ORIGINAL RESEARCH

Single Nucleotide Polymorphism Discovery in Cultivated Tomato via Sequencing by Synthesis

John P. Hamilton, Sung-Chur Sim, Kevin Stoffel, Allen Van Deynze, C. Robin Buell, and David M. Francis*

Abstract

Plant breeding is enhanced by the availability of molecular markers for rapid screening and selection in populations. Identification of polymorphic loci in cultivated tomato (*Solanum lycopersicum* L.) has been hampered by limited genome sampling across cultivated types. Whole transcriptome sequencing of six accessions that span cultivated market classes was performed using sequencing by synthesis. A total of 291,915,037 quality filtered reads representing 17 Gb of sequence were generated. Assembly of the reads resulted in 30.6 to 34.9 Mb of sequence for each of the six accessions and provided representation of 55.3 to 59.6% of the predicted tomato gene set with a wide range of molecular function Gene Ontologies (GOs) represented. A computational pipeline was developed to identify single nucleotide polymorphisms (SNPs). When coupled with two Sanger-derived expressed sequence tag datasets and a reference genome, 62,576 nonredundant putative SNPs in tomato were identified. The SNPs within the contigs were present within all of the GO molecular function categories. The computational pipeline had validation rates in SNP genotyping assays that ranged from 95 to 100%, and the utility of these SNPs for assessing genetic variation within cultivated and wild populations was demonstrated. Collectively, the transcript sequences and the annotated SNPs provide a resource to facilitate tomato genetics and breeding efforts.

Published in The Plant Genome 5:17–29.
doi: 10.3835/plantgenome2011.12.0033
© Crop Science Society of America
5585 Guilford Rd., Madison, WI 53711 USA
An open-access publication

TOMATO (*Solanum lycopersicum* L.) has undergone intensive selection through domestication and breeding (Miller and Tanksley, 1990). Although selection generally narrows genetic diversity relative to founding populations, there has been a long tradition of intraspecific hybridization in tomato breeding. This approach has contributed to higher coefficients of genetic distance and greater allelic richness in contemporary cultivated varieties relative to landraces and vintage varieties (Park et al., 2004; Sim et al., 2009, 2011; Williams and St. Clair, 1993). In addition, breeding for market specialization with a strong emphasis on distinct plant architecture and fruit characteristics has led to genetic differentiation within contemporary lineages (Sim et al., 2011).

Measures of genetic polymorphism, genetic distance, and population differentiation are important to the management of germplasm resources, crop improvement programs, and the success of association mapping. Tomato has provided a strong model for identifying genes that distinguish domestic and wild plants but has been explored less from the perspective of post-domestication selection. Mapping in wide crosses and the characterization of genes that affect specific traits have produced substantial understanding into the mechanisms of disease resistance (e.g., Jones et al., 1994; Martin et al., 1993), plant and fruit

J.P. Hamilton and C.R. Buell, Michigan State Univ., Dep. of Plant Biology, East Lansing, MI 48824; S. Sim and D.M. Francis, The Ohio State Univ., OARDC, Dep. of Horticulture and Crop Science, Wooster, OH 44691; K. Stoffel and A.V. Deynze, Univ. of California, Seed Biotechnology Center, Davis, CA 95616. Received 16 Dec. 2011. *Corresponding author (francis.77@osu.edu).

Abbreviations: EST, expressed sequence tag; F_{st}, the proportion of total genetic variance in a subpopulation relative to the total variance; GA2, Genome Analyzer II; GO, Gene Ontology; GOSlim, Gene Ontology Slim; NPGS, National Plant Germplasm System; PC, principal component; PCA, principal component analysis; RNA, ribonucleic acid; SNP, single nucleotide polymorphism; TAIR, The Arabidopsis Information Resource; TGI, Tomato Genome Initiative.

development (e.g., Frary et al., 2000; Pnueli et al., 1998; Xiao et al., 2009), and the regulation of biochemical pathways (e.g., Liu et al., 2003; Ronen et al., 2000). Our understanding of how selection influences cultivated populations has been limited by genomic resources that emphasize only a small number of accessions and are focused on biparental populations constructed from wide crosses. The shortage of genetic tools for investigation of diversity within cultivated lineages therefore limits our ability to ask important questions regarding human selection and restricts available tools for crop improvement.

The genomic resources available for tomato include 301,822 expressed sequence tags (ESTs) for *S. lycopersicum* (NCBI, 2011a) and draft genome sequences for *S. lycopersicum* (Heinz 1706) (SGN, 2011b) and *Solanum pimpinellifolium* L. (LA1589) (SGN, 2011a). The genomic resources have not sampled germplasm well enough to offer effective insight into the rich diversity conserved in germplasm collections or breeding programs. For example, 78.6% of the EST resources are derived from just two accessions, TA496 (116,711 ESTs) (NCBI, 2011a) and Micro-Tom (120,392 ESTs) (NCBI, 2011a). The accession TA496 is a processing tomato, E6203, with the addition of the *Tm-2*^{a} introgression (Tanksley et al., 1998) while Micro-Tom is a novelty dwarf variety (Scott and Harbaugh, 1989). These genomic resources do not capture the breadth of variation within *S. lycopersicum*. The availability of markers for genetic analysis within cultivated tomato has been limiting as many markers selected based on polymorphism in wide crosses are not polymorphic within relevant germplasm (Jimenez-Gomez and Maloof, 2009).

Single nucleotide polymorphisms (SNPs) are the most common type of sequence variation in plant species (Ching et al., 2002). In tomato, strategies to develop resources for cultivated germplasm include in silico analysis of EST databases (Labate and Baldo, 2005; Yang et al., 2004) and simple sequence repeats (Frary et al., 2005), oligonucleotide array hybridization (Sim et al., 2009), and sequencing introns of conserved orthologous set genes (Van Deynze et al., 2007). These strategies demonstrated the feasibility of discovering sequence variation within genetically restricted germplasm pools, including cultivated populations. Single nucleotide polymorphism discovery through sequencing therefore appears quite promising as a means to uncover variation in agriculturally relevant populations (Robbins et al., 2011; Shirasawa et al., 2010).

While Sanger sequencing (Sanger and Coulson, 1975) and SNP discovery has provided a useful starting point for detecting polymorphic loci in tomato, the number of polymorphisms that can be defined is restricted by the genotypes sequenced and the depth of sequencing performed. Alternative methods for sequencing and genotyping that rely on highly parallel reactions and detection systems have made it possible to cost-effectively sequence and genotype large numbers of individuals (Hamilton et al., 2011). The high-throughput sequencing systems, also known as "next generation sequencing," are characterized by higher error rates, higher levels of redundancy, and much higher throughput in most platforms. For example, emulsion-based pyrosequencing as implemented by 454 (Roche Inc.) can produce up to 1 million reads of 600 to 1,000 bp in a single run (Margulies et al., 2005) whereas emulsion polymerase chain reaction and sequencing by ligation technology as implemented by Applied Biosystems, Inc. (Life Technologies Inc.), produces 100 million reads of 50 bp (Valouev et al., 2008) and sequencing by synthesis as implemented by Illumina Hi-Seq (Illumina Inc.) produces 320 to 640 million reads of 150 bp (Bentley et al., 2008).

To increase the number of SNPs available for basic and applied tomato genetics, we sequenced the transcriptomes from six tomato accessions representing fresh market, processing, cherry, and *S. pimpinellifolium*, a close progenitor of cultivated tomatoes. We generated >2.7 Gb for each accession, representing an average of 32.5 Mb of unique transciptomic sequence per line. Using the transcriptome data coupled with the draft tomato genome sequence, we were able to identify a large collection of putative SNPs for use in high-throughput genotyping. Our polymorphism discovery was confirmed using high-throughput genotyping assays, with validation rates for SNP calls based on Genome Analyzer II (GA2) (Illumina Inc.) data greater than 95.8% and as high as 100%. These results demonstrate the potential of using high-throughput sequencing technologies to identify differences between cultivated plants and study the distribution of sequence variation within genes and crop lineages.

Materials and Methods

Plant Germplasm

Tomato accessions used for transcriptome sequencing were assembled from three public breeding programs across the United States and the USDA National Plant Germplasm System (NPGS) (Table 1). Sequencing efforts were designed to expand available genomic resources to fresh-market breeding efforts while also providing additional resources for processing tomatoes as well as comparisons to more distant species. Three of the accessions (NC84173, FL7600, and OH08-6405) represent germplasm relevant to fresh-market breeding efforts, including a parent of commercial hybrids (Gardner, 1992). Previous public genomic efforts have ignored the high-value fresh-market germplasm. The line OH9242 (Francis, 2002) was also included as a commercially relevant parent to maximize genetic variation within the processing germplasm as a complement to the EST resources for TA496 (Tanksley et al., 1998) and the genome sequence for Heinz 1706 (Ozminkowski, 2004). TA496 represents the California processing tomato germplasm pool while Heinz 1706, developed in Bowling Green, OH, represents one of two subpopulations of midwestern U.S. processing material (Sim et al., 2011). The cherry accession PI 114490 and the *S. pimpinellifolium* accession PI 128216 were chosen as close relatives for the cultivated tomatoes and because both accessions have contributed to contemporary breeding populations for fresh and processing market classes.

Table 1. Accessions and sequence datasets used in this study.

Accession	Species	Market class	Sequence data[†]	Comments
FL7600	*Solanum lycopersicum*	Fresh market	GA2 ESTs	Sequenced in this study
NC84173	*S. lycopersicum*	Fresh market	GA2 ESTs	Sequenced in this study
OH08-6405	*S. lycopersicum*	Fresh market	GA2 ESTs	Sequenced in this study
OH9242	*S. lycopersicum*	Processing	GA2 ESTs	Sequenced in this study
PI 114490	*S. lycopersicum*	Cherry	GA2 ESTs	Sequenced in this study
PI 128216	*S. pimpinellifolium*	Wild	GA2 ESTs	Sequenced in this study
Micro-Tom	*S. lycopersicum*	Novelty	Sanger ESTs	Novelty dwarf variety
TA496	*S. lycopersicum*	Processing	Sanger ESTs	E6203 genetic background
Heinz 1706	*S. lycopersicum*	Processing	Genome sequence	Used in Tomato Genome Initiative[‡]

[†] GA2, Genome Analyzer II; EST, expressed sequence tag.

[‡] Genome sequence of Heinz 1706 is available from the Sol Genomics Network (SGN, 2011b).

For validation of SNPs called from the transcriptome sequences, we used a core collection of 88 tomato accessions including the six accessions selected for sequencing. These accessions are described in Supplemental Table S1. Briefly, the collection included nine representatives of wild species, three Latin American cultivars (LA1216, LA2256, and LA2281), two unimproved breeding lines (Ha7981 and Ha7998), 18 vintage cultivars, 21 fresh-market varieties, and 35 processing varieties. These accessions were assembled from 10 breeding programs in North America (the United States and Canada), the NPGS, and the C.M. Rick Tomato Genetics Resource Center. The collection contained parents of populations utilized by the tomato research community such as segmental substitution lines (M82 and LA0716; Eshed and Zamir, 1995), parents of several important recombinant inbred and inbred backcross populations (Doganlar et al., 2002; Graham et al., 2004; Kabelka et al., 2002; Robbins et al., 2009; Yang et al., 2005), and a mutation library (Menda et al., 2004). During SNP validation eight accessions were duplicated for quality control purposes.

Sequencing, de novo Assembly, and Annotation of the Tomato Transcriptomes

Ribonucleic acid (RNA) was isolated from roots, callus, young leaves, flowers, and three stages of fruit development using the cetyltrimethylammonium bromide method (Chang et al., 1993). Callus was produced on Murashige and Skoog medium supplemented with Gamborg's B-5 (Gamborg et al., 1968; Murashige and Skoog, 1962). Fruit stages corresponded to green (including immature to mature green), breaker to turning (10 to 30% red), and ripe fruit (more than 90% of the surface was red) based on the USDA color classification. The isolated RNA was pooled in equimolar concentrations to synthesize normalized complementary DNA prepared as described previously (Hamilton et al., 2011). Libraries were sequenced on an Illumina GA2 with one paired end lane and two single end lanes of 47 to 84 bp for each accession (Supplemental Table S2). Sequence reads are available in the National Center for Biotechnology Information Sequence Read Archive (accession number SRP007969) (NCBI, 2011b).

The GA2 reads for each of the six tomato accessions were assembled separately into contigs using the Velvet assembler (Zerbino and Birney, 2008) using a k-mer length of 31 and a minimum contig length of 150 bp. The insert length of FL7600 fragments was 300 bp whereas all the other accessions had fragment lengths of 350 bp (Supplemental Table S2). The estimated sequence coverage for the transcriptomes are FL7600 (35.2x), NC84173 (34.6x), OH9242 (33.7x), PI 114490 (33.7x), PI 128216 (37.3x), and OH08-6405 (38.0x). For all downstream analyses with GA2-generated sequences, only contigs (and not singleton reads) were included due to the short nature of GA2-generated reads. Sanger-derived ESTs from TA496 and Micro-Tom were assembled as described previously (Hamilton et al., 2011) and for Sanger-derived sequences, contigs (≥150 bp) were included in downstream analyses. Gene Ontology Slim (GOSlim) associations were assigned to the contigs (GA2- and Sanger-derived) using a BLASTX search (cutoff E = 1×10^{-5}) against the *Arabidopsis thaliana* (L.) Heynh. proteome (The Arabidopsis Information Resource [TAIR] 10); TAIR, 2011) and transferring the TAIR GOSlim associations from the top hit. To assess representation of the complete tomato transcriptome, contigs (GA2- and Sanger-derived) were aligned with the Heinz 1706 Tomato Genome Initiative (TGI) gene model set (version iTAG2; ITAG, 2011) using Exonerate (Slater and Birney, 2005).

Single Nucleotide Polymorphism Discovery

Genome Analyzer II-Derived Sequences

The reads from each accession were mapped separately to the TGI v1.03 tomato scaffolds (SGN, 2011b) using Bowtie v0.12.3 (Langmead et al., 2009). Only read alignments where the read mapped uniquely on the genome were retained. Paired end reads were mapped separately from the single end reads and the paired and single end SAM alignment files were sorted and merged before further analysis. The merged SAM file was converted to BAM using SAMTools v 0.1.7 (Li et al., 2009) and SNPs were called using the SAMTools pileup tool.

The raw SNP calls were filtered using the samtools.pl varFilter script with a minimum depth of 20 reads, a maximum read depth of 240, and a maximum of one other SNP in a 100 bp flanking window, and the SNP had to be 50 bp from areas identified as indels by the pipeline. (samtools.pl varFilter-d 20-D 240-W 100-N 2-w

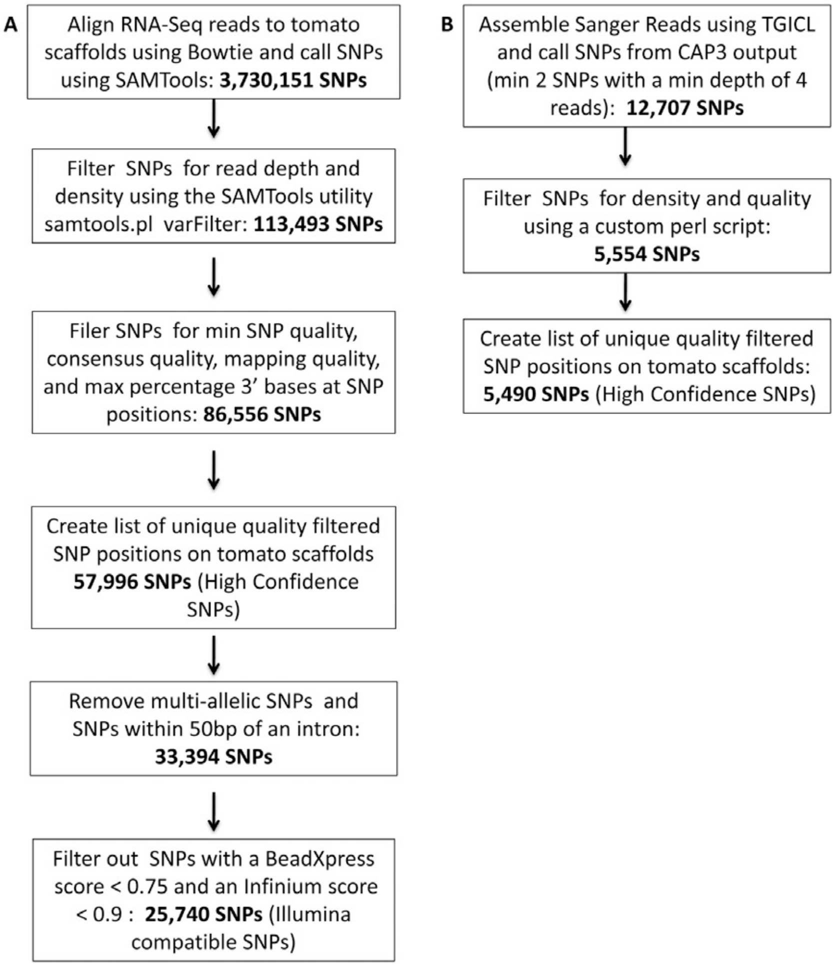
A
Align RNA-Seq reads to tomato scaffolds using Bowtie and call SNPs using SAMTools: **3,730,151 SNPs**
Filter SNPs for read depth and density using the SAMTools utility samtools.pl varFilter: **113,493 SNPs**
Filer SNPs for min SNP quality, consensus quality, mapping quality, and max percentage 3' bases at SNP positions: **86,556 SNPs**
Create list of unique quality filtered SNP positions on tomato scaffolds **57,996 SNPs** (High Confidence SNPs)
Remove multi-allelic SNPs and SNPs within 50bp of an intron: **33,394 SNPs**
Filter out SNPs with a BeadXpress score < 0.75 and an Infinium score < 0.9 : **25,740 SNPs** (Illumina compatible SNPs)
B
Assemble Sanger Reads using TGICL and call SNPs from CAP3 output (min 2 SNPs with a min depth of 4 reads): **12,707 SNPs**
Filter SNPs for density and quality using a custom perl script: **5,554 SNPs**
Create list of unique quality filtered SNP positions on tomato scaffolds: **5,490 SNPs** (High Confidence SNPs)

Figure 1. A. Single nucleotide polymorphism (SNP) discovery pipeline showing filtering steps to generate high confidence and Illumina-compatible SNPs for tomato from Genome Analyzer II transcriptome sequences. B. The SNP discovery pipeline for high confidence SNPs from Sanger-derived sequences. RNA-Seq, ribonucleic acid sequencing (reads generated from the Genome Analyzer II platform) (Wang et al., 2009); TGICL, The Institute for Genomic Research (TIGR) gene indices clustering tools (Pertea et al., 2003); CAP3, sequence assembly program (Huang and Madan, 1999).

50). The SNPs were further filtered using a custom Perl script that filtered out SNPs with a minimum consensus quality score of 20, a minimum SNP quality score 20, and a minimum mapping score 60 and SNPs where the reads aligning over the SNP position were composed of >10% of 3′ end base. The final filtered SNPs calls for each variety were then coalesced using a Perl script into a unified file of high confidence SNPs on the TGI scaffolds. The SNP positions and associated metadata were loaded into a PostgreSQL database (PostgreSQL Global Development Group, 2011) using a custom schema.

Sanger-Derived Sequences

Single nucleotide polymorphisms were called using the Sanger-derived TA496 ESTs in two phases based on availability of tomato genome sequence. Single nucleotide polymorphisms for the BeadXpress design (Illumina Inc.) were called using alignments of accession-specific assemblies to 1207 Heinz 1706 BAC sequences made available by the TGI (SolCAP, 2011). Upon availability of the draft tomato genome scaffold sequences by the TGI v1.03 (SGN, 2011b), a second phase of SNP calling was performed using Sanger-derived ESTs derived from both TA496 and Micro-Tom to augment SNP identification on a genome level using a total of eight transcriptome datasets. Sanger-derived SNPs used in the identification of high confidence SNPs required a minimum of two ESTs with the SNP (relative to Heinz 1706 genome sequence) with a minimum depth of four ESTs.

Single Nucleotide Polymorphism Validation

Single nucleotide polymorphisms were selected for validation using the BeadXpress platform in this study and the Illumina Infinium platform (Illumina Inc.) for future studies. To meet the Illumina design requirements, the contigs were aligned to the genome scaffolds using GMAP (Wu and Watanabe, 2005) to annotate the boundaries separating introns and exons. Single nucleotide polymorphisms located in exons that aligned at >95% identity without gaps were retained whereas SNPs within 50 bp of a boundary separating exons and introns and SNPs that were not biallelic across the accessions were removed (Fig. 1A). The remaining SNPs were scored for suitability for the BeadXpress and Infinium platforms by Illumina, and SNPs with a BeadXpress score <0.75, an Infinium score <0.90, or a fail code were removed. The SNPs that passed the filtering and design requirements described above were classified into six subgroups (subgroups 1–6) based on the presence of the SNP in combinations of market class, cherry, and the *S. pimpinellifolium* accession. Subgroup 1 SNPs have a SNP in the wild tomato and no SNPs in the cultivated or cherry accessions. Subgroup 2 SNPs have a SNP in the wild tomato and one cultivated variety. Subgroup 3 SNPs have a SNP in the wild tomato and a SNP in two, three, or four of the cultivated or cherry accessions. Subgroup 4 SNPs have a SNP in two, three, or four of the cultivated or cherry accessions and no SNP in the wild tomato. Subgroup 5 SNPs have one SNP in a cultivated or cherry variety. Subgroup 6 SNPs have a SNP in wild tomato and the cherry variety and no SNPs in the cultivated accessions. The subgroup membership of the SNPs was added to the SNP metadata. The GA2-derived SNPs were also classified based on sequence representation, that is, coverage within the six surveyed transcriptomes: class A SNPs contained data from all six varieties or accessions, class B from five, and class C from four. For the validation work in this study, 96 GA2-derived SNPs were selected for the BeadXpress assay and represent 60 class A SNPs, 24 class B SNPs, and 12 class C SNPs. To provide an assessment of SNP distribution among and between market classes, SNPs from the subgroups 1 through 6 were included: 12 SNPs for each of the subgroups 1 to 4 and 24 SNPs for each of subgroups 5 and 6.

For the Sanger-derived SNPs, 93 SNPs with a range of EST coverage were selected for a separate BeadXpress assay: 49 SNPs based on two ESTs, 18 SNPs based on three ESTs, six SNPs based on four ESTs, and 20 SNPs based on five or more ESTs. In addition, we included three SNPs for which calls were inconsistent among the EST sequences. For the BeadXpress assay, a total of 50 ng of genomic DNA per accession was used following the manufacturer's protocol. The BeadXpress raw data were processed using Illumina GenomeStudio software (genotyping module v1.7.4; Illumina Inc., 2010) for SNP calling.

To assess the utility of the SNPs for characterizing germplasm, SNP genotyping was performed using 88 tomato varieties with eight duplicated for quality control. The BeadXpress allele calls (Supplemental Table S1) were converted to numerical calls compatible with the Microsatellite Analyzer software v4.05 (Dieringer and Schlotterer, 2003) and to a proportional scoring (in which 2 is equal to homozygous for the common allele, 1 is equal to heterozygotes, and 0 is equal to homozygous for the rare allele) for principal component analysis (PCA) using R (R Development Core Team, 2011). Accessions were classified into subpopulations as described above. Analysis was conducted to determine the proportion of markers that were polymorphic within each subpopulation, standard allelic richness (El Mousadik and Petit, 1996), and Nei's genetic distance (Nei, 1978). The *p*-values for the pairwise distances were calculated based on 10,000 permutations and a Bonferroni correction was applied. Principal component analysis was conducted to visualize relationships and to identify SNPs that contribute to variance among the germplasm. Analysis of variance was conducted to determine whether there were significant differences between germplasm classes and principal components (PCs). Marker loadings were sorted to identify SNPs that contributed high positive and negative loadings to the eigenvector corresponding to PC1, PC2, and PC3.

Analysis of Loci under Selection

We investigated loci under positive selection using an outlier detection method as implemented in the LOSITAN workbench (Antao et al., 2008). The outlier detection method uses the available data to derive a distribution of genetic differentiation based on the proportion of total genetic variance in a subpopulation relative to the total variance (F_{st}, calculated according to Weir and Cockerham, 1984) and expected heterozygosity. Five simulations for each of three pairwise comparisons between three market classes (fresh market, processing, and vintage) were run with 10,000 iterations, a 95% confidence interval, and options for neutral and forced mean F_{st}. For the mutation model option, we used an infinite allele model. Loci that deviate from the expected distribution of neutral markers are identified based on excessively high or low F_{st}. Outliers suggest directional selection when F_{st} is higher than expected or balancing selection when F_{st} is lower than expected.

RESULTS AND DISCUSSION

Tomato Transcriptome Sequencing and Assembly

By sequencing normalized libraries from six accessions, we were able to generate 19.6 Gb (raw) and 17 Gb (quality filtered) of sequence using the Illumina GA2 platform (Table 2). For each of the six accessions, a similar number of reads were generated (51.9–59.4 million raw reads and 45.7–53.2 million quality filtered reads) suggesting that library construction and sequencing reactions were comparable among the RNA samples. De novo assembly of the GA2-derived reads resulted in 59,051 to 66,181 contigs representing 30.6 to 34.9 Mb across the six accessions further suggesting consistency of our sampling, sequencing, and assembly methods. We further examined the representation of our GA2-derived sequences of the complete tomato transcriptome by aligning these against the predicted gene set in the tomato genome. As shown in Table 3, not only is there substantial coverage of the predicted gene set (55.3–59.6%) but there also is substantial overlap between all six accessions based on alignment to the tomato gene model set as there were a limited number (185–497) of unique genes detected in each of the six accessions. Analysis of molecular function Gene Ontology (GO) associations demonstrated that the assembled transcriptomes for each of the six

Table 2. Tomato expressed sequence tag sequence and assembly statistics.

	Sanger			Genome Analyzer II						
	TA496	Micro-Tom	All Sanger	FL7600	NC84173	OH9242	OH08-6405	PI 114490	PI 128216	All Genome Analyzer II
Total no. sequences	131,308	120,462	251,770	54,162,444	52,539,617	51,954,487	59,348,840	52,727,224	57,699,707	328,432,319
Total no. bp sequences	55.3 Mb	63.6 Mb	118.9 Mb	3.1 Gb	3.2 Gb	2.9 Gb	3.4 Gb	3.1 Gb	3.9 Gb	19.6 Gb
No. sequences passed quality filters	101,154	117,562	218,716	49,053,794	45,741,571	47,425,783	53,160,164	46,228,186	50,305,539	291,915,037
No. of bp of sequences passed quality filters	55.0 Mb	62.7 Mb	117.7 Mb	2.8 Gb	2.8 Gb	2.7 Gb	3 Gb	2.7 Gb	3.0 Gb	17 Gb
No. contigs	12,349	13,570	25,919	59,581	60,534	59,051	60,031	61,310	66,118	366,625
No. Mb contigs	10.8	10.2	21.0	30.6	31.7	31	33.7	34.9	33.3	195.2
N50[†] contig size (bp)	879	761	794	863	850	880	1030	1016	812	908
Max. contig size (bp)	3,107	3,234	3,234	12,143	13,288	11,689	14,001	11,685	13,981	14,001
Min. contig size (bp)	150	102	102	150	150	150	150	150	150	150

[†] N50, a statistical measure related to the average length of a set of sequences. It is the minimum sequence size for which half of all sequences are equal to or larger.

Table 3. Coverage of the tomato genome gene complement by each transcriptome and single nucleotide polymorphism (SNP) discovery.

Tomato accession	No. contigs	No. tomato genes covered	Percent of tomato genes covered	Unique tomato gene hits[†]	Total SNPs	Accession-restricted SNPs
FL7600	59,581	19,816	55.3	265	8,132	1,900
NC84173	60,534	20,508	57.3	243	6,356	942
OH08-6405	60,031	20,632	57.6	247	7,972	2,202
OH9242	59,051	20,067	56.0	185	7,182	1,636
PI 114490	61,310	21,345	59.6	443	14,292	3,392
PI 128216	66,118	20,590	57.5	497	42,622	31,095
Micro-Tom	13,570	11,008	30.7	85	2,936	2,185
TA496	12,349	9,589	26.8	100	2,618	2,362

[†] Unique gene hits represent the number of tomato genes that aligned that were unique to that accession, that is, not detected with the other seven accessions.

accessions provide broad representation of the tomato transcriptome (294,176 total associations among the six accessions) and that the representation of genes encoding various molecular functions was similar among all six accessions (Fig. 2; Supplemental Table S3).

To provide historical reference to our GA2-generated sequences, we examined publicly available Sanger-derived ESTs from TA496 (Van der Hoeven et al., 2002) and Micro-Tom (Aoki et al., 2010). As shown in Table 2, substantially fewer ESTs were available using the Sanger platform. However, even though there are fewer reads, the length of the reads resulted in representation of 10.2 and 10.8 Mb of total assembled sequence for Micro-Tom and TA496, respectively. As anticipated, there was lower coverage of the tomato gene model set (Table 3) and fewer GO molecular function associations (Fig. 2; Supplemental Table S3) than observed with the GA2-derived contigs.

Single Nucleotide Polymorphism Discovery

In the first step of the SNP discovery pipeline (Fig. 1A), 3,730,151 SNPs were called against the tomato reference genome using the SAMTools (Li et al., 2009) SNP caller and GA2 data for the six sequenced accessions. Further filtering using SAMTools and custom Perl scripts was done on depth at the SNP position, SNP density, SNP quality, and consensus quality resulting in 86,556 SNPs that passed quality control criteria. The positions of the SNPs in the accessions were then collapsed into a set of high confidence SNPs (57,996 SNPs; Supplemental Table S4) with unique positions on the genome. The accessions with a SNP at each genomic position were tracked in a relational database. A parallel yet modified pipeline was invoked to identify SNPs in the two Sanger-generated EST assemblies, Micro-Tom and TA496, to enable a comparison of SNPs across these two sequencing pipelines and augment SNP discovery in tomato. Single nucleotide polymorphisms were called from the CAP3 sequence assembly (Huang and Madan, 1999) read alignment file using a custom Perl script yielding 12,707 SNPs. Filtering of the Sanger-derived SNPs based on base quality and density yielded 5490 high confidence SNPs (Fig. 1B; Supplemental Table S5). These SNPs were then mapped to a location on the TGI scaffolds by aligning the contigs with GMAP (Wu and Watanabe, 2005) and calculating the genomic position with a Perl script. Collectively, 62,576 nonredundant SNPs were identified in the tomato genome from the Sanger- and GA2-derived sequences. All downstream analyses were performed with the 57,996 high confidence SNPs from the six GA2-derived accessions and the 5490 high confidence SNPs from the two Sanger-derived accessions.

We identified high confidence SNPs in all eight accessions, ranging from 2618 in TA496 to 42,622 in PI 128216 (Table 3), attributable to the sequence coverage difference in Sanger- versus GA2-derived

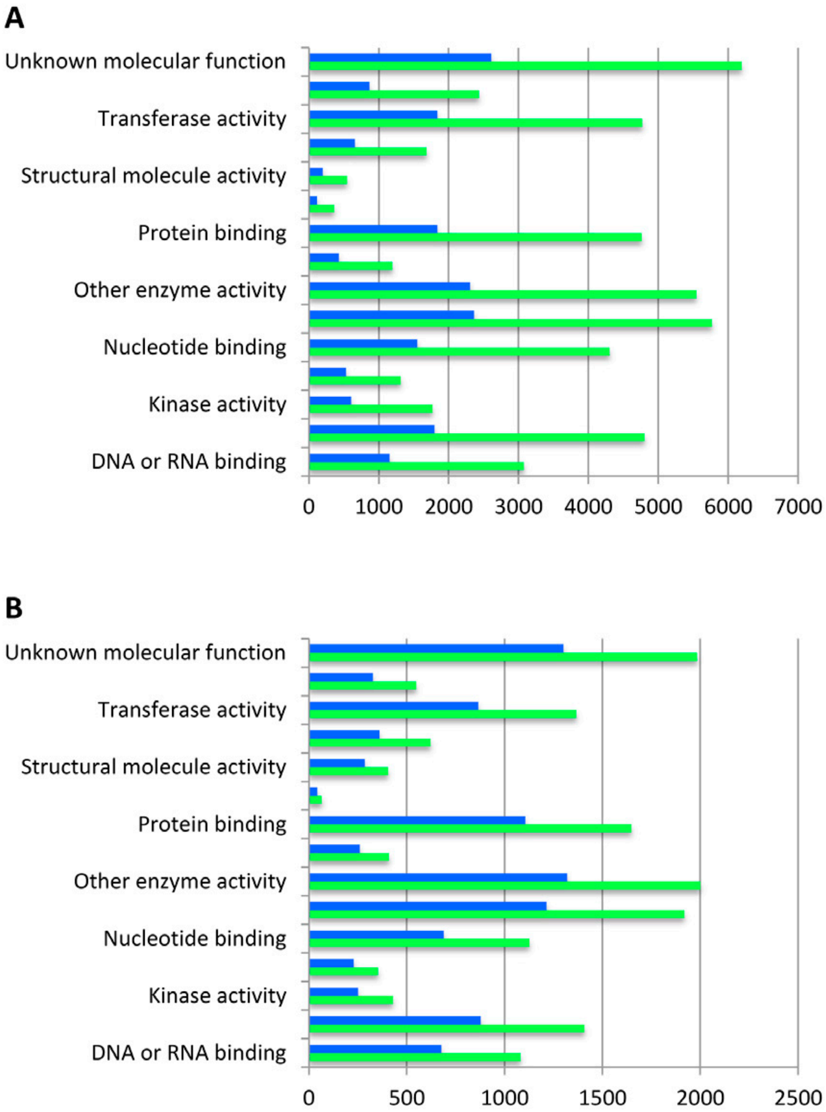

Figure 2. Molecular function Gene Ontology Slim (GOSlim) associations in FL7600 (A) and TA496 (B). Gene Ontology Slim associations were identified using the The Arabidopsis Information Resource (TAIR) GOSlim assignments for molecular function. Total GOSlim associations in each of the molecular function categories are shown in green. A subset of these associations has a high confidence single nucleotide polymorphism and is shown in blue. RNA, ribonucleic acid.

accessions and true genotype differences. A subset of these SNPs was restricted to a single accession ranging from 942 in NC84173 to 31,095 in PI 128216. Note that while we were able to identify accession-specific SNPs, this is an overestimate due to sampling limitations with transcriptome-derived SNP discovery and the stringent filtering imposed in our computational pipeline. Pairwise comparison of SNPs between all eight accessions revealed a range of total and accession-restricted SNPs (Table 4) with the fewest SNPs in any pairwise combination involving Micro-Tom and TA496, indicative of the reduced sampling of SNPs obtained through the Sanger platform. With respect to identifying SNPs within a market class, we were able to identify 2044 nonredundant SNPs unique to the fresh market class; of these, 134 SNPs were restricted to the three fresh market accessions (FL7600, NC84173, and OH08-6405; Table 5). Even fewer SNPs (89) were unique to the processing lineage, with five common to OH9242 and TA496 (for which sampling is reduced based on Sanger sequencing) relative to the reference genome Heinz 1706 (Table 5). Ample SNPs were apparent within the cherry and wild market classes (Table 5) reflective of the genetic diversity between these accessions.

Validation of Single Nucleotide Polymorphism Discovery using the BeadXpress Platform

Important components of SNP selection for genotyping are false positive rate, false negative rate, and informativeness, that is, is the SNP polymorphic in the population. We tested our SNP predictions using the BeadXpress platform. Further filtering to meet the design requirements of SNP detection platforms such as distance to the boundary between intron and exon and

Table 4. Pairwise analysis of accession-specific and accession-restricted single nucleotide polymorphisms (SNPs).

Accession 1	Accession 2	Total SNPs	Accession-restricted SNPs
FL7600	NC84173	3652	423
FL7600	OH08-6405	3164	280
FL7600	OH9242	3225	296
FL7600	PI 114490	3032	215
FL7600	PI 128216	2944	470
FL7600	Micro-Tom	158	7
FL7600	TA496	76	1
NC84173	OH08-6405	2747	162
NC84173	OH9242	3070	257
NC84173	PI 114490	2677	124
NC84173	PI 128216	2569	243
NC84173	Micro-Tom	134	8
NC84173	TA496	80	4
OH08-6405	OH9242	2992	258
OH08-6405	PI 114490	3388	508
OH08-6405	PI 128216	2639	394
OH08-6405	Micro-Tom	112	5
OH08-6405	TA496	50	0
OH9242	PI 114490	2621	139
OH9242	PI 128216	2668	492
OH9242	Micro-Tom	97	12
OH9242	TA496	89	5
PI 114490	PI 128216	8282	5534
PI 114490	Micro-Tom	259	23
PI 114490	TA496	70	7
PI 128216	Micro-Tom	569	336
PI 128216	TA496	148	62
Micro-Tom	TA496	64	33

Table 5. Single nucleotide polymorphisms (SNPs) by market class.

Market class	Total nonredundant SNPs[†]	Market class restricted[‡]
Fresh market[§]	2,044	134
Processing[¶]	89	5
Cherry[#]	14,292	3,395
Wild: PI 128216	42,622	31,095
Novelty: Micro-Tom	2,936	2,185

[†] The numbers reported are the numbers of nonredundant SNPs between the reference genome and the accessions within the market class.

[‡] Market class restricted refers to the total number of nonredundant SNPs detected in the surveyed accessions that are exclusive to the market class.

[§] FL7600, NC84173, and OH08-6405.

[¶] OH9242 and TA496.

[#] PI 114490.

allelic variants was performed using the high confidence GA2-derived SNP set to yield 25,740 SNPs compatible with both the Illumina BeadXpress and Infinium platforms (Fig. 1A; Supplemental Table S6). We first binned the Illumina-compatible SNPs into the three classes (class A, B, and C) based on coverage of that polymorphic base in the six GA2-sequenced accessions (Supplemental Table S7). We then binned the GA2-derived SNPs into six subgroups based on the presence of the SNP in each of the market classes (processing vs. fresh market), cherry, and the wild accession (*S. pimpinellifolium*) (Supplemental Table S7). This enabled us to validate our computational pipeline based on representation among the accessions and polymorphism across the genotypes. Validation tests with 96 SNPs derived from GA2-transcriptome sequences yielded a validation rate of 98% across all classes of SNPs suggesting we developed a robust computational pipeline for SNP discovery. High rates were obtained for class A (98.3%), class B (95.8%), and class C (100%), further supporting our filtering criteria for information content (Table 6).

To compare our computational pipeline with SNPs detected using the more conventional Sanger sequencing platform, we selected 96 SNPs from our alignments of TA496 ESTs to Heinz 1706 genome sequences. Validation rates for the SNPs called from the TA496 ESTs ranged from 61.3 to 95% (Table 7). The lowest rate was detected when EST coverage was two reads while the highest rate resulted from read coverage of five or greater, consistent with the notion that increasing depth of coverage improves concordance of SNPs calls. Three SNP calls based on inconsistent EST alignments were not validated, which also suggested that inconsistent calls were based on sequence error in the EST data, not heterozygosity within the source accession.

Examination of high confidence SNPs based on molecular function (Fig. 2) revealed broad coverage of the GOSlim molecular function categories among both the GA2- and the Sanger-derived sequences. However, a higher fraction of the Sanger-derived sequences had SNPs than the GA2-derived sequences.

We assessed how informative SNPs were for detection of polymorphism between and within germplasm classes. High levels of polymorphism were detected within all subclasses (Table 7). As a measure of SNP marker efficiency, the proportion of markers with no missing data was low with 86 to 99% of markers yielding no failed allele calls in each of the market classes of cultivated germplasm. We detected a higher level of missing data in the wild germplasm, a classification that consisted of *S. lycopersicum* var. *cerasiforme*, *S. pimpinellifolium*, *S. pennellii* Correll, and *S. habrochaites* S. Knapp & D. M. Spooner. The majority (74%) of missing SNP calls occurred in the *S. pennellii* and *S. habrochaites* accessions, suggesting that sequence divergence may have contributed to assay failure (Table 7; Supplemental Table S1). Levels of allelic richness (El Mousadik and Petit, 1996) for SNPs will fall between 1 and 2. Allelic richness in cultivated germplasm ranged from 1.28 to 1.39, indicating that the markers are informative (Table 7). The value of 1.54 in the wild germplasm reflected higher sequence diversity in the class. Between class genetic diversities were adjusted for sample sizes (Nei, 1978) and were consistent with between population subdivision detected in previous studies (Sim et al., 2011).

Use of PCA to visualize the distribution of germplasm based on SNP data revealed separation of the defined classes. Over 50% of the variance was explained by the first five PCs (out of 88). The first PC explained 22% of the total variance and separated wild germplasm from cultivated types ($p < 0.0001$). The second PC explained 11.5% variance and separated the cultivated classes (Fig. 3) with significant differences between vintage and landrace and both fresh market and processing ($p < 0.0004$). The two contemporary cultivated classes, processing and fresh market, were also significantly separated along the second PC ($p = 0.042$). Although the green-fruited accessions LA0407 and LA0716 were distinct from the cultivated lineages along PC1, these accessions clustered with the red-fruited *S. pimpinellifolium* accessions. This lack of discrimination may reflect some ascertainment bias as the more distant accessions were not sequenced and only polymorphisms shared with the sequenced accessions will be informative.

Inspection of the loadings for SNPs along each PC is a useful approach to identify markers and their associated chromosome segments that distinguish the germplasm. Four of the top seven SNPs (5% of total), based on high positive and negative loadings, were CL542Contig1, CL5590Contig1, CL6432Contig1, and sl_15930, from chromosome 4.

Candidate Loci under Positive Selection

Analysis of the distribution of SNP alleles within and among accessions representing distinct populations or market classes can provide insight into which areas of the genome might be under selection (Sim et al., 2009). We used an outlier detection method as implemented in the LOSITAN program (Antao et al., 2008) to investigate the distribution of validated SNPs. A total of 20 unique loci were detected from pairwise comparisons in three market classes (fresh market, processing, and vintage) as falling outside of the 95% confidence interval (Table 8). We identified four loci between fresh market and processing, eight loci between fresh market and vintage, and 12 loci between processing and vintage. Four loci overlapped between the pairwise comparisons. The F_{st} estimates of these 20 loci ranged from 0.28 to 0.88 (Table 8). The 20 loci were distributed on eight tomato chromosomes and a high portion of these loci were derived from chromosomes 2 (5 loci), 4 (5 loci), and 11 (3 loci). We inferred putative functions of these loci based on the corresponding UniRef100 and *A. thaliana* annotation (Table 8). A high proportion of these annotations are for genes involved in biotic or abiotic stress resistance, but caution is needed in interpreting the results of outlier detection as direct cause and effect. The decay of linkage disequilibrium in cultivated tomatoes occurs over centimorgan intervals (Robbins et al., 2011). The SNPs detected based on outlier detection point to regions of the genome likely to have been selected during the crop improvement process.

Table 6. Validation rates of single nucleotide polymorphism (SNP) calls.

	Genome Analyzer II-derived SNPs			
Group	Concordant	Discordant	Total	Validation rate (%)
Class A	59	1	60	98.3
Class B	23	1	24	95.8
Class C	12	0	12	100.0
	Sanger-derived SNPs			
EST[†] coverage	Concordant	Discordant	Total	Validation rate (%)
≥2	57	36	93	61.3
≥3	37	7	44	84.1
≥4	21	5	26	80.8
≥5	19	1	20	95.0

[†] EST, expressed sequence tag.

Interestingly, many of the same SNPs were identified based on F_{st} outlier detection and based on high loadings in the first and second PC. Chromosome 2 contains several genes involved in fruit size and shape whereas chromosome 11 contains a number of disease resistance genes that differentiate fresh-market varieties from vintage varieties. The second PC, which separates the cultivated varieties, contained the most overlap with the LOSITAN (Antao et al., 2008) outlier detection for SNPs on chromosome 4. The significance of chromosome 4 is less clear, compared to chromosomes 2 and 11, as there have been few genes from this chromosome that are characterized at the molecular level. A search of known morphological genes from this chromosome (TGRC,

Table 7. Informativeness of single nucleotide polymorphisms for characterizing germplasm.

				Nei's distance[†]			
Market class[‡]	Prop. PM[§]	Efficiency[¶]	Allelic richness[#]	Processing	Vintage	Latin American	Wild
Fresh market	0.68	0.86	1.31	0.051	0.050	0.111	0.242
Processing	0.67	0.93	1.39		0.082	0.138	0.210
Vintage	0.61	0.93	1.28			0.065	0.277
Latin American	0.37	0.99	1.33				0.258
Wild	0.80	0.54	1.54				

[†] Nei's standard genetic distance corrected for sample size (Nei, 1978).

[‡] Accessions and market class assignments are listed in Supplemental Table S1.

[§] Prop. PM, proportion of markers that are polymorphic.

[¶] Efficiency of markers scored in each market class is defined as the proportion of markers yielding no missing data.

[#] Standard allelic richness (total individuals – missing data) averaged across all loci (El Mousadik and Petit, 1996).

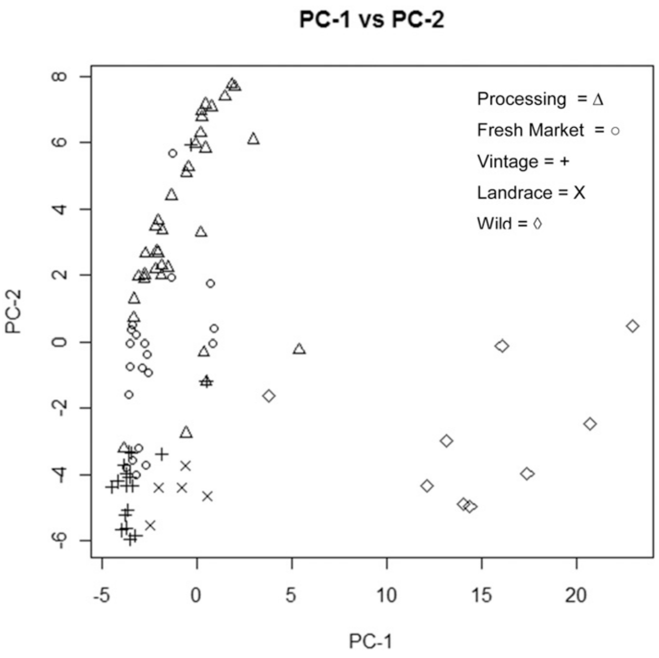

Figure 3. Principal component analysis based on the single nucleotide polymorphism validation set. Accessions from the processing market class are indicated as Δ; fresh market, ○; vintage, +; landrace, X; and Wild, ◊. PC, principal component.

2011) reveals numerous alleles affecting plant habit, leaf morphology, and sugar metabolism, which are all likely targets for modification through selection.

CONCLUSIONS

We generated extensive sequence data of transcriptomes of six tomato accessions representing fresh market (NC84173, FL7600, and OH08-6405), processing (OH9242), cherry (PI 114490), and *S. pimpinellifolium* (PI 128216). In conjunction with a draft sequence for the tomato genome, we were able to identify a large number of SNPs which were validated with high confidence. These polymorphisms complement a set identified in two Sanger-derived EST collections and in total we were able to identify 62,576 nonredundant SNPs in tomato. Validation rates of our computational pipeline for GA2 data were greater than 95.8% and as high as 100%, suggesting the set of high confidence and Illumina-compatible SNPs will be robust in genotyping assays. We demonstrate that these SNPs will be effective for characterization of cultivated and wild tomato populations. In addition, population level analysis with these SNPs appears to be a promising approach for identifying regions of the genome that are under selection due to crop improvement. As these SNPs were identified in elite germplasm

Table 8. Candidate loci under positive selection between three market classes of tomato germplasm.

		Fresh market vs. processing		Fresh market vs. vintage		Processing vs. vintage			
Locus	Chromosome	He[†]	F_{st}[‡]	He	F_{st}	He	F_{st}	UniRef100 and Arabidopsis annotation[§]	Gene Ontology term
solcap_snp_sl_12352	1	**0.58**	**0.56**	0.54	0.18	0.25	0.23	Protein transport protein sec23	endoplasmic reticulum (ER) to Golgi transport (GO:0006888)
solcap_snp_sl_13404	1	**0.57**	**0.47**	0.53	0.12	0.33	0.17	Stress-responsive protein	response to stress (GO:0006950)
CL6362Contig1	2	0.50	0.34	0.08	0.06	**0.50**	**0.49**	Unknown	Unknown
CL6523Contig1	2	0.62	0.50	0.08	0.06	**0.65**	**0.64**	Senescence-related gene 1	flavonoid biosynthesis (GO:0009813)
CL912Contig2	2	0.58	0.45	0.08	0.06	**0.60**	**0.59**	Unknown	Unknown
solcap_snp_sl_20325	2	0.32	0.31	**0.31**	**0.29**	0.43	0.00	Pro-resilin	Unknown
solcap_snp_sl_35770	2	0.37	0.36	0.37	0.36	**0.37**	**0.36**	Unknown	Unknown
solcap_snp_sl_34193	3	0.03	0.02	**0.28**	**0.26**	0.29	0.19	Unknown	Unknown
CL542Contig1	4	0.43	0.26	0.57	0.47	**0.88**	**0.86**	Annexin 1	oxidation-reduction process (GO:0055114)
CL5590Contig1	4	0.31	0.12	**0.70**	**0.62**	**0.88**	**0.86**	Exostosin family protein	biological process (GO:0008150)
CL6432Contig1	4	0.37	0.13	0.65	0.56	**0.85**	**0.83**	Serine carboxypeptidase-like 42	proteolysis (GO:0006508)
CL7515Contig1	4	0.32	0.08	**0.70**	**0.62**	**0.85**	**0.83**	Pathogenesis-related family protein	biological process (GO:0008150)
solcap_snp_sl_15930	4	0.50	0.37	0.50	0.38	**0.89**	**0.88**	Glucose-1-phosphate uridylyltransferase (UDP-glucose):glucosyltransferase	metabolic process (GO:0008152)
solcap_snp_sl_24440	6	0.37	0.06	0.15	0.13	**0.31**	**0.30**	Nucleic acid binding protein	regulation of transcription, DNA-dependent (GO:0006355)
CL2524Contig1	8	0.06	0.03	0.36	0.21	**0.34**	**0.32**	Haloacid dehalogenase-like hydrolase protein	metabolic process (GO:0008152)
CL657Contig1	8	0.17	0.16	**0.32**	**0.30**	0.38	0.04	Plant AT-rich sequence- and zinc-binding protein (PLATZ) transcription factor family protein	biological process (GO:0008150)
solcap_snp_sl_10976	11	**0.55**	**0.48**	0.54	0.34	0.13	0.03	L-ascorbate oxidase	oxidation-reduction process (GO:0055114)
solcap_snp_sl_21030	11	0.03	0.02	**0.58**	**0.57**	0.58	0.52	Disease resistance protein	defense response (GO:0006952)
solcap_snp_sl_21032	11	0.03	0.02	**0.86**	**0.86**	**0.84**	**0.82**	Unknown	Unknown
solcap_snp_sl_1567	12	**0.30**	**0.28**	**0.30**	**0.28**	0.00	0.00	Subtilisin-like protease	lateral root formation (GO:0010102)

[†] Pairwise estimates of expected heterozygosity (*He*) from the LOSITAN software (Antao et al., 2008). Bold indicates detection of loci at the 95% confidence level.

[‡] F_{st}, the proportion of total genetic variance in a subpopulation relative to the total variance, from the Lositan software (Antao et al., 2008). Bold indicates detection of loci at the 95% confidence level.

[§] Annotation was conducted using BLASTX with an E-value cutoff of 1×10^{-5} against both UniRef100 (Suzek et al. 2007; EBI, 2011) and Arabidopsis (TAIR, 2011) database.

and across market classes, they provide a resource for genetic diversity, genome-wide association studies, and marker-assisted selection in populations that are directly relevant to plant breeders.

Supplemental Information Available

Supplemental material is available at http://www.crops.org/publications/tpg.

Supplemental Table S1. List of germplasm used for single nucleotide polymorphism (SNP) validation using the BeadXpress assay.

Supplemental Table S2. Description of Genome Analyzer II (GA2) libraries sequenced.

Supplemental Table S3. Gene Ontology Slim (GOSlim) molecular function associations with eight tomato transcriptome assemblies and percentage of associations associated with single nucleotide polymorphisms (SNPs).

Supplemental Table S4. List of high confidence single nucleotide polymorphisms (SNPs) identified from Genome Analyzer II (GA2) sequences.

Supplemental Table S5. List of high confidence single nucleotide polymorphisms (SNPs) identified from Sanger sequences (Sanger and Coulson, 1975).

Supplemental Table S6. List of Illumina compatible single nucleotide polymorphisms (SNPs) identified from Genome Analyzer II (GA2) sequences.

Supplemental Table S7. Classification of Illumina-compatible Genome Analyzer II (GA2) single nucleotide polymorphisms (SNPs) based on sequence coverage and market class representation.

Acknowledgments

This work was supported by an award from the United States Department of Agriculture National Institute of Food and Agriculture (2008-55300-04757 and 2009-85606-05673). We are grateful to the Tomato Genome Initiative for the provision of the draft tomato genome sequence.

References

Antao, T., A. Lopes, R.J. Lopes, A. Beja-Pereira, and G. Luikart. 2008. LOSITAN: A workbench to detect molecular adaptation based on a Fst-outlier method. BMC Bioinformatics 9:323. doi:10.1186/1471-2105-9-323

Aoki, K., K. Yano, A. Suzuki, S. Kawamura, N. Sakurai, K. Suda, A. Kurabayashi, T. Suzuki, T. Tsugane, M. Watanabe, K. Ooga, M. Torii, T. Narita, I.T. Shin, Y. Kohara, N. Yamamoto, H. Takahashi, Y. Watanabe, M. Egusa, M. Kodama, Y. Ichinose, M. Kikuchi, S. Fukushima, A. Okabe, T. Arie, Y. Sato, K. Yazawa, S. Satoh, T. Omura, H. Ezura, and D. Shibata. 2010. Large-scale analysis of full-length cDNAs from the tomato (*Solanum lycopersicum*) cultivar Micro-Tom, a reference system for the Solanaceae genomics. BMC Genomics 11:210. doi:10.1186/1471-2164-11-210

The Arabidopsis Information Resource (TAIR). 2011. The Arabidopsis Information Resource. TAIR 10 database. Carnegie Institution of Washington, Stanford, CA. http://arabidopsis.org/ (accessed 1 Dec. 2011).

Bentley, D.R., S. Balasubramanian, H.P. Swerdlow, G.P. Smith, J. Milton, C.G. Brown, K.P. Hall, D.J. Evers, C.L. Barnes, H.R. Bignell, J.M. Boutell, J. Bryant, R.J. Carter, R. Keira Cheetham, A.J. Cox, D.J. Ellis, M.R. Flatbush, N.A. Gormley, S.J. Humphray, L.J. Irving, M.S. Karbelashvili, S.M. Kirk, H. Li, X. Liu, K.S. Maisinger, L.J. Murray, B. Obradovic, T. Ost, M.L. Parkinson, M.R. Pratt, I.M. Rasolonjatovo, M.T. Reed, R. Rigatti, C. Rodighiero, M.T. Ross, A. Sabot, S.V. Sankar, A. Scally, G.P. Schroth, M.E. Smith, V.P. Smith, A. Spiridou, P.E. Torrance, S.S. Tzonev, E.H. Vermaas, K. Walter, X. Wu, L. Zhang, M.D. Alam, C. Anastasi, I.C. Aniebo, D.M. Bailey, I.R. Bancarz, S. Banerjee, S.G. Barbour, P.A. Baybayan, V.A. Benoit, K.F. Benson, C. Bevis, P.J. Black, A. Boodhun, J.S. Brennan, J.A. Bridgham, R.C. Brown, A.A. Brown, D.H. Buermann, A.A. Bundu, J.C. Burrows, N.P. Carter, N. Castillo, E.C.M. Chiara, S. Chang, R. Neil Cooley, N.R. Crake, O.O. Dada, K.D. Diakoumakos, B. Dominguez-Fernandez, D.J. Earnshaw, U.C. Egbujor, D.W. Elmore, S.S. Etchin, M.R. Ewan, M. Fedurco, L.J. Fraser, K.V. Fuentes Fajardo, W. Scott Furey, D. George, K.J. Gietzen, C.P. Goddard, G.S. Golda, P.A. Granieri, D.E. Green, D.L. Gustafson, N.F. Hansen, K. Harnish, C.D. Haudenschild, N.I. Heyer, M.M. Hims, J.T. Ho, A.M. Horgan, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59. doi:10.1038/nature07517

C.M Rick Tomato Genetics Resource Center (TGRC). 2011. TGRC database. University of California, Davis, CA. http://tgrc.ucdavis.edu (accessed 1 Dec. 2011).

Chang, S., J. Puryear, and J. Cairney. 1993. A simple and efficient method for isolating RNA from pine trees. Plant Mol. Biol. Rep. 11:113–116. doi:10.1007/BF02670468

Ching, A., K.S. Caldwell, M. Jung, M. Dolan, O.S. Smith, S. Tingey, M. Morgante, and A.J. Rafalski. 2002. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet. 3:19. doi:10.1186/1471-2156-3-19

Dieringer, D., and C. Schlotterer. 2003. MICROSATELLITE ANALYSER (MSA): A platform independent analysis tool for large microsatellite data sets. Mol. Ecol. Notes 3:167–169. doi:10.1046/j.1471-8286.2003.00351.x

Doganlar, S., A. Frary, M.C. Daunay, R.N. Lester, and S.D. Tanksley. 2002. A comparative genetic linkage map of eggplant (*Solanum melongena*) and its implications for genome evolution in the solanaceae. Genetics 161:1697–1711.

El Mousadik, A., and R.J. Petit. 1996. High level of genetic differentiation for allelic richness among populations of the argan tree [*Argania spinosa* (L) Skeels] endemic to Morocco. Theor. Appl. Genet. 92:832–839. doi:10.1007/BF00221895

Eshed, Y., and D. Zamir. 1995. An introgression line population of *Lycopersicon pennellii* in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. Genetics 141:1147–1162.

European Bioinformatics Institute (EBI). 2011. UniRef database. Welcome Trust Genome Campus, Cambridge, UK. http://www.ebi.ac.uk/uniref (accessed 1 Dec. 2011).

Francis, D.M., T. Aldrich, K. Scaife, and W. Bash. 2002. 'Ohio OX 150' processing tomato. Tomato Genet. Coop. Rep. 52:36–37.

Frary, A., T.C. Nesbitt, S. Grandillo, E. Knaap, B. Cong, J. Liu, J. Meller, R. Elber, K.B. Alpert, and S.D. Tanksley. 2000. fw2.2: A quantitative trait locus key to the evolution of tomato fruit size. Science 289:85–88. doi:10.1126/science.289.5476.85

Frary, A., Y. Xu, J. Liu, S. Mitchell, E. Tedeschi, and S. Tanksley. 2005. Development of a set of PCR-based anchor markers encompassing the tomato genome and evaluation of their usefulness for genetics and breeding experiments. Theor. Appl. Genet. 111:291–312. doi:10.1007/s00122-005-2023-7

Gamborg, O.L., R.A. Miller, and K. Ojima. 1968. Nutrient requirements of suspension cultures of soybean root cells. Exp. Cell Res. 50:150–158. doi:10.1016/0014-4827(68)90403-5

Gardner, R.G. 1992. 'Mountain Spring' tomato; NC 8276 and NC 84173 tomato breeding lines. HortScience 27:1233–1234.

PostgreSQL Global Development Group. 2011. PostgreSQL. PostgreSQL Global Development Group. http://www.postgresql.org (verified 12/01/2011).

Graham, E.B., A. Frary, J.J. Kang, C.M. Jones, and R.G. Gardner. 2004. A recombinant inbred line mapping population derived from a *Lycopersicon esculentum* × *L. pimpinellifolium* cross. Tomato Genet. Coop. Rep. 54:22–25.

Hamilton, J.P., C.N. Hansey, B.R. Whitty, K. Stoffel, A.N. Massa, A. Van Deynze, W.S. De Jong, D.S. Douches, and C.R. Buell. 2011. Single nucleotide polymorphism discovery in elite North American potato germplasm. BMC Genomics 12:302. doi:10.1186/1471-2164-12-302

Huang, X., and A. Madan. 1999. CAP3: A DNA sequence assembly program. Genome Res. 9:868–877.

Illumina Inc. 2010. GenomeStudio software. Release 1.7.4. Illumina Inc., San Diego, CA.

International Tomato Annotation Group (ITAG). 2011. Files in ITAG2_release. Sol Genomics Network, Boyce Thompson Institute, Ithaca, NY. http://solgenomics.net/itag/release/2/list_files (accessed 1 Dec. 2012).

Jimenez-Gomez, J.M., and J.N. Maloof. 2009. Sequence diversity in three tomato species: SNPs, markers, and molecular evolution. BMC Plant Biol. 9:85. doi:10.1186/1471-2229-9-85

Jones, D.A., C.M. Thomas, K.E. Hammond-Kosack, P.J. Balint-Kurti, and J.D.G. Jones. 1994. Isolation of the tomato *Cf-9* gene for resistance to *Cladosporium fulvum* by transposon tagging. Science 266:789–792. doi:10.1126/science.7973631

Kabelka, E., B. Franchino, and D.M. Francis. 2002. Two loci from *Lycopersicon hirsutum* LA407 confer resistance to strains of *Clavibacter michiganensis* subsp. *michiganensis*. Phytopathology 92:504–510. doi:10.1094/PHYTO.2002.92.5.504

Labate, J.A., and A.M. Baldo. 2005. Tomato SNP discovery by EST mining and resequencing. Mol. Breed. 16:343–349. doi:10.1007/s11032-005-1911-5

Langmead, B., C. Trapnell, M. Pop, and S.L. Salzberg. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10:R25. doi:10.1186/gb-2009-10-3-r25

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. doi:10.1093/bioinformatics/btp352

Liu, J., B. Cong, and S.D. Tanksley. 2003. Generation and analysis of an artificial gene dosage series in tomato to study the mechanisms by which the cloned quantitative trait locus fw2.2 controls fruit size. Plant Physiol. 132:292–299. doi:10.1104/pp.102.018143

Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P. Irzyk, S.C. Jando, M.L. Alenquer, T.P. Jarvie, K.B. Jirage, J.B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, and J.M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380.

Martin, G.B., S.H. Brommonschenkel, J. Chunwongse, A. Frary, M.W. Ganal, R. Spivey, T. Wu, E.D. Earle, and S.D. Tanksley. 1993. Map-based cloning of a protein kinase gene conferring disease resistance in tomato. Science 262:1432–1436. doi:10.1126/science.7902614

Menda, N., Y. Semel, D. Peled, Y. Eshed, and D. Zamir. 2004. In silico screening of a saturated mutation library of tomato. Plant J. 38:861–872. doi:10.1111/j.1365-313X.2004.02088.x

Miller, J.C., and S.D. Tanksley. 1990. RFLP analysis of phylogenetic relationships and genetic variation in the genus *Lycopersicon*. Theor. Appl. Genet. 80:437–488.

Murashige, T., and F. Skoog. 1962. A revised medium for growth and rapid bioassays with tobacco tissue cultures. Physiol. Plant. 15:473–497. doi:10.1111/j.1399-3054.1962.tb08052.x

National Center for Biotechnology Information (NCBI). 2011a. EST database. National Institutes of Health, National Center for Biotechnology Information, Bethesda, MD. http://www.ncbi.nlm.nih.gov/nucest/ (accessed 1 Dec. 2011).

National Center for Biotechnology Information (NCBI). 2011b. Sequence read archive. National Institutes of Health, National Center for Biotechnology Information, Bethesda, MD. http://www.ncbi.nlm.nih.gov/sra/ (accessed 9 Feb. 2012).

Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a number of individuals. Genetics 89:583–590.

Ozminkowski, R. 2004. Pedigree of variety Heinz 1706. Tomato Genet. Coop. Rep. 54:26.

Park, Y.H., M.A. West, and D.A. St. Clair. 2004. Evaluation of AFLPs for germplasm fingerprinting and assessment of genetic diversity in cultivars of tomato (*Lycopersicon esculentum* L.). Genome 47:510–518. doi:10.1139/g04-004

Pertea, G., X. Huang, F. Liang, V. Antonescu, R. Sultana, S. Karamycheva, Y. Lee, J. White, F. Cheung, B. Parvizi, J. Tsai, and J. Quackenbush. 2003. TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19(5):651–652.

Pnueli, L., L. Carmel-Goren, D. Hareven, T. Gutfinger, J. Alvarez, M. Ganal, D. Zamir, and E. Lifschitz. 1998. The SELF-PRUNING gene of tomato regulates vegetative to reproductive switching of sympodial meristems and is the ortholog of CEN and TFL1. Development 125:1979–1989.

R Development Core Team. 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Robbins, M.D., A. Darrigues, S.C. Sim, M.A. Masud, and D.M. Francis. 2009. Characterization of hypersensitive resistance to bacterial spot race T3 (*Xanthomonas perforans*) from tomato accession PI 128216. Phytopathology 99:1037–1044. doi:10.1094/PHYTO-99-9-1037

Robbins, M.D., S.C. Sim, W. Yang, A. Van Deynze, E. van der Knaap, T. Joobeur, and D.M. Francis. 2011. Mapping and linkage disequilibrium analysis with a genome-wide collection of SNPs that detect polymorphism in cultivated tomato. J. Exp. Bot. 62:1831–1845. doi:10.1093/jxb/erq367

Ronen, G., L. Carmel-Goren, D. Zamir, and J. Hirschberg. 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. doi:10.1073/pnas.190177497

Sanger, F., and A.R. Coulson. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. 94(3):441–448.

Scott, J.W., and B.K. Harbaugh. 1989. Micro-Tom: A miniature dwarf tomato. Fla. Agr. Exp. Sta. Circ. S370:1–6.

Shirasawa, K., S. Isobe, H. Hirakawa, E. Asamizu, H. Fukuoka, D. Just, C. Rothan, S. Sasamoto, T. Fujishiro, Y. Kishida, M. Kohara, H. Tsuruoka, T. Wada, Y. Nakamura, S. Sato, and S. Tabata. 2010. SNP discovery and linkage map construction in cultivated tomato. DNA Res. 17:381–391. doi:10.1093/dnares/dsq024

Sim, S.C., M.D. Robbins, C. Chilcott, T. Zhu, and D.M. Francis. 2009. Oligonucleotide array discovery of polymorphisms in cultivated tomato (*Solanum lycopersicum* L.) reveals patterns of SNP variation associated with breeding. BMC Genomics 10:466. doi:10.1186/1471-2164-10-466

Sim, S.C., M.D. Robbins, A. Van Deynze, A.P. Michel, and D.M. Francis. 2011. Population structure and genetic differentiation associated with breeding history and selection in tomato (*Solanum lycopersicum* L.). Heredity 106:927–935. doi:10.1038/hdy.2010.139

Slater, G., and E. Birney. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31. doi:10.1186/1471-2105-6-31

Sol Genomics Network (SGN). 2011a. Genome: *Solanum pimpinellifolium*. Boyce Thompson Institute, Ithaca, NY. http://solgenomics.net/organism/Solanum_pimpinellifolium/genome (accessed 1 Dec. 2011).

Sol Genomics Network (SGN). 2011b. International tomato genome sequencing project. Boyce Thompson Institute, Ithaca, NY. http://solgenomics.net/organism/Solanum_lycopersicum/genome (accessed 1 Dec. 2011).

Solanaceae Coordinated Agricultural Project (SolCAP). 2011. Tomato intervarietal TA496 vs. Heinz1706 SNPs. Michigan State University, Dept. of Plant and Soil Science, East Lansing, MI. http://solcap.msu.edu/genomic_snps_ta496.shtml (accessed 1 Dec. 2011).

Suzek, B.E., H. Huang, P. McGarvey, R. Mazumder, and C.H. Wu. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23:1282–1288.

Tanksley, S.D., D. Bernachi, T. Beck-Bunn, D. Emmatty, Y. Eshed , S. Inai, J. Lopez, V. Petiard, H. Sayama, J. Uhlig, and D. Zamir. 1998. Yield and quality evaluations on a pair of processing tomato lines nearly isogenic for the Tm2(a) gene for resistance to the tobacco mosaic virus. Euphytica 99:77–83. doi:10.1023/A:1018320232663

Valouev, A., J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade, H. Peckham, K. Zeng, J.A. Malek, G. Costa, K. McKernan, A. Sidow, A. Fire, and S.M. Johnson. 2008. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. Genome Res. 18:1051–1063. doi:10.1101/gr.076463.108

Van der Hoeven, R., C. Ronning, J. Giovannoni, G. Martin, and S. Tanksley. 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. doi:10.1105/tpc.010478

Van Deynze, A.E., K. Stoffel, R.C. Buell, A. Kozik, J. Liu, E. van der Knaap, and D. Francis. 2007. Diversity in conserved genes in tomato. BMC Genomics 8:465. doi:10.1186/1471-2164-8-465

Wang, Z., M. Gerstein, and M. Snyder. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10(1):57–63.

Weir, B.S., and C.C. Cockerham. 1984. Estimating *F*-statistics for the analysis of population structure. Evolution 38:1358–1370.

Williams, C.E., and D.A. St. Clair. 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. doi:10.1139/g93-083

Wu, T.D., and C.K. Watanabe. 2005. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21:1859–1875. doi:10.1093/bioinformatics/bti310

Xiao, H., C. Radovich, N. Welty, J. Hsu, D. Li, T. Meulia, and E. van der Knaap. 2009. Integration of tomato reproductive developmental landmarks and expression profiles, and the effect of SUN on fruit shape. BMC Plant Biol. 9:49. doi:10.1186/1471-2229-9-49

Yang, W., X.D. Bai, E. Kabelka, C. Eaton, S. Kamoun, E. van der Knaap, and D. Francis. 2004. Discovery of single nucleotide polymorphisms in *Lycopersicon esculentum* by computer aided analysis of expressed sequence tags. Mol. Breed. 14:21–34. doi:10.1023/B:MOLB.0000037992.03731.a5

Yang, W., E.J. Sacks, M.L. Lewis Ivey, S.A. Miller, and D.M. Francis. 2005. Resistance in *Lycopersicon esculentum* intraspecific crosses to race T1 strains of *Xanthomonas campestris* pv. *vesicatoria* causing bacterial spot of tomato. Phytopathology 95:519–527. doi:10.1094/PHYTO-95-0519

Zerbino, D.R., and E. Birney. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821–829. doi:10.1101/gr.074492.107