Theor Appl Genet (2009) 118:1005–1014
DOI 10.1007/s00122-008-0957-2

ORIGINAL PAPER

EST, COSII, and arbitrary gene markers give similar estimates of nucleotide diversity in cultivated tomato (*Solanum lycopersicum* L.)

Joanne A. Labate · Larry D. Robertson · Feinan Wu · Steven D. Tanksley · Angela M. Baldo

Received: 3 June 2008 / Accepted: 20 December 2008 / Published online: 20 January 2009

Abstract Because cultivated tomato (*Solanum lycopersicum* L.) is low in genetic diversity, public, verified single nucleotide polymorphism (SNP) markers within the species are in demand. To promote marker development we resequenced approximately 23 kb in a diverse set of 31 tomato lines including TA496. Three classes of markers were sampled: (1) 26 expressed-sequence tag (EST), all of which were predicted to be polymorphic based on TA496, (2) 14 conserved ortholog set II (COSII) or unigene, and (3) ten published sequences, composed of nine fruit quality genes and one anonymous RFLP marker. The latter two types contained mostly noncoding DNA. In total, 154 SNPs and 34 indels were observed. The distributions of nucleotide diversity estimates among marker types were not significantly different from each other. Ascertainment bias of SNPs was evaluated for the EST markers. Despite the fact that the EST markers were developed using SNP prediction within a sample consisting of only one TA496 allele and one additional allele, the majority of polymorphisms in the 26 EST markers were represented among the other 30 tomato lines. Fifteen EST markers with published SNPs were more closely examined for bias. Mean SNP diversity observations were not significantly different between the original discovery sample of two lines (53 SNPs) and the 31 line diversity panel (56 SNPs). Furthermore, TA496 shared its haplotype with at least one other line at 11 of the 15 markers. These data demonstrate that public EST databases and noncoding regions are a valuable source of unbiased SNP markers in tomato.

The use of trade, firm, or corporation names in this publication is for the information and convenience of the reader. Such use does not constitute an official endorsement or approval by the United States Department of Agriculture or the Agricultural Research Service of any product or service to the exclusion of others that may be suitable.

Communicated by A. Schulman.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-008-0957-2) contains supplementary material, which is available to authorized users.

J. A. Labate (✉) · L. D. Robertson · A. M. Baldo
USDA-ARS Plant Genetic Resources Unit,
630 W. North Street, Geneva, NY 14456, USA
e-mail: joanne.labate@ars.usda.gov

F. Wu · S. D. Tanksley
Department of Plant Breeding and Genetics,
Cornell University, Ithaca, NY 14853, USA

Introduction

The majority of published tomato molecular markers and maps are based on polymorphisms between domesticated tomato and wild tomato species rather than within *Solanum lycopersicum* (Foolad 2007). Wild species are used extensively in crop improvement, and the ease of crossability for several species with the cultivar has facilitated saturation of interspecific genetic linkage maps. Increased efforts are being made to develop simple sequence repeat (SSR), insertion/deletion (indel) and single nucleotide polymorphism (SNP) markers within *S. lycopersicum* for marker assisted breeding, cultivar identification, and diversity studies (reviewed by Labate et al. 2007). SNP marker technology is particularly attractive because it overcomes limitations of SSRs such as inconsistency of scoring across platforms, irreproducibility due to polymerase chain reaction (PCR) artifacts, and homoplasy of alleles (Jones et al. 2007). Because individual SNPs are biallelic they are generally less informative than SSRs and must be applied

in greater numbers to obtain comparable resolution of polymorphism estimates. However, many nucleotide markers can in principle be converted into SNP markers, and SNPs can provide the densest maps within a genome. SNPs are currently not widely used in marker assisted breeding for *S. lycopersicum* intraspecific crosses although this will change as the opportunities for intraspecific marker development increase (SolCAP 2008). Small numbers of confirmed intraspecific SNP markers, on the order of dozens, are available in the Tomato Mapping Resource (www.tomatomap.net) (Francis et al. 2006) and Micro-Tom (MiBase, www.kazusa.or.jp/jsol/microtom/index.html) (Yamamoto et al. 2005) Databases, and at dbSNP of the National Center for Biotechnology Information (NCBI) (ss# 76883011–76883086, 77106585–77106606) (Labate and Baldo 2005). More recently, a large set (hundreds) of SNPs and indels was published and is available via FTP at the Solanaceae Genome Network (SGN, www.sgn.cornell.edu) (Van Deynze et al. 2007) (see below).

Expressed-sequence tag mining, intron mining, and oligonucleotide array hybridization are the primary methods used to predict large numbers (100–1,000s) of SNPs in the tomato genome. NCBI contains 258,830 ESTs from multiple cultivars (dbEST release 101708, 17 October 2008). Such sequences can be clustered into unigenes, aligned, and examined for SNPs using criteria that attempts to distinguish predicted SNPs from sequencing artifacts. Cultivars fixed for different alleles can then be targeted for SNP discovery. Yang et al. (2004) identified 101 candidate SNPs in 44 genes using this approach, and empirically verified 83% (43 SNPs) of polymorphisms tested. By applying a different algorithm to a broader set of cultivars, Labate and Baldo (2005) predicted 2,527 SNPs in 764 genes. Resequencing of 53 PCR amplicons from line TA496 and one other cultivar (Rio Grande PI 303784, Moneymaker PI 286255, or E6203) yielded 62 SNPs (27% of tested SNPs were verified). Using similar prediction methods MiBase has reported primer pairs for 1,995 candidate SNPs in 660 unigenes. By resequencing 15 amplicons, 26 SNPs were verified (69% of predicted) between lines Micro-Tom and E6203 (Yamamoto et al. 2005). This equaled one SNP per 121 bp.

The majority (35/62) of SNPs reported by Labate and Baldo (2005) were not predicted from ESTs but were instead found in unintentionally amplified introns. This implies that a second approach, intron mining, should be highly successful in tomato. The method generally applied is to align EST-derived unigenes or conserved ortholog set (COS) sequences with *Arabidopsis* genomic homologs and to design primers flanking predicted introns. SNPs are then discovered via resequencing two or more lines. The potential intron polymorphism (PIP) database has reported 1,003 tomato primer pairs predicted to flank introns (Yang et al. 2007) and Fukuoka et al. (2007) reported initial work on similar numbers of markers. To develop the largest confirmed set of intraspecific SNP markers to date, a set of 967 COS ESTs (UCD COS) were prescreened for polymorphism in pooled DNA representing 12 tomato lines (Van Deynze et al. 2007). In breeding germplasm 579 SNPs and 206 indels were verified in 162 and 122 loci, respectively. A set of COSII primer pairs designated as universal primers for asterid species (UPA) were designed to amplify a broad range of taxa from the euasterid I clade (Wu et al. 2006). Intronic UPA (iUPA) were designed to amplify a predicted intron, ideal markers for application within species or between closely related species. Currently 2,869 COSII genes are available on SGN, each of which is assigned as an *Arabidopsis* COSII. UPAs have been designed for over 1,600 of these, more than 1,400 of which are iUPAs.

A third approach for SNP prediction and discovery in tomato is to identify single feature polymorphisms (SFPs) by hybridization of labeled cDNA to an oligo-based microarray (Francis et al. 2006; Sim et al. 2007). Detected SFPs can represent any polymorphism between the target and probe sequences that affects hybridization. Using this approach with *S. lycopersicum* Ohio 7814 or *Solanum pimpinellifolium* LA1589 as the target, Francis et al. (2006) identified 1,296 putative SFPs of which 52% were verified by resequencing. The majority (86%) were found to be SNPs with the remaining 14% being indels.

DNA sequencing PCR amplified product of single copy loci for more than one tomato line, with no a priori expectation concerning polymorphism, is a direct approach to SNP detection. A tomato diversity study found only one SNP in 7 kb of sequence from each of four cultivars (Nesbitt and Tanksley 2002). This low level of polymorphism implies that this method may be cost-prohibitive for purposes of marker development. However, assembling a large set of diverse germplasm has contributed to the success of this approach in recent applications (Ganal et al. 2007; Luerssen et al. 2006; Robbins et al. 2007; Van Deynze et al. 2007).

Markers can reflect ascertainment bias because of arbitrary decisions made during sampling of individuals or loci. Bias may be inherent if only the most variable loci are retained, or if only a small panel of individuals is used for polymorphism discovery (Brumfield et al. 2003). For example, when confirmed markers are based on a sample of two individuals, moderate frequency alleles are most likely to be discovered while rare polymorphisms within the population can be missed (Clark et al. 2005). Moderate frequency alleles are potentially valuable, e.g., in backcross breeding schemes where it is desirable to maximize the number of polymorphic markers between two parental lines. However, unassayed, low frequency alleles can lead to missed lineages in phylogenetic reconstructions,

overestimates of mean diversity (Schlötterer and Harr 2002), underestimates of differentiation between populations (Smith et al. 2007), and spurious correlations in association mapping (Pritchard 2001).

The purpose of this study was to (1) increase the number of verified, published SNP markers within *S. lycopersicum*, (2) compare diversity of three classes of SNP markers—ESTs with predicted SNPs, COSII containing introns, and arbitrary genes known to be single copy, (3) test ascertainment bias in a set of previously verified SNP markers (Labate and Baldo 2005), and (4) obtain multilocus estimates of nucleotide diversity within this species. A set of germplasm known to be genetically diverse (Villand et al. 1998) was sampled to increase the probability of discovering nucleotide polymorphisms.

Materials and methods

Plant material

The Plant Genetic Resources Unit (PGRU) conserves upwards of 6,000 tomato accessions that serve as a resource to breeders and other researchers. The collections do not represent elite breeding material, but are often a source of novel alleles. A diversity panel of 30 PGRU tomato accessions (Table 1 and Supplementary Table S1) was assembled based on results of RAPD analyses (Villand et al. 1998). A highly diverse subset of the Villand et al. (1998) set, including 14 accessions from the primary center of diversity and 12 accessions from countries contiguous with the primary center, was selected for the current study. Accessions were grown in the field to verify cherry versus non-cherry tomato phenotype. Only accession PI 127825 was classified as cherry based on the tomato taxonomic key (Rick et al. 1990). Breeding line TA496, developed by introgressing Tobacco Mosaic Virus resistance gene *Tm-2*[a] into line E6203 (Tanksley et al. 1998; Yates et al. 2004) was also included in our study.

Table 1 *S. lycopersicum* diversity panel of 30 PGRU accessions that were assayed using EST, COSII/unigene, and arbitrary markers

Provenance	No. accessions[a]
Primary center of diversity (Chile, Ecuador, Peru), 1932–1974	14
Regions contiguous with primary center (Argentina, Bolivia, Brazil, Colombia, Costa Rica, Ecuador, Guatemala, Mexico, Nicaragua, Panama, Venezuela), 1932–1996	12
Secondary centers of diversity (Afghanistan, China, Cuba, Netherlands), 1932–1960	4

[a] Details on accession identity and provenance are in Supplementary Table S1

Markers

The 26 EST markers were originally developed at PGRU by sequencing PCR amplified genomic DNA in TA496 and one or two of cultivars Rio Grande (PI 303784), Moneymaker (PI 286255), and E6203 (Labate and Baldo 2005; Supplementary Table S2). All 26 were predicted to contain at least one SNP based on computational analyses of public *S. lycopersicum* unigenes downloaded from NCBI. In 15 of the markers SNPs were previously verified (175_1, 220_1, 241_2b, 296_1b, 437_2, 1260_2, 1287_1, 1909_2, 2325_3, 2486_1, 2534_1b, 2875_4b, 3155_3, 3300_2, 3332_3) while in 11 markers no SNPs were previously observed (1523_4, 1589_1, 1675_1, 1724_1, 1863_3, 2189_1, 2280_1, 2582_1, 2719_1, 2819_5, 4301_3). In the verified category four markers (241_2b, 296_1b, 2534_1b, 2875_4b) were redesigned for this experiment to yield shorter PCR amplicons and five markers (220_1, 437_2, 2325_3, 2486_1, 2534_1R) exhibited highly diverged alleles that were hypothesized to be introgressions from wild tomato species (Labate and Baldo 2005).

Ten COSII markers (Wu et al. 2006) (C2_At1g13380, C2_At1g14000, C2_At1g20050, C2_At1g32130, C2_At1g44575, C2_At1g50020, C2_At1g73180, C2_At2g15890, C2_At2g22570, C2_At2g36930) and four markers that were initially developed as COSII (C2_At1g11475, C2_At2g40600, C2_At2g41170, C2_At4g30950) but later re-designated as unigene markers (U146140, U221402, U318882, and U146437, respectively) were sampled for this study. All were amplified using iUPA and contained one or more introns.

Ten arbitrary markers were chosen to represent fragments of characterized, single copy genes in *S. lycopersicum* (Table 2). Primers were generally designed to amplify intron or UTR although some exonic regions were included. All genes except TG11, an anonymous RFLP marker (Nesbitt and Tanksley 2002), are known to influence fruit quality in tomato based on mutated phenotypes (see references in Table 2).

Sequencing

Seedlings were grown in a greenhouse and DNA was extracted from 50 mg young leaf tissue of one plant per accession using a modified CTAB protocol (Colosi and Schaal 1993). Primers used for PCR amplification were also used for DNA sequencing in separate forward and reverse reactions. Negative (no DNA) and positive (TA496) controls were included for every set of PCR reactions. PGRU's standard protocol for tomato PCR amplification is reported in GenBank dbSTS accessions BV448051–BV448073 (Labate and Baldo 2005). Occasionally a primer pair required minor modification of the thermoprofile for

Table 2 GenBank accessions used to design primers for arbitrary (published single-copy gene) tomato markers

Marker	Accession	Reference
hp2 exon 2	AJ224356	Mustilli et al. (1999)
hp2 3′ region	AJ224357	Mustilli et al. (1999)
Pds	X78271	Aracri et al. (1994)
Psy1	X60441	Ray et al. (1992)
fw 2.2	AY097181	Nesbitt and Tanksley (2002)
CRTISO	AF416727	Isaacson et al. (2002)
PTOX	AF177979	Josse et al. (2000)
TG11[a]	AY097137	Nesbitt and Tanksley (2002)
rin	AR580674	Vrebalov et al. (2002)
Cyc-B 5′ region[b]	AF254793	Hirschberg et al. (2001); Ronen et al. (2000)

[a] Primers from Nesbitt and Tanksley (2002)

[b] Primers designed based on US Patent 6,252,141

optimization of yield. Big Dye v. 3.1 (Applied Biosystems, CA, USA) cycle sequencing was used following manufacturer's instructions and data were collected on an ABI PRISM 3130 Genetic Analyzer. Mutation Surveyor software (SoftGenetics, PA, USA) and the phred, phrap, and Consed suite of software (Ewing and Green 1998; Ewing et al. 1998) were used to analyze trace files. Any primer pair that gave multiple PCR products or poor quality sequence in initial testing of one or two lines was excluded from additional development (Supplementary Table S2 and unpublished data). All retained markers were assumed to represent single loci, 24% of markers designed using EST databases were rejected during marker development based on multiple PCR bands or high proportions of heterozygous sites (Supplementary Table S2). Marker types COSII/unigene and arbitrary loci have been characterized in the literature and should correspond to single copy genes. Raw trace files were visually examined by two people independently at all polymorphic sites using the Consed graphical user interface. Results of Mutation Surveyor versus the Consed suite analysis packages were compared for consistency in scoring and discrepancies were resolved by visual inspection. Sequence data were trimmed of primer binding sites and low quality (phred <40) ends and prepared in the form of one FASTA formatted file containing 31 aligned genomic sequences for each marker.

Statistical analyses

PHASE v. 2.1 (Stephens and Donnelly 2003; Stephens et al. 2001) was used to infer probabilities of haplotypes when any plant was scored as heterozygous at more than one site within a marker. This occurred for 20 markers. Parameter default values of number of iterations = 100, thinning interval = 1, and burn-in = 100 were used with the value of the seed varied for each run. Four markers with $P < 0.95$ for at least one site were reanalyzed by PHASE using five runs and number of iterations = 1,000.

Population diversity measures 'number of SNPs', π (Nei 1987, equation 10.5), θ (Nei 1987, equation 10.3), and 'number of haplotypes' were estimated using DNAsp 4.10 (Rozas et al. 2003). Sequence sets were defined with and without TA496 in order to quantify contribution of this line to total numbers of SNPs, indels, and haplotypes for each of marker classes EST, COSII/unigene, and arbitrary. A one-way nonparametric ANOVA in the form of a Kruskal–Wallis test (Sokal and Rohlf 1981) was used to compare distributions of π and θ for three groups, i.e., the three classes of markers.

Principal components analysis of genotypes was performed using GenAlEx (Peakall and Smouse 2006). For this analysis, data were coded as diploid genotypes with each unique haplotype at a locus (set of correlated SNPs) treated as an allele. For 15 EST markers with previously reported SNPs (Labate and Baldo 2005) we examined ascertainment bias by comparing mean SNP diversity, h (Schlötterer and Harr 2002), for the 31 line sample at each locus by applying either pair-set SNPs (SNPs present using TA496 plus one other cultivar) or full-set SNPs (SNPs present using all 31 lines). This is a diversity measure appropriate for SNP-genotyping estimated as:

$$h = 1/m \sum_{i=1}^{m} 1 - (x^2 + y^2)$$

for m SNPs with allele frequencies x and y. Correction of h for number of sequenced chromosomes, n, was incorporated by multiplying by $n/(n-1)$. A sign test was used to test whether the median value of SNP diversity was likely to be greater, smaller, or unchanged in the full-set relative to the pair-set data. Eleven markers with no previously observed SNPs were excluded from this analysis because their h values inherently could not be scored as 'smaller'.

Annotation

FASTA files were viewed using BioEdit (Hall 1999) and markers were annotated for exons, introns, 5′ UTR, and 3′ UTR. For the arbitrary markers this was accomplished through alignment with the original annotated GenBank accession against which primers were designed (Table 2).

Conserved ortholog set II annotations were based on reference peptides reported in SGN which originated using alignments with *Arabidopsis* coding sequences. For annotating unigenes ESTScan (Iseli et al. 1999; Lottaz et al. 2003) results reported on SGN were used to predict peptide sequence.

Expressed-sequence tag markers were annotated using several software tools. NCBI's Spidey (http://www.ncbi.nlm.nih.gov) mRNA-to-genomic alignment program was used to identify introns and putative splice junctions. Genomic sequences were further annotated by (1) BLASTX searches of GenBank for protein-encoding homologs using our marker sequence as query, (2) paired alignment of our BLASTX query sequence to nucleic acid sequence of significant hit (subject), (3) alignment among query, subject, and amino acid sequence, and (4) application of splice site rules to intron/exon junctions using visual inspection. Alignment with amino acid sequences occasionally detected base-calling errors that caused a frame shift, observed when a site was dropped near sequence ends or within a short run of bases. When no protein-encoding homologs were found GeneSeqer (Usuka and Brendel 2000; Usuka et al. 2000) was used to predict exons based on plant EST databases.

TA496 genomic sequences for all markers are available in GenBank under accessions EU797528–EU797577. Observed polymorphisms and primer sequences for all markers are described in NCBI dbSNP accessions ss# 107751597–107751940.

Results

Scoring nucleotide polymorphisms by two independent methods gave a high degree of confidence to the data. Because Mutation Surveyor and phred use different base-calling algorithms they complement each other. In our experience this efficiently flagged base-calling errors made by one or the other program, or by human error in recording data, that were easily resolved by visual inspection of trace files. In 31 DNAs there were 188 polymorphisms (Table 3) for a total of 5,828 independent data points. The two methods were initially 99% congruent in the results; 60 discrepancies were observed and resolved to reach 100% congruity.

Inferring haplotypes

Of 20 markers where at least one plant was heterozygous at multiple sites, 14 markers gave $P = 1.0$ and two markers gave $1.0 > P > 0.95$ for inferred haplotypes at all polymorphic sites using PHASE software (Stephens and Donnelly 2003; Stephens et al. 2001). Markers 296_1, 1724_1, C2_At2g22570, and *rin* gave $P < 0.95$ for one to two heterozygous sites in one to two plants. Increased number of iterations did not improve *P*-values although the ranges of *P*-values were slightly tighter. The final range of *P*-values for the five plants involved were from 0.50 to 0.81 with results consistent across multiple runs. The uncertainty of these haplotypes did not affect results.

Diversity

Overall we observed 154 SNPs and 34 indels in 23,113 nucleotides, for a mean of one polymorphism per 125 nucleotides among the 31 tomato lines (Table 3). This raw estimate of variation was slightly higher in EST (99 polymorphisms in 9,187 bp, 1/93) relative to COSII/unigene (55 polymorphisms in 9,121 bp, 1/166) and arbitrary (34 polymorphisms in 4,805 bp, 1/141) markers. Of the 11 EST markers that had not previously contained a confirmed coding or discovered intronic SNP, six contained a SNP in the 31 line panel. This included three markers (1724_1, 2189_1, 4301_3) for which the original target SNP was verified.

Population parameters θ and π can be interpreted as SNP diversity per site (Nei 1987); the former is based on the number of sampled SNPs while the latter takes SNP frequency into account (i.e. expected heterozygosity). Similar mean values of π, θ, and number of haplotypes per locus

Table 3 Polymorphism in fragments from 50 loci resequenced in 31 *S. lycopersicum* lines

Marker class	No.	Regions sampled[a]	Mean no. of sequences sampled per locus	Nucleotides	SNPs[b]	Indels[b]	Mean π (maximum)	Mean θ (maximum)	Mean no. SNP haplotypes per locus[b] (maximum)
EST[c]	26	e, i, u	31.8	9,187	85; 62	14; 9	0.0017 (0.0066)	0.0021 (0.0064)	2.4; 2.2 (5; 5)
COSII/unigene	14	Mostly i	31.7	9,121	40; 38	15; 15	0.0012 (0.0092)	0.0015 (0.0086)	2.6; 2.5 (6; 5)
Arbitrary	10	Mostly i, u	32.4	4,805	29; 29	5; 5	0.0010 (0.0021)	0.0015 (0.0044)	2.8; 2.8 (5; 5)
Total	50			23,113	154; 129	34; 29			
Mean							0.00130	0.00170	2.6; 2.5 (5.3; 5)

[a] Annotations for all 50 markers are available in supplementary file sequences.fas.txt, *e* exon, *i* intron, *u* UTR

[b] First value includes TA496, second value excludes TA496

[c] All were predicted to contain SNPs based on public tomato unigene data (Labate and Baldo 2005)

were observed across the three marker classes with maximum values in the COSII/unigene class (Table 3). EST, COSII/unigene, and arbitrary markers had six, one, and one monomorphic markers respectively (1260_2, 1523_4, 1589_1, 1675_1, 2280_1, 2582_1, U146140, *Cyc-B*). A Kruskal–Wallis test statistic, H, showed that the distribution of diversity estimates was not significantly different across marker classes for π ($H = 2.37$, $df = 2$, $P = 0.306$) or θ ($H = 1.99$, $df = 2$, $P = 0.371$).

Percentage variation explained by the first three axes was 37.4, 16.2, and 13.4%, respectively, in PCoA of genotypes. Results supported Villand et al.'s (1998) observations that outliers in the scattergram originated from primary centers of diversity and occasionally from countries contiguous with primary centers (Supplementary Fig. S1). TA496 was moderately divergent from many accessions, falling within the larger cluster from primary centers and adjacent to clusters from contiguous countries and secondary centers.

Assuming our panel represents the broad diversity within the species, grand mean values of θ, π, and number of haplotypes per locus (Table 3) summarize multilocus estimates of available intraspecific polymorphism. Values of θ and π imply that approximately one to two SNPs per kb of sequence will be observed between a diverse pair of *S. lycopersicum* lines, on average.

Examination of ascertainment bias

Three of 15 EST markers with a priori observed SNPs (241_2, 2875_4, 3155_3) showed one, two, or one additional SNPs, respectively, in the 31 line panel relative to the original discovery pair of lines. For the other 12 EST markers with a priori observed SNPs no additional SNPs were found. For marker 1260_2 the single SNP in the pair-set was not observed among the 31 lines. This gain of four and loss of one SNP yielded totals of 56 and 53 SNPs in the full-set and pair-set datasets, respectively. A sign test was used to test the probability that the median value of SNP diversity would change if the full-set versus the pair-set SNPs had been genotyped in the panel. SNP diversity would be greater using the full-set SNPs if moderate frequency alleles were discovered, and smaller using the full-set SNPs if rare alleles were discovered. Loss of a SNP in marker 1260_2 in the full-set did not alter SNP diversity by resequencing versus SNP-genotyping, this marker is monomorphic in the panel using both methods. Therefore, it did not affect the sign test.

All four additional SNPs had a sample frequency of one (i.e., were singletons), so SNP diversity decreased for markers 241_2, 2875_4, 3155_3 in the full-set. This did not significantly alter the median value of SNP diversity as demonstrated by a sign test ($n - 3$, $n + 0$, $P \leq 0.25$).

Because line TA496 was specifically used to develop EST markers (Labate and Baldo 2005), its unique contribution to polymorphism was compared among the three marker classes. TA496 contributed proportionally more to total numbers of SNPs (27%), indels (26%), and haplotypes (8%) for the EST markers compared to the other two marker types (Table 3). No unique polymorphisms or haplotypes were observed in TA496 for the arbitrary markers, and 5, 0, and 4% of SNPs, indels, and haplotypes were uniquely contributed by TA496 to the COSII/unigene markers.

Annotation

For each EST marker annotation (Supplementary file sequences.fas.txt) we used the full-length tomato unigene subcluster sequence (Labate and Baldo 2005) to obtain better BLASTX matches compared to using the marker amplicon alone. Only EST marker 1287_1 did not give a significant match by our criterion ($E \leq 1 \times 10^{-25}$) to any GenBank NR protein. BLASTN found significant matches to tomato- and potato-specific expressed sequences (E values = 0.0–2 $\times 10^{-43}$). For this marker the longest open reading frame that maximized the number of synonymous, conservative polymorphisms was assumed for annotation. Marker TG11 was originally developed from a genomic DNA library. It did not significantly match any plant EST or peptide sequence so we did not assume it to be expressed and did not annotate it.

Discussion

Pioneering studies using restriction fragment length polymorphisms (RFLPs) (Miller and Tanksley 1990) and SNPs (Nesbitt and Tanksley 2002) showed domesticated tomato to be relatively low in nucleotide diversity. These observations supported a documented history of genetic bottlenecks, founder events, and directional selection within the species. Since expeditions by the US Department of Agriculture (USDA) to collect landraces and wild tomato species from centers of diversity in South America and Mexico beginning in the 1930s, wild species have been valuable sources of alleles for crop improvement (Stevens and Rick 1986). This has been especially true for various disease resistances. More recently it was demonstrated that favorable alleles for traits such as increased fruit size can be mined from wild germplasm (Tanksley et al. 1996) and techniques with which to incorporate wild alleles into modern cultivars continue to be refined (Monforte et al. 2001; Foolad 2007). Marker assisted breeding in tomato is greatly facilitated by high density genetic linkage maps such as an *S. lycopersicum* × *Solanum pennellii* F_2 map

(Fulton et al. 2002) that continues to be saturated with markers (Tomato-EXPEN 2000, SGN). There are increasing demands for intraspecific markers and maps of the *S. lycopersicum* genome for application in breeding and germplasm management (Baldo et al. 2007; Saliba-Colombani et al. 2000; Sim et al. 2007; Van Deynze et al. 2007). Substantive efforts are being directed towards confirmation of SNPs and small indels in *S. lycopersicum* and their subsequent mapping (SolCAP 2008; Ganal et al. 2007; Van Deynze et al. 2007).

We compared two approaches for SNP discovery in tomato—mining noncoding regions such as introns, and mining amplicons with predicted SNPs based on ESTs. Noncoding regions of fruit quality genes were largely chosen to represent arbitrary, single-copy sequences in order to potentially increase the utility of discovered polymorphisms. These 'arbitrary' markers were least variable in terms of mean and maximum values of population parameters π and θ, although they showed slightly more haplotypes per locus.

COSII containing introns are attractive markers for intraspecific or closely related species. The iUPA primer pairs are predicted to robustly amplify single copy loci across a range of taxa as diverse as tobacco, *petunia*, sweet potato, coffee, olive, mint, sesame, *Mimulus*, and *Antirrhinum* (Wu et al. 2006). Application of COSII markers rather than tomato-specific markers can allow a more comprehensive interpretation of data across disparate taxa when the same markers are applied. COSII markers are quickly being adopted for phylogenetic (Rodriguez et al. 2006), mapping (Crouzillat et al. 2006; Moncada et al. 2006) and diversity (Labate et al. 2006; Olarte et al. 2006) studies. Given the evolutionary conservation of COSII genes, we wanted to know if COSII intron mining would yield ample numbers of polymorphisms within *S. lycopersicum* to make SNP discovery efficient. We found the iUPA amplified COSII markers to be a rich source of polymorphisms within domesticated tomato even though they are under strong selection pressure for conservation of exons. An additional, nonoverlapping set of COS markers spanning introns (Van Deynze et al. 2007) supports this. Assays of ten lines representing US fresh market, processing, and heirloom tomatoes yielded one SNP per 1,647 bp and one indel per 4,624 bp in the UCD COS (Van Deynze et al. 2007).

Expressed-sequence tag libraries can be a valuable resource for SNP mining (Cogan et al. 2007; Hayes et al. 2007) even though they are generally not created for this purpose. When algorithms were applied to distinguish sequencing artifacts from true SNPs, many predicted polymorphisms were verified in tomato (Labate and Baldo 2005; Yang et al. 2004). Two factors seemingly contributed to more SNPs per bp in the EST markers relative to the other two marker classes in our study (1) PCR primers were designed to amplify regions with predicted SNPs in exons, (2) ten of the markers had previously observed SNPs in exons and introns, and (3) five of the markers were a priori known to be highly variable (possibly wild species' alleles) (Labate and Baldo 2005). However, the most variable marker in our study was C2_At1g73180 with nine SNPs and two indels in 266 bp. Five SNPs and both indels were in the 86 bp intron, while four SNPs were exonic. This level of variation may also be indicative of a wild tomato species allele.

Many statistical genetic properties of populations rely on SNP frequency, e.g., nucleotide diversity (π), population structure (F_{ST}), neutrality tests, and linkage disequilibrium (Clark et al. 2005). Development of molecular markers carries a risk of biasing these estimates in ways that can mimic selection or demographic (e.g., bottlenecks) scenarios (Chikhi 2008). Bias can stem from marker type, the ascertainment panel, or both. For example, a study comparing an array of available marker types (e.g., RFLPs, SSRs, SNPs) across autosomes, sex-chromosomes, and mitochondrial DNA in human populations found broad congruency in diversity measures (Jorde et al. 2000). However, autosomal RFLPs showed ascertainment bias, probably because they were originally developed based on heterozygosity in European individuals. Similarly, Smith et al. (2007) found concordance in results of estimates of Chinook salmon (*Oncorhynchus tshawytsscha*) broad-scale population structure in comparing SNPs, SSRs, and allozymes. In spite of this, bias was evident in comparing within-population diversity estimates; SNP diversities were relatively low for samples originating from outside the geographic range of the SNP ascertainment panel.

In the current study, marker bias was examined either indirectly, through comparing population diversity estimates among three marker classes, or directly, by analyzing pair-set versus full-set SNPs. The former examined marker type bias, while the latter pertained to the EST SNP ascertainment panel. There were compelling reasons to hypothesize bias among SNP markers from various classes. For example, all of the assayed EST markers were predicted to contain SNPs based on computational analysis (Labate and Baldo 2005). COSII primers were designed to amplify a broad range of taxa based on tomato, potato, pepper, and coffee orthologous sequences. Highly conserved sequences are under selection. Even neutral regions (i.e., introns) in such genes may have reduced variation within species due to 'background selection', a combination of purifying selection and hitchhiking of linked neutral regions (Andolfatto 2001). Alternatively, genes influencing fruit quality may have experienced fixation of alleles due to directional selection.

Such selective sweeps can reduce variation at linked neutral sites. These scenarios particularly apply to inbreeding species such as domesticated tomato, where opportunities for recombination are reduced (Charlesworth 2003) and linkage disequilibrium may be extensive. A high variance of polymorphism estimates was evident for all three groups of markers, which, when tested using a Kruskal–Wallis statistic, showed that differences among groups were not statistically significant. However, it cannot be ruled out that relatively small numbers of markers within groups, in conjunction with relatively low polymorphism, resulted in low statistical power.

To examine ascertainment panel bias, we asked whether ESTs with predicted SNPs based on tomato line TA496 showed evidence of bias. First, for the EST markers, what if we had performed genotyping using only the original pair-validated SNPs (for $n = 15$ loci) instead of resequencing the 31 line diversity panel? Remarkably, all informative SNPs were present in the discovery sample of two lines; four additional discovered SNPs were at a frequency of 1/31 lines (singletons). A sign test showed that SNP diversity estimates in the panel would not have been biased if we had SNP genotyped only known SNPs. This may simply reflect that there was no bias towards prediscovery of moderate frequency alleles in our study. The majority of minor alleles (defined as <0.50) at polymorphic sites were at rare frequencies (defined as ≤0.10) in the sample, even when TA496 was excluded. Only 12 of 129 SNPs (excludes TA496) were at moderate frequencies, i.e., ≥0.20 (results not shown). However, we did see a preponderance of singletons in the EST markers when TA496 was included versus excluded in the data set. EST mining predicted SNPs using *S. lycopersicum* ESTs deviated towards the identification of singletons. Singletons exert a small influence on population diversity parameters and are uninformative for taxonomic studies. The practical result of using SNPs based on TA496 ESTs may be that upwards of 27% of polymorphic sites will be monomorphic in a sample that does not include this line.

This raises the question as to whether TA496 should be avoided in SNP discovery panels. The pedigree of TA496 can be directly traced to processing-type E6203 (synonymous with FM6203 developed by Ferry Morris) (Tanksley et al. 1998; Yates et al. 2004). E6203 contains multiple disease resistance introgressions from wild tomato (see Labate and Baldo 2005 for references), as do most modern cultivars. This can inflate diversity of modern germplasm (Park et al. 2004). On the other hand, TA496 was not an extreme outlier and fell well within the range of our *S. lycopersicum* panel according to PCoA (Supplementary Fig. S1). In comparing SNP diversity among three *S. lycopersicum* types, processing tomato was the least variable, followed by fresh market, then cherry (M. Ganal, personal communication 2008). TA496 and E6203 should be particularly attractive for SNP mining in processing germplasm given their high representation in public EST databases (such as SGN, GenBank, MiBASE).

The final objective of our study was to obtain multilocus estimates of *S. lycopersicum* polymorphism. Our panel of 31 *S. lycopersicum* lines was approximately as polymorphic as one population of *S. pimpinellifolium* (mean $\theta = 1.6 \times 10^{-3}$), the closest wild relative (Roselius et al. 2005). Ten of the 50 loci in our study contained no SNPs, versus two of 15 loci in *S. pimpinellifolium* (Roselius et al. 2005). We conclude that there is a high variance in nucleotide diversity among *S. lycopersicum* loci. This includes a skewness towards low frequency SNPs, a moderate proportion of monomorphic loci, and some highly diverged (rare, possibly introgressed) alleles.

Acknowledgments We thank S. Sheffer, W. Lamboy, P. Kisly, T. Balch, and K. Timmer for excellent technical assistance. Dr. D. Spooner provided unpublished data for COSII markers, and Dr. J. Giovannoni provided primer sequences for *rin*. This work was funded by CRIS project 1907-21000-006-00D

References

Andolfatto P (2001) Adaptive hitchhiking effects on genome variability. Curr Opin Genet Dev 11:635–641

Aracri B, Bartley GE, Scolnik PA, Giuliano G (1994) Sequence of the *phytoene desaturase* locus of tomato. Plant Physiol 106:789

Baldo AM, Lamboy WF, Robertson LD, Sheffer SM, Labate JA (2007) The distribution of genetic variation in cultivated tomato. P170. Plant and Animal Genome XV, San Diego

Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. Trends Ecol Evol 18:249–256

Charlesworth D (2003) Effects of inbreeding on the genetic diversity of populations. Philos Trans R Soc Lond Ser B 358:1051–1070

Chikhi L (2008) Genetic markers: how accurate can genetic data be? Heredity 101(6):471–472. doi:10.1038/hdy.2008.106

Clark AG, Hubisz MJ, Bustamente CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15:1496–1502

Cogan N, Drayton M, Ponting R, Vecchies A, Bannan N, Sawbridge T, Smith K, Spangenberg G, Forster J (2007) Validation of in silico-predicted genic SNPs in white clover (*Trifolium repens* L.), an outbreeding allopolyploid species. Mol Genet Genom 277:413–425

Colosi JC, Schaal B (1993) Tissue grinding with ball bearing and vortex mixer for DNA extraction. Nucleic Acids Res 21:1051–1052

Crouzillat D, Wu F, Rigoreau M, Lin C, Mueller L, Tanksley S, Petiard V (2006) A synteny map for coffee based on COSII markers. Abstract 51. Solanaceae 2006, Madison

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186–194

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175–185

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

Francis D, Yang W, van der Knaap E, Hogenhout S, Deynze AV, Darrigues A (2006) DNA-microarray detection of single feature polymorphisms as a discovery tool for marker assisted selection within elite tomato populations. W299. Plant and Animal Genome XIV, San Diego

Fukuoka H, Miyatake K, Nunome T, Ohyama A, Negoro S, Kono I, Kanamori H, Yamaguchi H (2007) EST sequencing in eggplant and comparative sequence analysis for DNA marker development in *Solanum* species. P4. Plant and Animal Genome XV, San Diego

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

Ganal MW, Durstewitz G, Kulosa D, Luerssen H, Polley A, Wolf M (2007) Development of EST-derived SNP markers for plant breeding. W172. Plant and Animal Genome XV, San Diego

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser 41:95–98

Hayes B, Laerdahl JK, Lien S, Moen T, Berg P, Hindar K, Davidson WS, Koop BF, Adzhubei A, Hoyheim B (2007) An extensive resource of single nucleotide polymorphism markers associated with Atlantic salmon (*Salmo salar*) expressed sequences. Aquaculture 265:82–90

Hirschberg J, Ronen G, Zamir D (2001) Tomato gene *B* polynucleotides coding for *lycopene cyclase*. Patent number US 6252141

Isaacson T, Ronen G, Zamir D, Hirschberg J (2002) Cloning of *tangerine* from tomato reveals a carotenoid isomerase essential for the production of beta-carotene and xanthophylls in plants. Plant Cell 14:333–342

Iseli C, Jongeneel CV, Bucher P (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol, pp 138–148

Jones E, Sullivan H, Bhattramakki D, Smith J (2007) A comparison of simple sequence repeat and single nucleotide polymorphism marker technologies for the genotypic analysis of maize (*Zea mays* L.). Theor Appl Genet 115:361–371

Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, Batzer MA (2000) The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet 66:979–988

Josse E-M, Simkin AJ, Gaffe J, Labouré A-M, Kuntz M, Carol P (2000) A plastid terminal oxidase associated with carotenoid desaturation during chromoplast differentiation. Plant Physiol 123:1427–1436

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

Labate JA, Robertson LD, Sheffer SM, Lamboy WF, Baldo AM (2006) EST-based SNP markers: is there ascertainment bias in tomato? Annual meeting of the society for the study of evolution, Stony Brook

Labate JA, Grandillo S, Fulton T, Muños S, Caicedo AL, Peralta I et al (2007) Tomato. In: Kole C (ed) Genome mapping and molecular breeding in plants: vegetables. Springer, New York, pp 1–125

Lottaz C, Iseli C, Jongeneel C, Bucher P (2003) Modeling sequencing errors by combining Hidden Markov models. Bioinformatics 19:103–112

Luerssen H, Polley A, Ganal M (2006) SNP identification in tomato and pepper using comparative sequencing. W178. Plant and Animal Genome XIV, San Diego

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

Moncada P, Montoya JC, Lopez G, Gonzalez A, Iriarte G, Zarate LA, Cristancho M (2006) Advances in construction of tetraploid and diploid maps and in populations for QTL analysis in coffee. Abstract 56. Solanaceae 2006, Madison

Monforte AJ, Friedman E, Zamir D, Tanksley SD (2001) Comparison of a set of allelic QTL-NILs for chromosome 4 of tomato: deductions about natural variation and implications for germplasm utilization. Theor Appl Genet 102:572–590

Mustilli AC, Fenzi F, Ciliento R, Alfano F, Bowler C (1999) Phenotype of the tomato *high pigment-2* mutant is caused by a mutation in the tomato homolog of *DEETIOLATED1*. Plant Cell 11:145–157

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

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

Olarte A, Barrero LS, Lobo M, Tanksley S (2006) Use of COS markers for the Andean fruited species lulo and tree tomato. Abstract 385. Solanaceae 2006, Madison

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

Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Mol Ecol Notes 6:288–295

Pritchard JK (2001) Deconstructing maize population structure. Nat Genet 28:203–204

Ray J, Moreau P, Bird C, Bird A, Grierson D, Maunders M, Truesdale M, Bramley P, Schuch W (1992) Cloning and characterization of a gene involved in phytoene synthesis from tomato. Plant Mol Biol 19:401–404

Rick CM, Laterrot H, Philouze J (1990) A revised key for the *Lycopersicon* species. Tomato Genet Coop Rep 40:31

Robbins MD, Yang W, van der Knaap E, Francis D (2007) SNP variation and patterns of selection in lineages of cultivated tomato. P172. Plant and Animal Genome XV, San Diego

Rodriguez F, Wu F, Tanksley S, Spooner D (2006) A multiple single-copy gene phylogenetic analysis of wild tomatoes (*Solanum* L. section *Lycopersicon* (Mill.) Wettst.) and their outgroup relatives. Abstract 193. Solanaceae 2006, Madison

Ronen G, Carmel-Goren L, Zamir D, Hirschberg J (2000) An alternative pathway to beta-carotene formation in plant chromoplasts discovered by map-based cloning of *beta* and *old-gold* color mutations in tomato. Proc Natl Acad Sci USA 97:11102–11107

Roselius K, Stephan W, Stadler T (2005) The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. Genetics 171:753–763

Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496–2497

Saliba-Colombani V, Causse M, Gervais L, Philouze J (2000) Efficiency of RFLP, RAPD, and AFLP markers for the construction of an intraspecific map of the tomato genome. Genome 43:29–40

Schlötterer C, Harr B (2002) Single nucleotide polymorphisms derived from ancestral populations show no evidence for biased diversity estimates in *Drosophila melanogaster*. Mol Ecol 11:947–950

Sim S-C, Yang W, van der Knaap E, Hogenhout S, Xiao H, Francis D (2007) Microarray-based SNP discovery for tomato genetics and breeding. P173. Plant and Animal Genome XV, San Diego

Smith CT, Antonovich A, Templin WD, Elfstrom CM, Narum SR, Seeb LW (2007) Impacts of marker class bias relative to locus-specific variability on population inferences in chinook salmon: a comparison of single-nucleotide polymorphisms with short tandem repeats and allozymes. Trans Am Fish Soc 136:1674–1687

Sokal RR, Rohlf FJ (1981) Biometry: the principles and practice of statistics in biological research, 2nd edn. Freeman, New York

SolCAP (2008) USDA-CSREES awards $5.4 million to SolCAP. In: Zarka K (ed) SolCAP Newsl. Michigan State University, East Lansing, 5pp

Stephens M, Donnelly P (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162–1169

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989

Stevens M, Rick C (1986) Genetics and breeding. In: Atherton J, Rudich J (eds) The tomato crop. Chapman and Hall, NY, pp 35–109

Tanksley SD, Grandillo S, Fulton TM, Zamir D, Eshed Y, Petiard V, Lopez J, Beck-Bunn T (1996) Advanced backcross QTL analysis in a cross between an elite processing line of tomato and its wild relative *L. pimpinellifolium*. Theor Appl Genet 92:213–224

Tanksley SD, Bernachi D, Beck-Bunn Teresa, Emmatty D, Eshed Y, Inai S, Lopez J, Petiard V, Sayama H, Uhlig J, Zamir D (1998) Yield and quality evaluations on a pair of processing tomato lines nearly isogenic for the *Tm-2*^{a} gene for resistance to the tobacco mosaic virus. Euphytica 99:77–83

Usuka J, Brendel V (2000) Gene structure prediction by spliced alignment of genomic DNA with protein sequences: increased accuracy by differential splice site scoring. J Mol Biol 297:1075–1085

Usuka J, Zhu W, Brendel V (2000) Optimal spliced alignment of homologous cDNA to a genomic DNA template. Bioinformatics 16:203–211

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

Villand J, Skroch PW, Lai T, Hanson P, Kuo CG, Nienhuis J (1998) Genetic variation among tomato accessions from primary and secondary centers of diversity. Crop Sci 38:1339–1347

Vrebalov J, Ruezinsky D, Padmanabhan V, White R, Medrano D, Drake R, Schuch W, Giovannoni J (2002) A MADS-box gene necessary for fruit ripening at the tomato *ripening-inhibitor* (*Rin*) locus. Science 296:343–346

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 systematic 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 WC, Bai XD, 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 L, Jin G, Zhao X, Zheng Y, Xu Z, Wu W (2007) PIP: a database of potential intron polymorphism markers. Bioinformatics. doi: 10.1093/bioinformatics/btm1296

Yates HE, Frary A, Doganlar S, Frampton A, Eannetta NT, Uhlig J, Tanksley SD (2004) Comparative fine mapping of fruit quality QTLs on chromosome 4 introgressions derived from two wild tomato species. Euphytica 135:283–296