Genome-Wide Analysis of Intraspecific DNA Polymorphism in ‘Micro-Tom’, a Model Cultivar of Tomato (Solanum lycopersicum)

Masaaki Kobayashi1,8, Hideki Nagasaki2,8, Virginie Garcia3, Daniel Just3, Cécile Bres3, Jean-Philippe Mauxion3, Marie-Christine Le Paslier4, Dominique Brunel4, Kunihiro Suda5, Yohei Minakuchi6, Atsushi Toyoda6, Asao Fujiyama2, Yasukazu Nakamura2, Kentaro Yano1 and Koh Aoki5,7,*

1Faculty of Agriculture, Meiji University, 1-1-1 Higashi Mita, Tama-ku, Kawasaki, 214-8571 Japan
2Genome Informatics Laboratory, National Institute of Genetics, 1-111 Yata, Mishima, 411-8540 Japan
3INRA, Université de Bordeaux, UMR 1332 Biologie du Fruit et Pathologie, 33140 Villenave d’Ornon, France
4Unité Etude du Polymorphisme des Génomes Végétaux, CEA-Institut de Génomique-CNG, INRA, UR1279, 91057 Evry, France
5Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamata, Kisarazu, 292-0818 Japan
6Comparative Genomics Laboratory, National Institute of Genetics, 1111 Yata, Mishima, 411-8540 Japan
7Graduate School of Life and Environmental Sciences, Osaka Prefecture University, 1-1 Gakuen-cho, Naka-ku, Sakai, 599-8531 Japan
8These authors contributed equally to this work.

*Corresponding author: E-mail, kaoki@plant.osakafu-u.ac.jp; Fax: +81-72-254-9918

(Received October 8, 2013; Accepted November 26, 2013)

Tomato (Solanum lycopersicum) is regarded as a model plant of the Solanaceae family. The genome sequencing of the tomato cultivar ‘Heinz 1706’ was recently completed. To accelerate the progress of tomato genomics studies, systematic bioresources, such as mutagenized lines and full-length cDNA libraries, have been established for the cultivar ‘Micro-Tom’. However, these resources cannot be utilized to their full potential without the completion of the genome sequencing of ‘Micro-Tom’. We undertook the genome sequencing of ‘Micro-Tom’ and here report the identification of single nucleotide polymorphisms (SNPs) and insertions/deletions (indels) between ‘Micro-Tom’ and ‘Heinz 1706’. The analysis demonstrated the presence of 1.23 million SNPs and 0.19 million indels between the two cultivars. The density of SNPs and indels was high in chromosomes 2, 5 and 11, but was low in chromosomes 6, 8 and 10. Three known mutations of ‘Micro-Tom’ were localized on chromosomal regions where the density of SNPs and indels was low, which was consistent with the fact that these mutations were relatively new and introgressed into ‘Micro-Tom’ during the breeding of this cultivar. We also report SNP analysis for two ‘Micro-Tom’ varieties that have been maintained independently in Japan and France, both of which have served as standard lines for ‘Micro-Tom’ mutant collections. Approximately 28,000 SNPs were identified between these two ‘Micro-Tom’ lines. These results provide high-resolution DNA polymorphic information on ‘Micro-Tom’ and represent a valuable contribution to the ‘Micro-Tom’-based genomics resources.

Keywords: Genome • Indels • Micro-Tom • SNPs • Solanum lycopersicum • Tomato.

Abbreviations: BAC, bacterial artificial chromosome; EMS, ethylmethane sulfonate; indels, insertion and deletions; LTR, long terminal repeat; NBS-LRR, nucleotide-binding site leucine-rich repeat; SNP, single nucleotide polymorphism; UTR, untranslated region.

Introduction

Tomato (Solanum lycopersicum) is regarded as a model plant of the Solanaceae family, which comprises 1,000–2,000 species (Knapp 2002). Tomato is also regarded as a model plant for studying fruit development (Giovannoni 2004). Since many of the members of the Solanaceae family share highly syntenic genomes composed of 12 chromosomes, further determination of sequences will allow the implementation of rational strategies for improving Solanaceae crops that are important in the human food supply.
Sequencing of *S. lycopersicum* cultivar ‘Heinz 1706’ (Mueller et al. 2005) was completed in 2012 by an international collaboration (Tomato Genome Consortium 2012). Comparative analysis of the ‘Heinz 1706’ reference genomic sequence provided as SL2.40 and sequences of other model plants has revealed the evolutionary trajectory of the tomato genome (Tomato Genome Consortium 2012). Furthermore, the availability of the ‘Heinz 1706’ reference genomic sequence has made it possible to link DNA polymorphisms to phenotypic differences observed in cultivars of the species *S. lycopersicum*.

Among *S. lycopersicum* cultivars, ‘Micro-Tom’ is regarded as an emerging model system for tomato research. ‘Micro-Tom’ has phenotypic features that fulfill the requirements for a model organism, such as a short life cycle, small size and high efficiency of genetic transformation (Meissner et al. 2000, Matsukura et al. 2008). To improve the utility of ‘Micro-Tom’ as a model genomics system, several ‘Micro-Tom’-based genomic resources have been developed, including transposon tagging lines (Meissner et al. 2000), ethylmethane sulfonate (EMS)-mutagenized lines (Meissner et al. 2000), near-isogenic lines (Carvalho et al. 2011a) and a full-length cDNA collection (Aoki et al. 2010). However, the applicability of ‘Micro-Tom’ genomic resources to other commercial tomato cultivars may be limited by the genetic difference between ‘Micro-Tom’ and other commercial tomato cultivars (Shirasawa et al. 2010, Hirakawa et al. 2013). Using 7,054 single nucleotide polymorphism (SNP) markers, 1,624 loci showed polymorphisms between ‘Micro-Tom’ and ‘Heinz 1706’, and 455 loci showed polymorphisms among five lines of ‘Micro-Tom’ (Hirakawa et al. 2013).

‘Micro-Tom’ genomics resources have been independently developed in several countries (Meissner et al. 2000, Dan et al. 2007, Ariizumi et al. 2011, Carvalho et al. 2011b), including EMS-mutagenized lines developed by the National Biosource Project (NBRP) of Japan and by the French National Institute for Agricultural Research (INRA). Previous marker-based genotyping studies have demonstrated that the numbers of polymorphic loci detected between ‘Micro-Tom’ lines maintained in Japan and France cannot be ignored (Shirasawa et al. 2010, Hirakawa et al. 2013). Since the ‘Micro-Tom’ lines of Japan and France serve as the parental lines of EMS-mutagenized lines developed in each country, it is important to understand the DNA polymorphisms that are present between these parental lines as these differences may introduce unnecessary complexity into EMS-mutagenized lines shared among scientists. Thus, there has been demand from researchers to undertake a comprehensive study to identify the DNA polymorphisms found in the entire genome of ‘Micro-Tom’. However, prior to the present study, the genome of ‘Micro-Tom’ had not been sequenced.

In the present study, we sequenced the genome of ‘Micro-Tom’ using high throughput sequencing platforms, and we identified SNPs and insertions and deletions (indels) that correspond to SNPs and indels in the ‘Heinz 1706’ reference genome. We identified 1.23 million SNPs and 0.19 million indels. The density of SNPs and indels differed from chromosome to chromosome, being higher in the heterochromatic regions of chromosomes 2, 5 and 11, and lower in chromosomes 6, 8 and 10. The majority of the SNPs were localized in the intergenic regions and most were frequently found in retro-transposon-type repeat sequences. We also report polymorphic analysis of two ‘Micro-Tom’ lines independently maintained in Japan and France. This will be of help in identifying causal mutations of phenotypic changes observed in mutagenized lines of ‘Micro-Tom’.

**Results**

**Genome sequencing of ‘Micro-Tom’ and mapping onto the ‘Heinz 1706’ reference genome**

The ‘Micro-Tom’ genome was sequenced using both the Illumina and 454 platforms. Using the Illumina platform, we obtained 1,301,590,508 paired-end 100 bp reads. Using the 454 platform, we obtained 24,131,610 single-end reads and 9,174,806 paired-end reads. Altogether, we were able to sequence 145 Gbp, which accounts for approximately 185× the coverage of the ‘Heinz 1706’ reference genome sequence, or SL2.40. Approximately 52% of the total reads were mapped to a single unique position in SL2.40. The rest of the reads were mapped to multiple loci and were not used for the identification of SNPs and indels. The obtained reads covered a total of 99.1% (730.7 Mbp) of the readable bases of SL2.40 (737.6 Mbp). A consensus sequence of ‘Micro-Tom’ was obtained from this mapping.

**Identification of SNPs**

To clarify the number of reads sufficient for detecting nucleotide differences between ‘Micro-Tom’ and ‘Heinz 1706’, we examined the SNPs detected with a varied number of input reads. We found that the number of detected SNPs started to become saturated for an input of 80× reads (Supplementary Fig. S1). To reduce the rate of false-positive SNP detection, we split Illumina reads into two non-overlapping groups, and each of these was combined with the 454 reads to obtain two data sets (Fig. 1). SNPs were first detected using these two combined data sets separately, and a total of 1,231,191 SNPs were detected in both data sets (Supplementary Table S1; a vcf file is available as Supplementary Table S7).

To estimate the accuracy of SNP detection, we compared the base calls of 1,231,191 SNPs from ‘Micro-Tom’ bacterial artificial chromosome (BAC)-end sequences obtained using the Sanger sequencing method (Asamizu et al. 2012). The SNPs matched the base calls of the BAC-end sequences of 99.2% of the bases. We used this set of 1,231,191 SNPs (Supplementary Table S1) in further analyses.

**Distribution of SNPs**

The average density of SNPs was 1.62 SNPs kb⁻¹. The density of SNPs differed from chromosome to chromosome...
CDS SNPs were identified in 9,496 loci of 3,464 genes. SNP density in intragenic regions was higher in chromosomes 1, 4, 7, 8, 9, 10 and 12, whereas the SNP density in intergenic regions was higher than in the intragenic regions. Interestingly, the SNP-rich regions of euchromatin were located in pericentromeric heterochromatin. In contrast, the SNP-rich windows of chromosomes 2, 5 and 11 are located in gene-sparse regions of euchromatin.

**Annotation of SNPs**

We estimated the number and density of SNPs between 'Micro-Tom' and 'Heinz 1706' in intergenic and intragenic regions (Table 1). Most of the total SNPs were identified in intergenic regions. In chromosomes 2, 3, 5, 6 and 11, SNP density in the intergenic regions was higher than in the intragenic regions. On the other hand, chromosomes 1, 4, 7, 8, 9, 10 and 12 showed higher SNP density in intragenic regions.

Examination of SNP density in the 5'-untranslated regions (UTRs), exons, introns and 3'-UTRs of intragenic regions (Table 2) showed lower SNP density in exons than in introns. To estimate the extent to which SNPs in the exons caused amino acid substitutions, we investigated the number of synonymous and non-synonymous substitutions in coding sequences (CDS) that do not contain indels (Table 3). CDS SNPs were identified in 9,496 loci of 3,464 genes.

Non-synonymous SNP density per unit CDS length was higher than average on chromosomes 2, 4, 5 and 11. Functional description specifically enriched in genes containing non-synonymous SNPs included nucleotide-binding site leucine-rich repeat (NBS-LRR) resistance protein (NBS-LRR) resistance protein. The majority (16.4%) of the repeat sequences containing SNPs were classified as retrotransposons, such as gypsy (Schmidt 1999), copia (Schmidt 1999), long terminal repeats (LTRs) (Schmidt 1999) and RAM12 (Shankar and Jurka 2006a). Repeat classification also demonstrated that SNPs in retrotransposons were

**Table 1** Number and density of SNPs by chromosome between 'Micro-Tom' and 'Heinz 1706'

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Number of SNPs (density of SNPs, kb⁻¹)</th>
<th>Intergenic</th>
<th>Intragenic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>48,014 (0.53)</td>
<td>39,648 (0.53)</td>
<td>8,366 (0.54)</td>
</tr>
<tr>
<td>2</td>
<td>166,837 (3.34)</td>
<td>158,421 (4.05)</td>
<td>8,416 (0.78)</td>
</tr>
<tr>
<td>3</td>
<td>87,760 (1.35)</td>
<td>74,638 (1.39)</td>
<td>13,122 (1.20)</td>
</tr>
<tr>
<td>4</td>
<td>81,978 (1.28)</td>
<td>63,294 (1.16)</td>
<td>18,684 (1.99)</td>
</tr>
<tr>
<td>5</td>
<td>465,671 (7.16)</td>
<td>452,240 (7.83)</td>
<td>13,431 (1.84)</td>
</tr>
<tr>
<td>6</td>
<td>12,742 (0.28)</td>
<td>11,532 (0.31)</td>
<td>1,210 (0.13)</td>
</tr>
<tr>
<td>7</td>
<td>61,399 (0.94)</td>
<td>40,450 (0.71)</td>
<td>20,949 (2.58)</td>
</tr>
<tr>
<td>8</td>
<td>14,314 (0.23)</td>
<td>11,583 (0.21)</td>
<td>2,731 (0.33)</td>
</tr>
<tr>
<td>9</td>
<td>23,835 (0.35)</td>
<td>17,351 (0.29)</td>
<td>6,484 (0.81)</td>
</tr>
<tr>
<td>10</td>
<td>8,534 (0.13)</td>
<td>6,157 (0.11)</td>
<td>2,377 (0.35)</td>
</tr>
<tr>
<td>11</td>
<td>187,476 (3.51)</td>
<td>174,224 (3.80)</td>
<td>13,252 (1.75)</td>
</tr>
<tr>
<td>12</td>
<td>36,187 (0.55)</td>
<td>31,972 (0.55)</td>
<td>4,215 (0.59)</td>
</tr>
<tr>
<td>13</td>
<td>36,444 (1.67)</td>
<td>35,571 (1.72)</td>
<td>873 (0.77)</td>
</tr>
<tr>
<td>Total</td>
<td>1,231,191 (1.62)</td>
<td>1,177,081 (1.66)</td>
<td>114,110 (1.05)</td>
</tr>
</tbody>
</table>

Chromosome 00 is a scaffold that is not anchored to tomato chromosomes.

Non-synonymous SNP density per unit CDS length was higher than average on chromosomes 2, 4, 5 and 11. Functional description specifically enriched in genes containing non-synonymous SNPs included nucleotide-binding site leucine-rich repeat (NBS-LRR) resistance protein (NBS-LRR) resistance protein. The majority (16.4%) of the repeat sequences containing SNPs were classified as retrotransposons, such as gypsy (Schmidt 1999), copia (Schmidt 1999), long terminal repeats (LTRs) (Schmidt 1999) and RAM12 (Shankar and Jurka 2006a). Repeat classification also demonstrated that SNPs in retrotransposons were

**Table 2** Distribution of SNPs in the Micro-Tom genome sequence

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Total Intergenic Intragenic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>48,014 (0.53)</td>
</tr>
<tr>
<td>2</td>
<td>166,837 (3.34)</td>
</tr>
<tr>
<td>3</td>
<td>87,760 (1.35)</td>
</tr>
<tr>
<td>4</td>
<td>81,978 (1.28)</td>
</tr>
<tr>
<td>5</td>
<td>465,671 (7.16)</td>
</tr>
<tr>
<td>6</td>
<td>12,742 (0.28)</td>
</tr>
<tr>
<td>7</td>
<td>61,399 (0.94)</td>
</tr>
<tr>
<td>8</td>
<td>14,314 (0.23)</td>
</tr>
<tr>
<td>9</td>
<td>23,835 (0.35)</td>
</tr>
<tr>
<td>10</td>
<td>8,534 (0.13)</td>
</tr>
<tr>
<td>11</td>
<td>187,476 (3.51)</td>
</tr>
<tr>
<td>12</td>
<td>36,187 (0.55)</td>
</tr>
<tr>
<td>13</td>
<td>36,444 (1.67)</td>
</tr>
<tr>
<td>Total</td>
<td>1,231,191 (1.62)</td>
</tr>
</tbody>
</table>
more abundant in chromosomes 2, 5 and 11 than in the other chromosomes ($\chi^2$ test, $P < 0.01$) (Fig. 4). On the other hand, the percentage of DNA transposons was higher in chromosomes 7, 9 and 10 ($\chi^2$ test, $P < 0.01$) (Fig. 4).

**Distribution analysis and annotation of indels**

We detected 117,354 insertions and 70,124 deletions in ‘Micro-Tom’ (Table 4; Supplementary Table S3; a vcf file is available as Supplementary Table S8). We classified the location of indels into intergenic regions, 5'-UTRs, exons, introns and 3'-UTRs (Supplementary Table S4). Like SNPs, indels were abundant in the intergenic regions of chromosomes 2, 5 and 11. The exons of chromosomes 4 and 7 contained a relatively large number of indels. The indels caused changes in the amino acid sequences of 1,284 genes (Supplementary Table S3). Functional categories of genes that underwent amino acid changes were enriched with NBS-LRR resistance protein ($\chi^2$ test, $P < 0.01$). A survey of the indel length distribution (Supplementary Fig. S3) showed that the length of indels ranged from a 56 bp deletion to a 53 bp insertion. The majority of the indels were short; 69% of the indels had a length of 1 bp, and indels shorter than 5 bp accounted for $>90\%$ of the indels.
Presence of SNPs and indels that determine the phenotype of 'Micro-Tom'

Previous reports identified mutations in three genes that determine the morphological phenotype of 'Micro-Tom' (Pnueli et al. 1998, Marti et al. 2006, Powell et al. 2012): DWARF, SELF-PRUNING and SlGLK2. We confirmed that these mutations are present in the 'Micro-Tom' genome sequences obtained in this study. The dwarf mutation was identified in the DWARF gene (Solyc02g089160) that encodes Cyt P450 protein, or deoxocastasterone oxidase, as a single base substitution in the 3'0 end of intron 8 (Marti et al. 2006). The corresponding locus, SL2.40ch02:45624583, is a T in 'Micro-Tom' and an A in 'Heinz 1706' (Supplementary Table S1).

Cyt P450 protein, or deoxocastasterone oxidase, as a single base substitution in the 3' end of intron 8 (Marti et al. 2006). The corresponding locus, SL2.40ch02:45624583, is a T in 'Micro-Tom' and an A in 'Heinz 1706' (Supplementary Table S1). The self-pruning mutation was identified in the SELF-PRUNING gene (Solyc06g074350) as a non-synonymous single
Since these mutation-sparse regions are conserved between 'Micro-Tom' and 'Micro-Tom MM', this observation is consistent with the mutations being introgressed into 'Micro-Tom' by recent breeding. The uniform ripening mutation was identified as a single base insertion in the SIGL2 gene. The corresponding locus, SL2.40ch10:2292346, is an A in 'Micro-Tom', but is absent in dark green shouldered varieties, or varieties having green patches at the top of the fruit (Powell et al. 2012). 'Heinz 1706' has an insertion of an A since 'Heinz 1706' also carries the uniform ripening mutation.

SNPs between Japanese and French 'Micro-Tom' lines

To clarify the extent to which Japanese and French 'Micro-Tom' sequences differ, we performed genome-wide detection of SNPs in common between 'Micro-Tom' and 'Micro-Tom MM'. A whole-genome sequencing of 'Micro-Tom MM' was performed using the Illumina platform. We obtained 378,542,690 paired-end 101 bp reads and 198,384,555 single 101 bp reads, representing 58 Gbp. Of these, 496,683,927 reads were reliably mapped to the consensus (mapping quality > 17). We detected 28,358 SNPs in common between 'Micro-Tom' and 'Micro-Tom MM' (Supplementary Table S5). SNPs were detected more frequently on chromosomes 4, 6 and 7 than the other chromosomes (Table 5), and the intragenic SNP density was 2.8-fold higher than the intergenic SNP density. Occurrences of SNPs between 'Micro-Tom' and 'Micro-Tom MM' have shown a bias toward intragenic regions (Table 5, Supplementary Table S5), and the intragenic SNP density was 2.8-fold higher than the intergenic SNP density. Occurrences of SNPs between 'Micro-Tom' and 'Micro-Tom MM' have shown a bias toward intragenic regions (Table 5, Supplementary Table S5), and the intragenic SNP density was 2.8-fold higher than the intergenic SNP density. Occurrences of SNPs between 'Micro-Tom' and 'Micro-Tom MM' have shown a bias toward intragenic regions (Table 5, Supplementary Table S5), and the intragenic SNP density was 2.8-fold higher than the intergenic SNP density.
regions than in other parts of the ‘Micro-Tom’ genome where the ratio of SNPs to retrotransposons, including gypsy, copia, LTRs, TOPSCOTCH and TS2 (Shankar and Jurka 2006d), was higher ($\chi^2$ test, $P < 0.01$) (Supplementary Fig. S5).

**Discussion**

We performed two analyses of the genome sequences of ‘Micro-Tom’ in comparison with other *S. lycopersicum* varieties: one was cultivar ‘Heinz 1706’ and the other was the ‘Micro-Tom’ variety maintained in France, namely ‘Micro-Tom MM’ (Dan et al. 2007, Shirasawa et al. 2010). Comparison of DNA sequences on a genome-wide scale provides an opportunity to approach the question of to what extent the ‘Heinz 1706’ reference genome sequence can be utilized when working with ‘Micro-Tom’ genomics resources.

The comparison of ‘Micro-Tom’ with the ‘Heinz 1706’ genome sequence demonstrated that there are >1.23 million SNPs and 0.19 million indels. The comparison with ‘Micro-Tom MM’, on the other hand, identified 28,358 SNPs. This confirms the hypothesis proposed previously (Shirasawa et al. 2010) that multiple lines might have been selected as ‘Micro-Tom’ from the original breeding population before the genotype of each line became fixed. Both ‘Micro-Tom’ and ‘Micro-Tom MM’ were descended from the original Micro-Tom bred at Florida University (Scott and Harbaugh 1989). In Japan and France, ‘Micro-Tom’ and ‘Micro-Tom MM’ were maintained by self-fertilization. Thus, the presence of 28,358 SNPs may indicate the degree of heterozygosity in the original Micro-Tom line.

Approximately 97% of the SNPs and indels between ‘Micro-Tom’ and ‘Heinz 1706’ were detected in non-coding regions, including intergenic regions, introns, 5’-UTRs and 3’-UTRs (Tables 1, 2). This is consistent with the results of SNP identification using ‘Micro-Tom’ BAC-end sequences mapped onto the ‘Heinz 1706’ genome SL2.40 (Asamizu et al. 2012). Approximately 72% of all SNPs (858,825 SNPs) were identified in the repeat sequences. Classification of the repeat type revealed that, although 77% of the SNPs were localized in repeat sequences with no matches to known repeat patterns, repeat SNPs were enriched in retrotransposons, such as gypsy, copia and LTRs. The SNPs in retrotransposons were concentrated in the heterochromatic regions of chromosomes 2, 5 and 11. On the other hand, SNPs and indels in the CDS caused changes in the amino acid sequences of 4,082 genes. Amino acid changes were most frequently predicted in genes coding for the NBS-LRR resistance protein, which is involved in the detection of diverse biological attacks on plants, such as those of pathogens, fungi, oomycetes, nematodes, parasitic plants and insects (DeYoung and Innes 2006). The LRR domain is involved in the perception of ligands, and tends to be highly variable (McHale et al. 2006, Marone et al. 2013). Thus, the presence of non-synonymous SNPs in genes encoding NBS-LRR resistance proteins confers divergence with respect to ligand perception.

We previously compared the chromosomal SNP distribution profile of ‘Micro-Tom’ with those of 40 domesticated tomato cultivars (Hirakawa et al. 2013). SNP density values were normalized by the total number of SNPs in each chromosome (Supplementary Fig. S6). The comparison of the SNP distribution profile between ‘Micro-Tom’ and ‘Heinz 1706’ (present study) showed maximum differences in different chromosomal positions found in the comparisons of cultivars other than ‘Micro-Tom’ or ‘Heinz 1706’ (Hirakawa et al. 2013) (Supplementary Fig. S6). In many regions, SNP density between ‘Micro-Tom’ and ‘Heinz 1706’ was found to be high in regions where SNP density in comparisons between other cultivars was low, or vice versa. Although this discrepancy is partly due to the fact that Hirakawa et al. (2013) used SolCAP SNPs (Hamilton et al. 2012) based on RNA sequencing, it illustrates that the degree of polymorphism between ‘Micro-Tom’ and ‘Heinz 1706’ is limited compared with the total diversity observed among all of cultivated tomatoes. This also suggests that the development of new SNP markers is possible in regions where DNA markers are currently few.

Three genes that code for known mutations that characterize the ‘Micro-Tom’ phenotype, dwarf, self-pruning and uniform ripening, are localized in chromosomal regions where SNP density and indel density are relatively low (Fig. 5). An SNP-dense region can be regarded as being a genetically distant region, since it takes a long time to accumulate single-nucleotide mutations. Three mutations that are localized in genetically close regions between ‘Micro-Tom’ and ‘Heinz 1706’ thus are thought to be new mutations. This is possibly attributed to the fact that the mutations were introgressed into ‘Micro-Tom’ recently during the breeding of the cultivar.

In the comparison between ‘Micro-Tom’ and ‘Micro-Tom MM’, 91% of the SNPs identified were found in non-coding

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Total SNPs (density of SNPs, kb$^{-1}$)</th>
<th>Intergenic SNPs</th>
<th>Intragenic SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of SNPs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chromosome 1</td>
<td>263 (0.0029)</td>
<td>218 (0.0029)</td>
<td>45 (0.0029)</td>
</tr>
<tr>
<td>Chromosome 2</td>
<td>1,296 (0.026)</td>
<td>1,050 (0.0027)</td>
<td>246 (0.023)</td>
</tr>
<tr>
<td>Chromosome 3</td>
<td>1,291 (0.020)</td>
<td>811 (0.015)</td>
<td>480 (0.043)</td>
</tr>
<tr>
<td>Chromosome 4</td>
<td>4,631 (0.072)</td>
<td>2,707 (0.050)</td>
<td>1,924 (0.20)</td>
</tr>
<tr>
<td>Chromosome 5</td>
<td>2,810 (0.043)</td>
<td>2,423 (0.042)</td>
<td>387 (0.053)</td>
</tr>
<tr>
<td>Chromosome 6</td>
<td>3,740 (0.081)</td>
<td>2,415 (0.065)</td>
<td>1,325 (0.15)</td>
</tr>
<tr>
<td>Chromosome 7</td>
<td>9,275 (0.14)</td>
<td>6,131 (0.11)</td>
<td>3,144 (0.39)</td>
</tr>
<tr>
<td>Chromosome 8</td>
<td>384 (0.0061)</td>
<td>285 (0.0052)</td>
<td>99 (0.012)</td>
</tr>
<tr>
<td>Chromosome 9</td>
<td>155 (0.0022)</td>
<td>114 (0.0019)</td>
<td>41 (0.0051)</td>
</tr>
<tr>
<td>Chromosome 10</td>
<td>764 (0.012)</td>
<td>715 (0.012)</td>
<td>49 (0.0073)</td>
</tr>
<tr>
<td>Chromosome 11</td>
<td>575 (0.011)</td>
<td>499 (0.011)</td>
<td>76 (0.010)</td>
</tr>
<tr>
<td>Chromosome 12</td>
<td>2,358 (0.036)</td>
<td>1,405 (0.024)</td>
<td>953 (0.13)</td>
</tr>
<tr>
<td>Chromosome 13</td>
<td>816 (0.037)</td>
<td>769 (0.037)</td>
<td>47 (0.0042)</td>
</tr>
<tr>
<td>Total</td>
<td>28,358 (0.037)</td>
<td>19,542 (0.029)</td>
<td>8,816 (0.081)</td>
</tr>
</tbody>
</table>
regions (Table 5) and 50% of the SNPs were localized in repeat sequences. The repeat SNPs were enriched in DNA transposons. The percentage of DNA transposons that contained SNPs was 5.1% of all SNPs containing repeats, which was higher than the ratio of the DNA transposon number to the total repeat number (1.9%). These SNPs in DNA transposons were concentrated in the euchromatic regions of chromosomes 4, 6, 7 and 12 (Supplementary Fig. S4). This result suggests that DNA transposons contributed to the genetic heterogeneity in the original ‘Micro-Tom’ line.

‘Micro-Tom’ is regarded as a model system for tomato research, and ‘Micro-Tom’-based mutagenized lines have been developed in several research communities. Among these, the EMS-mutagenized lines developed by the NBRP Tomato (Japan) and the French INRA, Bordeaux (France) collectively serve as the largest mutant resources for tomato genomics. The DNA polymorphism detected in this study between ‘Micro-Tom’ (the Japanese standard line) and ‘Micro-Tom MM’ (the French standard line) will provide fixed reference points for the identification of genes responsible for important traits in tomato using these ‘Micro-Tom’ mutant resources.

Materials and Methods

Plant materials

Tomato (Solanum lycopersicum) cultivar ‘Micro-Tom’ (TOMJPF00001) seeds were obtained from the NBRP Tomato (MEXT, Japan, http://tomato.nbrp.jp/indexEn.html) (Ariizumi et al. 2011). The original Micro-Tom inbred line (Scott and Harbaugh 1989) was self-fertilized for six generations to obtain ‘Micro-Tom’ (TOMJPF00001). ‘Micro-Tom’ plants were grown on culture medium [1/2× MS medium (Murashige and Skoog 1962), 3% (w/v) sucrose, 0.8% (w/v) agar] in sterile conditions with a photoperiod of 16 h light (80 μmol m−2 s−1)/8 h dark at 25°C (Iijima et al. 2008). ‘Micro-Tom MM’ is descended from the original Micro-Tom bred by Scott and Harbaugh (1989). ‘Micro-Tom MM’ (INRA Micro-Tom) seeds were obtained from the French Tomato Genetic Resource Center at INRA Avignon. It was self-fertilized for seven generations before use in the present analysis.

DNA extraction

DNA was extracted from mature leaves of ‘Micro-Tom’ and young leaves of ‘Micro-Tom MM’ using a DNeasy Plant Mini Kit (QIAGEN, http://www.qiagen.com) according to the manufacturer instructions.

DNA sequencing

Sequencing libraries of ‘Micro-Tom’ for the Illumina platform were prepared using the Paired End DNA Sample Prep Kit (Illumina, http://www.illumina.com). Single-end sequencing libraries for the 454 platform were prepared using a GS Titanium Rapid Library Preparation Kit (Roche, http://www.roche.com). Paired-end libraries were prepared according to the Paired End Library Preparation Method Manual-3 kb Span (October 2009 version, Roche). DNA sequencing on the Illumina platform (Harris et al. 2008) was performed using a HiSeq2000 system (Illumina), and that on the 454 platform (Margulies et al. 2005) was performed using a Genome Sequencer FLX system (Roche). Raw sequence data are available in the DDBJ Sequence Read Archive (DRA) (Kaminuma et al. 2010) (study ID DRP000312) and in the Sequence Read Archive (SRA) of NCBI (Leinonen et al. 2011) (accession Nos. DRX000454, DRX000455 and DRX000482). Sequencing libraries of ‘Micro-Tom MM’ were prepared using the Illumina TruSeq® DNA Sample prep Kit and sequenced on the Hiseq 2000 platform. Raw data are available through the European Nucleotide Archive from EMBL-EBI (http://www.ebi.ac.uk/ena/data/view/PRJE4585).

Mapping reads onto the ‘Heinz 1706’ genome sequence

The reads generated using the HiSeq2000 system were split into two groups, and each group, containing approximately 80-fold coverage, was combined with the reads generated using the Genome Sequencer FLX system to obtain two data sets to be mapped onto the ‘Heinz 1706’ genome sequence. The reads generated by the HiSeq2000 system were mapped to the ‘Heinz 1706’ genome sequence (Tomato Genome Consortium 2012) (SL2.40, obtained from ftp://ftp.sgn.cornell.edu/genomes/Solanum_lycopersicum/) provided by the International Tomato Genome Sequencing Project using BWA software (Li and Durbin 2009) with the default settings, except for selection of the option for long reads (bwasw) (Li and Durbin 2010). Results of the mapping were filtered to obtain reads that were uniquely mapped to SL2.40 based on BWA output (either or both of the paired reads had the annotation code ‘XT:A:U’ in the output). The uniquely mapped reads were used for identification of SNPs and indels. The reads from each sequencer run were merged to the respective chromosome and piled up using SAMtools software (Li et al. 2009).

Identification of SNPs and indels

Candidate SNPs and indels between piled-up sequences of the ‘Micro-Tom’ and ‘Heinz 1706’ genomes were detected using SAMtools software. We used two functions of SAMtools for SNP detection: the SAMtools ver. 0.1.16 (Li et al. 2009) pileup function and the SAMtools ver. 0.1.18 (Li et al. 2009) mpileup function. The command lines used in the analysis are provided in the Supplementary text. SNPs were detected independently from the two data sets. Candidate SNPs and indels were filtered by the following conditions: Phred-scaled likelihood (fifth column in the pileup file generated by SAMtools) > 0 for SNP detection and Phred-scaled likelihood > 100 with depth > 4 for indel detection. These conditions were adapted from supporting materials for the SAMtools package (http://sourceforge.
net/apps/mediawiki/samtools/index.php?title=SAM_FAQ). To reduce false positives in the identification of candidates obtained from each data set, we selected SNPs that were detected from both data sets for further analyses. SNPs found in common between ‘Micro-Tom’ and ‘Micro-Tom MM’ were identified by mapping Illumina HiSeq2000 reads from ‘Micro-Tom MM’ onto a piled-up sequence of ‘Micro-Tom’ using the same SNP detection parameters.

‘Micro-Tom’ BAC-end sequences (FT272487-FT321168) were obtained from the DNA Data Bank of Japan (DDBJ) (ftp://ftp.ddbj.nig.ac.jp/ddbj_database/mass/Solanum_lycopersicum_GSS/).

**Adding annotations to SNPs and indels**

ITAG2.3 (Tomato Genome Consortium 2012) annotations (http://solgenomics.net/organism/Solanum_lycopersicum/-genome) released by the International Tomato Annotation Group (ITAG) were used to obtain the positions of the genes, gene identifiers, functional descriptions and positions of exons, introns and UTRs. The files ITAG2.3_gene_models.gff3, ITAG2.3_repeats.gff3, ITAG2.3_protein_functional.gff3 and ITAG2.3_cDNA.fasta were downloaded from the Sol Genomics Network FTP site (http://solgenomics.net/itag/release/2.3/list_files). Synonymous and non-synonymous substitutions were identified using an in-house detection tool. Statistical analysis was performed using functions provided in Microsoft Excel software (ver. 2010, Microsoft, http://www.microsoft.com).

**Data availability**

Bulk downloading of SNP and indel data is available as Supplementary Tables S1, S2, S3 and S5, and is also available in the database ‘TOMATOMICS’ (http://bioinf.mind.meiji.ac.jp/tomatomics/). The database ‘TOMATOMICS’ also implements a search function of SNPs in both ‘Micro-Tom’ and ‘Heinz 1706’ strains.

**Supplementary data**

*Supplementary data* are available at PCP online.

**Funding**

This work was supported by the Ministry of Education, Culture, Sports, Science and Technology [National Bioresource Project Genome Information Upgrading Program ‘Micro-Tom genome sequencing’ (to K.A. and A.T.), Grant-in-Aid for Scientific Research on Innovative Areas (to K.Y.) (grant No. 24113518), Grant-in-Aid for Scientific Research (A) (to K.A.) (grant No. 23248005)]; INRA AIP Bioresource Program [to C.R. and D.B.]; CREST, Japan Science and Technology Agency [Core Research for Evolutional Science and Technology (to K.Y.)]; Kazusa DNA Research Foundation [to K.A.].

**Acknowledgments**

We are grateful for and acknowledge the use of the draft tomato genome sequence generated by the International Tomato Genome Sequencing Consortium (http://solgenomics.net/tomato/). We are also grateful for and acknowledge the use of the tomato SSM data set (http://www.kazusa.or.jp/tomato/), which was generated by the Kazusa DNA Research Institute. We thank Dr. Kaworu Ebana (National Institute of Agrobiological Sciences) for helpful discussions. Computer time for the ‘Micro-Tom MM’ study was provided by the computing facilities MCIA (Mésocentre de Calcul Intensif Aquitain) of the Université de Bordeaux and of the Université de Pau et des Pays de l’Adour.

**Disclosures**

The authors have no conflicts of interest to declare.

**References**


