs, and the number of Ts was
higher in the 3'-UTRs than in the 5'-UTRs. These results
are similar to those observed in Arabidopsis full-length

---

**Table 2:** Length of cDNA inserts, CDSs, and UTRs, and comparison with full-length cDNAs from other plants

| Plant       | cDNA insert | CDS | 5'-UTR | 3'-UTR | Reference |
|-------------|-------------|-----|--------|--------|-----------|
| Tomato      | average     | 1418| 938    | 175    | 257       | (Present study) |
|             | median      | 1324| 876    | 106    | 220       |
| Arabidopsis | average     | 1445| n.a.   | n.a.   | n.a.      | [24]          |
|             | median      | 1459| 1097   | 88     | 184       |
| Soybean     | average     | 1539| 1042   | 123    | 248       | [23]          |
|             | median      | n.a.| 933    | 75     | 233       |
| Poplar      | average     | 1045| 649    | 109    | 228       | [25]          |
|             | median      | 990 | 558    | 77     | 209       |
| Rice        | average     | n.a.| 993    | 260    | 398       | [24]          |
|             | median      | 1548| 947    | 123    | 279       |
| Maize       | average     | n.a.| 799    | 99     | 206       | [26]          |
|             | median      | n.a.| 741    | 80     | 228       | [27]          |

n.a.: Not available.

---

**Table 3:** Nucleotide composition for different parts of tomato full-length cDNAs

| Part         | Frequency (%) |
|--------------|---------------|
|              | A  | T  | G  | C  |
| Full-length cDNA | 29.3 | 30.4 | 21.7 | 18.6 |
| 5'-UTR       | 30.2 | 33.3 | 17.1 | 19.4 |
| CDS          | 29.1 | 28.1 | 23.4 | 19.4 |
| 3'-UTR       | 29.3 | 37.0 | 18.3 | 15.4 |

---

**Figure 4:** Scheme for detection of alternatively spliced transcripts. Each nrFLcDNA was searched against nrFLcDNA dataset, DFCI tomato TCs, and SGN tomato unigenes using BLASTN. Query nrFLcDNAs were designated as "target nrFLcDNAs." Sequences that matched the target nrFLcDNA by a threshold E-value of 1e-50 were grouped into a multi-sequence group. Intron detection was performed using the est2genome program by setting the target nrFLcDNA as "est" and member nrFLcDNA, unigene, or TC as "genome," and vice versa. Consequently, splicing variants were detected for 1,206 target nrFLcDNAs. Out of the 1,206 nrFLcDNAs, 388 nrFLcDNAs were aligned to SGN tomato BAC sequences (bacs.v374.seq.20081128091837). Classification of alternative splicing events was carried out manually using the 388 multi-sequence groups.
cDNAs [24]. However, nucleotide compositions of the nrFLcDNA was more AT-rich than those of maize transcripts [27]. The most frequently used stop codon is TGA (occur 40.6% of all nrFLcDNAs) followed by TAA (35.6%) and TAG (23.8%). Frequency of stop codon is similar to that occurred in Arabidopsis (TGA, 44%; TAA, 36%; and TAG, 20%), but different from those occurred in rice (TGA, 43%; TAA, 27%; and TAG, 30%) and maize (TGA, 51%; TAA, 19%; and TAG, 30%) [27].

Alternative splicing and retained intron
As reported for Arabidopsis and rice full-length cDNA collections, the collection of full-length cDNAs contains cDNAs generated from alternatively spliced transcripts [29,30]. To identify nrFLcDNAs derived from alternatively spliced transcripts, we first generated multi-sequence groups using the non-redundant 12,106 HTCs, SGN tomato unigenes, and DFCI tomato TCs according to the similarity to the target non-redundant HTCs (Figure 4). All pairs of target non-redundant HTCs and member sequences were subjected to intron detection, and we confirmed that 1206 target non-redundant HTCs had splicing variants. The percentage of non-redundant HTCs with splicing variants was 10%, which is slightly lower than that of Arabidopsis (11.6%) [29] or rice (13.1%) [30].

Out of these 1206 non-redundant HTCs, 388 HTCs were aligned with SGN tomato BAC sequences (bacs.v374.seq.20081128091837 downloaded from the SGN ftp site http://sgn.cornell.edu/bulk/input.pl?mode=ftp), which allowed us to classify the types of 434 alternative splicing events (see Additional file 2: Clone number list). The retained intron-type was the most frequent (55% of all events). The alternative acceptor site- and alternative donor site-types also occurred frequently, comprising 20% and 16% of all events, respectively. The exon skip-type and other types (e.g., alternative terminal exon) were rather rare. The occurrence of alternative acceptor site- and alternative donor site-types was slightly higher than that reported previously for tomato [31]. nrFLcDNAs with retained introns encode proteins with alternative amino acid sequences. We found that 509 nrFLcDNAs may contain retained full- or partial-length introns (see Additional file 2: Clone number list). Larger scale identification of alternative splicing events using the draft of tomato whole-genome sequence is currently underway.

Functional annotation of tomato full-length cDNA
Next, putative functions of nrFLcDNAs were assessed by BLASTP searches against the TAIR9, RAP-DB, and NCBI nr protein datasets. To identify functions of the nrFLcDNAs, we used gene ontology (GO) annotations for Arabidopsis genes that matched nrFLcDNAs. The results from the case using Arabidopsis proteins that

![Figure 5 Profile of GO annotations for nrFLcDNAs](http://www.biomedcentral.com/1471-2164/11/210)  
*Figure 5 Profile of GO annotations for nrFLcDNAs.* To obtain GO annotations for nrFLcDNAs, similarity between the amino acid sequences predicted from nrFLcDNAs and Arabidopsis proteins was assessed using BLASTP. Arabidopsis genes corresponding to the top hit Arabidopsis protein to each nrFLcDNA was chosen at threshold E-values of 1e-50 (gray bars). GO annotations for nrFLcDNAs were then retrieved by subjecting the list of Arabidopsis genes to a TAIR GO annotation search http://www.arabidopsis.org/tools/bulk/go/index.jsp. GO annotations for all Arabidopsis genes (black bars) were retrieved from the TAIR GO annotation search. No statistically significant difference in the frequency was observed in all categories.
match nrFLcDNAs at E-value smaller than 1e-50 is shown (see Additional file 2: Clone number list). This revealed that nrFLcDNAs covered functional categories as broadly as all Arabidopsis genes (Figure 5).

To obtain insights into the metabolism-related genes, the nrFLcDNAs were assigned to the LycoCyc pathways http://sgn.cornell.edu/tools/solcyc/. Based on the best matches to SGN tomato unigenes, 448 nrFLcDNAs (see Additional file 2: Clone number list) were assigned to LycoCyc pathways, and 1117 pathway annotations were obtained. Comparison of the pathway-classification patterns with those for all SGN unigenes demonstrated that the relative abundance of annotations assigned to the pathways “Amino Acids Degradation”, “Fermentation”, and “TCA cycle” was higher in the nrFLcDNAs than in all SGN unigenes (Figure 6). This demonstrated that genes related to primary metabolism are slightly overrepresented in the nrFLcDNA set compared to genes related to secondary metabolism.

We then attempted to identify transcription factors that are crucial for the transcriptional regulation of various biological processes. We searched for transcription factor domains described in AGRIS AtTFDB [32] using the InterPro database by querying the CDS of the nrFLcDNAs. This search demonstrated that 311 nrFLcDNAs (see Additional file 2: Clone number list) had 417 InterPro entries classified into 34 transcription factor families (Table 4). The G2-like and MYB families were the most numerous (the nrFLcDNAs assigned to the MYB family were completely overlapped with those in the G2-like family), followed by the Homeobox family. This result was slightly different from those of Arabidopsis [33] and rice [30], in which the AP2/EREBP family and zinc finger family were predominant, respectively. Difference in the distribution of transcription factors between tomato nrFLcDNA and Arabidopsis implies that the nrFLcDNA set failed to contain transcription factors whose expression level is low or cell type-specific.

Comparative analysis with genes of other plants
To assess the similarity of nrFLcDNAs with genes of other plants, deduced peptide sequences were compared with protein databases of Arabidopsis (TAIR9) and rice (RAP-DB) using BLASTP, and with the DFCI Gene Indices of barley, wheat, maize, pine, spruce, poplar, Lotus japonicus, Medicago truncatula, soybean, orange, apple, grape, tobacco, and potato using tBLASTN (see Additional file 5: Datasets used for comparison with other plants). nrFLcDNAs has high similarity to two solanaceae plants, potato and tobacco (Figure 7A). Percentages of nrFLcDNA that did not match to potato and tobacco TCs (E-value > 1e-10) were 8.4% and 8.5%, respectively, which are the lowest among these 16 species. On the other hand, percentages of nrFLcDNA that did not match (E-value > 1e-10) to barley, rice, wheat, maize, pine, spruce, and L. japonicus were higher than the average over these 16 species (12.6%). Distribution patterns of nrFLcDNAs in varied E-value range between 1e-180 and 1e-10 were classified into two groups (Figure 7B). Distribution curves of nrFLcDNAs matched potato and tobacco TCs have peaks at 1e-100 and are similar to the distribution of nrFLcDNAs matched DFCI tomato TCs. On the other hand, distribution curves of
nrFLcDNAs matched TCs of other 14 plants have peaks at 1e-50 or larger. These results suggest that tomato transcript sequences are more closely related to transcripts from potato and tobacco than to other plant species.

To obtain insight into functions of nrFLcDNAs that were not found in other plants, we extracted 694 nrFLcDNAs that matched none of the other plant genes (see Additional file 2: Clone number list). More than 90% of the 694 nrFLcDNAs matched the prerelease of tomato genome shotgun sequence with E-value smaller than 1e-120 and only 27 nrFLcDNAs did not match genome sequence with E-value threshold of 1e-10, suggesting that most of the 694 nrFLcDNAs are tomato-derived transcripts. By searching the InterPro database, out of the 694 nrFLcDNAs, 54 nrFLcDNAs were assigned with 75 protein domains. We also identified known protein domains by searching the similarity with the tomato SBM protein datasets. Out of 48 nrFLcDNAs matched (E-value < 1e-10) SBM proteins, 19 nrFLcDNA had known protein domains. Domains such as Cyclin-like F-box (IPR001810), Aldehyde dehydrogenase (IPR015590), and Polynucleotidyl transferase, Ribonuclease H fold (IPR012337) were identified in both searches. However, known functional protein domains were not found in most (640 out of 694) of the nrFLcDNAs that were not found in other 16 plants.

| Family         | IPR entry       | Description of IPR entry                                                                 | No. of tomato nrFLcDNAs |
|----------------|-----------------|------------------------------------------------------------------------------------------|-------------------------|
| G2-like        | IPR006447       | Myb-like, SHAQKYF class                                                                   | 52a                     |
| MYB            | IPR015495       | Myb transcription factor                                                                  | 50a                     |
| Homeobox       | IPR001356       | Homeobox                                                                                  | 40                      |
| C2H2           | IPR007087       | Zinc finger, C2H2-type                                                                    | 31                      |
| bHLH           | IPR011598       | Helix-loop-helix DNA-binding                                                               | 22                      |
| CCAAT-HAP5, HAP3, and HAP2 | IPR003958   | Transcription factor CBF/NF-Y/archaeal histone                                           | 19                      |
| PHD            | IPR001065       | Zinc finger, PHD-type                                                                     | 18                      |
| AP2-EREBP      | IPR001471       | Pathogenesis-related transcription factor and ERF, DNA-binding                            | 18                      |
| WRKY           | IPR003657       | DNA-binding WRKY                                                                          | 18                      |
| bZIP           | IPR011616       | bZIP transcription factor, bZIP-1                                                          | 17                      |
| HSF            | IPR000232       | Heat shock factor (HSF)-type, DNA-binding                                                  | 15                      |
| C3H            | IPR000571       | Zinc finger, CCCH-type                                                                    | 13                      |
| MADS           | IPR002100       | Transcription factor, MADS-box                                                            | 11                      |
| MYB-related    | IPR01588        | Myb-related protein P, C-terminal                                                          | 11                      |
| C2C2-YABBY     | IPR006780       | YABBY protein                                                                             | 11b                     |
| ARID           | IPR009071       | High mobility group box                                                                   | 11b                     |
| NAC            | IPR003441       | No apical meristem (NAM) protein                                                          | 10                      |
| TCP            | IPR005333       | Transcription factor, TCP                                                                  | 8                       |
| GRAS           | IPR005202       | GRAS transcription factor                                                                  | 7                       |
| C2C2-GATA      | IPR00679        | Zinc finger, GATA-type                                                                    | 6                       |
| AB3VP1         | IPR003340       | Transcriptional factor B3                                                                  | 6                       |
| EIL            | IPR006957       | Ethylene insensitive 3                                                                     | 6                       |
| BBR/BPC        | IPR010409       | GAGA binding-like                                                                         | 2                       |
| TUB            | IPR000007       | Tubby, C-terminal                                                                         | 2                       |
| C2C2-Dof       | IPR003851       | Zinc finger, Dof-type                                                                     | 2                       |
| C2C2-CO-like   | IPR000315       | Zinc finger, B-box, “IPR002926 Zinc finger, CONSTANS-type” was deleted from IPR entry. | 2                       |
| ZF-HD          | IPR006456       | ZF-HD homeobox protein Cys/His-rich dimerisation region                                   | 2                       |
| GeBP           | IPR007592       | Protein of unknown function DUF573                                                        | 2                       |
| ARR-B          | IPR001789       | Response regulator receiver                                                               | 2                       |
| NLP            | IPR003035       | Plant regulator RWP-RK                                                                    | 2                       |
| BZR            | IPR008540       | BZR1, transcriptional repressor                                                           | 1                       |
| JUMONJI        | IPR013129       | Transcription factor jumonji                                                              | 1                       |
| ZIM            | IPR007853       | Zinc finger, Zim17-type                                                                   | 1                       |
| Whirly         | IPR009044       | ssDNA-binding transcriptional regulator                                                   | 1                       |

*aMember nrFLcDNAs of MYB were completely overlapped with those of G2-like. bMember nrFLcDNAs of C2C2-YABBY and ARID completely overlapped.*
Mapping full-length cDNAs onto tomato genomic sequence

Mapping of full-length cDNA sequences onto the genome sequence provides insights into tomato genomic structure. Out of 11,502 nrFLcDNAs, full-length regions of 10,544 nrFLcDNAs (see Additional file 2: Clone number list) were mapped onto the prerelease of tomato genome shotgun sequence (S_lycopersicum_scaffolds_20091201). The rest, 958 nrFLcDNAs, failed to match genome sequence with nucleotide identity larger than 90% or with alignment length more than 90% of the nrFLcDNA, or failed to match in full-length. Based on this full-length mapping, the average exon number per gene was estimated to be 4.9 (median value, 4 exons per gene) (Figure 8A). Exon length ranged from 14 to 4528 bp. The average and median exon lengths were estimated to be 274 bp and 154 bp, respectively (Figure 8B, black curve). Intron lengths showed larger variation than exon lengths, ranging from 12 to 9664 bp. The average and median intron lengths were estimated to be 594 bp and 264 bp, respectively (Figure 8B, gray curve). The median intron length was longer than those of Arabidopsis (100 bp) and rice (145 bp). To compare exon lengths with those of Arabidopsis and rice, we classified exons into four types: initial, internal, terminal, and single exons. The distribution profiles of exon lengths were similar to those observed in Arabidopsis (Figure 8C) [24]. The median lengths of the initial, internal, and terminal exons were 244, 109, and 415 bp, respectively, which were comparable to those of Arabidopsis (274, 112, and 402 bp) and rice (284, 113, and 476 bp). Finally, we estimated the entire length of each gene corresponding to nrFLcDNA. The average and median lengths of tomato genes were estimated to be 3735 bp and 3281 bp, respectively (Figure 8D).

To investigate the occurrence of introns in UTRs, the positions of translation start sites and translation termination sites were mapped to the exons. The frequency of the first methionine in the internal exons was 19.9%. On the other hand, the frequency of stop codons in the internal exons was 5.7%. This result indicates that intron mapping to a UTR was more likely to occur in 5'-UTRs than in 3'-UTRs.

This mapping also provides insights into the single nucleotide polymorphism frequency between Micro-Tom and the genome-sequencing cultivar, Heinz 1706.

**Figure 7** Comparison of nrFLcDNAs with genes of other plants. (A) Percentage of nrFLcDNAs with very high similarity (E-value < 1e-180, black) and very low similarity (E-value > 1e-10, gray) to genes of other plants. (B) Distribution of the number of nrFLcDNAs matched plant unigenes with intermediate similarity between 1e-180 and 1e-10.
To estimate frequency of nucleotide mismatch, we used nrFlcDNA-scaffold pairs with identity in the BLAST output equal to or larger than 99.5%. Resulting set of nrFlcDNA-scaffold pairs contained 90% of the total pairs. Frequency of nucleotide mismatches in exon regions was estimated to be 0.061% (i.e., one difference in 1640 nt). This is nearly comparable to the previous result based on the comparison of ESTs between Micro-Tom and other *S. lycopersicum* cultivars [9]. This result suggests that the Micro-Tom cDNA sequence serves as a good reference for the tomato genome sequence. We note that the results presented here will be revised when annotation of the tomato genome sequence is finished in near future.

**A tomato full-length cDNA database: KaFTom**

Information regarding the 13,227 full-length cDNA sequences including annotations and the results of similarity searches are distributed from the KaFTom database http://www.pgb.kazusa.or.jp/kaftom/. KaFTom also provides the results of mapping full-length sequences onto SGN tomato BAC sequences and predictions of exons and introns. It directly links to the clone request site, and all of the full-length cDNA clones (89,872 clones) are available from the National Bioresource Project Tomato http://tomato.nbrp.jp.

**Significance of Micro-Tom full-length cDNA in tomato genomics**

Full-length cDNAs serves as a valuable tool that will accelerate tomato genomics in several respects. First, the sequence information of the full-length cDNAs will be integrated into the whole-genome sequence information of tomato, and it will help genome annotation and the identification of regulatory elements in UTRs. Second, sequences of the full-length cDNAs are valuable information for generating molecular markers. Third, full-length cDNAs help our understanding of the functions of tomato genes through gain-of-function and loss-of-function analyses. Fourth, full-length cDNAs promote the use of “targeting induced local lesions in genome” (TILLING) screening. By combining the use of full-length cDNA sequence information with EMS-mutagenized lines of Micro-Tom [16], the TILLING approach will provide an efficient way to screen non-transgenic mutant lines as parental germplasm favorable for breeding programs.

**Conclusion**

We developed a set of 13,227 full-length cDNAs from the model tomato cultivar Micro-Tom, and then we generated a set of 11,502 nrFlcDNAs, each of which represents a non-redundant transcript coding a full-length CDS. Analysis of the 5′-UTRs, CDSs, and 3′-UTRs demonstrated that tomato transcripts have longer 5′- and 3′-UTRs than other plant species. Classification of functions of deduced proteins according to the GO annotation revealed that the nrFlcDNA set covers a broad range of proteins. Comparison of the nrFlcDNAs with genes of other plants facilitated the identification of tomato cDNAs that have very low similarity to 16 other plants tested, most of which did not have known protein domains. Mapping of the nrFlcDNAs onto
tomato genome sequences revealed that tomato genes have longer introns than Arabidopsis and rice. The frequency of single nucleotide polymorphisms between exons of Micro-Tom and the genome-sequencing cultivar, Heinz 1706, was estimated to be 0.061%. In combination with other resources, the Micro-Tom full-length cDNAs will bridge the gap between basic and applied studies, providing a valuable tool not only for tomato whole-genome annotation but also for genomics studies and molecular breeding.

Methods

Plant materials

The miniature tomato cultivar, Micro-Tom (Solanum lycopersicum cv Micro-Tom), was used to construct full-length cDNA libraries. Total RNA samples were prepared from leaves, fruits, and roots of Micro-Tom. The tissues used for RNA preparation are summarized in Table 1. For a more detailed description of tissue samples, see Additional file 1: Micro-Tom tissues used for RNA preparation.

Leaves were treated with various pathogens. For fungal and bacterial inoculation, Micro-Tom plants were grown in sterilized soil at 20-28°C with a photoperiod of 14-16 h light (7000 lux, or 85 μmol/m²-s)/8-10 h dark. For virus inoculation, Micro-Tom plants were grown in a mixture of vermiculite and perlite (1:1 (v/v)) at 23-25°C with a photoperiod of 16 h light (8000 lux, or 97 μmol/m²-s)/8 h dark, and irrigated with a nutrient medium every three days. Pathogen treatments were performed as described previously [34].

Fruits were harvested at four ripening stages, mature green, breaker, turning, and red ripe stages. We used two harvests of Micro-Tom fruit for fruit RNA preparation. In the first harvest, Micro-Tom plants were grown in the year 2003 and fruit pericarp was harvested as described previously [35]. From the 5'-end sequences, vector-derived sequences and low-quality regions (Phred quality value < 30) were trimmed using a combination of the cross_match program http://www.phrap.org and the phred program [38]. Sequence data of lengths shorter than 50 bases were not included in further data processing. If the repetition of a single nucleotide was longer than 10% of a sequence, such sequences were not included in further data processing. The 5'-end sequences were combined with previously sequenced Micro-Tom ESTs [9] and ESTs registered in the SGN database (file name: tomato_species_2008_10_21.seq, obtained from the SGN ftp site ftp://ftp.sgn.cornell.edu/est_sequences/species/Tomato/), then clustered using the phrpd program http://www.phrap.org/phredphrapconsed.html[39]. Clusters containing 5'-end sequences of clones derived from the FC and LEFL libraries were chosen. Based on the sequence alignment within each cluster, an FC or LEFL clone carrying the insert that has the longest extension in the 5'-direction was selected as a representative of the cluster. The representative clones were sent for full-length sequencing. All of the 5'-end sequences were

RNA extraction

Total RNA was extracted from the tissues using an acid guanidinium thiocyanate-phenol-chloroform method [37]. Sugars were further removed by a sodium acetate-precipitation method [35].

Construction of the full-length-enriched cDNA library

Equal amounts of total RNA (approximately 60 μg each) from 42 leaf tissue types were mixed and then used to construct the LEFL1 cDNA library. Equal amounts of total RNA (approximately 2400 μg each) from four fruit tissues harvested in the year 2003 were mixed and then used to construct the LEFL2 cDNA library. Likewise, equal amounts of total RNA (approximately 230 μg each) from 12 root tissue types were mixed and then used to construct the LEFL3 cDNA library. Construction of the above-mentioned libraries was performed by the biotinylated CAP-trapper method as described previously [18]. For the RNA mixture from fruits harvested in the year 2003, a full-length-enriched cDNA library, namely the FC library, was constructed by the vector-capping method described previously [35]. Finally, cDNA inserts were cloned into the pFLCIII vector or the pGCApzF3 vector.

5'-end sequencing and clustering

cDNA clones randomly collected from the libraries were single-pass sequenced from the 5'-end as described previously [35]. From the 5'-end sequences, vector-derived sequences and low-quality regions (Phred quality value < 30) were trimmed using a combination of the cross_match program http://www.phrap.org and the phred program [38]. Sequence data of lengths shorter than 50 bases were not included in further data processing. If the repetition of a single nucleotide was longer than 10% of a sequence, such sequences were not included in further data processing. The 5'-end sequences were combined with previously sequenced Micro-Tom ESTs [9] and ESTs registered in the SGN database (file name: tomato_species_2008_10_21.seq, obtained from the SGN ftp site ftp://ftp.sgn.cornell.edu/est_sequences/species/Tomato/), then clustered using the phrpd program http://www.phrap.org/phredphrapconsed.html[39]. Clusters containing 5'-end sequences of clones derived from the FC and LEFL libraries were chosen. Based on the sequence alignment within each cluster, an FC or LEFL clone carrying the insert that has the longest extension in the 5'-direction was selected as a representative of the cluster. The representative clones were sent for full-length sequencing. All of the 5'-end sequences were
registered to the DDBJ http://www.ddbj.nig.ac.jp/indexe.html with accession numbers BW684914-BW692959, DB678259-DB727670, and FS179211-206426.

**Full-length sequencing**

The clones selected as the representatives of the clusters were re-arrayed in 96-well plates. A small aliquot of each clone was inoculated to 2 ml of LB medium containing 50 μg/ml ampicillin, and grown overnight at 37°C in 96-deep-well plates. Plasmid DNA was prepared from the overnight culture using CosMC Prep (Agen-court Bioscience, Beverly, MS, USA) according to the manufacturer’s protocol. By using the plasmid DNA as templates, full-length sequencing was performed by a clone-by-clone primer walking method using a Model 3730 × l sequence analyzer (Applied Biosystems, Forster City, CA, USA). Sequencing reaction was performed using the BigDye Terminator version 3.1 cycle sequencing kit (Applied Biosystems). Sequences were assembled using the phrap program http://www.phrap.org/phredphrapconsed.html. Based on the assembled sequences, primers corresponding to the terminal sequences were designed using the Primer3 program [40]. The primer walking procedure was repeated until a poly(A) tail or a vector-derived sequence appeared. The full-length sequences were finished by trimming the vector-derived sequences from both ends using the cross_match program. Full-length cDNA sequences (13,227 sequences) obtained by this procedure were registered to the high-throughput cDNA sequence (HTC) division of the DDBJ with accession numbers AB211519-211522, AB211526, AK224591-AK224910, AK246135-AK248077, and AK319176-330134.

**Selection of a non-redundant set of full-length cDNAs**

To identify redundant HTCs, pair-wise similarity searches between members of the 13,227 HTC sequences were performed using BLASTN. Redundant HTCs were identified using the following criteria; E-value < 1e-180, nucleotide identity ≥ 97%, and the percentage of alignment length exceeded 80% of the length of each HTC. The resulting non-redundant HTCs (12,105 HTCs) were then checked for whether they contain clones derived from non-coding RNA, clones derived from pathogen transcripts, whether they contain chimeric clones, and whether they contain clones with retained introns. HTCs derived from non-coding RNAs were identified by searching for similarity of the HTCs against the “ncrna_NONCODE[v2.0].fasta” dataset obtained from the NONCODE download site http://www.noncode.org/download.htm[19]using BLASTN with a threshold E-value of 1e-180. To search HTCs derived from pathogen-derived transcripts, nucleotide sequences from Alternaria alternate, Cladosporium fulvum, Corynespora cassicola, Cucumber mosaic virus, Fusarium oxysporum, Pseudomonas syringae, and Tomato mosaic virus were retrieved from NCBI nucleotide database (see Additional file 2: Clone number list for accession numbers of pathogen-derived sequences). Similarity of the HTCs to pathogen nucleotide sequences were searched using BLASTN with threshold E-value of 1e-30. When the E-value was smaller than 1e-30, the nrFLcDNA was further checked for whether or not they have similar SGN tomato unigenes or DFCI tomato TCs with smaller E-values. If so, the nrFLcDNA was regarded as cDNA derived from Micro-Tom, not from pathogens. To identify chimeric clones, the full-length cDNA sequences were searched against SGN tomato unigenes ‘draft_Solanum_lycopersicum_transcript_assembly_2009_08_05.fasta’ with cut-off E-value of 1e-50. HTCs that matched more than two unigenes having different functional descriptions were identified as candidate chimeric clones. Out of these candidates, if the candidate clone matched to two or more scaffolds of the prerelease of tomato genome shotgun sequence (S_lycopersicum_scaffolds_20091201.fa) downloaded from the SGN ftp site ftp://ftp.solgenomics.net/tomato_genome/wgs/assembly, the candidate was regarded as a chimeric clone. Out of the remaining candidates, if a candidate HTC had either a BamHI or XhoI site (which was used for ligation of cDNA to vectors) in between regions matched to unigenes, the candidate was regarded as a chimeric clone. To identify HTCs containing retained intron, HTC sequences were subjected to similarity searches against the HTC itself using BLASTN with threshold E-value of 1e-50. To detect introns, pair-wise alignments of each target nrFLcDNA with members of that multi-sequence group were carried out using est2genome [41]. HTCs derived from non-coding RNAs, pathogen-derived transcripts, chimeric clones, and intron-containing clones were excluded and non-redundant 11,597 HTCs were subjected to the prediction of coding sequences.

Amino acid sequences were predicted in sense orientation using three methods. First, the CDS coding the longest amino acid sequence was predicted. Second, CDSs were predicted using FrameDP http://iant.toulouse.inra.fr/FrameDP/[20]. Parameters were set as follows: framed_minimum_peptide_length, 10; reference protein database, TAIR9_pep_20090619; e-value cut-off for considering ncbi-blastx hits, 1e-3; the method used for the first classification, GC3; the maximum number of models, 3; the maximum number of iteration, 3; cut-offs for considering ncbi-blastx HSPs, e-value 1e-4 and length 100. Third, CDSs were predicted using GeneMark.hmm eukaryotic version 3.3 [21]. The model used was a_thaliana.mod which was supported by the program. Amino acid sequences of all possible frames were then subjected to BLASTP search (threshold E-value was 1e-10) against
NCBI nr and the protein dataset (protein_sequence) of Tomato SBM Database http://www.kazusa.or.jp/tomato/. CDSs with the smallest E-value against proteins registered in either nr or SBM proteins were selected. If CDS with the smallest E-value did not cover full-length protein, CDS with the second smallest E-value were selected. If E-values of all CDSs were the same, the longest CDSs were selected. Whether or not HTCs have full-length CDS was assessed by checking if the amino acid sequences contain both start and stop codons. Resulting set of 11,502 non-redundant HTCs encoding full-length proteins was referred to as ‘nrFLcDNA.’

**UTR identification**

UTRs were identified according to the selected CDS. 5’-UTRs were defined as the nucleotide sequence upstream of the start codon. 3’-UTRs were defined as the nucleotide sequence downstream of the stop codon to the poly (A) tail. Nucleotide composition analysis was performed essentially by counting A, T, C, G, N, and X using an in-house Perl program.

**Identification of retained introns and alternative splicing**

After removing poly(A) tails, a set of 12,106 non-redundant HTC sequences were subjected to similarity searches against non-redundant HTCs, the SGN tomato unigenes (Tomato_200607_build_1, ftp://ftp.sgn.cornell.edu/unigene/builds), and the DFCI Tomato Gene Index (release 12.0, http://compbio.dfci.harvard.edu/tgi/) using BLASTN with threshold E-value of $1 \times 10^{-50}$. Searches for the protein domains in the amino acid sequences predicted from six open reading frames were performed using InterProScan [42]. Gene ontology (GO) annotations [43] for Arabidopsis genes with the highest similarity to nrFLcDNAs were retrieved from a TAIR GO annotation search http://www.arabidopsis.org/tools/bulk/go/index.jsp. GO annotation documents for InterProScan domains including “interpro2go,” “entry.list,” and “ParentChildTreeFile.txt” were obtained at the InterPro public ftp site ftp://ftp.ebi.ac.uk/pub/databases/interpro/ in May 2009. Data files of Arabidopsis transcription factors (file named AtTFDB.zip) were downloaded from AGRIS AtTFDB http://arabidopsis.med.ohio-state.edu/AtTFDB/. The difference in the frequency of GO terms was tested by Fisher’s exact probability test ($P < 0.01$).

**Pathway annotations**

The SGN tomato unigene (Tomato_200607_build_1, downloaded from SGN ftp site ftp://ftp.sgn.cornell.edu/unigene/builds/) with the highest similarity to each nrFLcDNA was searched for using BLASTN, and then SGN tomato unigenes that matched nrFLcDNAs at the threshold E-value of $< 1 \times 10^{-180}$ were assigned to tomato metabolic pathways according to LycoCyc version 1.0.1.1 ftp://ftp.sgn.cornell.edu/pathways/lycocyc.dump.txt. Pathway hierarchy was retrieved on May 1, 2009 from “Hierarchical pathway ranking for Solanum lycopersicum” http://solcyc.sgn.cornell.edu/LYCO/hierarchical.html. The version of LycoCyc used in this study provides pathway annotations to 1532 SGN unigenes. The difference in the frequency of pathway annotation terms was tested by Fisher’s exact probability test ($P < 0.05$).

**Similarity to genes of other plants**

The similarity of nrFLcDNAs to genes of other plants was estimated using tBLASTN with E-value cut-off of $1e-10$ against DFCI Gene Indexes of Hordeum vulgare (barley), Triticum aestivum (wheat), Zea mays (maize), Pinus (pine), Picea (spruce), Populus (poplar), Lotus japonicus, Medicago truncatula, Glycine max (soybean), Citrus sinensis (orange), Malus × domestica (apple), Vitis vinifera (grape), Nicotiana tabacum (tobacco), and Solanum tuberosum (potato) (see Additional file 5: DataSets used for comparison with other plants). GO annotations for DFCI tomato TCs (“LGIGO.071608”) were downloaded from the DFCI public ftp site. The difference in the frequency of GO terms was tested by Fisher’s exact probability test ($P < 0.05$).
Mapping full-length cDNA sequences to tomato genome sequence
Poly(A) tails of the nrFLcDNA sequences were excised, and this set of nrFLcDNAs without poly(A) tails was used for similarity searches against the Prerelease of Tomato Genome Shotgun Sequence (S_lycopersicum_scaffolds_20091201.fa, ftp://ftp.solgenomics.net/tomato_genome/wgs/assembly). Tomato genome scaffold sequences with the highest similarity to a given nrFLcDNA sequence were identified using BLASTN (threshold E-value, 1e-50). Each pair of scaffold and nrFLcDNA sequences was then submitted to exon and intron prediction using the est2-genome program [41] by setting scaffold and nrFLcDNA sequences as “genome” and “est,” respectively. To analyze numbers and lengths of exons and introns, cut-off value for the alignment length was set as 90% of nrFLcDNA length. Cut-off value for nucleotide identity was set as 90%. To estimate the single nucleotide polymorphisms (SNPs) between nrFLcDNA and scaffold sequences, the number of nucleotide mismatches was counted in the nrFLcDNA-scaffold pairs having ‘identity’ in BLAST output equal to or larger than 99.5%.

Database
The data presented in this study, including full-length cDNA sequences of the 13,227 HTCs, the results of similarity searches against nr, UniProt, TAIR, RAP-DB, SGN tomato unigenes, DFCI tomato TCs, and InterProScan, and mapping of full-length sequences against SGN tomato BAC sequences are available at the tomato full-length cDNA database KaFTom http://www.pgb.kazusa.or.jp/kaftom/. KaFTom is accessible directly or via the National Bioresource Project Tomato website http://tomato.nbrp.jp.

Acknowledgements
We are grateful for and acknowledge use of the draft tomato genome sequence, which was generated by the International Tomato Genome Sequencing Consortium http://solgenomics.net/tomato/. We are also grateful for and acknowledge use of the tomato SBD dataset, which was generated by Kazusa DNA Research Institute. We thank Hideki Hirakawa and Shinobu Nakayama (Kazusa DNA Res. Inst.) for assistance in GeneMark.hmm analysis. We thank Shusen Sato (Kazusa DNA Res. Inst.) for critical reading of the manuscript, and Kenta Shirasawa (Kazusa DNA Res. Inst.) for helpful discussion. We also thank Tsugumi Isotsuki, Miyuki Inde, and Tsutae Aoyama (Kazusa DNA Res. Inst.) for plant care and lab assistance. We thank Hiroshi Otani (Tottori Univ.), Wataru Hasama (Oita Pref. Agric. Res. Center) and Hiroshi Shiomi (Taki and Co. Ltd.) for providing the Conympesara cassicola isolate. This work was supported by National Bioresource Project, Genome program, “Enhancing tomato resources by sequencing Micro-Tom full-length cDNA” (2008, MEXT, Japan) to KA, by the Japan Solanaceae Consortium (JSOL), a grant from Meiji Univ. to KY, and a grant from the Kazusa DNA Res. Inst. to DS and KA.

Author details
1Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, 292-0818, Japan. 2Meiji University, 1-1-1 Higashi-mita, Tama-ku, Kawasaki, 214-8571, Japan. 3Chiba Prefectural Agriculture and Forestry Research Center, 808 Daizenno-cho, Medori-ku, Chiba, 266-0006, Japan. 4National Institute of Genetics, Yata 1111, Mishima, 411-8540, Japan. 5Tohoku University, 4-101 Koyama-minami, Tottori, 680-8553, Japan. 6Kakuyama University, 1-1-1 Tsuchima-naka, Kita-ku, Okayama, 700-8530, Japan. 7The University of Tokyo, Komaba, Meguro-ku, 153-8902, Japan. 8Tottori University, 4-101 Koyama-minami, Tottori, 680-8553, Japan. 9Tokyo University of Agriculture and Technology, 3-5-8 Saiwaicho, Fuchu, 183-8509, Japan. 10Institute of Biological Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, 305-8571, Japan. 11Gene Research Center, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, 305-8571, Japan.

Authors’ contributions
This study was conceived and directed by KA and DS. RNA samples were prepared by NS, KS, TT, MW, HT, YW, ME, MK, YI, MK, SF, AO, TA, YS, KY, SS, TO, and HE. cDNA libraries were constructed by KA, SS, HE, and DS. Sequencing was carried out by KS, AK, TS, TT, MW, MT, TN, TS, and YK. Assembly of the full-length sequences was performed by KS, AK, TS, and MW under the direction of KY. Sequence analyses and other bioinformatics were directed by KY and performed by AS, SK, NY, and KA. Database construction was directed by KY and performed by AS, SK, and K. KA wrote the paper. All authors approved the final manuscript.

Received: 23 September 2009 Accepted: 30 March 2010
Published: 30 March 2010

References
1. Knaap S: Tobacco to tomatoes: a phylogenetic perspective on fruit diversity in the Solanaceae. J Exp Bot 2002, 53:2001-2022.
2. Giovannoni JJ: Genetic regulation of fruit development and ripening. Plant Cell 2004, 16(Suppl)S170-180.
3. Pedley KF, Martin GB: Molecular basis of Pto-mediated resistance to bacterial speck disease in tomato. Annu Rev Phytopathol 2003, 41:215-243.
4. Mueller LA, Tankley SD, Giovannoni JJ, van Eck J, Stack S, Choi D, Kim BD, Chen M, Cheng Z, Li C, et al: The Tomato Sequencing Project, the First Cornerstone of the International Solanaceae Project (SOL). Comp Funct Genomics 2005, 6:153-158.
5. Mueller LA, Solow TH, Taylor N, Skwarecki B, Bueler R, Binns J, Lin C, Wright NH, Ahrens R, Wang Y, et al: The SOL Genomics Network: a comparative resource for Solanaceae biology and beyond. Plant Physiol 2005, 138:1310-1317.
6. Hoeven Van der R, Ronning C, Giovannoni J, Martin G, Tankley S: Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. Plant Cell 2002, 14:1441-1456.
7. Alba R, Fei Z, Payton P, Liu Y, Moore SL, Debbie P, Cohn J, DAscenzo M, Gordon JS, Rose JK, et al ESTs, cDNA microarrays, and gene expression

Additional file 1: Tom tissues used for RNA preparation. Detailed description of Micro-Tom tissues used for RNA preparation including tissue positions, treatments, and age of plants when tissues were harvested.

Additional file 2: Clone number list. Lists of clone numbers of nrFLcDNAs-sets appeared in the text.

Additional file 3: cDNA derived from non-coding RNA. Clone numbers and annotations of nrFLcDNAs that matched to non-coding RNAs registered in NONCODE database.

Additional file 4: Predicted CDS of 11,597 non-redundant HTCs. A list of amino acid sequences predicted for 11,597 non-redundant HTCs and results of BLASTP search against protein datasets of nr and tomato SBD database http://www.kazusa.or.jp/tomato/.

Additional file 5: Datasets used for comparative analysis with other plants. Name and version of protein- and tentative consensus-datasets used for comparison of nrFLcDNAs with gene of other plants.

Abbreviations
HTC: high throughput cDNA sequence; (nrFLcDNA): non-redundant full-length cDNA; (CDS): coding sequence; (UTR): untranslated region; (DFCI): Dana-Farber Cancer Institute, (SGN): SOL genomics network; (TC): tentative consensus.

BMC Genomics 2010, 11:210
http://www.biomedcentral.com/1471-2164/11/210
profiling: tools for dissecting plant physiology and development. Plant J 2004, 39:697-714.
8. Fei Z, Tang X, Alba RM, White JA, Ronning CM, Martin GB, Tanksley SD, Giovannini JJ. Comprehensive EST analysis of tomato and comparative genomics of fruit ripening. Plant J 2004, 40:47-59.
9. Yamamoto N, Tsugane T, Watanabe M, Yano K, Maeda F, Kusawa C, Torki M, Ban Y, Nishimura S, Shibata D. Expressed sequence tags from the laboratory-grown mini tomato cultivar (Lycopersicon esculentum) Micro-Tom and mining for single nucleotide polymorphisms and insertions/deletions in tomato cultivars. Gene 2005, 356:127-134.
10. Yano K, Tsugane N, Yamamoto N, Tsugane T, Aoki K, Sakurai N, Shibata D. MiBASE: a database of a mini tomato cultivar Micro-Tom. Plant Biotechnol 2006, 23:195-198.
11. Dong Q, Lawrence CJ, Schlueuter SD, Willerson MD, Kurtz S, Lushbough C, Brendel V. Comparative plant genome resources at PlantGDB. Plant Physiol 2005, 139:610-618.
12. D’Agostino N, Aversano M, Fruscante L, Chusano ML. TomEST database: in silico exploitation of EST data to explore expression patterns in tomato species. Nucleic Acids Res 2007, 35:D901-905.
13. Scott JW, Harbaugh BK. Micro-Tom-a miniature dwarf tomato. Florida Agr Exp Sta Circ 1989, 370:1-6.
14. Meissner R, Jacobson Y, Melamed S, Levy AA. A new model system for tomato genetics. Plant J 1997, 12:1465-1472.
15. Shibata D. Genome sequencing and functional genome approaches in tomato. J Gen Plant Biol 2005, 71:1-7.
16. Matsuura C, Aoki K, Fukuda N, Mizoguchi T, Asamizu E, Saito T, Shibata D, Ezura H. Comprehensive resources for tomato functional genomics based on the mini model tomato model micro-tom. Curr Genomics 2008, 9:435-443.
17. Kato S, Ohto ko K, Ohtake H, kimura T. Vector-capping: a simple method for preparing a high-quality full-length cDNA library. DNA Res 2005, 12:53-62.
18. Caminci P, Kvar C, Kitamura A, Ohsumi T, Okazaki Y, Itoh M, Kamiya M, Shibata K, Sasaki N, Iwasa M, et al. High-throughput full-length cDNA cloning by biotinylated CAP trapper. Genomics 1996, 37:337-336.
19. He S, Liu C, Skogerbo G, Zhao H, Wang J, Liu T, Bai B, Zhao Y, Chen R. NONCODE v2.0: decoding the non-coding. Nucleic Acids Res 2008, 36:D170-172.
20. Gouzy J, Carrere S, Schiex T. Vector-capping: a simple method for preparing a high-quality full-length cDNA library. DNA Res 2005, 12:1465-1472.
21. Gouzy J, Carrere S, Schiex T. FrameDP: sensitive peptide detection on noisy matured sequences. Bioinformatics 2009, 25:670-671.
22. Lomazade A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res 2005, 33:6494-6506.
23. Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HC, Kim C, Nguyen M, et al. Empirical analysis of transcriptional activity in the Arabidopsis genome. Science 2003, 302:842-845.
24. Umemura T, Sakurai T, Totoki Y, Aoki K, Shiba T, Ishihara A, Akiyama K, Kurotani A, Yoshida T, Mochida K, et al. Sequence analysis and expression of approximately 40,000 soybean cDNA clones from a full-length-enriched cDNA library. DNA Res 2008, 15:333-346.
25. Alexandrov NN, Troubhan ME, Brover VV, Tatarinova T, Flavell RB, Feldmann KA. Features of Arabidopsis genes and genome discovered using full-length cDNAs. Plant Mol Biol 2006, 60:669-685.
26. Ralph SG, Chun H, Cooper D, Kirttapat R, Kolosova N, Gunter L, Tuskan GA, Douglas CJ, Holt RA, Jones SJ, et al. Analysis of 4,664 high-quality sequence-finished poplar full-length cDNA clones and their utility for the discovery of genes responding to insect feeding. BMC Genomics 2008, 9:57.
27. Jia J, Fu J, Zheng J, Zhou X, Huai J, Wang J, Wang M, Zhang Y, Chen X, Zhang J, et al. Annotation and expression profile analysis of 2073 full-length cDNAs from stress-induced maize (Zea mays L.) seedlings. Plant J 2006, 48:710-727.
28. Alexandrov NN, Brover VV, Freidin S, Troubhan ME, Tatarinova TV, Zhang H, Swaller TJ, Lu YP, Bouck J, Flavell RB, et al. Insights into corn genes derived from large-scale cDNA sequencing. Plant Mol Biol 2009, 69:179-194.
29. Mignone F, Gissi C, Liuni S, Pesole G. Untranslated regions of mRNAs. Genome Biol 2002, 3:REVIEWS5004.
30. Iida K, Seki M, Sakurai T, Satou M, Akiyama K, Toyoda T, Konagaya A, Shinozaki K. Genome-wide analysis of alternative pre-mRNA splicing in Arabidopsis thaliana based on full-length cDNA sequences. Nucleic Acids Res 2004, 32:5096-5103.
31. Kikuchi S, Satoh K, Nagata T, Kawagishira N, Doi K, Kishimoto N, Yatsuki J, Ishikawa M, Yamada H, Okka H, et al. Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. Science 2003, 301:376-379.
32. Nan-Gon H, Levitan N, Rubin E, Fluhir R. Comparative cross-species alternative splicing in plants. Plant Physiol 2007, 144:1632-1641.
33. Dauvillier RV, Sun H, Palanisamy SK, Matthews N, Molina C, Kurtz M, Grotewold E. AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. BMC Bioinformatics 2003, 4:35.
34. Rechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adair L, Pineda O, Ratcliffe DJ, Samaha RR, et al. Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science 2000, 290:2105-2110.
35. Takahashi H, Shiina A, Arie T, Rosmalawati S, Fukushima S, Kikuchi M, Hikita Y, Kanda A, Takahashi A, Kiba A, et al. Catalog of Micro-Tom tomato responses to common fungal, bacterial, and viral pathogens. J Gen Plant Pathol 2005, 71:18-22.
36. Tsugane T, Watanabe M, Yano K, Sakurai N, Suzuki H, Shibata D. Expressed sequence tags of full-length cDNA clones from the mini tomato (Lycopersicon esculentum) cultivar Micro-Tom. Plant Biotechnol 2005, 22:161-165.
37. Iijima Y, Nakamura Y, Ogata Y, Tanaka K, Sakurai N, Suda K, Suzuki T, Suzuki H, Okazaki K, Katayama M, et al. Metabolite annotations based on the integration of mass spectral information. Plant J 2008, 54:949-962.
38. Chomczynski P, Sacchi N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem 1987, 162:156-159.
39. Ewing B, Hillier L, Wendle MC, Green P. Base-calling of automated sequencing traces using phred. I. Accuracy assessment. Genome Res 1998, 8:175-185.
40. Green P. Phrap documentation. 1996.
41. Rozen S, Skalicky H. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 2000, 132:365-386.
42. Mott R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput Appl Biosci 1997, 13:477-478.
43. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R, Proctor Scan: protein domains identifier. Nucleic Acids Res 2005, 33:W116-120.
44. Bansell B, Omer D, Hurstley RP, Binns D, O’Donovan C, Apweiler R. The GOA database in 2009—an integrated Gene Ontology Annotation resource. Nucleic Acids Res 2009, 37:D396-403.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at www.biomedcentral.com/submit