Heredity (2009) 103, 257–267
© 2009 Macmillan Publishers Limited All rights reserved 0018-067X/09 $32.00
www.nature.com/hdy

ORIGINAL ARTICLE

Multilocus sequence data reveal extensive departures from equilibrium in domesticated tomato (*Solanum lycopersicum* L.)

JA Labate, LD Robertson and AM Baldo
USDA-ARS Plant Genetic Resources Unit, Geneva, NY, USA

Limited genetic variation has been observed within tomato (*Solanum lycopersicum* L.), although no studies have extensively surveyed single nucleotide polymorphism (SNP) diversity among tomato landraces. We estimated intraspecific DNA sequence variation by analyzing 50 gene fragments (23.2 kb) per plant in a 31 plant diversity panel. The majority of loci (80%) were polymorphic with the minor allele at a frequency of 10% or less for most (141 of 155) SNPs. Mean diversity as estimated by θ and π was approximately 1.5 SNPs per kb. Significant linkage disequilibrium was observed between 19% of locus pairs, and within-locus population recombination estimates were negligible. We also sequenced 43 gene fragments from wild tomato *Solanum arcanum* Peralta as an outgroup. Various statistical tests rejected a neutral equilibrium model of molecular evolution at 10 of 50 loci. Rare, highly diverged alleles were observed, involving at least seven tomato lines and five loci. Some of these may represent introgressions that originated both from natural hybridization with *Solanum pimpinellifolium* L. and from crosses with *S. pimpinellifolium* and additional wild relatives for crop improvement. The former was reported from classical field studies carried out by CM Rick; the latter has been extensively documented in the crop, particularly for transfer of disease resistance alleles. Extensive introgression and frequent bottlenecks within *S. lycopersicum* will pose a challenge to reconstructing the genetic bases of domestication and selection using methods that rely on patterns of molecular polymorphism.

Heredity (2009) **103,** 257–267; doi:10.1038/hdy.2009.58; published online 13 May 2009

Keywords: population genetics; nuclear SNP; linkage disequilibrium; introgression; selection; haplotype network

Introduction

Many thousands of open-pollinated populations and inbred lines of domesticated tomato (*Solanum lycopersicum* L.) are conserved in genebanks such as the USDA, ARS Plant Genetic Resources Unit (PGRU), AVRDC-The World Vegetable Center in Taiwan and the Centre for Genetic Resources (CGN), The Netherlands (Robertson and Labate, 2007). These collections provide a publicly available resource for experimentation and breeding to a broad community of users around the world. Genetic diversity within cultivated tomato is understood to be low (Nesbitt and Tanksley, 2002) and has never been extensively surveyed in these large assemblages of accessions. Cost-effective molecular markers such as simple sequence repeats (SSRs) and single nucleotide polymorphisms (SNPs) are increasingly available for *S. lycopersicum* (He *et al*., 2003; Yang *et al*., 2004; Labate and Baldo, 2005; Ruiz *et al*., 2005; Van Deynze *et al*., 2007). Such markers can make the large sample sizes required for extensive surveys of genebank collections tractable.

Alongside a successful tradition of crop improvement via interspecific hybridization and backcrossing (Stevens and Rick, 1986; Causse *et al*., 2000; Bai and Lindhout, 2007) there is growing interest to mine new alleles from the intraspecific gene pool of *S. lycopersicum* (Van Deynze *et al*., 2007). Intraspecific breeding saves time because wild tomato species are generally poor in fruit quality and key agronomic traits, therefore, many backcross generations are required. Landraces collected from centers of diversity in Latin America are potentially a rich source of novel alleles. Although a Mexican versus South American origin of domestication for tomato remains unclear (Peralta and Spooner, 2007), it is assumed that the crop went through multiple bottlenecks during domestication and subsequent intercontinental dispersal via explorers of the New World. Two post-domestication founder events, first from South America to Europe, then from Europe to the United States, occurred relatively recently; prior bottlenecks have been hypothesized (Rick and Fobes, 1975). Recovering alleles that were lost as a consequence of small effective population size is highly feasible using genebank collections.

A pioneering survey of restriction fragment length polymorphism (RFLP) in *S. lycopersicum* showed up to threefold within-accession variation within landraces relative to cultivars (Miller and Tanksley, 1990). The same study looked at restriction fragments that were

Correspondence: Dr JA Labate, USDA-ARS Plant Genetic Resources Unit, 630 W North Street, Geneva, NY 14456, USA.
E-mail: joanne.labate@ars.usda.gov

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

Received 24 November 2008; revised 10 April 2009; accepted 14 April 2009; published online 13 May 2009

absent from cultivars but found in wild tomato species and landraces. Less than 5% of these were unique to landraces, whereas the two most diverse species, *Solanum peruvianum* L. and *Solanum pennellii* Correll, together accounted for more than 50% of the fragments.

Population genetics approaches, that is, strategically sampling and genotyping germplasm without making crosses, have provided a basis for discovering allelic diversity by linkage disequilibrium (LD) mapping and selective sweep mapping (also known as genome scanning) in crops (Ross-Ibarra *et al*., 2007; Burger *et al*., 2008). One benefit of the latter is that it does not require phenotypic data. This 'bottom-up' strategy relies on interpreting molecular polymorphism patterns to discover candidate alleles that show evidence of natural or artificial selection (Ross-Ibarra *et al*., 2007). Additional genetic analyses involving bioinformatic, molecular and phenotypic characterizations are then applied to dissect gene function. Selective sweep mapping has been applied in several crop species, for example, maize (*Zea mays* L.; Wright and Gaut, 2005; Yamasaki *et al*., 2007), Asian cultivated rice (*Oryza sativa* L.; Olsen *et al*., 2006; Caicedo *et al*., 2007; Zhu *et al*., 2007), sorghum (*Sorghum bicolor* (L.) Moench; Casa *et al*., 2005; Hamblin *et al*., 2006) and sunflower (*Helianthus annuus* L.; Burke *et al*., 2005) and will likely become more accessible with the generation of large amounts of crop sequence data (Paterson, 2006).

Design of LD mapping and selective sweep mapping studies requires basic knowledge of population parameters such as gene diversity, LD, recombination and introgression. The objective of this study was to estimate these parameters in a diversity panel largely composed of tomato landraces collected from the early 1930s through the 1970s and now conserved at PGRU.

Most of the extant genetic variation within *S. lycopersicum* has been suggested to reside in subsp. *cerasiforme* (cherry tomato; Williams and St Clair, 1993; Nesbitt and Tanksley, 2002). We chose not to focus on cherry tomato because it represents only a small fraction of *S. lycopersicum* germplasm conserved at PGRU. In addition, subsp. *cerasiforme* is not recognized in the recent taxonomic revision of genus *Lycopersicon* into *Solanum* L. section *Lycopersicon* (Peralta *et al*., 2008) and may represent an admixture of *S. lycopersicum* and *S. pimpinellifolium* L. (Nesbitt and Tanksley, 2002). Cherry tomato cultivars were shown to be genetically distinct from beef and round tomato cultivars using amplified fragment length polymorphism markers (van Berloo *et al*., 2008). Only one accession in our panel was classified as cherry.

Diversity of 50 gene fragments was estimated and tested for nonneutrality using wild tomato (*Solanum arcanum* Peralta) data when available. This is the first study to extensively survey SNP diversity in tomato landraces. Results of this initial multilocus sampling can help inform population based mapping approaches and intraspecific allele mining in the crop.

Materials and methods

Germplasm

Sampled material (Supplementary Table S1) consisted of a panel of 30 USDA, ARS PGRU accessions (LR1–LR30) largely composed of landraces. The majority of these were selected based on a random amplified polymorphic DNA (RAPD) marker study of AVRDC—The World Vegetable Center tomato collection (Villand *et al*., 1998). The Villand *et al*. (1998) study randomly chose 96 of 2000 morphologically characterized *S. lycopersicum* accessions using a stratified proportional sampling scheme based on geographic regions. For this study, we chose 28 of the 96 based on extensive RAPD diversity. Three additional *S. lycopersicum* lines were sampled. These were 'Ailsa Craig' PI 262995, 'Ccoilo-Chuma' PI 127825 and TA496. Sequences from three lines used for previous marker development were included when available (TA209, Rio Grande PI 303784 and Moneymaker PI 286255 assayed for 5, 29 and 19 loci, respectively; Labate and Baldo, 2005). Narrative descriptions from the Germplasm Resources Information Network illustrate the broad phenotypic diversity of our panel. Only accession PI 127825 was classified as cherry tomato based on fruit size and number of locules when grown in the field in Geneva, NY. Sequences from all lines were included in analyses except when otherwise noted. One self-compatible (SC), highly homozygous line of *S. arcanum* (formerly classified as *Lycopersicon peruvianum* L.) was sampled as an outgroup. *L. peruvianum* was initially chosen for this study because its breadth of diversity serves as a central point of reference for tomato species. For example, it was reported to contain most of the nucleotide variation found within or between five tomato species in a multilocus survey (Baudry *et al*., 2001). Each line was represented by DNA extracted from a single plant. All samples were assumed to accurately represent gene pools from original collection populations. Tomato seed is highly viable during long-term storage, therefore, most of the accessions have experienced only one or two regenerations *ex situ* using sample sizes of 24 plants.

Markers

The 50 amplicons (Table 1) represented three types of markers—expressed sequence tag (EST), Solanaceae Genome Network (SGN) Conserved Ortholog Set II (COSII) or unigene and arbitrary genes. COSII are single-copy conserved orthologous genes in Asterid species (Wu *et al*., 2006). General attributes of the markers were described in Labate *et al*. (2009) including TA496 reference sequences, SNP identities available in dbSNP at the National Center for Biotechnology Information (NCBI), and our annotations of coding regions (Labate *et al*., 2009 Supplementary material). Chromosome locations were available from the literature for each of the 10 arbitrary loci (see Table 2; Labate *et al*., 2009 for source NCBI accessions). COSII/unigene marker tomato map positions were available from SGN except for C2_At1g32130 and C2_At1g73180, which had been mapped in Arabidopsis only, and U318882 for which no map position was available. Nine EST markers had been previously virtually mapped using high confidence BLAST scores ($E = $zero) to mapped markers from SGN (Labate and Baldo, 2005). We attempted to virtually map the three unmapped COSII and the 17 unmapped EST-based markers by using the BLAST tool at SGN (NCBI BLAST version 2.2.15), performing separate BLASTn searches of databases 'All SGN marker

Table 1 Estimates of polymorphism within *S. lycopersicum* and divergence between *S. lycopersicum* and *S. peruvianum* of 50 loci; all indel regions were ignored

Chr	*Marker*	n	*nt*[a] *(nt coding)*	S[b]	θ[c]	π[d]	*Tajima's* D[e]	*Mean pairwise no. of SNPs between species*	*Div*[f]
10	437_2	33	655 (298)	8	3.22	0.79	−2.241***	3	0.52
4	3300_2	35	536 (426)	1	0.45	0.39	−0.211	NA[g]	NA
4	4301_3	32	445 (445)	1	0.56	0.14	−1.142	5	1.14
5	3155_3	33	862 (209)	5	1.43	1.03	−0.747	9	1.08
6	1260_2	33	377 (377)	1	0.65	0.16	−1.140	5	1.37
7	3332_3	32	151 (4)	1	1.69	2.40	0.643	2	1.59
9	2534_1b	33	656 (17)	9	3.47	1.67	−1.577*	10	1.51
	2325_3	32	419 (419)	5	2.96	1.44	−1.376	7	1.74
	1287_1	36	145 (111)	1	1.66	3.55	1.668*	3	1.76
1	1523_4	32	248 (88)	0	0.00	0.00	ND[h]	NA	NA
	1589_1	32	142 (37)	0	0.00	0.00	ND[h]	2	1.64
	1675_1	31	149 (149)	0	0.00	0.00	ND[h]	3	2.17
10	1724_1	35	232 (149)	3	3.14	2.42	−0.518	12	5.34
6	175_1	36	123 (123)	2	3.92	6.75	1.401	NA	NA
6	1863_3	32	727 (102)	5	1.71	1.59	−0.195	NA	NA
	1909_2	33	163 (61)	1	1.51	1.88	0.371	3	2.77
12	2189_1	34	249 (159)	3	2.95	5.11	1.670*	2	1.00
	220_1	32	150 (150)	2	3.31	0.83	−1.504	NA	NA
	2280_1	32	141 (141)	0	0.00	0.00	ND[h]	5	5.00
	241_2b	33	272 (0)	2	1.81	1.49	−0.354	NA	NA
	2486_1	34	198 (141)	5	6.37	2.57	−1.576*	2	1.10
	2582_1	32	127 (127)	0	0.00	0.00	ND[h]	1	0.79
	2719_1	33	153 (0)	1	1.62	3.31	1.580	3	2.33
1	2819_5	32	656 (125)	16	6.07	1.53	−2.509***	4	0.70
	2875_4b	34	582 (0)	7	3.25	2.47	−0.692	17	3.22
	296_1b	32	659 (0)	3	1.13	1.37	0.485	4	0.87
6	U146140	33	570 (59)	0	0.00	0.00	ND[h]	13	2.33
5	C2_At1g13380	33	684 (155)	1	0.36	0.09	−1.140	15	2.42
5	C2_At1g14000	34	716 (202)	3	1.03	0.90	−0.285	7	0.97
6	C2_At1g20050	34	899 (151)	5	1.44	1.26	−0.332	12	1.49
	C2_At1g32130	34	430 (225)	1	0.57	0.39	−0.483	4	0.95
6	C2_At1g44575	34	739 (182)	5	1.74	1.08	−1.002	48	7.31
1	C2_At1g50020	35	1218 (119)	0	0.00	0.00	ND[h]	6	1.22
	C2_At1g73180	34	266 (180)	9	8.43	8.74	0.111	NA	NA
1	C2_At2g15890	34	769 (163)	2	0.64	0.23	−1.277	11	1.52
11	C2_At2g22570	35	1082 (199)	4	0.90	0.46	−1.212	7	2.48
9	C2_At2g36930	33	274 (182)	0	0.00	0.00	ND[h]	3	1.14
5	U221402	33	601 (75)	2	0.83	0.20	−1.502*	9	1.51
	U318882	33	582 (109)	0	0.00	0.00	ND[h]	6	1.07
7	U146437	33	352 (235)	5	3.54	2.05	−1.124	6	1.84
1	*hp2* exon 2	32	347 (347)	1	0.72	0.35	−0.783	5	1.42
3	*Pds*	30	584 (36)	1	0.43	0.11	−1.147	6	1.05
3	*Psy1*	31	592 (57)	3	1.27	0.33	−1.731**	2	0.39
6	*Cyc-B* 5′ region	31	436 (5)	0	0.00	0.00	ND[h]	14	3.24
1	*hp2* 3′region	31	490 (215)	1	0.51	0.13	−1.145	15	3.06
2	*fw 2.2*	34	482 (23)	7	3.57	1.88	−1.364	16	3.34
10	TG11	32	559 (0)	4	1.78	1.82	0.063	12	2.06
10	*CRTISO*	34	449 (0)	1	0.60	0.64	0.085	9	2.18
5	*rin*	36	415 (0)	3	1.75	2.14	0.502	17	5.89
11	*PTOX*	33	456 (273)	5	2.71	1.49	−1.198	6	1.32
Total or mean		33	23 209 (7050)	145	1.71	1.34	−0.573	351	2.04

Abbreviations: NA, not available; ND, not defined; SNP, single nucleotide polymorphism.

[a] Length based on NCBI aligned PopSet, may include indels.

[b] Number of segregating sites.

[c] Diversity based on *S* per kb (Watterson, 1975).

[d] Mean pairwise diversity per kb (Nei and Li, 1979).

[e] $*P \leqslant 0.05$, $**P \leqslant 0.01$, $***P \leqslant 0.001$.

[f] Net nucleotide divergence (Nei, 1987) × 100.

[g] Not available, *S. arcanum* homologous sequence was not isolated.

[h] Undefined because there was no polymorphism.

sequences', 'Tomato BAC sequences' and 'Tomato BAC-end sequences'. Map location for a BAC or BAC-end sequence in the SGN database is sometimes available, for example, through associations with overgo probes (a sequenced marker with a known map location). NCBI databases nt, nr, EST, gss and tomato Entrez genome were also searched using BLAST algorithms for sequences that were significantly similar to the 20 unmapped markers.

Generation of sequences

Methods for DNA extraction, PCR, two-pass sequencing and phasing of haplotypes were reported in Labate *et al.*

Table 2 Multilocus HKA test performed iteratively

Iteration	*Sum of deviations*[a]	P[b]	*Maximum observed deviation*	*Locus*	S[c] *observed*	*S expected*	P[d]
1	91.8444	0.0000	16.1562	2819_5	16	5.00	0.0000
2	73.1191	0.0001	10.6429	437_2	8	2.39	0.0057
3	61.5654	0.0007	6.6827	2486_1	5	1.55	0.0921
4	54.3426	0.0086	5.3270	2534_1b	9	3.76	>0.10
5	49.6263	0.0238	3.9795	2189_1	3	0.96	>0.10
6	45.5949	0.0475	3.8216	*Psy1*	3	0.98	>0.10

The most deviant locus for each iteration is reported; the most deviant locus was removed and the test was rerun until the overall test statistic was nonsignificant.
[a] Deviation = (observed–expected)^2/variance.
[b] P = proportion of 10 000 simulated runs with $\geqslant$sum of deviations.
[c] Number of segregating sites within *S. lycopersicum*.
[d] P = proportion of 10 000 simulated runs with $\geqslant$maximum observed deviation.

Table 3 McDonald–Kreitman test of polymorphism and divergence of synonymous and nonsynonymous variation; pooled for 27 loci

	Syn	*Nonsyn*
Within *S. lycopersicum*	14	13
Between *S. lycopersicum* and *S. arcanum*	29	13
P value	0.151	

(2009). *S. arcanum* PCR reactions were performed in two replicates each of which was independently sequenced using forward and reverse primers. This gave high quality consensus sequences based on minimum phred score of 40 (Ewing *et al*., 1998). All *S. lycopersicum* and *S. arcanum* sequences have been deposited into NCBI as 50 popsets (accessions EU935868–EU937513).

Statistical analyses

Diversity parameters S (number of segregating sites), θ (Watterson, 1975), π (Nei and Li, 1979) and divergence between species including mean pairwise number of SNPs and Nei's (1987) estimator were generated using SITES software (Hey and Wakeley, 1997). Run parameters were set to exclude sites with more than two nucleotides and indel regions based on alignments both within *S. lycopersicum* and between *S. lycopersicum* and *S. arcanum*. Reported *S. lycopersicum* results (Table 1) varied slightly from those reported in Labate *et al*. (2009) (Table 3) because interspecific indels were not previously considered, and Rio Grande, Moneymaker and TA209 data were not previously included.

SNP allele frequencies were estimated within *S. lycopersicum* based on the number of chromosomes carrying the rarest base divided by the number of sampled chromosomes. Minimum spanning networks (Templeton *et al*., 1992) were generated for each marker using TCS software (Clement *et al*., 2000) to visualize relationships among alleles. Gaps were treated as a fifth character and gaps at contiguous sites were interpreted as a single mutation.

Neutrality tests

Initial input for HKA software (http://lifesci.rutgers.edu/~heylab/HeylabSoftware.htm#HKA) consisted of 43 loci for which polymorphism and divergence data were collected. Ten thousand coalescent simulations were conducted and the resultant sum of deviations was checked for statistical significance, that is, the sum of deviations was $\geqslant$0.95 of simulated values. The locus with the maximum observed deviation was removed from the analysis and the test was rerun. Using this method, six iterations were performed before the sum of deviations was nonsignificant ($P>0.05$). Observed Tajima's D values (Tajima, 1989) were tested for each locus against simulated distributions using HKA software. A maximum-likelihood-ratio test was applied using MLHKA software (Wright and Charlesworth, 2004) to test a selection model of the nine fruit quality markers. Five replicate runs with different random starting seeds and chain length of 500 000 were performed for neutral and selection models. Input starting values of θ and divergence time parameter (T) were obtained from HKA results. Fu and Li's test (Fu and Li, 1993) with an outgroup was performed on 34 markers with intraspecific polymorphism and an *S. arcanum* homolog using DNAsp (Rozas *et al*., 2003).

For a test of adaptive evolution of proteins (McDonald and Kreitman, 1991), 27 loci with polymorphism and divergence data in amino-acid coding regions were available. Numbers of synonymous and nonsynonymous polymorphic sites within *S. lycopersicum* and between *S. lycopersicum* and *S. arcanum* were tallied based on the SITES (Hey and Wakeley, 1997) output. Ambiguous cases were scored as single observations. For example, CCT (pro) in *S. lycopersicum* to CTC (leu) in *S. arcanum* was scored as one nonsynonymous replacement even though there were two mutated sites. DNAsp (Rozas *et al*., 2003) was used to perform a *G*-test of the four classes of observations using pooled numbers of sites.

Ratios of mean pairwise differences per nonsynonymous site to synonymous site (π_a/π_s) were estimated for five loci with non-zero values within *S. lycopersicum* (175_1, 220_1, 437_1, 2325_3, C2_At1g73180) by analyzing replacement or synonymous polymorphisms, respectively, using SITES (Hey and Wakeley, 1997). Similarly, mean pairwise divergence between *S. lycopersicum* and *S. arcanum* per nonsynonymous site to synonymous site (K_a/K_s) was estimated for eight loci with non-zero values (1260_2, 1287_1, 1724_1, 2280_1, 2325_3, 4301_3, U146437, *hp2* exon 2).

Linkage disequilibrium and recombination within loci

LD analysis was performed using DNAsp (Rozas *et al*., 2003) excluding singleton alleles. A total of 21 loci and

160 pairwise comparisons were analyzed. The resultant r^2 for each pair of sites was plotted against the average number of nucleotides that separated the pair. Microsoft Excel was used to fit a logarithmic curve to the plot. The population recombination parameter γ ($4Nc$) (Rozas *et al*., 2003) and the minimum set of recombination intervals R_{min} (Hudson and Kaplan, 1985) were estimated for the same 21 loci using SITES.

Linkage disequilibrium between loci

As our data showed extensive intralocus LD it was necessary to control for nonindependence of linked sites when estimating interlocus LD (Morrell *et al*., 2005). We did so by treating each intralocus haplotype as an allele. DNAsp (Rozas *et al*., 2003) was used to generate haplotype distributions and the number of haplotypes at each locus was ascertained. Polymorphism tables output from SITES (Hey and Wakeley, 1997) were used to create a file of diploid genotypes for 41 loci with more than one haplotype within *S. lycopersicum*. Data were formatted as homozygous or heterozygous diploid genotypes for input into GenePop (Raymond and Rousset, 1995). For each locus pair, GenePop reports an unbiased estimate of the *P* value of the observed genotypic contingency table.

Results

Sample size (number of alleles sequenced per marker, treating homozygotes as a single allele) within *S. lycopersicum* ranged from 30 to 36 with a mean of 33 (Table 1). At 24 loci, one to five individuals were heterozygous at one or more sites including indels. Heterozygotes were observed in 13 of 34 *S. lycopersicum* lines. Three accessions from Ecuador were unusually heterozygous; PI 129026, PI 390510 and PI 129142 were heterozygous at 13, 9 and 8 loci, respectively. One to three heterozygous loci per accession were observed in the 10 other cases. Across the 50 markers, sequence length ranged from 123 to 1218 nt (mean = 464 nt) for a total 23 209 nt. For 7 of the 50 markers, the *S. arcanum* homolog was not successfully isolated.

Significant matches were obtained for 10 of the 20 unmapped markers in BLAST searches against SGN and NCBI databases. Three were associated with chromosomal locations by using additional information available in SGN. These were 175_1 ($E = 1 \times 10^{-47}$), 1724_1 ($E = 1 \times 10^{-128}$), and 1863_3 ($E = 0.00$).

The majority of markers have been mapped (Table 1) and additional map data will soon be available with the results from the International Genome Sequencing Project (Mueller *et al*., 2005). All markers were assumed to represent single loci, 25% of markers designed using EST databases were rejected during marker development based on multiple PCR bands or high proportions of heterozygous sites (Labate *et al*., 2009). Marker types COSII/unigene and arbitrary loci have been characterized in the literature (see Labate *et al*., 2009 for references) and should correspond to single copy genes. TA496 alleles at arbitrary loci were 100% identical to primer source sequences in GenBank (Labate *et al*., 2009; Table 2) with the exception of *rin* (97%, $e = 0.0$) and *Psy1* (94%, $e = 0.0$; results not shown). A *Psy1* allele from Ohio 9242 (GenBank accession EF157835) was 100% identical to our TA496 allele in the fully aligned fragment.

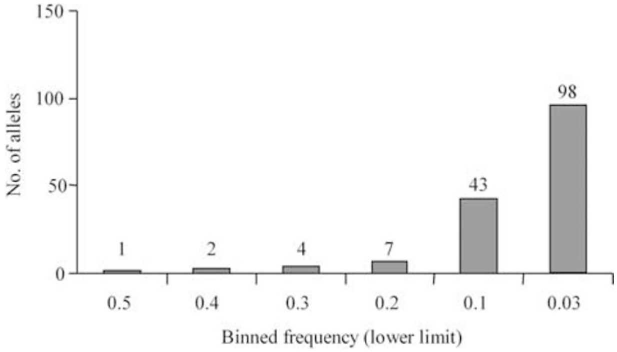

Figure 1 Frequency of minor allele at 50 *Solanum lycopersicum* loci. Mean frequency of the minor allele across loci equaled 0.10.

A possible observation of paralogs in our study was marker C2_At1g44575. This marker showed an unusually large mean pairwise number of SNPs between species (48, compared to 1–17 for the remaining 42 markers; Table 1). This marker yielded the largest estimate of interspecific divergence (Table 1) but did not reject neutrality tests. If C2_At1g44575 was omitted, mean divergence equaled 1.92 compared to 2.04 if retained. As it did not overly skew any results and there was no definitive reason to reject it as homologous, it was retained.

Ignoring interspecific data (that is, indels and triallelic sites) so that the maximum number of SNPs were counted, 155 SNPs were observed at 50 loci within *S. lycopersicum*. The majority (141) were at rare frequencies ($\leqslant 0.10$) and only 1 SNP was at 0.50 frequency (Figure 1). Non-zero estimates of θ ranged from 0.36 to 8.43 per kb (mean = 1.71), whereas π ranged from 0.09 to 8.74 per kb (mean = 1.34; Table 1). Divergence (Nei, 1987) between *S. lycopersicum* and *S. arcanum* was on average 2.04 per 100 nucleotides (Table 1). Marker C2_At1g44575 showed large mean pairwise numbers of SNPs between species (48 in 739 nt). This region was 76% intronic, with 182 nucleotides of exon in which the only two variable sites were two synonymous fixed differences between species.

Fu and Li's estimator (Fu and Li, 1993) tests whether the number of SNPs on external branches (η_e) of a genealogy fits neutral expectations based on the total number of SNPs (S) in a sample. Markers 437_2 ($S = 8$, $\eta_e = 7$), 2819_5 ($S = 14$, $\eta_e = 12$) and *Psy1* ($S = 3$, $\eta_e = 3$) gave significant results in this test (Supplementary Figure S1). Of 50 markers, 8 (437_2, 2534_1b, 1287_1, 2189_1, 2486_1, 2819_5, U221402 and *Psy1*) rejected the null hypothesis for the Tajima's test (Table 1; Supplementary Figure S1). Six of the values were negative (excess number of rare SNPs) and two were positive (excess moderate frequency SNPs). Five markers (2819_5, 437_2, 2486_1, 2534_1b and 2189_1) rejected the null hypothesis in both Tajima's and multilocus HKA tests (Tables 1 and 2; Supplementary Figure S1). One additional marker, *Psy1*, contributed to significant multilocus HKA tests (Table 2).

Labate and Baldo (2005) observed five markers (220_1, 437_2, 2325_3, 2486_1, 2534_1R) with 1.6–13-fold higher θ values relative to 18 other markers based on resequencing two or three tomato lines. These markers were hypothesized to carry rare, introgressed alleles from wild tomato species. In this study, these 5 markers remained among the most diverse within the set of 50; all fell within the 12 highest θ values (Supplementary Figure S1). Three of the five (437_2, 2486_1, 2534_1b) rejected neutrality in both Tajima's and HKA tests (Tables 1 and 2; Supplementary Figure S1) whereas one (437_2) rejected Fu and Li's test (Supplementary Figure S1).

A visual inspection of the distribution of θ values in this study showed three distinct ranges: 0–1.81 (36 markers), 2.71–3.92 (11 markers) and 6.07–8.43 (3 markers; Table 1; Supplementary Figure S1). All 7 markers with significant HKA tests, 5 of 8 markers with significant Tajima's tests and 2 of 4 markers with significant Fu and Li's tests were among the 14 markers within the $\theta = 2.71$–8.43. All markers were examined for evidence of highly divergent alleles by visual inspections of haplotype networks (Supplementary Figure S2). Highly divergent alleles were observed at markers 437_2, 2534_1b, 2486_1, 2819_5 and C2_At1g73180 (Supplementary Figure S2). Of these, only C2_At1g 73180 did not reject neutrality based on the HKA and Tajima's tests. Values of θ and π were similar for C2_At1g73180, and the *S. arcanum* homolog necessary for an HKA test was not isolated. Fu and Li's test was performed for this locus without an outgroup and was not significant ($S = 9$, number of singletons $= 0$, $D^* = 1.36$).

A McDonald–Kreitman test (McDonald and Kreitman, 1991) indicated a potential excess of amino acid changes within *S. lycopersicum*, although a *G*-test was not significant (Table 3). A single locus method to detect positive selection is to examine the d_N/d_S ratio, where d_N is the number of nonsynonymous substitutions per nonsynonymous site and d_S is the ratio of synonymous substitutions per synonymous site (Nielsen, 2001). Ratios of intraspecific π_a/π_s and interspecific K_a/K_s fell close to the neutral expectation (ratio $= 1.0$) for 13 loci tested. The largest ratio $K_a/K_s = 4.25$ was based on only one synonymous and two nonsynonymous SNPs at marker 1724_1.

Five outlier markers (hypothetical introgressions 2325_3, 437_2, 2534_1b, 2819_5, 2486_1) (Supplementary Figure S1) were removed from analysis before applying a maximum-likelihood-ratio test that utilizes interspecific data (Wright and Charlesworth, 2004). First, a neutral model was fit for 38 loci. Then, a model was fit that allowed for selection on the nine fruit quality genes. Mean values of ML, *T* were -139.68, 14.39 and -135.52, 12.62 for neutral and selection models, respectively. This gave a likelihood-ratio statistic of 8.33, which was not significant (d.f. $= 9$, $P > 0.5$).

LD (r^2) among linked sites was extensive and decayed very little over distance, with the log trend showing a plateau at $r^2 > 0.6$ (Figure 2). Of 160 pairwise comparisons at 21 loci, 119 (0.74) were in significant LD using Fisher's test with Bonferroni correction. Population recombination parameter γ (Hey and Wakeley, 1997) either equaled zero (15 loci) or could not be estimated (6 loci). R_{min} estimates (Hudson and Kaplan, 1985) equaled zero or were not available with the exception of C2_At2g22570 for which $R_{\mathrm{min}} = 1$. This recombination was evident in the minimum spanning network at sites 325 and 935 (Supplementary Figure S2).

Given 41 loci with more than one haplotype within *S. lycopersicum*, there were 820 possible pairwise comparisons to estimate two-locus LD. Due to some missing loci for some individuals, 797 pairs were tested. Significant LD ($P \leqslant 0.5$) was observed in 152 (0.191) of the pairs according to both Fisher's exact tests and *G*-tests. For 75 of the 152 pairs (0.49) in LD, the chromosomal map position was known for both markers. Of these, eight pairs (0.107) were located on the same chromosome.

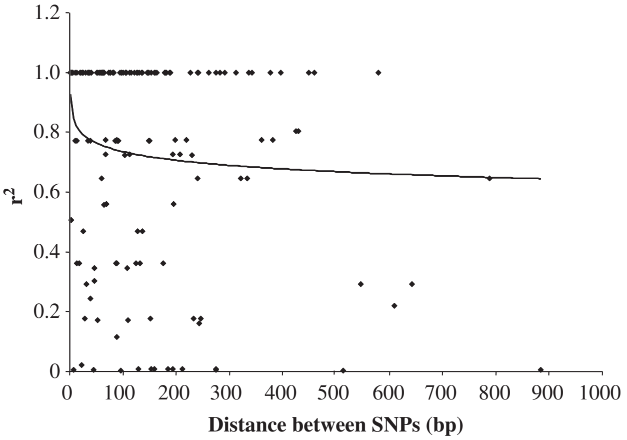

Figure 2 Decay of intralocus linkage disequilibrium over distance for 21 loci, 160 pairwise comparisons. The line represents a logarithmic trend line fit to the data.

Discussion

Our results showed that although low, number of SNPs per bp within domesticated tomato was higher than previously estimated (Nesbitt and Tanksley, 2002; Van Deynze *et al*., 2007). This can be attributed to our targeted sampling of landraces. Variant haplotypes tended to be rare and sometimes highly divergent. The source of observed polymorphism likely stemmed from wild and human mediated introgression as well as native diversity of the species. Characterizing the patterns of genetic diversity will establish an important basis for future work involving associations with phenotype.

Diversity estimates

The accessions sampled for our study were mainly composed of landraces from centers of diversity (Chile, Ecuador, Peru) and contiguous countries with the expectation that they would give accurate estimates of the latent diversity within *S. lycopersicum*. Most accessions were homozygous for most loci, as expected for a self-fertilizing species. Three accessions from Ecuador were outliers in that 16–25% of markers were heterozygous. Two (PI 129026, PI 129142) were originally collected in 1938. The third (PI 390510) was originally collected in 1974 and was described as wild. In addition, PI 129026 and PI 390510 carried a highly divergent allele at marker C2_At1g73180. All three accessions showed at least a low frequency of slightly or moderately exserted stigmas when grown in the field (JA Labate *et al.*, manuscript in preparation). Diverse populations of *S. lycopersicum* from Ecuador and Peru were hypothesized to have hybridized with sympatric *S. pimpinellifolium* based on allozyme evidence and lack of reproductive barriers (Rick and Fobes, 1975).

Raw numbers of SNPs discovered in our study (1/158 bp) were large relative to 1/1647 bp observed in a more elite sample representing *S. lycopersicum* fresh

Table 4 Nucleotide diversity estimates for various crops and wild tomato species

Species	*Mating system*[a]	θ	π	*Reference*
Wild barley	S	0.0081	0.0075	Morrell *et al.* (2005)
Asian rice	S	0.0021	0.0023	Caicedo *et al.* (2007)
Sorghum	S	0.0022	0.0022	Hamblin *et al.* (2006)
Soybean	S	0.0010	0.0011[a]	Zhu *et al.* (2003)
S. lycopersicum	S	0.0017	0.0013	This study
S. pimpinellifolium	S	0.0016[b]		Roselius *et al.* (2005)
S. chmielewskii	S	0.0003[b]		Roselius *et al.* (2005)
S. habrochaites	SI	0.0040		Städler *et al.* (2005)
S. chilense	SI	0.0126	0.0110	Arunyawat *et al.* (2007)
S. peruvianum	SI	0.0181	0.0129	Arunyawat *et al.* (2007)

Abbreviations: S, self-fertilizing; SI, self-incompatible.

[a] $\pi_{\text{noncoding}}$.

[b] θ_{silent}.

market, processing and heirloom types (952.5 kb, 10 lines; Van Deynze *et al.*, 2007). Compared to studies in other crops that used similar sample sizes, numbers of SNPs were small in comparison to 1/24 bp in outcrossing but vegetatively propagated potato (*Solanum tuberosum* L.; >25 kb, 47 accessions; Simko *et al.*, 2006), similar to 1/148 bp in self-fertilized sorghum (*S. bicolor* (L.) Moench; 29.2 kb, 27 accessions; Hamblin *et al.*, 2004), and large relative to 1/327 bp observed in a sample representing North American soybean (*Glycine max* L. Merr.; 76.3 kb, 25 lines; Zhu *et al.*, 2003).

Comparison of diversity among various crops and wild tomato relatives are summarized in Table 4. *S. lycopersicum* mean diversity fell within the lower end of estimated means relative to other selfing species such as wild barley (*Hordeum vulgare* ssp. *spontaneum*), cultivated Asian rice (*O. sativa* L.), sorghum and soybean. *S. lycopersicum* mean θ was similar to estimated diversity within a single population of *S. pimpinellifolium* but greater than that observed within a single population of *Solanum chmielewskii* (CM Rick, Kesicki, Fobes and M Holle) DM Spooner, GJ Anderson and RK Jansen; all three species are primarily selfing. Self-incompatible (SI) wild tomato species yielded higher estimates of diversity, for example, *Solanum habrochaites* S Knaap and DM Spooner (one accession), *Solanum chilense* (Dunal) Reiche (four populations) and *S. peruvianum* (four populations). Differences observed among wild tomato species in mean diversity have arisen through a variety of factors including mating system, demography and population structure (Baudry *et al.*, 2001; Roselius *et al.*, 2005; Städler *et al.*, 2005; Arunyawat *et al.*, 2007).

Minor alleles were skewed toward rare frequencies in *S. lycopersicum* to a much greater degree than that reported in soybean (Zhu *et al.*, 2003) and *Arabidopsis thaliana* (Nordborg *et al.*, 2005). This is inconsistent with bottlenecks due to domestication and migration, as rare alleles go extinct during those processes. Excess rare alleles and the associated negative Tajima's *D* values can be a consequence of population structure, recent population expansion, or purifying selection (Tajima, 1993). We hypothesize that some of the rare SNPs observed in this study originated from introgression.

Neutrality tests

A large fraction (20%) of loci rejected neutrality tests in our study. Predicted protein products for these are reported in Supplementary Figure S2. In potato, 9% of 66 surveyed loci rejected neutrality using Tajima's test; all showed positive *D* values indicative of excess moderate frequency alleles (Simko *et al.*, 2006). In sorghum, 17% of 160 estimated Tajima's *D* values were significant (Hamblin *et al.*, 2006). The 10 most negative of these showed a pattern of many singletons restricted to a few lineages, possibly due to population structure or introgression from a divergent population. Interspecific introgression is likely one cause of rejection of neutral equilibrium in this study. Introgression has been apparent in many morphological and/or molecular studies of domesticated tomato (Rick, 1950, 1958, 1971; Rick and Fobes, 1974, 1975; Rick *et al.*, 1974; Williams and St Clair, 1993; Villand *et al.*, 1998).

A pattern consistent with slightly deleterious amino acid changes escaping purifying selection as a consequence of small effective population size, as would be predicted for *S. lycopersicum*, was observed. Sampled numbers of polymorphisms in coding regions were low so statistical power was weak. Nevertheless, nonsynonymous changes were approximately twofold greater than predicted based on comparison with *S. arcanum*. Of 17 predicted amino-acid polymorphisms within *S. lycopersicum*, 7 were nonconservative, 4 of which were singletons (Supplementary Figure S2). Excess replacement polymorphism illustrating a load of slightly deleterious mutations was reported in *A. thaliana* (Nordborg *et al.*, 2005) and sorghum (Hamblin *et al.*, 2004). Additional examination of amino-acid evolution based on ratio tests π_a/π_s and K_a/K_s (Nielsen, 2001) revealed values to be scattered close to the neutral expectation of one (mean values equaled 0.993 and 1.202, respectively). Ratios less than one are often interpreted as purifying selection (Simko *et al.*, 2006 and examples therein). Mean *S. lycopersicum* ratios supported a small effective population size, that is, selection coefficients that were too small to eliminate deleterious amino-acid mutations in this species.

An alternative hypothesis is that the observed amino-acid polymorphisms have undergone positive selection. *Psy1* displayed evidence of excess high frequency derived SNPs in Fu and Li's test as can result from a selective sweep (Fu and Li, 1993). However, a selection model of the nine fruit quality markers, all likely candidates, was rejected. Markers 1287_1 and 2189_1 were significant for Tajima's test due to excess moderate frequency alleles as can result from balancing selection. This was based on small numbers of SNPs (one and three, respectively) so may be a spurious result of sampling. Positive Tajima's *D* values can also result from population subdivision or a bottleneck (Tajima, 1993). However, marker 2189_1 ranked as the fifth most deviant in iterative HKA tests, so deserves further study.

Linkage disequilibrium

LD between polymorphic sites can result from various forces such as low recombination rate, population structure, selection, inbreeding and drift (Hartl and Clark, 1989). LD estimates are valuable for understanding whether association mapping will be efficient within a species. For this purpose LD patterns have been increasingly reported in crop species, for example, barley (*H. vulgare* L.; Malysheva-Otto *et al*., 2006), potato (Simko *et al*., 2006), soybean (Hyten *et al*., 2007), rice (Mather *et al*., 2007), maize (Yu *et al*., 2008), sorghum (Casa *et al*., 2008), sugarcane (*Saccharum* spp.; Raboin *et al*., 2008), sunflower (*H. annuus*; Fusari *et al*., 2008), tomato (Mazzucato *et al*., 2008; van Berloo *et al*., 2008) and wheat (*Triticum* spp.; Somers *et al*., 2007). In *S. peruvianum* and *S. chilense* intragenic LD decayed within 150 and 750 bp, respectively, suggestive of high recombination rates and large effective population sizes in these species (Arunyawat *et al*., 2007). Mean R_{min} across eight loci (40–46 sequenced alleles per locus per species) was 12 in *S. peruvianum* and 9 in *S. chilense* (Arunyawat *et al*., 2007). Sampled germplasm in this study showed extensive LD and very little evidence of recombination within loci. Only marker C2_At2g22570 revealed a recombination event using the four-gamete test. Two additional markers, 1724_1 and *rin*, showed all four gametes when either *S. arcanum* or indels were included (Supplementary Figure S2). The majority of pairs of loci were in linkage equilibrium. Similarly, interlocus LD was rare in a study of European greenhouse tomato cultivars (van Berloo *et al*., 2008). In a study of Italian tomato landraces 0.04 of SSR loci pairs were in LD, mostly due to interchromosomal LD (Mazzucato *et al*., 2008).

Some LD is required for association mapping to identify candidate functional polymorphisms (Rafalski, 2002). Commercial tomato cultivars showed strong intrachromosomal LD extending up to 15–20 cM (van Berloo *et al*., 2008). *S. lycopersicum* commercial cultivars and landraces are potentially useful for broad genome scans although the low frequency of most polymorphisms presents a disadvantage. This can be overcome by selecting markers that are at least at 0.20 frequency in the experimental population (Ganal *et al*., 2007). Wild tomato species populations will be attractive for fine-scale association mapping (Arunyawat *et al*., 2007).

Diversity patterns

The most striking result was the observation of a large proportion of highly divergent alleles within *S. lycopersicum*. For example, the current results supported a previous observation of diverged alleles at three EST markers (Labate and Baldo, 2005). The most probable candidate was 2534_1 that we inferred to map within 1 cM of *Tm-2^a* (Labate and Baldo, 2005). *Tm-2^a* was recorded to have been introgressed into line TA496 for improved disease resistance (Tanksley *et al*., 1998; Yates *et al*., 2004). Therefore, we hypothesized that locus 2534_1 contained an introgression as a consequence of linkage drag. Larger sample sizes in this study have provided more precise estimates of θ, but it remains impossible to unequivocally recognize an introgressed allele. For example, hypothetical introgression 2325_3 (Labate and Baldo, 2005) did not reject neutrality tests in this study but showed two clades of $n=30$ and $n=2$. The $n=2$ clade members were TA496 and PI 258478, both of which showed evidence of highly divergent alleles at other loci.

Five markers (see Results) and seven lines (see below) showed highly divergent alleles based on θ and minimum spanning networks. Four of the five were the four most deviant markers in multilocus HKA tests. Three of the four indicated TA496 as carrying a highly diverged allele in haplotype networks (Supplementary Figure S2). Marker C2_At1g73180 is of interest for its very high (and similar) values of θ and π. The haplotype network showed two clades of $n=5$ and $n=29$ separated by eight or nine SNPs, including a predicted nonconservative thr/lys polymorphism, and two indels (Supplementary Figure S2). As the wild species homolog was not isolated at this locus it was not possible to qualitatively compare intraspecific to interspecific mutations. However, accession PI 196297 in the $n=4$ taxon (Supplementary Figure S2) was reported 50 years ago to show evidence of introgression from *S. pimpinellifolium* based on carrying the *cm* allele (Rick, 1958). This allele was believed to have been transferred to *S. lycopersicum* from *S. pimpinellifolium* in coastal Ecuador where the two species were sympatric (Rick, 1958). The broad geographic and temporal range of accessions in which the two minor alleles of C2_At1g73180 were found imply that this represents an introgression that occurred in nature previous to 1938. Based on our results upwards of 10% of *S. lycopersicum* loci may carry highly divergent alleles. This may be an overestimate because the EST markers were originally designed based on predicted polymorphism with reference to line TA496 (Labate and Baldo, 2005).

In addition to TA496, two landraces collected in 1938 (PI 129026, PI 129128) and four collected during the 1950s to 1970s (PI 196297, PI 258474, PI 258478, PI 390510) showed evidence of highly divergent alleles. These could represent segregating ancestral alleles within the recently domesticated crop, interspecific gene flow in native habitats or wild species genes that were transferred into *S. lycopersicum* during crop improvement. Multiple domestication events can also generate highly divergent alleles that mimic introgression (Burger *et al*., 2008). This is a plausible, alternative explanation for observed diversity patterns in this crop (Peralta and Spooner, 2007).

Introgression

Rick (1950, 1958) discussed the likelihood of hybridization between *S. lycopersicum* (cultivated and wild 'escape' populations) and wild tomato species. Outcrossing rates in experimental plots of *S. lycopersicum* male-sterile (*ms*) plants were high in Peru, from 10% to greater than 40% of flowers set fruit (Rick, 1950). Study of local insect pollinators suggested that *S. lycopersicum* and *S. pimpinellifolium* shared certain vectors that were largely excluded from sympatric *Solanum corneliomuelleri* JF Macbr., *S. peruvianum* and *S. habrochaites* (Rick, 1950). Furthermore, wild forms of *S. lycopersicum* in Peru and Ecuador usually had exserted stigmas. Natural hybrids between *S. lycopersicum* and *S. pimpinellifolium* were observed in these field studies. Additional morphological observations in the field led Rick (1958) to report a lack of evidence for gene flow between *S. lycopersicum* and sympatric populations of *S. chilense*, *S. habrochaites* and *S. peruvianum* in Chile, Ecuador and Peru. Unlike these green-fruited species, the closely related red-fruited, sympatric *S. pimpinellifolium* yielded highly fertile hybrids and showed extensive evidence of cross-hybridization. The one additional near-relative and only other color-fruited tomato species, *Lycopersicon cheesma-*

niae (now classified into species *Solanum cheesmaniae* (L Riley) Fosberg and *Solanum galapagense* S. Darwin and Peralta), evolved in geographical isolation so could not form interspecific hybrids in nature (Rick and Fobes, 1975). Studies of geographical distributions of various marker alleles further supported the hypothesis of frequent hybridization between *S. lycopersicum* and *S. pimpinellifolium* (Rick, 1971; Rick *et al*., 1974; Rick and Fobes, 1975).

The extent of linkage drag of cryptic wild species alleles in improved lines is assumed to be low but has not often been quantified. As many as 30 backcross generations did not break the linkage between allozyme *Aps*^{1} and nematode resistance gene *Mi* that were introgressed together from *S. peruvianum* (Rick and Fobes, 1974). A fourfold increase in percent polymorphic loci in modern relative to vintage cultivars was attributed to interspecific introgression during crop improvement (Williams and St Clair, 1993). However, relatively high levels of recombination between *S. peruvianum* and *S. lycopersicum* during several backcrosses closely monitored with RFLPs suggested that linkage drag was not impervious (Fulton *et al*., 1997).

Implications

SNP markers within *S. lycopersicum* can be utilized for evolutionary studies, genetic mapping and allele mining. Domestication and improvement of crops is associated with a decrease in genetic variation along a continuum of wild progenitors, landraces, improved cultivars and elite germplasm (McCouch, 2004). This depletion of variation can hinder utilization of molecular markers for improvement. Although minor alleles are rare within the *S. lycopersicum* gene pool they are likely to be fixed within particular lines, facilitating their isolation.

Historical or recent gene flow among wild tomato species (Nesbitt and Tanksley, 2002; Städler *et al*., 2005; Arunyawat *et al*., 2007) and between wild species and the cultivar imposes challenges for evolutionary studies. Notably, a Mexican versus a South American domestication origin continues to be debated (Labate *et al*., 2007; Peralta and Spooner, 2007). The argument of greater *S. lycopersicum* diversity in South America must be weighed in light of both interspecific hybridization and widespread adoption of improved lines in that region (Rick, 1958). Historically it has been difficult to establish authenticity of native, local varieties (Rick, 1971). By the 1950s there was evidence of several US varieties being grown alongside native types in Peru (Rick, 1958). Some of these were likely to have been improved by interspecific introgression for disease resistance (Stevens and Rick, 1986). Gene flow from introduced cultivars into native populations can be promoted because the latter are adapted to outcrossing (Williams and St Clair, 1993).

Saturating *S. lycopersicum* intraspecific genetic maps poses a challenge in the face of interspecific introgression because polymorphic markers may tend to cluster (Villand *et al*., 1998; Yang *et al*., 2004). In population genomics studies of domestication and selection (Nordborg *et al*., 2005; Wright *et al*., 2005; Ross-Ibarra *et al*., 2007), complex demographic processes (for example, bottlenecks, global exchange of germplasm and introgression from multiple wild species), must be incorporated. This study indicated patterns of tomato genetic diversity that will inform strategies associating genotype with phenotype, and foster more efficient utilization of the intraspecific gene pool.

Acknowledgements

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

References

Arunyawat U, Stephan W, Stadler T (2007). Using multilocus sequence data to assess population structure, natural selection, and linkage disequilibrium in wild tomatoes. *Mol Biol Evol* **24**: 2310.

Bai Y, Lindhout P (2007). Domestication and breeding of tomatoes: what have we gained and what can we gain in the future? *Ann Bot* **100**: 1085–1094.

Baudry E, Kerdelhué C, Innan H, Stephan W (2001). Species and recombination effects on DNA variability in the tomato genus. *Genetics* **158**: 1725–1735.

Burger JC, Chapman MA, Burke JM (2008). Molecular insights into the evolution of crop plants. *Am J Bot* **95**: 113.

Burke JM, Knapp SJ, Rieseberg LH (2005). Genetic consequences of selection during the evolution of cultivated sunflower. *Genetics* **171**: 1933–1940.

Caicedo AL, Williamson SH, Hernandez RD, Boyko A, Fledel-Alon A, York TL *et al*. (2007). Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet* **3**: 1745–1756.

Casa AM, Mitchell SE, Hamblin MT, Sun H, Bowers JE, Paterson AH *et al*. (2005). Diversity and selection in sorghum: simultaneous analyses using simple sequence repeats. *Theor Appl Genet* **111**: 23–30.

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.

Causse M, Caranta C, Saliba-Colombani V, Moretti A, Damidaux R, Rousselle P (2000). Enhancement of tomato genetic resources via molecular markers. *Cah Agric* **9**: 197–210.

Clement M, Posada D, Crandall KA (2000). TCS: a computer program to estimate gene genealogies. *Mol Ecol* **9**: 1657–1660.

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

Fu YX, Li WH (1993). Statistical tests of neutrality of mutations. *Genetics* **133**: 693–709.

Fulton TM, Nelson JC, Tanksley SD (1997). Introgression and DNA marker analysis of *Lycopersicon peruvianum*, a wild relative of the cultivated tomato, into *Lycopersicon esculentum*, followed through three successive backcross generations. *Theor Appl Genet* **95**: 895–902.

Fusari CM, Lia VV, Hopp HE, Heinz RA, Paniego NB (2008). Identification of single nucleotide polymorphisms and analysis of linkage disequilibrium in sunflower elite inbred lines using the candidate gene approach. *BMC Plant Biol* **8**: 7.

Ganal MW, Durstewitz G, Kulosa D, Luerssen H, Polley A, Wolf M (2007). *Development of EST-Derived SNP Markers for Plant Breeding. W172*. Plant and Animal Genome XV: San Diego, CA.

Hamblin MT, Casa AM, Sun H, Murray SC, Paterson AH, Aquadro CF *et al*. (2006). Challenges of detecting directional selection after a bottleneck: lessons from *Sorghum bicolor*. *Genetics* **173**: 953–964.

Hamblin MT, Mitchell SE, White GM, Gallego J, Kukatla R, Wing RA *et al*. (2004). Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of *Sorghum bicolor*. *Genetics* **167**: 471–483.

Hartl D, Clark A (1989). *Principles of Population Genetics*, 2nd edn. Sinauer Associates Inc.: Sunderland, MA.

He C, Poysa V, Yu K (2003). Development and characterization of simple sequence repeat (SSR) markers and their use in determining relationships among *Lycopersicon esculentum* cultivars. *Theor Appl Genet* **106**: 363–373.

Hey J, Wakeley J (1997). A coalescent estimator of the population recombination rate. *Genetics* **145**: 833–846.

Hudson RR, Kaplan NL (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**: 147–164.

Hyten DL, Choi IY, Song Q, Shoemaker RC, Nelson RL, Costa JM *et al*. (2007). Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics* **175**: 1937–1944.

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

Labate JA, Grandillo S, Fulton T, Muños S, Caicedo AL, Peralta I *et al*. (2007). Tomato. In: Kole C (ed). *Genome Mapping and Molecular Breeding in Plants: Vegetables*. Springer: New York, Vol. 5, pp 1–125.

Labate JA, Robertson LD, Wu F, Tanksley SD, Baldo AM (2009). EST, COSII, and arbitrary gene markers give similar estimates of nucleotide diversity in cultivated tomato (*Solanum lycopersicum* L.). *Theor Appl Genet* **118**: 1005–1014.

Malysheva-Otto LV, Ganal MW, Röder MS (2006). Analysis of molecular diversity, population structure and linkage disequilibrium in a worldwide survey of cultivated barley germplasm (*Hordeum vulgare* L.). *BMC Genet* **7**: 6.

Mather KA, Caicedo AL, Polato NR, Olsen KM, McCouch S, Purugganan MD (2007). The extent of linkage disequilibrium in rice (*Oryza sativa* L.). *Genetics* **177**: 2223–2232.

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.

McCouch S (2004). Diversifying selection in plant breeding. *PLoS Biol* **2**: e347.

McDonald JH, Kreitman M (1991). Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**: 652–654.

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

Morrell PL, Williams-Coplin TD, Lattu AL, Bowers JE, Chandler JM, Paterson AH (2005). Crop-to-weed introgression has impacted allelic composition of johnsongrass populations with and without recent exposure to cultivated sorghum. *Mol Ecol* **14**: 2143–2154.

Mueller LA, Tanksley SD, Giovannoni JJ, Van Eck J, Stack S, Choi D *et al*. (2005). The Tomato Sequencing Project, the first cornerstone of the International Solanaceae Project (SOL). *Comp Funct Genomics* **6**: 153–158.

Nei M (1987). *Molecular Evolutionary Genetics*. Columbia University Press: New York.

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

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

Nielsen R (2001). Statistical tests of selective neutrality in the age of genomics. *Heredity* **86**: 641–647.

Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng HG *et al*. (2005). The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol* **3**: e196.

Olsen KM, Caicedo AL, Polato N, McClung A, McCouch S, Purugganan MD (2006). Selection under domestication: evidence for a sweep in the rice *Waxy* genomic region. *Genetics* **173**: 975–983.

Paterson AH (2006). Leafing through the genomes of our major crop plants: strategies for capturing unique information. *Nat Rev Genet* **7**: 174–184.

Peralta IE, Knapp S, Spooner D (2008). The taxonomy of tomatoes: a revision of wild tomatoes (*Solanum* section *Lycopersicon*) and their outgroup relatives (*Solanum* sections *Juglandifolium* and *Lycopersicoides*). *Syst Bot Monogr* **84**: 1–186.

Peralta IE, Spooner DM (2007). History, origin and early cultivation of tomato. In: Razdan MK, Mattoo AK (eds). *Genetic Improvement of Solanaceous Crops*. Science Publishers: Enfield, NH, Vol. 2: Tomato, pp 1–24.

Raboin LM, Pauquet J, Butterfield M, D'Hont A, Glaszmann JC (2008). Analysis of genome-wide linkage disequilibrium in the highly polyploid sugarcane. *Theor Appl Genet* **116**: 701–714.

Rafalski A (2002). Applications of single nucleotide polymorphisms in crop genetics. *Curr Opin Plant Biol* **5**: 94–100.

Raymond M, Rousset F (1995). GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *J Hered* **86**: 248–249.

Rick C (1971). The tomato *Ge* locus: linkage relations and geographic distribution of alleles. *Genetics* **67**: 75–85.

Rick CM (1950). Pollination relations of *Lycopersicon esculentum* in native and foreign regions. *Evolution* **4**: 110–122.

Rick CM (1958). The role of natural hybridization in the derivation of cultivated tomatoes of western South America. *Econ Bot* **12**: 346–367.

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

Rick CM, Fobes JF (1975). Allozyme variation in the cultivated tomato and closely related species. *Bull Torrey Bot Club* **102**: 376–386.

Rick CM, Zobel RW, Fobes JF (1974). Four peroxidase loci in red-fruited tomato species: genetics and geographic distribution. *Proc Natl Acad Sci USA* **71**: 835–839.

Robertson LD, Labate JA (2007). Genetic resources of tomato (*Lycopersicon esculentum* Mill.) and wild relatives. In: Razdan MK, Mattoo AK (eds). *Genetic Improvement of Solanaceous Crops*. Science Publishers: Enfield, NH. Vol. 2: Tomato, pp 25–75.

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

Ross-Ibarra J, Morrell PL, Gaut BS (2007). Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proc Natl Acad Sci USA* **104** (suppl 1): 8641–8648.

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

Ruiz JJ, García-Martínez S, Picó B, Gao M, Quiros C (2005). Genetic variability and relationship of closely related Spanish traditional cultivars of tomato as detected by SRAP and SSR markers. *J Am Soc Hortic Sci* **130**: 88–94.

Simko I, Haynes KG, Jones RW (2006). Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers. *Genetics* **173**: 2237–2245.

Somers DJ, Banks T, DePauw R, Fox S, Clarke J, Pozniak C *et al*. (2007). Genome-wide linkage disequilibrium analysis in bread wheat and durum wheat. *Genome* **50**: 557–567.

Städler T, Roselius K, Stephan W (2005). Genealogical footprints of speciation processes in wild tomatoes: demography and evidence for historical gene flow. *Evolution* **59**: 1268–1279.

Stevens M, Rick C (1986). Genetics and breeding. In: Atherton J, Rudich J (eds). *The Tomato Crop*. Chapman and Hall: New York, NY, pp 35–109.

Tajima F (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–596.

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

Tanksley SD, Bernachi D, Beck-Bunn T, Emmatty D, Eshed Y, Inai S *et al.* (1998). Yield and quality evaluations on a pair of processing tomato lines nearly isogenic for the *Tm-2^a* gene for resistance to the tobacco mosaic virus. *Euphytica* **99**: 77–83.

Templeton AR, Crandall KA, Sing CF (1992). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* **132**: 619–633.

van Berloo R, Zhu A, 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.

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

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

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.

Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD *et al.* (2005). The effects of artificial selection on the maize genome. *Science* **308**: 1310–1314.

Wright SI, Charlesworth B (2004). The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. *Genetics* **168**: 1071–1076.

Wright SI, Gaut BS (2005). Molecular population genetics and the search for adaptive evolution in plants. *Mol Biol Evol* **22**: 506–519.

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

Yamasaki M, Wright SI, McMullen MD (2007). Genomic screening for artificial selection during domestication and improvement in maize. *Ann Bot* **100**: 967–973.

Yang WC, 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.

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

Yu J, Holland JB, McMullen MD, Buckler ES (2008). Genetic design and statistical power of nested association mapping in maize. *Genetics* **178**: 539–551.

Zhu Q, Zheng X, Luo J, Gaut BS, Ge S (2007). Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: severe bottleneck during domestication of rice. *Mol Biol Evol* **24**: 875–888.

Zhu YL, Song QJ, Hyten DL, Tassell CPV, Matukumalli LK, Grimm DR *et al.* (2003). Single-nucleotide polymorphisms in soybean. *Genetics* **163**: 1123–1134.

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