Heredity (2011) 106, 927–935

www.nature.com/hdy

ORIGINAL ARTICLE

Population structure and genetic differentiation associated with breeding history and selection in tomato (*Solanum lycopersicum* L.)

S-C Sim[1], MD Robbins[1,4], A Van Deynze[2], AP Michel[3] and DM Francis[1]
[1]*Department of Horticulture and Crop Science, The Ohio State University, Ohio Agricultural Research and Development Center, Wooster, OH, USA;* [2]*Seed Biotechnology Center, University of California, Davis, CA, USA and* [3]*Department of Entomology, The Ohio State University, Ohio Agricultural Research and Development Center, Wooster, OH, USA*

Tomato (*Solanum lycopersicum* L.) has undergone intensive selection during and following domestication. We investigated population structure and genetic differentiation within a collection of 70 tomato lines representing contemporary (processing and fresh-market) varieties, vintage varieties and landraces. The model-based Bayesian clustering software, STRUCTURE, was used to detect subpopulations. Six independent analyses were conducted using all marker data (173 markers) and five subsets of markers based on marker type (single-nucleotide polymorphisms, simple sequence repeats and insertion/deletions) and location (exon and intron sequences) within genes. All of these analyses consistently separated four groups predefined by market niche and age into distinct subpopulations. Furthermore, we detected at least two subpopulations within the processing varieties. These subpopulations correspond to historical patterns of breeding conducted for specific production environments. We found no subpopulation within fresh-market varieties, vintage varieties and landraces when using all marker data. High levels of admixture were shown in several varieties representing a transition in the demarcation between processing and fresh-market breeding. The genetic clustering detected by using the STRUCTURE software was confirmed by two statistics, pairwise F_{st} (θ) and Nei's standard genetic distance. We also identified a total of 19 loci under positive selection between processing, fresh-market and vintage germplasm by using an F_{st}-outlier method based on the deviation from the expected distribution of F_{st} and heterozygosity. The markers and genome locations we identified are consistent with known patterns of selection and linkage to traits that differentiate the market classes. These results demonstrate how human selection through breeding has shaped genetic variation within cultivated tomato.
Heredity (2011) **106,** 927–935; doi:10.1038/hdy.2010.139; published online 17 November 2010

Keywords: breeding; domestication; population structure; selection

Introduction

Crop species were first domesticated from wild species about 10 000 years ago (Tanksley and McCouch, 1997). Domestication has led to dramatic changes in agronomic traits of interest such as non-shattering seeds, loss of germination inhibition, compact growth habit and increased size of fruit (Tanksley and McCouch, 1997; Doebley *et al*., 2006). During this process, however, a progressive genetic bottleneck reduced genetic diversity in cultivated plants relative to their wild ancestors (Tanksley and McCouch, 1997). After domestication, plant breeding has proceeded with the competing goals of selecting for favorable alleles and introducing new variation. As the genetic base of breeding populations narrows through selection and fixation of specific alleles, heritability declines and limits breeding progress. Thus, plant breeding also addresses the importance of maintaining a diverse genetic base and has developed methodologies to introduce new variation into cultivated plants.

Tomato (*Solanum lycopersicum* L.) has undergone intensive selection through domestication and breeding and cultivated varieties have a narrow genetic base relative to other crops (Miller and Tanksley, 1990; Williams and St. Clair, 1993; Park *et al*., 2004). For traits like fruit size and shape, however, cultivated forms show far greater phenotypic variation than their wild progenitors. Since 1930, introgression of genes for biotic stress resistance from wild species has been practised and has broadened the genetic diversity in modern varieties relative to landraces and vintage varieties (Williams and St. Clair, 1993; Park *et al*., 2004; Sim *et al*., 2009). In addition to the reintroduction of genetic variation, tomato breeding for fresh-market and processing varieties diverged with a strong emphasis on distinct ideotypes reinforced by the initiation of mechanical harvest. Efforts to develop tomatoes specifically for mechanical harvest were initiated in 1943, but did not produce acceptable varieties until the mid 1960s (Rasmussen, 1968). This market specialization has led to genetic differentiation in contemporary tomato varieties (Sim *et al*., 2009).

Correspondence: Dr DM Francis, Department of Horticulture and Crop Science, The Ohio State University, Ohio Agricultural Research and Development Center, 1680 Madison Ave, Wooster, OH 44691, USA.
E-mail: francis.77@osu.edu
[4]*Current address: Forage & Range Research Laboratory, USDA ARS, 690 N. 1100 E. Logan, UT 84322, USA*
Received 4 May 2010; revised 14 August 2010; accepted 27 September 2010; published online 17 November 2010

Understanding how genetic variation is distributed within and among populations is important to germplasm management, crop breeding and association mapping. Population structure can be inferred using pedigree information of varieties. This approach has a limitation where pedigree information is missing for certain varieties and because assumptions must be made about the relationship of progenitors in the pedigree. The use of DNA-based markers offers another approach for population level genetic analysis. When coupled to advances in software and computational power, marker analysis can offer insight into the forces that shape populations (Excoffier and Heckel, 2006).

Model-based clustering has been developed to detect underlying population structure in a collection of individuals genotyped with multiple markers. An advantage of the analysis implemented with the software STRUCTURE (Pritchard *et al*., 2000), relative to other methods of quantifying subdivision, is its ability to estimate the proportion of the genome of an individual that belongs to each inferred population (admixture). STRUCTURE's quantitative clustering method uses a Bayesian approach and has been utilized in numerous genetic diversity and association mapping studies in plant species including rice (Garris *et al*., 2005), Arabidopsis (Schmid *et al*., 2006), wheat (Breseghello and Sorrells, 2006), sorghum (Casa *et al*., 2008) and tomato (Mazzucato *et al*., 2008; van Berloo *et al*., 2008).

In this study, we investigated population structure and genetic differentiation within a collection of tomato germplasm representing contemporary (processing and fresh-market) varieties, vintage varieties and landraces using genome-wide single-nucleotide polymorphisms (SNPs), simple sequence repeats (SSRs) and insertion/deletions (InDels) derived from exon and/or intron sequences. The variation in allele frequency between subpopulations was quantified by calculating the pairwise F_{st} for each of 173 markers in order to identify genes that distinguish subpopulations. An F_{st}-outlier method was used to identify which loci may be under positive selection and therefore might be linked to regions of the genome responsible for phenotypic variation present in cultivated tomato germplasm. We believe that this approach is scalable to whole-genome analysis, and will prove to be useful for allele mining and functional characterization. We demonstrate that human selection has resulted in distinct subpopulations as a result of specialization for market types and differences between breeding programs that may correlate with environmental adaptation.

Materials and methods

Plant material

The germplasm panel of cultivated tomato used in this study consisted of 28 contemporary processing varieties, 19 contemporary fresh-market varieties, 19 vintage varieties and 4 representatives of Latin American cultivars (landraces) (Supplementary Table S1). In describing these varieties we followed the definitions adopted by Williams and St. Clair (1993), where vintage refers to varieties released before the 1960s. We prefer the description 'contemporary' to 'modern' so as to avoid confusion with the artistic and architectural movement dating to early in the 20th century. Processing and fresh-market germplasm represent contemporary varieties, and parents that were developed for specific market niches and adapted to specific environments. These varieties were selected from public breeding efforts that release commercially relevant parents and hybrids. Several processing lines were donated directly by seed companies. In addition, selected inbred lines were obtained through self-pollination and sequential single-seed descent of commercial hybrids. These selections represent a sample of alleles present in commercial hybrids, though they do not recreate the parents themselves. Two main regions are represented in the processing germplasm, material adapted to the arid conditions of California and material adapted to the humid conditions of the United States and Canada surrounding the Great Lakes. The vintage germplasm included cultivars that were developed before application of Mendelian principles and are sometimes referred to as heirlooms, and those that represent early tomato improvement efforts. The four landraces represent early domesticates from Costa Rica (LA1215), Panama (LA1216), Yucatan, Mexico (LA1462) and San Martin, Peru (LA2256) and were selected to represent geographical diversity near the center of domestication (Williams and St. Clair, 1993; Sim *et al*., 2009). Although we included a few landraces in the germplasm panel, the focus was on the contemporary and vintage varieties that reflect the history of tomato breeding and are widely used as breeding materials.

Genotyping

Genomic DNA was isolated from fresh, young leaves using the modified CTAB method described by Kabelka *et al*. (2002). A total of 300 markers including 149 markers from transcribed (exon) sequences (Yang *et al*., 2004, 2005a) and 151 markers from intron sequences (Van Deynze *et al*., 2007) were used to genotype the collection of cultivated tomato varieties. We used multiple genotyping platforms including agarose gel electrophoresis for SNP detection as cleaved amplified polymorphic sequences (CAPS) and the LUMINEX 200 (Luminex, Corp., Austin, TX, USA) for detection of SNPs using allele-specific primer extension (ASPE) (Lee *et al*., 2004). The LICOR IR2 system (LICOR, Lincoln, NE, USA) was used for fragment analysis of SSR and InDel markers. PCR reactions were conducted in a total volume of 10 μl containing 10 mM Tris–HCl (pH 9.0 at room temperature), 50 mM KCl, 1.5 mM $MgCl_2$, 50 μM of each dNTP, 0.1 μM of each forward and reverse primers, 20 ng of DNA template and 1 unit of *Taq* DNA polymerase. To visualize the amplicons on the LICOR system, PCR reactions were conducted using a set of three primers including an additional IRD 700 or 800 dye-labeled M-13 forward primer (LICOR), an M-13-tailed forward primer and a reverse primer. Amplification was performed in a thermal cycler (MJ Research, Inc., Watertown, MA, USA) programmed for 3 min at 94 °C followed by 40 cycles of 45 s at 94 °C, 45 s at a suitable annealing temperature between 45 and 60 °C, and 1 min 45 s at 72 °C, followed by an extended incubation for 6 min at 72 °C. Amplicons of SSR markers were mixed with formamide loading buffer (95% formamide, 20 mM EDTA pH 8.0 and 0.08% bromophenol blue) and were separated on 6–7% denaturing acrylamide gels. Detection of InDel markers was performed using the LICOR system and 2–4% agarose

gels. For SNP markers that were easily and cost-effectively scored as CAPS, amplification was performed in 20 μl reactions, amplicons were digested with restriction enzymes and separated on 2–4% agarose gels.

The SNP markers that could not be screened as CAPS were scored in an ASPE assay. We multiplexed 10–15 markers per ASPE assay and detected SNPs with the Luminex 200 (Luminex, Corp.). In detail, the target regions containing SNPs were first amplified using a pair of flanking forward and reverse primers. For multiplexing, 5 μl of each amplicon were pooled and precipitated by adding two volumes of PR solution (90% ethanol, 0.2 M sodium acetate pH 7.0), thoroughly mixing, centrifuging at 2500 g for 30 min and discarding the supernatant. The precipitate was washed with 70% ethanol, centrifuged at 2500 g for 10 min, and allowed to air dry after the supernatant was removed. The purified PCR products were then rehydrated in 8 μl ddH_2O and 4 μl were used as a template in 10 μl ASPE reactions including 1.25 mM $MgCl_2$, 5 μM each of dATP, dGTP and dTTP, 5 μM biotin-14-dCTP (Invitrogen Corporation, Carlsbad, CA, USA), 25 nm of each ASPE primer and 1 unit of Platinum GenoType *Tsp* DNA Polymerase (Invitrogen Corporation) in 1X supplied buffer. The ASPE primers were designed for each marker using Primo SNP 3.4 (Chang Bioscience; www.changbioscience.com/primo/primosnp.html) or BatchPrimer3 (You *et al*., 2008). Cycling conditions for the ASPE reactions were 2 min at 96 °C followed by 30 cycles of 30 s at 94 °C, 1 min at 55 °C, and 2 min at 74 °C.

Data analysis

The collection of 70 tomato lines was genotyped using 300 markers including SNPs, SSRs and InDels derived from both exon and intron sequences. We subsequently excluded 127 markers because they showed no polymorphism in the selected germplasm. The 173 polymorphic markers were grouped based on type (SNPs, SSRs and InDels) and location of polymorphism (exon and intron sequences) for genetic structure analyses in the tomato germplasm (Supplementary Table S2). We therefore analyzed six data sets, based on marker type and location, and including combined analysis. Population structure was first inferred using the model-based clustering program STRUCTURE v2.2 (Pritchard *et al*., 2000; http://pritch.bsd.uchicago.edu/structure.html). The STRUCTURE model we used in this study allows for admixture and correlated allele frequencies. To find the best K (number of clusters) for each data set, we first tested a continuous series of Ks (1–10) in five independent runs for each K with a burn-in of 10 000 iterations and a run length of 100 000 iterations. The five log likelihood values for each K were then plotted to find the Ks around a plateau of the likelihood values. The Ks selected as candidates were further tested in 20 independent runs for each K with a burn-in of 500 000 iterations and a run length of 750 000. To determine the best K, the log likelihood values from these 20 runs were subjected to analysis using nonparametric Wilcoxon (Rosenberg *et al*., 2001) and Kruskal–Wallis tests as implemented in SAS (SAS software, Cary, NC, USA). We also identified the best K using the delta K method (Evanno *et al*., 2005).

In order to validate the STRUCTURE clustering, we estimated pairwise θ (Weir and Cockerham, 1984), hereafter referred to as F_{st}, and Nei's standard genetic distance (Nei, 1978) using the Microsatellite analyzer v4.05 (Dieringer and Schlotterer, 2003). The *P*-value for the pairwise F_{st} was calculated based on 10 000 permutations and a Bonferroni correction was applied. Allelic richness of polymorphic markers was calculated as a sample size adjusted measure of diversity (Hurlbert 1971) as implemented in the Microsatellite analyzer program.

We identified loci under positive selection between processing, fresh-market and vintage varieties using an F_{st}-outlier detection method as implemented in the LOSITAN workbench (Beaumont and Nichols, 1996; Antao *et al*., 2008). The outlier detection method uses the available data to derive a distribution of F_{st} and expected heterozygosity. Five simulations for each of three pairwise comparisons were run for 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 on the basis of 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

Genotyping with SNPs, SSRs and InDels derived from exon and intron sequences

Of the 300 markers screened, we detected at least two alleles from 173 markers (86 SNPs, 52 SSRs and 35 InDels) within the 70 tomato lines tested (Supplementary Table S2). The polymorphic SNPs were derived from both exon sequences (36 SNPs) and intron sequences (50 SNPs), whereas the polymorphic SSRs were from exon sequences only and the InDels were from intron sequences only. Allelic richness did not vary greatly between germplasm groups or marker types, though there was a slight trend toward a greater number of alleles for SSR markers, markers from exon sequences and for landrace varieties (Table 1).

Model-based Bayesian clustering analysis

Six data sets were independently used for the model-based clustering method as implemented in STRUCTURE. These data sets included all 173 markers, 89 exon-sequence markers, 84 intron-sequence markers, 87 SNP markers, 52 SSR markers and 34 InDel markers, respectively. The best K (number of clusters) was tested using non-parametric analysis (Rosenberg *et al*., 2001) and the delta K method (Evanno *et al*., 2005). With all six data sets, the best K ranged from five to six, differentiating the four predefined groups based on market niche and age of variety (Figure 1 and Supplementary Tables S1). The two statistical methods suggested the best K was six for data sets consisting of all markers and intron markers. Both methods suggested the best K was five for the data set consisting of exon markers. For the data set consisting of SSR markers, non-parametric analysis determined $K = 5$, whereas the delta K method suggested $K = 6$. The delta K method failed to identify a best K for the SNP and InDel data sets, whereas the non-parametric analysis suggested that $K = 6$.

Table 1 Allelic richness of polymorphic markers in four predefined groups of tomato germplasm

Market class	*Allelic richness*						
	Exon sequence marker			*Intron sequence marker*			*Total*
	SNP	*SSR*	*Total*	*SNP*	*InDel*	*Total*	
Processing	1.40 (21)[a]	1.68 (37)	1.54 (58)	1.52 (35)	1.46 (22)	1.49 (70)	1.52 (128)
Fresh-market	1.37 (22)	1.67 (38)	1.52 (60)	1.43 (38)	1.31 (17)	1.37 (55)	1.44 (115)
Vintage	1.27 (15)	1.75 (36)	1.51 (51)	1.17 (17)	1.26 (20)	1.22 (37)	1.36 (88)
Landrace	1.32 (12)	2.15 (41)	1.74 (53)	1.43 (21)	1.45 (14)	1.44 (35)	1.59 (88)

Abbreviations: InDel, insertion/deletions; SNP, single-nucleotide polymorphisms; SSR, simple sequence repeats.
[a]Number of polymorphic markers.

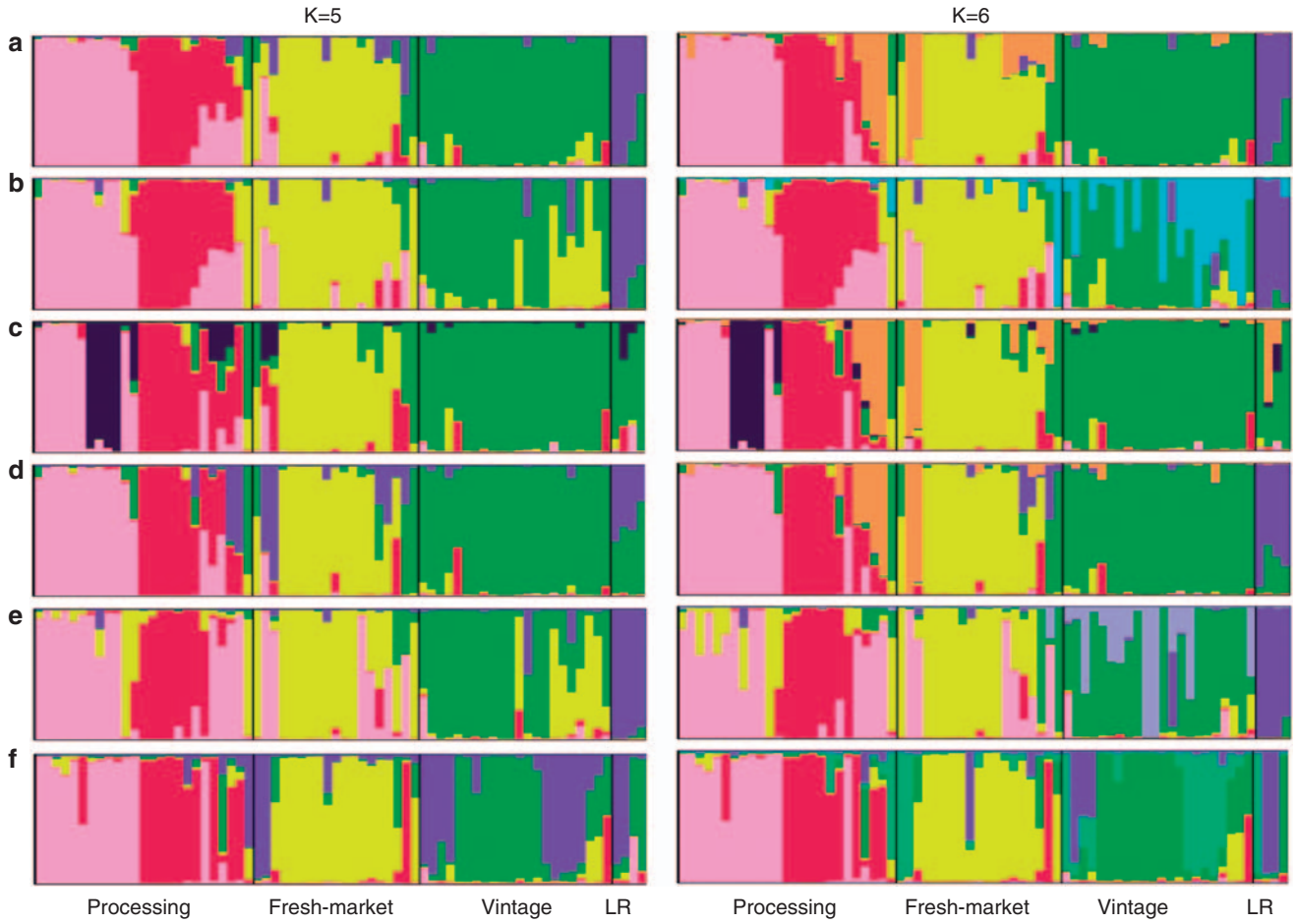

Figure 1 Inferred population structure in a collection of tomato germplasm (28 processing varieties, 19 fresh-market varieties, 19 vintage varieties and 4 landraces) using the model-based program STRUCTURE (Pritchard *et al*., 2000). Results shown are for K = 5 and 6 subpopulations. *y*-axis in figure indicates the estimated membership coefficients for each individual. Each variety's genome is represented by a single vertical line, which is partitioned into colored segments in proportion to the estimated membership in the five or six subpopulations. Black line separates individuals of four predefined groups. The landrace is represented by LR in figure. (**a**) all 173 markers, (**b**) 89 expressed-sequence markers, (**c**) 84 intron-sequence markers, (**d**) 87 SNP markers, (**e**) 52 SSR markers and (**f**) 34 InDel markers.

The processing varieties were divided into at least two and upto four clusters. The first cluster (pink color, Figure 1) included all of the varieties with pedigrees tracing to California including M82, which is a selection from the inbred line UC82 (Lawson *et al*., 1997). This group also contained upto six of the varieties with pedigrees tracing to the Ohio breeding program. The second cluster (red color, Figure 1) consisted of the other varieties from Ohio and those from Canada. The geographic regions are distinguished by arid and humid growing conditions for California and the Great Lakes (Ohio, United States and Ontario, Canada), respectively. The intron sequence markers further divided the first cluster by separating varieties from humid environments (purple color, Figure 1) from the varieties derived from arid environments (Figure 1C). Three of four Canadian varieties from Ontario (orange color, Figure 1) were separated from these clusters when K = 6, the best K for

the three data sets consisting of all 173 markers (Figure 1a), 84 intron sequence markers (Figure 1c) and 87 SNP markers (Figure 1d). In addition to four clusters, several processing varieties showed relatively high levels of admixture. For example, the varieties developed by Campbell's seed for processing into soups and juices, 'Campbell 28' (C28; LA 3317) and 'Campbell 1327' (C1327; PI 341132) both showed high levels of admixture. C28 is the older of the two varieties and shares the fresh-market ideotype with respect to fruit morphology and vine size. This variety represents the period when the processing crop was hand harvested and there were few phenotypic differences between fresh-market and processing varieties. C1327, released in the 1960s, represents a transitional variety to mechanical harvest types as fruit shape, core size and vine type were modified through selection to fit changing production methods.

In contrast to the processing varieties, a single cluster was detected at $K=5$ and $K=6$ within the fresh-market varieties regardless of type and location of marker polymorphism (Figure 1). High levels of admixture, however, were detected from several fresh-market varieties including 'Rio Grande' and 'NC99471-3'. Both of these varieties are 'Roma' style tomatoes representing a fresh-market niche consisting of fruit with the slightly elongated shape characteristic of processing tomatoes. As such, these likely share genetic background with processing varieties. For the vintage varieties, the analyses with three data sets (all 173 markers, the intron markers and the SNP markers) detected a single cluster and low levels of admixture (Figures 1a, c, and d). The other data sets (the exon markers, the SSR markers and InDel markers) detected two clusters and high levels of admixture (Figures 1b, e, and f). Four landraces were grouped together and separately from processing and fresh-market varieties (Figure 1).

Pairwise F_{st} (θ) and Nei's standard genetic distance (D) analysis

We tested the hypothesis that the groups defined by STRUCTURE represent statistically supported subpopulations using pairwise F_{st} and D (Table 2). The parameter F_{st} has been used as a standard measure of differentiation, though this use is not without controversy (Hedrick, 2005; Jost, 2008). Results from the two methods were correlated well for our data set ($R^2=0.72$, $P<0.05$), probably because heterozygosity in tomato populations does not approach the high values where Ht $\sim$ Hs. Thus, for our collection bootstrapping of F_{st} values provides an independent and objective test of population structure. The processing varieties represented a distinct subpopulation relative to fresh-market varieties ($F_{st}=0.22$, $P<0.001$), vintage varieties ($F_{st}=0.28$, $P<0.001$) and landraces ($F_{st}=0.28$, $P<0.001$) (Table 2). Similarly, the fresh-market varieties represented a distinct subpopulation relative to vintage varieties ($F_{st}=0.23$, $P<0.001$) and landraces ($F_{st}=0.31$, $P<0.001$) varieties. The vintage varieties were also genetically differentiated from the landraces ($F_{st}=0.20$, $P<0.001$). This genetic differentiation between the four defined groups was supported by analyses using all subsets of the marker data with the exception of the InDel markers, which indicated no significant genetic differentiation of the landraces from the processing and vintage varieties. This result is most likely because of small sample size of the landrace collection.

The analysis with STRUCTURE further suggested at least two and as many as four clusters of processing varieties. Genetic differentiation between the first cluster, which consisted of varieties from California and the second cluster, which consisted of varieties from Ohio was well supported ($F_{st}=0.59$, $P<0.005$; $D=0.22$). However, the third and fourth clusters consisting of varieties from Ohio and Canada, respectively, were not supported (Table 2). Two clusters of vintage varieties were detected using SSR and InDel markers when $K=6$. Genetic differentiation of these clusters was not supported (Table 2). The analysis with STRUCTURE detected subpopulations that were consistent with predefined groups based on market niche and variety age. Further subdivision of the processing varieties into two clusters is supported by pairwise F_{st} and D.

Candidate loci under positive selection

We conducted further analysis to identify candidates for loci that are under positive selection between three germplasm groups based on market class or variety age including processing, fresh-market and vintage varieties.

Table 2 Pairwise estimates of F_{st} (θ) and Nei's standard genetic distance (D) between predefinded groups and between STRUCTURE clusters

Predefind groups	*Processing*				*Fresh-market*	*Vintage*		*Landrace*
Processing					0.22****	0.28****		0.28****
Fresh-market	0.08					0.23****		0.31****
Vintage	0.10				0.06			0.20****
Landrace	0.15				0.13	0.06		
STRUCTURE clustering	PC1	PC2	PC3	PC4	Fresh-market	VC1	VC2	Landrace
Processing cluster 1 (PC1)		0.59***	0.39^{NS}	0.37^{NS}	0.48***	0.53***	0.48*	0.45^{NS}
Processing cluster 2 (PC2)	0.22		0.68^{NS}	0.45^{NS}	0.52***	0.48***	0.54*	0.48^{NS}
Processing cluster 3 (PC3)	0.13	0.22		0.45^{NS}	0.58**	0.56***	0.59^{NS}	0.44^{NS}
Processing cluster 4 (PC4)	0.15	0.10	0.15		0.40^{NS}	0.35*	0.34^{NS}	0.18^{NS}
Fresh-market	0.18	0.15	0.23	0.12		0.36***	0.33**	0.42*
Vintage_cluster 1 (VC1)	0.21	0.14	0.22	0.11	0.10		0.08^{NS}	0.21*
Vintage_cluster 2 (VC2)	0.19	0.14	0.22	0.11	0.09	0.02		0.18^{NS}
Landrace	0.26	0.18	0.25	0.14	0.16	0.06	0.08	

Pairwise estimates of θ (Weir and Cockerham, 1984) appears above the diagonal and D (Nei, 1978) appears below the diagonal. For pairwise estimates of F_{st}, P-value was calculated by 10 000 permutations with Bonferroni correction. NS, not significant, $^{*}P<0.05$, $^{**}P<0.01$, $^{***}P<0.005$, and $^{****}P<0.001$.

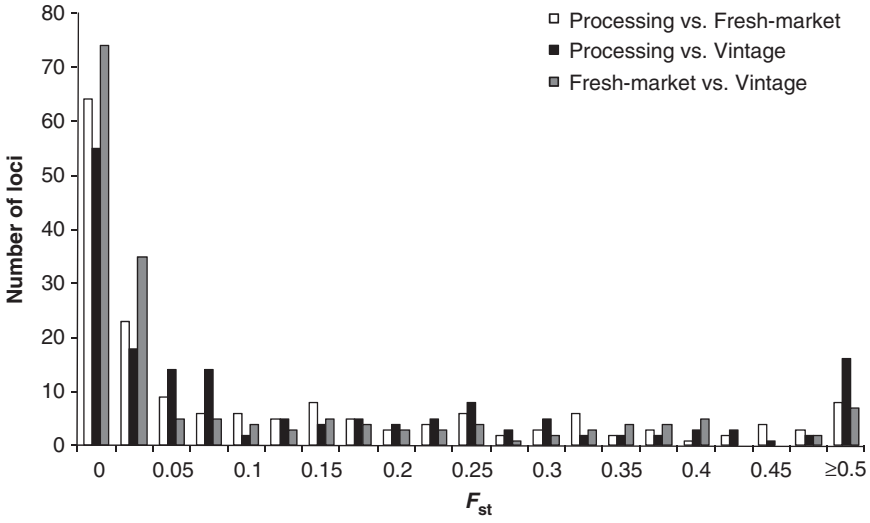

Figure 2 Distribution of pairwise F_{st} (θ) for 173 loci between three subpopulations of cultivated tomato varieties.

These loci represent candidates for further functional analysis in order to identify the loci underlying phenotypic differences between varieties. All processing varieties were used as a group for this analysis because of small sample size of each subpopulation. We also excluded the landrace group that consisted of four accessions. The pairwise F_{st} on a single locus basis was first calculated for all 173 markers using the Microsatellite analyzer (Dieringer and Schlotterer, 2003). We observed considerable variation for F_{st} among the 173 loci between the groups (Figure 2). A high portion of loci ranging 32–43% was identified with $F_{st} = 0$, suggesting allele fixation or equal allele frequencies between the groups. We also found a number of loci with fairly high levels of F_{st} ($\geqslant 0.3$): 32 loci (19%) between processing and fresh-market, 36 loci (21%) between processing and vintage and 27 loci (16%) between fresh-market and vintage.

Analysis of F_{st} on a locus-by-locus basis provides no statistical cutoff for identifying loci that may be under positive selection. Therefore, we used an outlier detection method as implemented in the LOSITAN program (Beaumont and Nichols, 1996; Antao *et al*., 2008). We identified 8 loci between processing and fresh-market, 15 loci between processing and vintage and 5 loci between fresh-market and vintage (Table 3). A total of 19 unique loci were detected as falling outside of the 95% confidence interval. Nine loci overlapped between three pairwise comparisons. The F_{st} values of these 19 loci ranged from 0.27 to 0.97 (Table 3). A high portion of these loci (47%) were derived from chromosomes 2 and 5. Among the 19 loci, the self-pruning gene, which controls determinate growth habit in tomato (Pnueli *et al*., 1998) was also identified as under selection between contemporary varieties (both fresh-market and processing) and vintage varieties. We inferred putative functions of the other loci based on the corresponding Sol Genomics Network unigene annotation (Table 3).

Discussion

We investigated population structure and genetic differentiation within cultivated tomato germplasm. The model-based clustering method implemented in the STRUCTURE program revealed that four predefined groups of tomato (processing varieties, fresh-market varieties, vintage varieties and landraces) represented distinct subpopulations regardless of marker type (SNPs, SSRs and InDels) and location of the SNP within a gene (exon and intron sequences). This finding was supported by bootstrap analysis of estimates of pairwise F_{st} (θ) at $P < 0.05$ with a Bonferroni correction. These results are consistent with previous studies demonstrating separation between vintage and processing varieties (Williams and St. Clair, 1993; Park *et al*., 2004), and genetic differentiation among market niches of cultivated tomato (Sim *et al*., 2009).

We detected at least two subpopulations within the processing varieties. Subpopulations associated with distinct fruit morphologies were observed within commercial European greenhouse cultivars (van Berloo *et al*., 2008) and within Italian tomato landraces (Mazzucato *et al*., 2008). The processing varieties we examined show little variation for fruit morphology. Rather, the population structure appears to reflect breeding history and possibly selection for different environments. Historically, the breeding of tomatoes for processing in North America has been conducted in California, the Midwest of the United States, the East Coast of the United States and Ontario, Canada. These programs drew from different founder parents and selection was conducted under arid conditions in California and humid conditions in other locations. A review of the Ohio breeding program pedigree records verifies that gene exchange occurred commonly between programs selecting under humid conditions, and more rarely between programs selecting for arid environments (for example, Berry *et al*., 1992, 1993). Thus, the subpopulation structure observed within the processing varieties reflects both selection for environment and breeding history. In contrast, we observed no subpopulation structure within the fresh-market germplasm representing varieties developed from two major breeding programs in the United States (University of Florida and North Carolina State University). Both of these programs select under humid conditions.

Table 3 Candidate loci under positive selection between three predefined groups of tomato germplasm

Marker	*Chromosome*	*Map location (cM)*[a]	*Exon/Intron*	*SGN unigene ID*[b]	*Expected heterozygosity/F_{st}*[c]			*SGN annotation*
					Processing vs Fresh-market	*Processing vs Vintage*	*Fresh-marvet vs Vintage*	
SL20105	1	14.7	Intron	SGN-U582267	0.48/0.35	*0.48 /0.47*[d]	0.06/0.03	Unknown
SL10153	2	1.2	Intron	SGN-U577497	0.47/0.00	*0.31/0.30*	0.41/0.39	Unknown
SL10649	2	6.4	Intron	SGN-U567848	*0.48/0.47*	*0.48/0.47*	0.00/0.00	Urease accessory protein (UREG)
SL10050	2	43.5	Intron	SGN-U576990	*0.50/0.49*	*0.50/0.49*	0.00/0.00	Glycosyl transferase family 4 protein
TOM188	2	Unknown	Exon	Unknown	*0.31/0.30*	*0.31/0.30*	0.00/0.00	Unknown
SL20023	3	50.5	Intron	Unknown	*0.33/0.32*	*0.33/0.32*	0.00/0.00	Unknown
SL10136	4	107.9	Intron	SGN-U564154	0.34/0.17	*0.31/0.30*	0.06/0.03	Unknown
SL10238	5	3.6	Intron	SGN-U573605	*0.52/0.51*	*0.52/0.51*	0.00/0.00	MutT/nudix family protein
SL20210i	5	45.5	Intron	SGN-U566861	0.54/0.21	*0.78/0.75*	0.44/0.36	Acidic endochitinase (CHIB1)
LEOH63	5	47.0	Exon	SGN-U574808	*0.52/0.51*	*0.52/0.51*	0.00/0.00	kelch repeat-containing protein
SL10591	5	76.3	Intron	SGN-U575430	*0.91/0.90*	*0.97/0.97*	0.12/0.00	LMBR1 integral membrane family protein
Rx3-L1	5	80.0	Exon	SGN-U575600	0.66/0.48	*0.75/0.74*	0.18/0.15	CBL-interacting protein kinase 1 (CIPK1)
SP	6	66.0	Exon	SGN-U600210	0.18/0.15	*0.79/0.78*	0.69/0.53	Self-pruning protein
SL20017	7	3.1	Intron	SGN-U566360	0.36/0.00	*0.73/0.72*	*0.81/0.81*	Vacuolar ATP synthase subunit H family protein
LEOH8.4	9	10.0	Exon	SGN-U581554	0.38/0.03	*0.31/0.30*	0.18/0.15	Plasma membrane intrinsic protein 2A (PIP2A)
SL10024	9	50.9	Intron	SGN-U567408	0.29/0.27	0.00/0.00	*0.29/0.27*	Tetratricopeptide repeat (TPR)-containing protein
SL10120	11	34.8	Intron	SGN-U564640	0.45/0.11	0.57/0.39	*0.76/0.67*	NADH-ubiquinone dehydrogenase
SL10781	11	48.5	Intron	SGN-U575465	0.53/0.08	0.39/0.38	*0.63/0.61*	Armadillo/beta-catenin repeat family protein
SL20074i	Unknown	Unknown	Intron	SGN-U581949	*0.32/0.30*	0.00/0.00	*0.32/0.30*	Proteasome-related protein

[a]Map location based on Tomato Mapping Resource Database (TMRD:http://www.tomatomap.net).
[b]Sol Genomics Network (SGN:http://sgn.cornell.edu) unigene ID.
[c]Pairwise estimates of expected heterozygosity and F_{st} were obtained using the Lositan software (Beaumont and Nichols, 1996; Antao *et al*., 2008).
[d]Italics indicate detection of loci at the 95% confidence level.

We detected little evidence for subpopulation structure within the vintage varieties when using all marker data as well as two subsets of markers (the intron sequence markers and the SNP markers). An inconsistent pattern of clustering was observed when using the exon sequence markers, the SSR markers or the InDel markers alone. The instances of inconsistent clustering that we occasionally observed are most likely due to the population and marker sampling, which is referred to as ascertainment bias (Chikhi, 2008; Romero *et al*., 2009). Our failure to detect subpopulations within vintage varieties contrasts with past studies that detected differentiation within landraces and vintage varieties (Park *et al*., 2004; Mazzucato *et al*., 2008). These differences likely reflect the germplasm and markers sampled as subpopulations are determined with respect to genotypes within the collection chosen for analysis (Chikhi, 2008; Romero *et al*., 2009).

We also compared markers based from intron sequences with those from transcribed sequences for population structure analysis. Significant differences in the ability of markers to discriminate and assign individuals to subpopulations were not observed because of marker position when analyzing the predetermined groups. Differences in the number of subpopulations within processing germplasm were detected between intron and exon markers, but the additional groups identified with intron markers were not supported on the basis of statistical analysis of F_{st}. In the comparison of vintage varieties, markers in exons detected two subpopulations, whereas intron markers detected only one. Again, the additional subpopulation was not supported by F_{st}. Thus, position of the polymorphism within a gene does not appear to affect the performance of a marker for population level analysis.

There exists some precedence to suggest that the type of marker may affect performance in population level analysis because of allelic richness (Liu *et al*., 2005). Although we observed slightly more allelic diversity with SSRs relative to SNPs, the ability to distinguish subpopulations was comparable between marker types. SSRs appeared to detect more subpopulations than SNPs within the vintage varieties, but these added subpopulations were not supported by subsequent F_{st} analysis. Thus, as with position within a gene, marker type does not appear to have an affect on the conclusions of population level analysis for cultivated tomato.

The genetic structure we observed in this study suggests that selection for market specialization within cultivated tomato has left a genetic signature. Identifying which loci are responsible for this signature will identify the regions of the genome that have experienced selection during domestication and breeding. Previous efforts to identify regions of plant genomes under selection have focused on detecting bias in the ratio of synonymous to non-synonymous mutations (Ingvarsson, 2010), identifying reduced polymorphism as evidence for strong directional selection or 'selective sweeps' (Wang *et al*., 1999; Clark *et al*., 2004), and the identification of elevated polymorphism with respect to divergence as evidence for balancing selection (Zhao *et al*., 2008). Low levels of polymorphism within tomato and the inbred mating system make these approaches problematic. A method based on the statistic F_{st} has been used to quantify variation in SNP-allele frequencies between human populations as a means of detecting signatures of natural selection (Lewontin and Krakauer, 1973; Rana *et al*., 1999; Hollox *et al*., 2001; Akey

et al., 2002; Hamblin *et al*., 2002). This F_{st} approach has been applied to the inbred crop, common bean, where AFLP polymorphisms that deviated from the neutral expectation were detected. In bean, as many as 16% of the markers showed departure from expectation, and these tended to map to regions of the genome with known genes and quantitative trait loci related to domestication (Papa *et al*., 2007). Applying F_{st} analysis to comparisons of germplasm groups that have been shaped by human selection has the potential to identify both regions of the genome that are fixed because of selection and regions of the genome that remain polymorphic and explain existing phenotypic variation. The LOSITAN workbench provided an effective approach to establish objective criteria for F_{st} values that indicate selection between germplasm groups including processing, fresh-market and vintage varieties.

We identified 19 candidate loci under positive selection based on F_{st} values that fall outside of the 95% confidence interval established for the distribution (Antao *et al*., 2008). These loci may be directly under selection, but more likely mark regions of the genome that have been selected during breeding. Linkage disequilibrium decays over 3–16 cM within cultivated tomato (Robbins *et al*., 2010), and it is likely that many polymorphic markers will remain associated with the actual genes that are under selection when comparisons are made between groups of cultivated varieties. The loci we identified have a disproportional bias with 47% mapping to chromosomes 2 and 5. This observation suggests that there are 'hot spots' for directional selection through breeding in tomato. Genes on chromosome 2 that were segregating in our collection include the fruit shape gene, *ovate* (Liu *et al*., 2002) and loci affecting locule number (Barrero and Tanksley 2004). Chromosome 5 contains multiple bacterial disease-resistance genes including *Pto* (Martin *et al*., 1993) and *Rx3* (Yang *et al*., 2005b), and genes that affect plant morphology, which were previously shown to distinguish processing from fresh-market tomatoes (Jones *et al*., 2007). In addition, the *self-pruning* gene on chromosome 6 appears to be under selection when comparing contemporary with vintage varieties. The markers and genome locations we identified as outliers are consistent with known patterns of selection and linkage to traits that differentiate the market classes. These results suggest that the use of objective approaches to identify polymorphisms with extreme values of F_{st} will reveal portions of the genome that are under selection. Such objective assessment will provide a scalable means for comprehensive assessments of genetic variation within cultivated tomato as emerging sequence data and improved genotyping platforms lead to larger data sets.

Tomato is an excellent model to investigate the effect of breeding for shaping genetic variation in cultivated varieties. In this study, we defined genetic structure within a collection of cultivated tomato using the Bayesian model-based clustering method implemented in the STRUCTURE program. Major clusters were confirmed on the basis of F_{st} and Nei's standard genetic distance. The genetic differentiation we observed demonstrates that breeding for specialized markets has resulted in genetically distinct subpopulations within cultivated tomato. This differentiation has occurred recently in history. Population structure detected within the processing germplasm suggests that breeding efforts for varieties adapted to different target environments has also led to further genetic differentiation. We cannot yet distinguish the effects of different founder parents from ecological adaptation without designing experiments to test the effects of specific genome regions for adaptation to humid or arid conditions. Our results can be used to accelerate tomato improvement by addressing the patterns of genetic variation within cultivated tomato and helping breeders maximize variation within market classes in order to improve genetic gain.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We would like to thank Tea Meulia, Tarek Joobeur and Jody Whittier of The Ohio State University, OARDC, Molecular and Cellular Imaging Center for access to equipment for Luminex genotyping. We also thank Esther van der Knaap and Gustavo Rodriguez of The Ohio State University, OARDC for comments and valuable suggestions on the manuscript, and Paul Gepts and Myounghai Kwak of the University of California, Davis for helpful suggestions on the data analysis. Research was supported by the Ohio Plant Biotechnology Consortium Competitive Grant 2007-025 and USDA/AFRI 2009-85606-05673 to DM Francis, and USDA/NRI 2007-35300-18316 to MD Robbins.

References

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002). Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* **12**: 1805–1814.

Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G (2008). LOSITAN: A workbench to detect molecular adaptation based on a F_{st}-outlier method. *BMC Bioinformatics* **9**: 323.

Barrero LS, Tanksley SD (2004). Evaluating the genetic basis of multiple-locule fruit in a broad cross section of tomato cultivars. *Theor Appl Genet* **109**: 669–679.

Beaumont MA, Nichols RA (1996). Evaluating loci for use in the genetic analysis of population structure. *Proc R Soc London Ser B* **263**: 1619–1626.

Berry SZ, Wiese KL, Gould W (1992). Ohio 7983 Processing Tomato. *HortScience* **27**: 939.

Berry SZ, Wiese KL, Aldrich TS (1993). Ohio 8556 Processing Tomato. *HortScience* **28**: 751.

Breseghello F, Sorrells ME (2006). Association analysis as a strategy for improvement of quantitative traits in plants. *Crop Sci* **46**: 1323–1330.

Casa AM, Pressoir G, Brown PJ, Mitchell SE, Rooney WL, Tuinstra MR *et al.* (2008). Community resources and strategies for association mapping in sorghum. *Crop Sci* **48**: 30–40.

Chikhi L (2008). Genetic markers: How accurate can genetic data be? *Heredity* **101**: 471–472.

Clark RM, Linton E, Messing J, Doebley JF (2004). Pattern of diversity in the genomic region near the maize domestication gene tbl. *Proc Natl Acad Sci USA* **101**: 700–707.

Dieringer D, Schlotterer C (2003). MICROSATELLITE ANALYSER (MSA): a platform independent analysis tool for large microsatellite data sets. *Mol Ecol Notes* **3**: 167–169.

Doebley JF, Gaut BS, Smith BD (2006). The molecular genetics of crop domestication. *Cell* **127**: 1309–1321.

Evanno G, Regnaut S, Goudet J (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* **14**: 2611–2620.

Excoffier L, Heckel G (2006). Computer programs for population genetics data analysis: a survival guide. *Nat Rev Genet* **7**: 745–758.

Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S (2005). Genetic structure and diversity in Oryza sativa L. *Genetics* **169**: 1631–1638.

Hamblin MT, Thompson EE, Di Rienzo A (2002). Complex signatures of natural selection at the Duffy blood group locus. *Am J Hum Genet* **70**: 369–383.

Hedrick P (2005). A standardized genetic differentiation measure. *Evolution* **59**: 1633–1638.

Hollox EJ, Poulter M, Zvarik M, Ferak V, Krause A, Jenkins T *et al*. (2001). Lactase haplotype diversity in the Old World. *Am J Hum Genet* **68**: 160–172.

Hurlbert SH (1971). The nonconcept of species diversity: a critique and alternative parameters. *Ecology* **52**: 557–586.

Ingvarsson PK (2010). Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in *Populus tremula*. *Mol Biol Evol* **27**: 650–660;.

Jones CM, Rick CM, Adams D, Jernstedt J, Chetelat RT (2007). Genealogy and fine mapping of *obscuravenosa*, a gene affecting the distribution of chloroplasts in leaf veins, and evidence of selection during breeding of tomatoes (*Lycopersicon esculentun*; Solanaceae). *Am J Bot* **94**: 935–947.

Jost L (2008). G_{ST} and its relatives do not measure differentiation. *Mol Ecol* **17**: 4015–4026.

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

Lawson DM, Lunde CF, Mutschler MA (1997). Marker-assisted transfer of acylsugar-mediated pest resistance from the wild tomato, *Lycopersicon pennellii*, to the cultivated tomato, *Lycopersicon esculentum*. *Mol Breeding* **3**: 307–317.

Lee SH, Walker DR, Cregan PB, Boerma HR (2004). Comparison of four flow cytometric SNP detection assays and their use in plant improvement. *Theor Appl Genet* **110**: 167–174.

Lewontin RC, Krakauer J (1973). Distribution of gene frequency as a test of theory of selective neutrality of polymorphisms. *Genetics* **74**: 175–195.

Liu J, Van Eck J, Cong B, Tanksley SD (2002). A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. *Proc Natl Acad Sci USA* **99**: 13302–13306.

Liu N, Chen L, Wang S, Oh C, Zhao H (2005). Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. *BMC Genetics* **6**: S26.

Martin GB, Brommonschenkel SH, Chunwongse J, Frary A, Ganal MW, Spivey R *et al.* (1993). Map-based cloning of a protein kinase gene conferring disease resistance in tomato. *Science* **262**: 1432–1436.

Mazzucato A, Papa R, Bitocchi E, Mosconi P, Nanni L, Negri V *et al*. (2008). Genetic diversity, structure and marker-trait associations in a collection of Italian tomato (*Solanum lycopersicum* L landraces. *Theor Appl Genet* **116**: 657–669.

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

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

Papa R, Bellucci E, Rossi M, Leonardi S, Rau D, Gepts P *et al*. (2007). Tagging the signatures of domestication in common bean (*Phaseolus vulgaris*) by means of pooled DNA samples. *Ann Bot* **100**: 1039–1051.

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.

Pnueli L, CarmelGoren L, Hareven D, Gutfinger T, Alvarez J, Ganal M *et al.* (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.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Rana BK, HewettEmmett D, Jin L, Chang BHJ, Sambuughin N, Lin M *et al.* (1999). High polymorphism at the human melanocortin 1 receptor locus. *Genetics* **151**: 1547–1557.

Rasmussen WD (1968). Advances in American Agriculture: The Mechanical Tomato Harvester as a Case Study. *Technol Cult* **9**: 531–543.

Robbins MD, Sim S, Yang W, Van Deynze A, van Knaap E, Joobeur T *et al.* (2010). *Genome-wide analysis reveals different patterns of linkage disequilibrium among market classes of tomato P442*. Plant and Animal Genome XVIII: San Diego, CA.

Romero IG, Manica A, Goudet J, Handley LL, Balloux F (2009). How accurate is the current picture of human genetic variation? *Heredity* **102**: 120–126.

Rosenberg NA, Burke T, Elo K, Feldmann MW, Freidlin PJ, Groenen MAM *et al.* (2001). Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. *Genetics* **159**: 699–713.

Schmid K, Torjek O, Meyer R, Schmuths H, Hoffmann MH, Altmann T (2006). Evidence for a large-scale population structure of *Arabidopsis thaliana* from genome-wide single nucleotide polymorphism markers. *Theor Appl Genet* **112**: 1104–1114.

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

Tanksley SD, McCouch SR (1997). Seed banks and molecular maps: Unlocking genetic potential from the wild. *Science* **277**: 1063–1066.

van Berloo R, Zhu AG, Ursem R, Verbakel H, Gort G, van Eeuwijk FA (2008). Diversity and linkage disequilibrium analysis within a selected set of cultivated tomatoes. *Theor Appl Genet* **117**: 89–101.

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

Wang RL, Stec A, Hey J, Lukens L, Doebley J (1999). The limits of selection during maize domestication. *Nature* **398**: 236–239.

Weir BS, Cockerham CC (1984). Estimating F-statistics for the analysis of populatoin structure. *Evolution* **38**: 1358–1370.

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

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

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

Yang W, Sacks EJ, Ivey MLL, Miller SA, Francis DM (2005b). Resistance in *Solanum lycopersicum* intraspecific crosses to race T1 strains of *Xanthomonas campestris* pv. *vesicatoria* causing bacterial spot of tomato. *Phytopathology* **95**: 519–527.

You F, Huo N, Gu Y, Luo M, Ma Y, Hane D *et al.* (2008). BatchPrimer3: A high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics* **9**: 253–265.

Zhao Q, Thuillet AC, Uhlmann NK, Weber A, Rafalski JA, Allen SM *et al.* (2008). The role of regulatory genes during maize domestication: Evidence from nucleotide polymorphism and gene expression. *Genetics* **178**: 2133–2143.

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)