Theor Appl Genet (2010) 121:1199–1207
DOI 10.1007/s00122-010-1381-y

ORIGINAL PAPER

Discovery of intron polymorphisms in cultivated tomato using both tomato and Arabidopsis genomic information

Yuanyuan Wang · Jia Chen · David M. Francis · Huolin Shen · Tingting Wu · Wencai Yang

Received: 26 November 2009 / Accepted: 3 June 2010 / Published online: 16 June 2010
© Springer-Verlag 2010

Abstract A low level of genetic variation has limited the application of molecular markers for characterizing important traits in cultivated tomato. To detect polymorphisms in tomato conserved ortholog sets (COS), expressed sequence tags (ESTs) were searched against tomato and Arabidopsis genomic sequences to define the positions of introns. Introns were amplified from 12 different accessions of tomato by polymerase chain reaction and nucleotide sequences were determined by sequencing. Results indicated that there was a possibility of 71% to amplify introns from tomato genomic DNA through this approach. A total of 201 introns were sequenced from 86 COS unigenes. The intron positions and numbers were conserved between tomato and Arabidopsis, but average intron length was three times longer in tomato than in Arabidopsis. A total of 307 single nucleotide polymorphisms (SNPs) and 75 indels were detected in introns of 57 COS unigenes among 12 tomato lines. Within cultivated tomato germplasm 172 SNPs and 47 indels were detected in introns of 33 COS unigenes. In addition, 41 SNPs were identified in the exons of 27 COS unigenes. The frequency of SNPs was 2.4 times higher in introns than in exons in the 22 COS unigenes having both intronic and exonic polymorphisms. These results indicate that intronic regions may contain sufficient variation to develop sufficient marker resources for genome-wide analysis in cultivated tomato.

Communicated by T. Close.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-010-1381-y) contains supplementary material, which is available to authorized users.

Y. Wang · J. Chen · H. Shen · T. Wu · W. Yang (✉)
Department of Vegetable Science, College of Agronomy and Biotechnology, China Agricultural University, Beijing 100193, China
e-mail: yangwencai@cau.edu.cn

D. M. Francis
Department of Horticulture and Crop Science, The Ohio State University/OARDC, 1680 Madison Ave, Wooster, OH 44691, USA

Introduction

Tomato (*Solanum lycopersicum* L.) is an important vegetable crop as well as an excellent model plant for genetic analysis. Molecular marker development for tomato has been on-going for more than 30 years (Rick and Fobes 1974). Since the first high-density linkage map of tomato was constructed using restriction fragment length polymorphism (RFLP) markers (Bernatzky and Tanksley 1986), efforts have been made to discover and develop markers through a variety of methods including random amplified polymorphic DNA (RAPD), amplified fragment length polymorphisms (AFLPs), simple sequence repeats (SSRs), cleaved amplified polymorphisms (CAPs), and conserved ortholog sets (COS) (reviewed by Chen et al. 2007). To date, more than ten molecular maps are available for tomato. Almost all maps are constructed using populations derived from crosses between wild species and cultivated tomato (Foolad 2007). This approach maximizes the polymorphisms for map construction and has lead to the discovery of new genes. However, less than 5% of markers developed for mapping purposes in wide crosses are polymorphic within cultivated tomato (Miller and Tanksley 1990). The emphasis on wide crosses has left a void in our ability to manipulate many traits of agricultural importance within elite breeding populations.

Efforts to discover new markers now focus on single nucleotide polymorphisms (SNPs) within *S. lycopersicum*. SNPs are widely distributed and constitute the most abundant molecular markers in the genome. By mining the tomato expressed sequence tags (ESTs) database, Yang et al. (2004) discovered 101 candidate SNPs in 44 genes between two genotypes, TA496 and Rio Grande. Experimental verification suggested that 83% of the candidate SNPs were real polymorphisms. Furthermore, SNPs discovered between these two genotypes had a high probability (53.5%) of detecting SNPs among other *S. lycopersicum* genotypes (Yang et al. 2004). Using a different algorithm and a broader set of tomato accessions, Labate and Baldo (2005) predicted 2,527 SNPs in 764 genes. Resequencing of 53 polymerase chain reaction (PCR) amplicons indicated that only 27% of the predicted SNPs could be verified. Meanwhile Yamamoto et al. (2005) predicted 1,995 SNPs in 660 genes using a similar approach. Subsequent resequencing of PCR amplicons from two genotypes (Micro-Tom and E6203) verified 69% of predicted SNPs. Based on SNP mining in EST databases for tomato, the frequency of SNPs in coding regions is low with approximately one SNP every 7–8.5 kb. The sequence divergence in three genes averaged 0.61 for non-coding regions versus 0.28 for coding sites between *S. lycopersicum* and seven related tomato species (Nesbitt and Tanksley 2002). Thus, the discovery of polymorphisms in non-coding regions is likely to be more efficient.

Two types of intron polymorphisms, intron single nucleotide polymorphisms (ISNPs) and intron length polymorphisms (ILPs), have been reported. In a genome-wide search of potential intron polymorphisms (PIPs) using the draft genomic sequences, Yang et al. (2007) identified 4,645 PIPs between two rice cultivars, 93–11 (*indica*) and Nipponbare (*japonica*), and 14,258 PIPs between two Arabidopsis ecotypes, Columbia and Landsberg. Examination by electronic PCR suggested approximately 70% of candidate PIPs were either ILPs or ISNPs. The frequency of ISNPs was about three times that of ILPs (Yang et al. 2007). The ILP markers proved useful to molecular map construction in rice (Zhao and Wu 2008). Although genomic information has been created for many plant species, discovery of intron polymorphisms is limited by the fact that most crops have genomic sequences for only one cultivar or ecotype. In species with a lack of genomic information, development of intron-flanking markers through conserved genes is an alternative approach.

Exon-primed intron-crossing (EPIC)-PCR is an approach developed to amplify non-coding introns using primers designed from highly conserved exon sequences (Lessa 1992). This method uncovers substantial genetic variability, mainly from ILPs, and has been successfully used in several population genetic surveys (Daguin and Borsa 1999; Bierne et al. 2000; Hassan et al. 2003). In tomato, four sets of COS have been reported. Fulton et al. (2002) defined the first COS (SGN COS) of 1,025 sequences with high homology between tomato and Arabidopsis genome. Four years later, Wu et al. (2006) identified the second COS (SGN COSII) shared by most euasterid plant species and Arabidopsis. SGN COS II contains 2,869 single-copy orthologs with some duplication in the first set. Using different approaches, Van Deynze et al. (2007) identified 2,185 sequences as a set of COS between Arabidopsis and tomato, lettuce, sunflower, soybean, and maize, of which 1,704 are represented in tomato. The fourth is the PIP database containing 1,003 tomato primer pairs predicted to flank introns (Yang et al. 2007). These four databases overlap and represent a complementary resource to increase the number of polymorphisms for *S. lycopersicum*.

By sequencing products amplified from genomic DNA of 10 *S. lycopersicum* inbred lines, Van Deynze et al. (2007) identified 579 SNPs and 206 indels from introns of 162 and 122 loci, respectively. The frequency of SNPs in introns was 1/1,467 bp. Our initial work on sequencing amplicons from 23 loci identified 13 SNPs in 33 introns among three genotypes, PI 114490, FL 7600 and OH 9242. The frequency of polymorphisms was 1 SNP every 1,234 bp (Yang et al. 2005). Both studies suggested promise for developing intron-flanking molecular markers in cultivated tomatoes. The purposes of this study were to (1) develop more intron-flanking markers that are polymorphic within *S. lycopersicum*, (2) compare intron prediction precision in the SGN COS and PIP databases, and (3) compare the intron position, number and size in conserved genes between tomato and Arabidopsis. The experience gained in this study will provide a potential approach to marker development in species lacking whole genomic DNA sequence information.

Materials and methods

Plant materials and DNA isolation

Twelve tomato lines were used to detect intron polymorphisms in this study. These inbred lines were selected to represent a diverse collection including a *S. pimpinellifolium* line (PI128216), a *S. lycopersicum* var. *cerasiforme* line (PI114490), seven fresh-market cultivars (Rio Grande, Moneymaker, Micro-Tom, Baiguoqiangfeng, Shijifeng, Meifen 1, and Zhongshu 5), and three processing cultivars (OH9242, OH88119, and Liger 87-5). PI114490 is a cherry tomato with yellow fruit and resistance to tomato bacterial spot (Scott et al. 2003; Yuan et al. 2008). OH9242 and OH88119 were developed in the USA, while Liger 87-5 is a major processing tomato cultivar growing in China.

Moneymaker is a greenhouse cultivar. Micro-Tom is a tomato genotype with miniature size and short life cycle (Scott and Harbaugh 1989). Baiguoqiangfeng, Shijifeng, Meifen 1, and Zhongshu 5 are cultivars developed by distinct institutes and grown in four environmentally unique regions of China. Of the 12 lines used in this study, PI114490, OH9242 and OH88119 have been used for marker development by Van Deynze et al. (2007). For DNA isolation, seedlings were grown in 288 Square Plug Tray Deep (Taizhou Longji Yuanyi Cailiao Co., Ltd, Zhejiang, China) filled with a mixture of peat soils and vermiculite (3:1) in the greenhouse. Genomic DNA was isolated from fresh-collected young leaves of at least eight plants for each line using the modified CTAB method described by Kabelka et al. (2002).

Prediction of intron position and primer design

Existing DNA sequence resources for tomato and Arabidopsis were used to estimate intron positions and develop primers. Two hundred and one COS unigenes (Tomato 200607#1) from the SOL Genomics Network (SGN, http://sgn.cornell.edu/) and PIP database (http://ibi.zju.edu.cn/pgl/pip/) were subjected to intron prediction through two approaches. These unigenes were selected by comparing them with published tomato SNP markers to ensure that they have not been used in previous studies. Sequences of all unigenes were first searched against the tomato bacterial artificial chromosome (BAC) sequences and BAC ends sequence database at SGN. If a BAC or BAC end containing the unigene was identified, then the unigene was aligned to the genomic DNA sequence to determine intron positions. Otherwise, the sequence of the unigene was used to search the Arabidopsis genome sequence database (http://www.arabidopsis.org) using the method described by Yang et al. (2005). Positions of introns were determined by identifying the gaps in tomato ESTs.

Primers flanking intron regions were designed using Primer 3 (Rozen and Skaletsky 2000). For the unigenes with available tomato genomic DNA sequence, primers were designed based on the genomic DNA sequence. Otherwise, primers were designed based on tomato EST sequences. Multiple pairs of primers were designed for some unigenes, particularly for those with known genomic DNA sequences. Thus, a total of 305 pairs of primers were designed for the 201 COS unigenes, of which 34 pairs of primers for 34 unigenes were from the PIP database. All primers were synthesized by Sunbiotech (Beijing, China).

Primer screening and sequence analysis

Amplification of primers was first tested on two lines, PI114490 and OH9242. PCR was conducted in a 20-μl reaction solution (Yang et al. 2004). Primers that successfully amplified a product were randomly selected for polymorphism detection through sequencing PCR products amplified from the genomic DNA of the 12 lines.

For sequencing purposes, PCR reactions were conducted in a 50 μl volume consisting of 10 mM Tris–HCl (pH 9.0 at room temperature), 50 mM KCl, 1.5 mM $MgCl_2$, 50 μM of each dNTP (Vigorous Biotechnology Beijing Co., Beijing, China), 0.4 μM primers, 30 ng genomic DNA template, and 2 units of Taq DNA polymerase (TaKaRa Biotechnology Dalian Co., Dalian, China). Reactions were heated at 94°C for 3 min followed by 36 cycles of 1 min at 94°C, 1 min at suitable annealing temperature (54°C for most primer pairs), and a 3-min extension at 72°C with final extension reaction at 72°C for 5 min. PCR products were separated on 1% agarose gel, purified using a Cycle-Pure Kit (Omega Bio-Tek, Inc., GA, USA), and sequenced for both forward and reverse directions using a ABI 3730 (Applied Biosystems, Foster City, CA, USA). Sequences were analyzed and assembled using Sequencer 4.0 (Gene Codes Corporation, Ann Arbor, MI, USA). SNPs and indels were identified by aligning sequences of the same unigene from 12 tomato lines. Average number of nucleotide substitutions per thousand sites between populations (*Dxy*) (Nei 1987), number of segregating sites (*S*), θ (Watterson 1975), π (Nei and Li 1979) were generated using the DnaSP v5.10 software package (Rozas and Rozas 1999).

Results

Success of intron amplification

Two hundred and seventeen (71%) pairs of primers had PCR products from either PI114490 or OH9242 or both, and 212 pairs of primers amplified a single band (Table 1). As expected, primers designed based on tomato genomic sequences had a higher probability of producing PCR products than those designed based on EST sequences. For primers designed using tomato genomic sequences, 65 of 74 had single PCR products and three had multiple bands. However, only 64.5% of the primers designed based on tomato ESTs could amplify products from tomato genomic DNA. Based on the size estimation on agarose gels using the 100 bp standard molecular marker (Tiangen Biotech Co., Beijing, China), 95.4% of the PCR products were larger than the length of corresponding EST sequences suggesting that at least one intron was amplified. One hundred and nineteen pairs of primers from 86 unigenes were selected to amplify the genomic DNA of the 12 lines for sequencing.

The success rate for intron amplification using primers based on tomato ESTs depended on the length of the EST

Table 1 Summary statistics for primer design, PCR amplification and sequencing

Sequences used for primer design	No. of genes	No. of primer pairs designed	No. of successful PCR	% of PCR success	No. of primer pairs with single band	No. of PCR products sequenced
Tomato genomic DNA	23	74	68	91.9	65	45
Tomato EST	178	231	149	64.5	147	74
Total	201	305	217	71.1	212	119

sequence flanking the predicted intron(s) included in the amplicons. As the length of EST sequences increased, the success rate of PCR amplification decreased (Fig. 1). Greater than 78% of primer pairs generated PCR products when the lengths of EST sequences were less than 300 bp. The success rate of intron amplification was slightly higher than 50% when the lengths of EST sequences were between 400 and 800 bp. However, the rate dropped to 38.5% (5 of 13 pairs of primers) when the primers were separated in EST sequences by more than 800 bp.

Both the PCR success rate and intron prediction rate were slightly higher using the PIP database than using SGN COS. Of the 34 pairs of primers from the PIP database, 28 (82.4%) successfully amplified products from genomic DNA and all PCR products contained one intron as predicted. However, 69.7% of primer pairs amplified products and 93.7% of PCR products had at least one intron in SGN COS.

Comparison of introns in conserved genes between tomato and Arabidopsis

One hundred and eight of 201 unigenes were randomly selected for intron position, number, and size comparison between tomato and Arabidopsis. Tomato genomic DNA sequences of these unigenes were from either the SOL Genomics Network or this study. Results indicated that conserved unigenes also had conserved intron positions and numbers. Most unigenes (78.7%) had the same number of introns in both species, while 13.9% of the unigenes had more introns in Arabidopsis than in tomato. The 108 unigenes contained 300 introns in tomato and 310 introns in Arabidopsis with a range 1–14 in both species (Supplementary Table S1).

The lengths of introns in tomato were much larger than in Arabidopsis. Intron lengths varied from 64 to 4,546 bp in tomato and from 67 to 1,736 bp in Arabidopsis. Average intron length was 466 bp in tomato and 147 bp in Arabidopsis. Of the comparable introns, 77.3% were longer in tomato than in Arabidopsis, and only 2.6% of the introns were the same size in both species (Supplementary Table S1). In tomato, 50.2% of the introns were less than 200 bp and 19.0% were larger than 800 bp (Fig. 2). In Arabidopsis most introns (83.5%) were less than 200 bp, and only 0.6% were larger than 1,000 bp.

Intron polymorphisms in tomato

Genomic DNA sequences for 86 COS unigenes were obtained from the 12 tomato lines. The sequenced regions contained 203 introns. Numbers of introns in the 86 unigenes ranged from 1 to 15 with an average of 2.4 introns per unigene. More than half (54.7%) of the unigenes had more than one intron in sequenced regions. Intron lengths ranged from 69 to 2,599 bp, and 10.2% of the introns were larger than 1,000 bp.

Of the 86 unigenes sequenced, 57 (66.3%) contained 94 introns with polymorphisms among the 12 tomato lines.

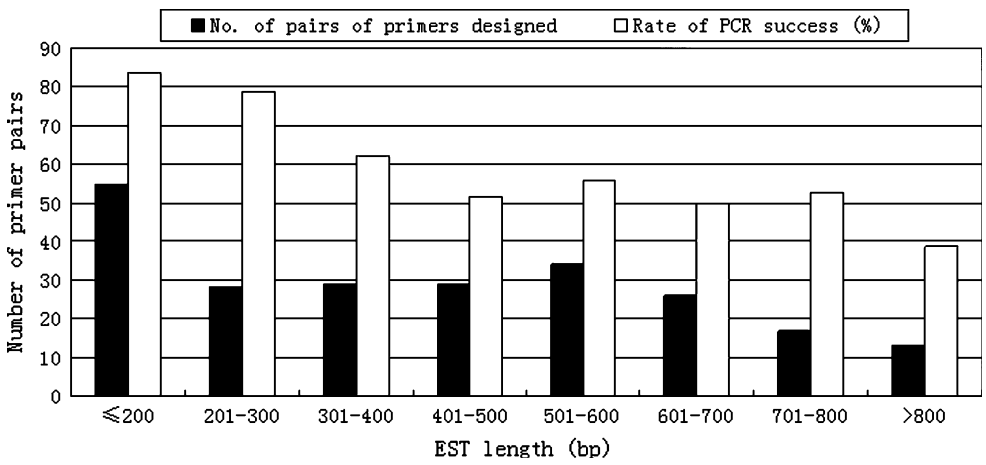

Fig. 1 Distribution of success rate for PCR amplification using 305 pairs of primers designed based on 201 COS unigenes in tomato

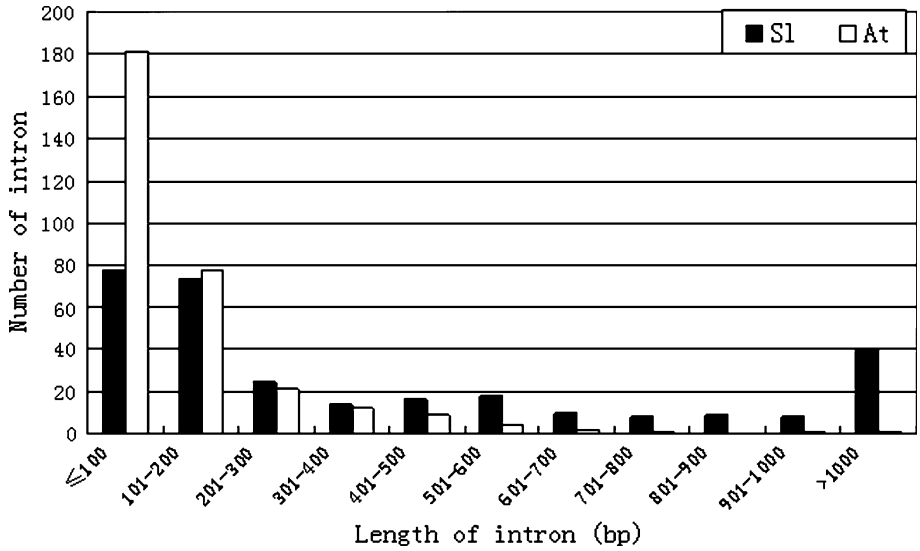

Fig. 2 Distribution of intron sizes of 86 COS unigenes in tomato (Sl) and Arabidopsis (At)

A total of 307 SNPs in 56 unigenes and 75 indels in 36 unigenes were identified (Table 2, Supplementary Table S2). The overall frequency was 1 SNP in 252 bp and 1 indel in 1,033 bp. Of the 57 unigenes, 61.4% of the unigenes had both SNPs and indels, 36.8% of the unigenes only had SNPs, and 1.8% of the unigenes only had indels. Thirty-three COS unigenes had polymorphic introns among 11 *S. lycopersicum* lines. One hundred and seventy-two SNPs in 30 unigenes and 47 Indels in 19 unigenes were identified with frequencies of 1 SNP in 451 bp and 1 Indel in 1,649 bp (Tables 2, 3).

Nucleotide sequences were more polymorphic between *S. pimpinellifolium* PI128216 and *S. lycopersicum* lines than within 11 *S. lycopersicum* lines. Almost all of the 62 polymorphic COS genes had polymorphisms between PI128216 and at least one of the 11 *S. lycopersicum* lines. Numbers varied from 192 to 264 for SNPs and 50 to 61 for indels between PI128216 and 11 *S. lycopersicum* lines (Supplementary Table S3). Pairwise comparisons between *S. lycopersicum* lines and PI 128216 suggested 1 SNP every 294–404 bp and one indel every 1,271–1,550 bp (Supplementary Table S4).

Exon polymorphisms in tomato

Polymorphisms in exon regions were also detected in this study and the frequency of polymorphisms was lower than in introns. Of the 62 unigenes, only five had polymorphisms in exons, while 22 had polymorphisms in both coding and non-coding regions. In these 27 unigenes, 41 SNPs were identified in exons between *S. pimpinellifolium* PI128216 and *S. lycopersicum* lines. But no indels were detected (Table 2). The frequency of SNP (1 SNP in 857 bp) was 3.4 times lower in exons than in introns. Of the 41 SNPs, 13 were polymorphic within *S. lycopersicum* lines (Table 3) with a frequency of 1 SNP in 2,703 bp. Five SNPs were detected in processing lines and nine were detected in fresh-market lines (Table 2). In the 22 unigenes that had polymorphisms in both introns and exons, the number of SNPs was 4.4 times higher in introns than in

Table 2 Summary statistics for single nucleotide polymorphisms (SNP) and indels in 86 conserved genes in 12 tomato lines

	Intron										Exon				
	Loci with SNPs	No. of SNPs	% Loci with SNPs	SNPs/ locus	Bases/ SNP	Loci with indels	No. of indels	% Loci with indels	Indels/ locus	Bases/ indel	Loci with SNPs	No. of SNPs	% Loci with SNPs	SNPs/ ocus	Bases/ SNP
S. lycopersicum	30	172	34.9	5.7	451	19	47	22.1	2.5	1,649	11	13	12.8	1.2	2,703
Fresh-market[a]	20	143	23.3	7.2	542	16	41	18.6	2.6	1,890	7	9	8.1	1.3	3,904
Fresh-market[b]	12	42	14.0	3.5	1,845	8	13	9.3	1.6	5,962	5	6	5.8	1.2	5,856
Processing	11	83	12.8	7.5	934	7	15	8.1	2.1	5,167	4	5	4.7	1.3	7,027
All 12 tomato lines	56	307	65.1	5.5	252	36	75	41.9	2.1	1,033	27	41	31.4	1.5	857

[a] Line Micro-Tom was included

[b] Line Micro-Tom was excluded

Table 3 Map positions, number of SNPs and indels in 36 COS unigenes in 11 *Salanum lycopersicum* lines

Cos marker	Chromosome	Map position (cM) on Tomato-EXPEN 2000	Intron		No. of SNPs in exon
			No. of SNPs	No. of indels	
T0005	9	Unknown	3	1	0
T0035	Unknown	Unknown	2	0	0
T0055	10	37.00	1	0	0
T0128	Unknown	Unknown	2	2	0
T0161	9	50.30	3	3	2
T0187	Unknown	Unknown	3	2	0
T0196	3	96.00	11	2	1
T0201	Unknown	Unknown	2	2	0
T0217	Unknown	Unknown	6	0	1
T0244	6	10.00	9	3	0
T0256	7	1.20	6	3	0
T0283	1/10	165.00/36.00	1	0	0
T0291	10	30.50	6	0	0
T0347	2	111.00	0	0	1
T0360	4	131.00	6	0	0
T0489	Unknown	Unknown	9	2	1
T0577	3	120.00	4	1	0
PIP88	Unknown	Unknown	2	0	0
PIP114	Unknown	Unknown	1	0	0
PIP447	Unknown	Unknown	9	1	0
PIP564	Unknown	Unknown	1	0	1
PIP881	8	Unknown	3	0	0
T1784	12	96.00	0	0	1
T0266	2	67.00	2	2	0
T0586	5	65.00	1	0	0
T1366	7	71.00	25	10	0
T1400	2	119.00	0	1	0
T0133	Unknown	Unknown	1	0	0
T0114	2	0.00	0	1	1
T0292	4	85.00	1	0	0
T0343	1	18.00	0	1	1
T0761	3	133.00	44	8	2
T1014	11	67.00	4	1	0
T1155	3	74.00	0	0	1
T1546	4	98.00	1	0	0
T1716	4	119.50	3	1	0
Total			172	47	13

exons (Table 4). Across all COS unigenes we sequenced, the frequency of SNPs was 2.4 times higher in introns than in exons.

Sequence divergence and molecular population genetics analysis in tomato

Sequences of the 62 COS unigenes showing nucleotide polymorphisms between/within species were used to analyze sequence divergence and population divergence. Sequence divergence was higher between wild species and cultivated tomato than within cultivated tomato. Most (61%) of the 62 unigenes with sequence variation between *S. pimpinellifolium* and *S. lycopersisum* were not polymorphic within *S. lycopersisum*. The average number of nucleotide substitutions per thousand sites between populations (*Dxy*) varied from 0 to 22.85 in the 62 unigenes (Supplementary Table S5). Mean *Dxy* based on 62

Table 4 SNPs and indels in intronic and exonic regions in 22 COS unigenes in 12 tomato lines

	No. of intron/exon	Length (bp)	No. of SNPs	Bases/SNP	No. of indels	Bases/indel
Intron	68	22,959	146	157	31	741
Exon	94	12,487	33	378	0	–

unigenes between *S. pimpinellifolium* and *S. lycopersisum* was almost three times higher than that between *S. lycopersicum* var. *cerasiforme* PI 1144890 and *S. lycopersisum*, and approximately five times higher than that between processing and fresh-market tomato lines.

Highly divergent alleles were observed at 11 unigenes in all 12 tomato lines and at five unigenes in cultivated tomato including both processing and fresh-market lines based on the neutrality in Tajima's *D* tests (1993). Only one unigene, CosOH18, showed high divergence in all populations (Supplementary Table S6). The θ values were not consistent with the neutrality test. Values of θ for nine unigenes (CAU15, CAU41, CAU61, CAU77, PIP447, CosOH2, CosOH18, CosOH50, and CosOH51) in 12 tomato lines were greater than 3. However, five of them (CAU15, CAU77, CosOH2, CosOH50, and CosOH51) did not reject neutrality based on Tajima's *D* test. In contrast the neutrality test was rejected for seven unigenes with θ values less than 3 (CAU6, CAU9, CAU32, CAU42, CAU50, CAU62, and CosOH3). Values of θ and π were the same for all unigenes in the processing tomato population. Tajima's *D* test could not be conducted because only three lines were used in this study.

Discussion

Precise prediction and successful amplification of introns from genomic DNA are critical for intron polymorphism detection in a species without genomic DNA sequence information. An approach for defining intron position is to use genomic DNA sequence information of conserved genes in other species. Genomic DNA sequences of conserved genes in Arabidopsis and rice have been used to predict intron positions in other species (Yang et al. 2007). Yang et al. (2005) and Van Deynze et al. (2007) also used this approach to predict intron positions in tomato. Accurate prediction of intron position does not guarantee that PCR amplification will be successful. In the current study, 64.5% pairs of primers designed using ESTs successfully amplified product(s) from genomic DNA, which was consistent with other findings (Temesgen et al. 2001; Wei et al. 2005; Van Deynze et al. 2007). Although positions of introns can be predicted by aligning tomato COS with Arabidopsis genomic DNA sequences, the sizes of introns are much larger in tomato than in Arabidopsis (Van Deynze et al. 2007; this study). Recently, we rechecked the availability of genomic DNA sequences for unigenes used in this study and found predicted PCR product sizes for 15 of 23 pairs of primers that failed in PCR amplification were greater than 2.2 kb (2,261–7,680 bp, data not shown), suggesting that amplicon size limited the ability of Taq DNA polymerase to complete the reaction. Thus, amplicon size was the major reason for PCR failure in this study. Minimizing the length of EST in amplicons surrounding predicted introns might therefore improve success.

The frequency of polymorphisms is higher in intronic regions than in exonic regions in tomato. The expected frequency of polymorphisms in coding regions is one SNP per 7–8.5 kb (Nesbitt and Tanksley 2002; Yang et al. 2004). SNPs appear 1.6–5.2 times more frequently in non-coding regions than in coding regions (Yang et al. 2005; Van Deynze et al. 2007; Jiménez-Gómez and Maloof 2009). Similarly, the frequency is 2.6–8.7 times higher for indels in untranslated regions than in coding sequences (Jiménez-Gómez and Maloof 2009). A recent study analyzing 50 gene fragments amplified from 31 *S. lycopersicum* landraces found an average of 1 SNP per 158 bp with the minor allele at a frequency of 10% (Labate et al. 2009). The data presented in this paper based on a more diverse collection of germplasm. The number of unigenes (57 unigenes) with sequence variation in introns was two times higher than that of unigenes (27 unigenes) with sequence variation in exons, and the frequency of SNPs was 3.2–7.5 times higher in introns than in exons. In addition, a total of 75 indels were identified in introns but no indel was detected in exons. Direct comparison of the frequencies of SNPs in introns and exons in 22 unigenes also suggested that intronic regions had more diversity than exonic regions.

Previous studies using RFLP, RAPD, inter-simple sequence repeat (ISSR), AFLP, and SSR markers indicated that genetic variation was limited in *S. lycopersicum* (Miller and Tanksley 1990; Williams and St Clair 1993; Archak et al. 2002; Park et al. 2004; Garcia-Martinez et al. 2005; Tam et al. 2005; Chen et al. 2009). The lack of polymorphisms prevents the detailed molecular study of traits with agricultural importance in cultivated tomato. This limitation may be overcome by more fully utilizing sequence variation through marker systems such as SNPs. Mining SNPs through EST database indicated that approximately 3.5% of the genes had polymorphisms

between the two genotypes TA496 and Rio Grande (Yang et al. 2004). A later investigation estimated that 3.8–7.9% of all loci differed in 10 accessions (Van Deynze et al. 2007). Using an oligonucleotide array, Sim et al. (2009) discovered 279 SNPs and 27 indels in 111 loci from three lines, FL7600, OH9242 and PI114490. In the current study, polymorphic unigenes ranged from 7.0 to 10.5% among three processing tomato lines and from 1.2 to 11.6% among six elite fresh-market lines (Supplementary Table S7). Taken together, these studies suggest that sufficient polymorphisms exist in cultivated tomato to support genetic analysis within elite populations. Based on the estimate of 35,000 genes in tomato (Van der Hoeven et al. 2002) and frequency of polymorphisms between any two *S. lycopersicum* lines, we may expect between 402 and 4,070 unigenes with polymorphisms that can be used as markers in most pair-wise comparisons of cultivated tomato.

EPIC-PCR has been widely used to amplify introns in conserved genes. In this study, unigenes randomly picked from two COS databases were used to predict and amplify intronic regions. There were no significant difference for intron prediction and amplification using ESTs from PIP and COS SGN databases when the amplicons contained EST lengths less than 300 bp. The amplicons based on EST sequence greater than 300 bp are expected to have a reduced success rate for PCR amplification. We successfully amplified 300 introns from 108 unigenes. Sequence analysis of 203 introns identified 307 SNPs and 75 indels between *S. pimpinellifolium* and *S. lycopersicum*, of which 55% were polymorphic between *S. lycopersicum* lines. The unigenes used in this study are different from those previously reported and provide novel markers for tomato genetics and breeding. All genomic DNA sequences obtained in this study are available from NCBI with Genbank accession numbers of GS598741 through GS598774. As tomato genomic DNA sequences grow (Mueller et al. 2009), it will become easier to discover intron polymorphisms.

Acknowledgments The authors thank the Tomato Genetic Resource Center at the University of California Davis (California, USA) for providing seeds of some tomato lines. We also thank Dr. David B. Weaver from Auburn University for his critical review on the manuscript. The work was supported by National Natural Science Foundation of China (30671425) and the Program for New Century Excellent Talents in University (NCET-08-0531).

References

Archak S, Karihaloo JL, Jain A (2002) RAPD markers reveal narrowing genetic base of Indian tomato cultivars. Curr Sci 82:1139–1143

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

Bierne N, Lehnert SA, Bédier E, Bonhomme F, Moore SS (2000) Screening for intron-length polymorphism in penaeid shrimps using exon-primed intron-crossing (EPIC)-PCR. Mol Ecol 9:233–235

Chen J, Shen HL, Yang WC (2007) Development of tomato molecular markers. Mol Plant Breed 5(6S):130–138

Chen J, Wang H, Shen HL, Chai M, Li JS, Qi MF, Yang WC (2009) Genetic variation in tomato populations from four breeding programs revealed by single nucleotide polymorphism and simple sequence repeat markers. Sci Hortic 122:6–16

Daguin C, Borsa P (1999) Genetic characterization of *Mytilus galloprovincialis* Lmk. in North West Africa using nuclear DNA markers. J Exp Mar Biol Ecol 235:55–65

Foolad MR (2007) Genome mapping and molecular breeding of tomato. Int J Plant Genomics. doi:10.1155/2007/64358

Fulton TM, Van der Hoeven R, Eannetta NT, Tanksley SD (2002) Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell 14:1457–1467

Garcia-Martinez S, Andreani L, Garcia-Gusano M, Geuna F, Ruiz JJ (2005) Evolution of amplified length polymorphism and simple sequence repeats for tomato germplasm fingerprinting: utility for grouping closely related traditional cultivars. Genome 49:648–656

Hassan M, Lemaire C, Fauvelot C, Bonhomme F (2003) Seventeen New EPIC-PCR amplifiable introns in fish. Mol Ecol 2:334–340

Jiménez-Gómez JM, Maloof JN (2009) Sequence diversity in three tomato species: SNPs, markers, and molecular evolution. BMC Plant Biol 9:85

Kabelka E, Franchino B, Francis DM (2002) Two loci from *Lycopersicon hirsutum* LA407 confer resistance to strains of *Clavibacter michiganensis* subsp. michiganensis. Phytopathology 92:504–510

Labate JA, Baldo AM (2005) Tomato SNP discovery by EST mining and resequencing. Mol Breed 16:343–349

Labate JA, Robertson LD, Baldo AM (2009) Multilocus sequence data reveal extensive departures from equilibrium in domesticated tomato (*Solanum lycopersicum* L.). Heredity 103:257–267

Lessa EP (1992) Rapid survey of DNA sequence variation in natural populations. Mol Biol Evol 9:323–330

Miller JC, Tanksley SD (1990) RFLP analysis of phylogenetic relationships and genetic variation in the genus Lycopersicon. Theor Appl Genet 80:437–448

Mueller LA, Lankhorst RK, Tanksley SD et al (2009) A snapshot of the emerging tomato genome sequence. Plant Genome 2:78–92

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76:5269–5273

Nesbitt TC, Tanksley SD (2002) Comparative sequencing in the genus Lycopersicon: implications for the evolution of fruit size in the domestication of cultivated tomatoes. Genetics 162:365–379

Park YH, West MAL, St. Clair DA (2004) Evaluation of AFLPs for germplasm fingerprinting and assessment of genetic diversity in cultivars of tomato (*Lycopersicon esculentum* L.). Genome 47:510–518

Rick CM, Fobes JF (1974) Association of an allozyme with nematode resistance. Rpt Tomato Genet Coop 24:25

Rozas J, Rozas R (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15:174–175

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132:365–386

Scott JW, Harbaugh BK (1989) Micro-Tom—a miniature dwarf tomato, Circular S-370, Florida Agricultural Experiment Station, pp 1–6

Scott JW, Francis DM, Miller SA, Somodi GC, Jones JB (2003) Tomato bacterial spot resistance derived from PI 114490; inheritance to race T2 and relationship across three pathogen races. J Am Soc Hortic Sci 128:698–703

Sim SC, Robbins MD, Chilcott C, Zhu T, Francis DM (2009) Oligonucleotide array discovery of polymorphisms in cultivated tomato (*Solanum lycopersicum* L.) reveals patterns of SNP variation associated with breeding. BMC Genomics 10:466

Tajima F (1993) Statistical analysis of DNA polymorphism. Jpn J Genet 68:567–595

Tam SM, Mhiri C, Vogelaar A, Kerkveld M, Pearce SR, Grandbastien MA (2005) Comparative analyses of genetic diversities within tomato and pepper collections detected by retrotransposon-based SSAP, AFLP and SSR. Theor Appl Genet 110:819–831

Temesgen B, Brown GR, Harry DE, Kinlaw CS, Sewell MM, Neale DB (2001) Genetic mapping of expressed sequence tag polymorphism (ESTP) markers in loblolly pine (*Pinus taeda* L.). Theor Appl Genet 102:664–675

Van der Hoeven R, Ronning C, Giovannoni J, Martin G, Tanksley S (2002) 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 14:1441–1456

Van Deynze A, Stoffel K, Buell CR, Kozik A, Liu J, van der Knapp E, Francis D (2007) Diversity in conserved genes in tomato. BMC Genomics 8:465

Watterson GA (1975) On the number of segregating sites in geneticalmodels without recombination. Theor Popul Biol 7:256–276

Wei H, Fu Y, Arora R (2005) Intron-flanking EST–PCR markers: from genetic marker development to gene structure analysis in Rhododendron. Theor Appl Genet 111:1347–1356

Williams CE, St Clair DA (1993) Phenetic relationships and levels of variability detected by restriction fragment length polymorphism and random amplified polymorphic DNA analysis of cultivated and wild accessions of *Lycopersicon esculentum*. Genome 36:619–630

Wu F, Mueller LA, Crouzillat D, Petiard V, Tanksley SD (2006) Combining bioinformatics and phylogenetics to identify large sets of single copy, orthologous genes (COSII) for comparative, evolutionary and systematics studies: a test case in the Euasterid plant clade. Genetics 174:1407–1420

Yamamoto N, Tsugane T, Watanabe M, Yano K, Maeda F, Kuwata C, Torki M, Ban Y, Nishimura S, Shibata D (2005) Expressed sequence tags from the laboratory-grown miniature tomato (*Lycopersicon esculentum*) cultivar Micro-Tom and mining for single nucleotide polymorphisms and insertions/deletions in tomato cultivars. Gene 356:127–134

Yang W, Bai X, Kabelka E, Eaton C, Kamoun S, van der Knaap E, Francis D (2004) Discovery of single nucleotide polymorphisms in *Lycopersicon esculentum* by computer aided analysis of expressed sequence tags. Mol Breed 14:21–34

Yang W, Miller SA, Scott JW, Jones JB, Francis DM (2005) Mining tomato genome sequence databases for molecular markers: application to bacterial resistance and marker assisted selection. Acta Hortic 695:241–250

Yang L, Jin GL, Zhao XQ, Zheng Y, Xu ZH, Wu WR (2007) PIP: a database of potential intron polymorphism markers. Bioinformatics 23:2174–2177

Yuan DJ, Chen J, Shen HL, Yang WC (2008) Genetics of flesh color and nucleotide sequence analysis of phytoene synthase gene 1 in a yellow-fruited tomato accession PI114490. Sci Hortic 118:20–24

Zhao XQ, Wu WR (2008) Construction of a genetic map based on ILP markers in rice. Hereditas (Beijing) 30(2):225–230