Theor Appl Genet (2013) 126:567–581
DOI 10.1007/s00122-012-2002-8

ORIGINAL PAPER

Phenotypic diversity and association mapping for fruit quality traits in cultivated tomato and related species

Jiaxin Xu · Nicolas Ranc · Stéphane Muños · Sophie Rolland ·
Jean-Paul Bouchet · Nelly Desplat · Marie-Christine Le Paslier ·
Yan Liang · Dominique Brunel · Mathilde Causse

Received: 12 April 2012 / Accepted: 6 October 2012 / Published online: 4 November 2012

Abstract Association mapping has been proposed as an efficient approach to assist in the identification of the molecular basis of agronomical traits in plants. For this purpose, we analyzed the phenotypic and genetic diversity of a large collection of tomato accessions including 44 heirloom and vintage cultivars (*Solanum lycopersicum*), 127 *S. lycopersicum* var. *cerasiforme* (cherry tomato) and 17 *Solanum pimpinellifolium* accessions. The accessions were genotyped using a SNPlex$^{\text{TM}}$ assay of 192 SNPs, among which 121 were informative for subsequent analysis. Linkage disequilibrium (LD) of pairwise loci and population structure were analyzed, and the association analysis between SNP genotypes and ten fruit quality traits was performed using a mixed linear model. High level of LD was found in the collection at the whole genome level. It was lower when considering only the 127 *S. lycopersicum* var. *cerasiforme* accessions. Genetic structure analysis showed that the population was structured into two main groups, corresponding to cultivated and wild types and many intermediates. The number of associations detected per trait varied, according to the way the structure was taken into account, with 0–41 associations detected per trait in the whole collection and a maximum of four associations in the *S. lycopersicum* var. *cerasiforme* accessions. A total of 40 associations (30 %) were co-localized with previously identified quantitative trait loci. This study thus showed the potential and limits of using association mapping in tomato populations.

Introduction

Genetic dissection of quantitative traits in plants is a major goal for plant breeding. Quantitative trait loci (QTLs) were first mapped in bi-parental populations using linkage mapping approach (Paterson et al. 1991; Saliba-Colombani

Communicated by M. Havey.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-012-2002-8) contains supplementary material, which is available to authorized users.

J. Xu · Y. Liang
College of Horticulture, Northwest A&F University,
Yang Ling 712100, Shaanxi,
People's Republic of China

J. Xu · N. Ranc · S. Muños · S. Rolland · J.-P. Bouchet ·
N. Desplat · M. Causse (✉)
Unité de Génétique et Amélioration
des Fruits et Légumes, INRA, UR1052,
84143 Avignon, France
e-mail: mathilde.causse@avignon.inra.fr

Present Address:
N. Ranc
Syngenta Seeds SAS, 12, chemin de l'Hobit,
B.P. 27, 31790 Saint-Sauveur, France

Present Address:
S. Muños
Laboratoire des Interactions Plantes Micro-organismes (LIPM),
CNRS-INRA, UMR 2594/441,
31326 Castanet-Tolosan Cedex, France

Present Address:
S. Rolland
Amélioration des Plantes et Biotechnologies Végétales,
INRA-AgroCampus Ouest-Université Rennes1, UMR118,
BP 35327, 35653 Le Rheu Cedex, France

M.-C. Le Paslier · D. Brunel
Unité Etude du Polymorphisme des Génomes Végétaux,
CEA-Institut de Génomique-CNG, INRA,
UR1279, 91057 Evry, France

et al. 2001; Wang et al. 2006; Szalma et al. 2007; Orsini et al. 2012). This approach has several advantages: (1) no population structure in the mapping population, (2) segregating alleles are at balanced frequency, and (3) it allows the detection of rare alleles and epistasis. However, the linkage mapping approach has several limitations: (1) restricted allelic variation in bi-parental mapping population and (2) low precision due to limited recombination within the population (Hall et al. 2010).

Nowadays, association mapping or linkage disequilibrium (LD) mapping is proposed as an alternative approach. On one hand, association mapping has several advantages over QTL mapping: (1) it is based on occurring variation in collections of natural genetic resources, (2) it is more precise because of the recombination events resulting from many lineages, and (3) if LD is sufficiently low, it allows the discovery of the gene controlling the trait of interest. On the other hand, it has several limitations: (1) unbalanced allele frequency in the population, (2) it is not efficient for the detection of rare alleles, and (3) it requires large population sizes and an efficient control of the population structure. Many plant association studies have been published to date for several traits, such as flowering time and pathogen resistance in Arabidopsis (Aranzana et al. 2005), yield and its components in rice (Agrama et al. 2007), leaf architecture in maize (Tian et al. 2011), and iron deficiency chlorosis in soybean (Mamidi et al. 2011). Association mapping approach requires the knowledge of the genetic structure and the extent of LD of the samples studied. Population structure can lead to false positives and must be taken into account (Yu et al. 2006). Several statistical methods have been developed to deal with structured samples in order to control false associations (Pritchard et al. 2000; Price et al. 2006; Yu et al. 2006). The mixed linear model has been shown to be efficient in maize (Yu et al. 2006) and Arabidopsis (Atwell et al. 2010). LD is the non-random association between alleles at several loci. The extent of LD over the genome will influence association mapping strategy. If LD is high, the resolution will be low, but fewer markers will be required and a whole genome scan approach may be performed. If LD is low the resolution will be higher as the number of marker required and a candidate gene analysis may be conducted (Rafalski 2002). LD is expected to be higher in average in autogamous species than in allogamous species, because the higher homozygosity at a given locus leads to lower rate of efficient recombination than in allogamous species (Flint-Garcia et al. 2003). Nevertheless, a large range of variation in the rate of recombination per Mb exists along the chromosomes (Sim et al. 2012). Thus, the resolution of association mapping in autogamous species is expected to be lower than in allogamous species. Therefore, association mapping was first used in allogamous species or species with wide range of genetic diversity.

Tomato (*Solanum lycopersicum*, formerly *Lycopersicon esculentum*) is a highly autogamous species. It was domesticated from its wild relative *Solanum pimpinellifolium* with the first domesticated form presumably represented by *S. lycopersicum* var. *cerasiforme* (i.e., the cherry tomato, hereafter *S. l. cerasiforme*). Cultivated tomato shows a low genetic diversity, but higher phenotypic diversity compared to *S. pimpinellifolium* (Miller and Tanksley 1990) due to intensive human selection. The higher molecular diversity and the genetic admixture of *S. l. cerasiforme* genome (being a mosaic of cultivated and wild tomato genomes, Ranc et al. 2012) may be useful to overcome the high LD in this autogamous species. Several association studies have been carried out to dissect morpho-physical and fruit traits in tomato. Mazzucato et al. (2008) studied associations between 29 simple sequence repeat (SSR) markers and 15 morpho-physiological traits in 50 tomato landraces. Nesbitt and Tanksley (2002) investigated associations between fruit size and genomic sequence of the fw2.2 region which controls fruit weight (Frary et al. 2000) in a collection of 39 cherry tomato accessions. Unfortunately, they failed to find any association, but they demonstrated that the genome of cherry tomato accessions is a mosaic composed of polymorphisms of *S. pimpinellifolium* and *S. lycopersicum*. Munos et al. (2011) used association analysis to identify two single-nucleotide polymorphisms (SNP) located in a small region of chromosome 2 involved in the control of locule number of tomato fruit.

In a previous pilot study focused on one chromosome and 90 tomato accessions (Ranc et al. 2012), we showed that association mapping was possible in tomato. In the present work, we developed a SNPlex^{TM} assay (De La Vega et al. 2005; Tobler et al. 2005) of 192 SNPs selected from re-sequencing experiments or from databases (Van Deynze et al. 2007). A large germplasm collection including cultivated, cherry type and wild accessions were characterized for both genetic diversity using the SNPlex^{TM} assay, 20 SSR markers (Ranc et al. 2008) and ten phenotypic traits. We first describe the phenotypic diversity of the accessions, then the genetic structure of the collection based on SSR and SNP markers and finally the association mapping results. Associations are compared to previously mapped QTL. Our work is the first example of an association study carried out using a broad sample of cultivated, cherry type and wild tomato accessions.

Materials and methods

Plant materials

Tomato accessions were selected from a germplasm collection maintained and characterized at INRA Avignon (France) to maximize both genetic and phenotypic diversity. The sample consisted of 127 cherry type tomato accessions, 44 large fruited accessions (*S. lycopersicum* var. *esculentum*, hereafter named *S. lycopersicum*) and 17 *S. pimpinellifolium* accessions (Supplemental Table S1). Accessions were obtained from the Tomato Genetics Resource Center (Davis, USA), the Centre for Genetic Resources (Wageningen, The Netherlands), the North Central Regional Plant Introduction Station (Ames, Iowa, USA) and from the N.I. Vavilov Research Institute of Plant Industry (St Petersburg, Russia). Genomic DNA of the 188 accessions was isolated from 50 to 100 mg young leaves. After freeze-drying, the leaf material was ground and DNA extraction was performed using the DNeasy 96 plants Mini Kit (Qiagen, Valencia, USA) according to the manufacturer's protocol.

SNPlex$^{\text{TM}}$ assay design

Allele-specific probes and optimized multiplexed assays using SNPs of interest were designed by an automated multi-step pipeline (Applied Biosystems, Foster city, USA). The ABI probe design prevents self-complementarity and dimerization, and annealing efficiencies are optimized for ligation. Furthermore, the optimal combination of SNPs to produce the highest yield per multiplex reaction is determined (De La Vega et al. 2005; Tobler et al. 2005). Four SNPlex each carrying 48 SNPs were developed. Among the 192 SNPs used in the SNPlex$^{\text{TM}}$ assay, 131 SNPs were chosen from Sanger resequencing experiments of candidate genes mostly located on chromosome 2 (69 SNPs), 4 (22 SNPs) and 9 (30 SNPs) (Ranc et al. 2012). The remaining 61 SNPs were chosen from published information for covering the whole genome (Van Deynze et al. 2007). Supplemental Table S2 presents the characteristics of all the SNPs analyzed.

Genotyping was carried out on fragmented gDNA at a final concentration ranging from 45 to 225 ng and a final volume of 12.5 μl arrayed into 384-well plates according to the manufacturer's instructions. In each plate, six negative controls (water) and 18 positive controls (mixed DNA of known genotypes) were included. The allelic discrimination was detected using GeneMapper® Analysis Software v3.7 (Applied Biosystems Foster City, CA, USA) based on the SNPlex_Rules_3730 method. SNP markers with minimum allele frequencies (MAF) lower than 10 % and more than 10 % missing data were discarded from statistical analysis, which were thus performed on 121 markers.

Linkage disequilibrium analysis

The LD extent was calculated on two sets of genotypes, the whole collection and the subset of accessions of *S. l. cerasiforme*. GGT 2.0 (van Berloo et al. 2008) software was used to calculate the squared correlation coefficients r^2 (Zhao et al. 2005) between 121 markers throughout the genome. The decay of LD over genetic distance was investigated by plotting pair-wise r^2 values against the distances at the whole genome level and on chromosome 2 (covered by the largest number of SNPs) for all accessions, then for *S. l. cerasiforme* accessions separately.

Inference of population structure

Structure v2.1 program (Pritchard et al. 2000) was used to estimate the number of sub-populations in the complete set of accessions using admixture model for the ancestry of individuals and correlated allele frequencies. Population structure was modeled with a burning of 2.5×10^5 cycles followed by 10^6 Markov Chain Monte Carlo (MCMC) repeats. Evanno transformation method was then used to infer the most likely number of populations (K) (Evanno et al. 2005). Structure analysis was obtained on two sets of markers: 121 informative SNPs selected from SNPlex$^{\text{TM}}$ assay and compared to the structure obtained with 20 SSR markers (Ranc et al. 2008). Distruc1.1 program (Rosenberg 2004) was used to display the graphics of population structure. The kinship matrix was generated by SPAGeDi (Hardy and Vekemans 2002) software based on the two set of markers: 121 informative SNPs and 20 SSR markers. Diagonal of the matrix was set to two and negative values were set to zero, according to Yu et al. (2006).

Phenotyping

All tomato accessions (4 plants per accession) were grown in a plastic greenhouse in Avignon (south of France) during summers 2007 and 2008. Three harvests of ten ripe fruits per accession were used as repeats in the phenotypic analysis. Ten fruits were evaluated for fruit weight (FW), firmness (FIR), soluble solids content (SSC), sugar content (SUG), locule number (LCN), pH, titratable acidity (TA), and color components: lightness (L), color on red to green (a*) and on yellow to blue (b*) scales with a Konica Minolta CR-300 chromameter. All measurements were performed as described in Saliba-Colombani et al. (2001).

Statistical analysis

The heritabilities were estimated on the collection of homozygous lines. Heritabilities h^2 were calculated as

$h^2 = \sigma^2_g/(\sigma^2_g + \sigma^2_{gy} + \sigma^2_e)$ with σ^2_g, σ^2_{gy} and σ^2_e the genetic, genetic by environment interaction and residual variance, respectively. σ^2_g, σ^2_{gy} and σ^2_e were estimated by (MSc-MScy)/ry, (MScy-MSe)/r and MSe, respectively. MSc represents the accessions mean square, MScy represents the mean square of genotype by year interaction, MSe the residual mean square. *r* and *y* represents the number of replicates and the number of years. Associations were tested using the adjusted means of accessions calculated by general linear model. The Pearson correlation coefficients were calculated for all pairs of variables. Analyses were carried out with the R program (R Development Core Team 2005).

Association mapping

The association analyses were performed on the whole collection of 188 accessions and on 127 *S. l. cerasiforme* accessions by performing mixed linear model (MLM, K + Q model) as described by Yu et al. (2006) and implemented in Tassel 2.1 software (Bradbury et al. 2007). Two MLM models were used, with structure and kinship based on 20 SSR marker (Model A), and on 121 SNP markers (Model B). *p* values were corrected following the standard Bonferroni procedure. Significant associations were detected with corrected *p* value lower than 0.005 (an arbitrary choice of threshold due to the poor correction of the structure) for Model A, and 0.05 for Model B. After obtaining the significant markers, a general linear model with all fixed-effect terms was used to estimate R^2, the amount of phenotypic variation explained by each marker. The Pearson correlation coefficients were calculated between Q1 value (the probability that an individual belongs to the first subpopulation defined by 20 SSR markers and 121 SNPs) and each phenotypic trait.

Surrounding sequences of the 121 SNPs used in association analysis were blasted against Tomato whole genome sequence Chromosomes (SL2.40) database and International Tomato Annotation Group (ITAG) release 2.3 database (http://www.solgenomics.net) in order to map them on the physical map and get the annotation of surrounding genes. Physical map locations of the 121 SNPs were converted to genetic map on the tomato EXPEN 2000 map (http://www.solgenomics.net) based on the closest mapped markers and on the assumption of a local linear relationship between physical and genetic distance. Markers which have been previously mapped in QTL studies (Goldman et al. 1995; Grandillo and Tanksley 1996; Saliba-Colombani et al. 2001) were also aligned using BLAST on the genome sequence or directly mapped on the EXPEN 2000 map to get their corresponding genetic position. SNPs significantly associated with a trait were considered as co-localizing with previous QTLs, if they were located in a window of 20 cM surrounding the QTLs. Genetic map with associations and previously identified QTLs were plotted using MapChart v2.1 software (Voorrips 2002).

Results

Phenotypic variation and correlations of fruit quality traits

All the traits showed a large range of phenotypic variation in the whole collection, and within the three groups composed of 44 *S. lycopersicum*, 127 *S. l. cerasiforme*, and 17 *S. pimpinellifolium* (Table 1). Heritabilities were high (ranging from 0.57 to 0.85) in the whole collection with lower values for L, a*, LCN and pH in the *S. pimpinellifolium* group which exhibited a lower genetic variability. Phenotypic correlations among the fruit quality traits in the whole collection and within groups are detailed in Supplemental Table S3. Color components L, a* and b* were highly correlated with each other, and moderately correlated with the other traits in the whole collection, and in the *S. l. cerasiforme* and *S. pimpinellifolium* groups. However, a* was not significantly correlated with b* in the *S. lycopersicum* group. Few significant correlations were observed between FIR and the nine other traits in the whole collection and in the *S. l. cerasiforme* group. Correlations were higher between FIR and the nine other traits in *S. lycopersicum* and *S. pimpinellifolium* groups than in the *S. l. cerasiforme* group. FW was strongly positively correlated with LCN, and negatively correlated with SSC and TA in the whole collection and in each group. It was also negatively correlated with SUG in the whole collection and in the *S. l. cerasiforme* and *S. lycopersicum* groups. SSC and SUG were highly significantly correlated with each other in all the groups. pH was negatively correlated with TA in the whole collection and in the three groups. It was significantly correlated with SUG in *S. l. cerasiforme* and *S. pimpinellifolium* group.

Molecular polymorphism

The 188 accessions were genotyped with 192 SNPs, combined in four 48-plex panels, among which 139 SNPs (73 %) were successfully scored. Three SNPs with more than 10 % missing data and 15 SNPs with MAF lower than 10 % were removed from further analysis. Finally, 121 informative SNPs were used for polymorphism and association analysis. The results were very similar among the four SNPlex^{TM} panels (Supplemental Table S4). The distributions of the MAF were different among the three

Table 1 Phenotypic variation of fruit quality traits in the whole collection (all), *S. lycopersicum* (lyco), *S. l. cerasiforme* (cera) and *S. pimpinellifolium* (pimp) group, respectively

Trait	Phenotypic variation				Heritability			
	All ($N^a = 188$) mean ± SD	lyco ($N^a = 44$) mean ± SD	cera ($N^a = 127$) mean ± SD	pimp ($N^a = 17$) mean ± SD	All ($N^a = 188$)	lyco ($N^a = 44$)	cera ($N^a = 127$)	pimp ($N^a = 17$)
a*	15.47 ± 4.98***	18.51 ± 4.64***	14.32 ± 4.90***	16.16 ± 2.63***	0.81	0.77	0.82	0.54
b*	13.72 ± 5.47***	14.81 ± 4.47***	13.75 ± 5.88***	10.69 ± 3.33***	0.76	0.69	0.79	0.66
L	42.67 ± 3.4***	43.35 ± 2.28***	42.94 ± 3.58***	38.94 ± 2.06***	0.61	0.47	0.61	0.36
FIR	53.22 ± 7.43***	52.90 ± 8.58***	52.12 ± 6.30***	62.21 ± 6.10***	0.72	0.73	0.66	0.75
FW	33.03 ± 46.56***	96.29 ± 61.80***	15.19 ± 8.40***	2.65 ± 1.03***	0.83	0.75	0.86	0.87
LCN	3.22 ± 2.24***	5.19 ± 3.87***	2.69 ± 0.71***	2.09 ± 0.11***	0.85	0.81	0.78	0.34
pH	4.08 ± 0.13***	4.11 ± 0.12***	4.08 ± 0.12***	4.06 ± 0.15***	0.57	0.58	0.68	0.26
SSC	7.38 ± 1.34***	6.37 ± 0.84***	7.42 ± 1.09***	9.69 ± 1.18***	0.73	0.62	0.62	0.58
SUG	4.02 ± 1.54***	2.99 ± 0.92**	4.04 ± 1.22***	6.49 ± 2.12*	0.63	0.55	0.55	0.56
TA	11.10 ± 2.50***	9.44 ± 2.05***	11.30 ± 2.23***	13.84 ± 2.59***	0.75	0.73	0.70	0.73

Traits are described by mean, standard deviation (SD), significant level in group and heritability

a*, b*, L, color; *FIR* firmness, *FW* fruit weight, *LCN* locule number, *pH* pH, *SSC* soluble solids content, *SUG* sugar content, *TA* titratable acidity

* $p < 0.05$

** $p < 0.01$

*** $p < 0.001$

[a] Number of accessions

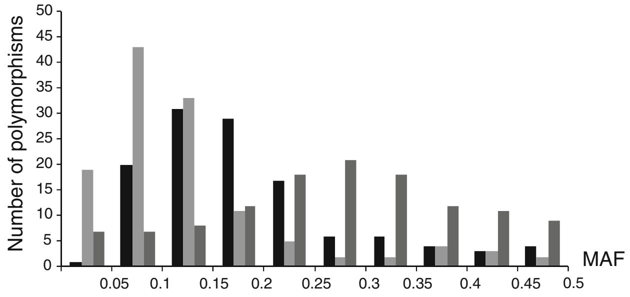

Fig. 1 Distribution of minimum allele frequencies (MAF) of the 121 SNPs in the three tomato groups [*S. l. cerasiforme* ($N = 127$) represented in *black*, *S. lycopersicum* ($N = 44$) in *light gray* and *S. pimpinellifolium* ($N = 17$) in *dark gray*]. Polymorphisms with MAF lower than 0.10 in the whole collection were previously discarded

groups of accessions (Fig. 1). The average MAF was 0.26, 0.18 and 0.12 for *S. pimpinellifolium*, *S. l. cerasiforme* and *S. lycopersicum*, respectively.

LD decay was analyzed separately for all markers and for the 50 markers on chromosome 2 (carrying the largest number of markers) for the 188 accessions and for 127 *S. l. cerasiforme* accessions. Pairwise r^2 was plotted according to genetic distance between loci and non-linear regression fitted the decay of LD over genetic distance. LD on the whole genome for all accessions extended on average over 18 cM for $r^2 = 0.3$ (Fig. 2a), with the same pattern, if only 50 markers of chromosome 2 were analyzed (Fig. 2b). The *S. l. cerasiforme* accessions had lower LD (reaching $r^2 = 0.3$ for 10 cM) as illustrated for chromosome 2 (Fig. 2c). The LD plot for *S. l. cerasiforme* accessions on the whole genome level showed the same pattern, with a few loci in strong LD with several markers responsible for high LD values.

Population structure

The structure of a collection of 360 accessions, including the 188 accessions studied here, was assessed with 20 SSR markers spread over the genome by Ranc et al. (2008). Four groups were detected, but the subset of 188 accessions studied here revealed two main groups, a cultivated and a wild group and many intermediate types (Fig. 3a). The genetic structure of the 188 accessions assessed with the 121 SNP markers revealed the same pattern with two groups (Fig. 3b). A similar genetic structure was obtained when combining SNPs and SSR markers (data not shown). The Q1 values of each accession (corresponding to the probability to belong to the "cultivated" group) estimated by the two types of markers were correlated ($r^2 = 0.63$), but several accessions were clustered in different groups. We thus compared both structure patterns in the mixed linear models to detect associations. L, FIR, FW, SSC and SUG strongly participated to the structure as shown by the highly significant correlations between these traits and the probability to belong to the cultivated group, with *r* values ranging from 0.30 to 0.59 (Table 2).

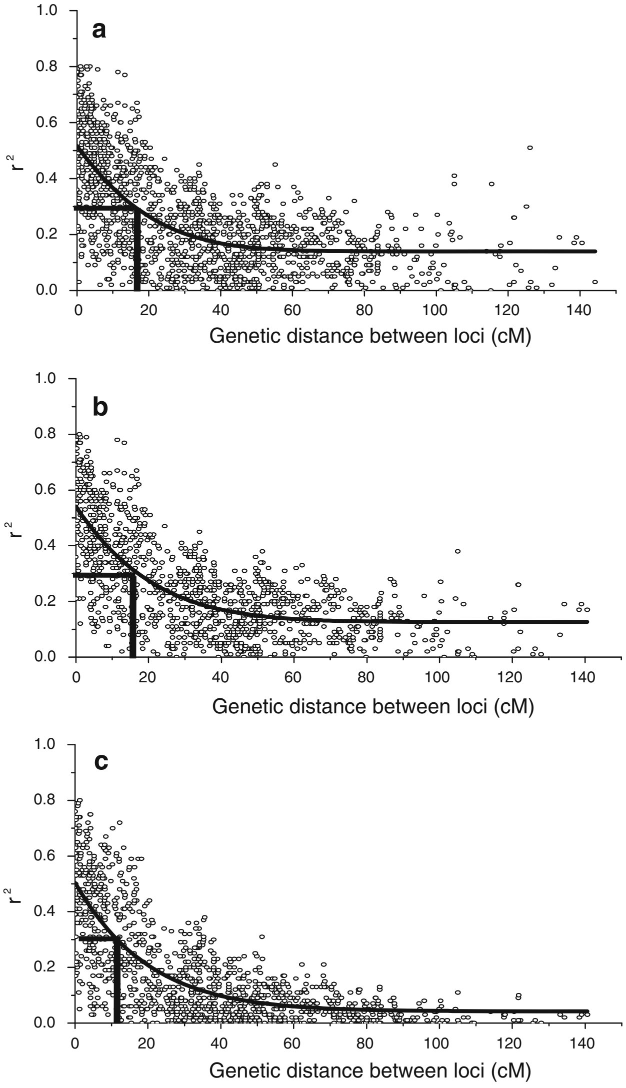

Fig. 2 Decay of linkage disequilibrium (r^2) over genetic distances **a** on all chromosomes for all accessions, **b** on chromosome 2 (50 markers) for all accessions, **c** on chromosome 2 for *S. l. cerasiforme* accessions. Each plot of r^2 over genetic distance is fitted by non-linear regression (*black curve*). Genetic distance corresponding to $r^2 = 0.3$ is indicated

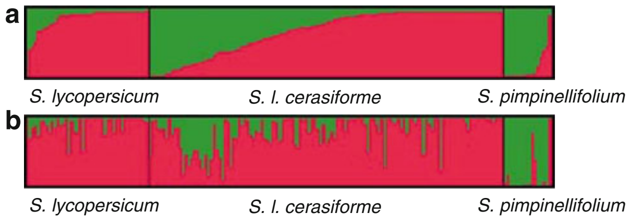

Fig. 3 Comparison of population structure generated from the genotypes with two different types of markers **a** 20 SSR markers, **b** 121 SNP markers. Accessions are in the same order for the two types of markers

Association mapping

Associations between polymorphisms and fruit quality traits were determined by taking into account structure and kinship in a K + Q mixed linear model (MLM) model first on the whole collection (Table 2; Supplemental Table S5; Fig. 4; Supplemental Fig. S1). The comparison of the probabilities obtained with either a simple linear model or K + Q model with Q based on SSR markers (Model A) or SNP markers (Model B) showed that Model A was intermediate between the simple linear model and Model B for a*, b*, FIR and TA, and was similar to the simple linear model for L, FW, LCN, and SSC. The three models provided very close results for pH (Fig. 5; Supplemental Fig. S2). Model B thus corrected better for the structure than the other models and revealed much less significant associations. To reduce the false positive associations, we thus described the associations with Model A using a threshold of the corrected p value of 0.005, while we used a threshold of 0.05 for Model B. For Model A, 132 significant associations were detected for six traits (L, FIR, FW, LCN, SSC and SUG), with a maximum of 41 (for SUG) and a minimum of 7 (for L) associations per trait. No association was detected for a*, b*, pH and TA. A total of 74 SNPs spread on almost all chromosomes except chromosome 7 and 8 were involved in at least one association. The percentage of phenotypic variation explained by the SNPs ranged from 4 to 24 %. For Model B, ten associations were significant for seven traits. Six associations were significant with both models. They concerned FIR on chromosome 4, FW, LCN and SSC on chromosome 2 and SUG on chromosome 10. In order to compare the associations with QTL previously mapped, markers showing significant association and linked in less than 10 cM were grouped together. Thus, 31 groups of association (with 1–27 associations) were defined (Table 2; Supplemental Table S5; Fig. 4; Supplemental Fig. S1).

For L, six groups of association (involving seven SNPs) were identified with Model A on chromosomes 2 (two SNPs), 4 (two SNPs) and 9 (three SNPs). The two most significantly associated markers TD160-458 in group 4.2 on chromosome 4 and Z1475-87 in group 9.2 on chromosome 9 explained 11 and 10 % of the phenotypic variation, respectively. No significant association was detected using Model B. For a*, a single association was detected with Model B on chromosome 2 with TD083-685.

For FIR, 16 groups of association (involving 30 SNPs) were detected using Model A on chromosomes 1, 2 (four SNPs), 3 (three SNPs), 4 (six SNPs), 5 (six SNPs), 6, 9, 10, 11 (six SNPs) and 12. The two most significant associations involved markers TD212-247 in group 4.4 on chromosome 4 and CON176-455 in group 10.1, responsible for

Table 2 Significant associations for color (a*), color (L), firmness (FIR), fruit weight (FW), locule number (LCN), soluble solid content (SSC), sugar content (SUG) and titratable acidity (TA) estimated with K + Q models on 188 accessions

Trait (correlation with Q1[a])	SNP group[b]	Chromosome	Locus	Location[c]	Model A			Model B
					Corrected *p* value[d]	R^2[e]	MAF	Corrected *p* value[d]
a* (0.02–0.11)	2.5	2	TD083-685	133.1	0.042 ns	0.09	0.47	0.011
L (0.29–0.43)	2.2	2	TD049-528	72.4	0.003	0.09	0.36	ns
	2.4	2	Z2117-98	120.5	0.002	0.09	0.22	ns
	4.1	4	TD200-317	6.6	0.004	0.09	0.19	ns
	4.2	4	TD160-458	27.7	4.53×10^{-04}	0.11	0.31	ns
	9.2	9	Z1475-87	51.0	6.47×10^{-04}	0.10	0.17	ns
	9.4	9	TD168-241	99.5	7.78×10^{-04}	0.09	0.13	ns
FIR (0.3–0.49)	1.1	1	CON203-643	44.7	1.86×10^{-04}	0.11	0.22	ns
	2.1	2	TD091-657	49.5	0.002	0.07	0.22	ns
	2.4	2	TD113-132	121.6	9.74×10^{-06}	0.14	0.14	ns
	3.2	3	CON174-206	102.3	8.59×10^{-06}	0.13	0.21	ns
	4.3	4	Z1703-106	65.5	1.51×10^{-07}	0.17	0.12	0.194 ns
	4.4	4	TD212-247	82.2	3.27×10^{-11}	0.24	0.17	0.011
	4.5	4	CON219-313	122.0	2.20×10^{-06}	0.15	0.22	ns
	4.5	4	CON105-290	131.7	0.001	0.10	0.12	ns
	5.1	5	CON173-501	68.3	1.16×10^{-07}	0.16	0.23	ns
	5.2	5	CON222-388	75.8	2.72×10^{-04}	0.11	0.19	ns
	6.1	6	TD025-87	20.8	0.002	0.09	0.12	ns
	9.2	9	TD167-449	59.8	6.84×10^{-04}	0.10	0.16	ns
	10.1	10	CON176-455	59.0	1.01×10^{-07}	0.17	0.12	0.117 ns
	11.1	11	TD247-57	6.4	3.10×10^{-05}	0.13	0.15	ns
	11.2	11	TD251-230	35.4	6.66×10^{-05}	0.12	0.12	ns
	11.3	11	CON141-576	54.8	7.38×10^{-07}	0.16	0.18	ns
	11.4	11	CON50-294	62.3	9.55×10^{-07}	0.16	0.13	ns
	12.2	12	TD156-314	93.7	0.004	0.08	0.12	ns
Log (FW) (0.47–0.53)	1.2	1	TD011-260	97.3	2.68×10^{-04}	0.08	0.13	ns
	2.3	2	TD133-395	83.8	7.30×10^{-06}	0.12	0.34	ns
	2.4	2	TD116-707	122.2	7.38×10^{-08}	0.16	0.34	0.007
	2.5	2	TD083-685	133.1	2.42×10^{-04}	0.12	0.47	0.157 ns
	3.1	3	CON57-121	63.5	0.003	0.07	0.13	ns
	3.3	3	TD152-159	167.1	0.002	0.09	0.17	ns
	4.1	4	TD200-317	6.6	0.001	0.08	0.19	ns
	4.3	4	CON300-472	68.6	1.000 ns	0.01	0.47	0.045
	4.4	4	CON130-112	91.8	9.21×10^{-04}	0.10	0.23	0.077 ns
	6.1	6	TD025-87	20.8	0.001	0.08	0.12	ns
	9.1	9	Z1707-100	38.0	0.003	0.08	0.17	ns
	9.2	9	Z1475-87	51.0	1.08×10^{-04}	0.11	0.17	ns
	9.3	9	Z2305-99	69.4	1.55×10^{-05}	0.11	0.22	ns
	9.4	9	TD243-38	114.4	2.32×10^{-06}	0.13	0.31	0.206 ns
	10.1	10	CON369-191	52.6	9.34×10^{-04}	0.09	0.34	ns
	12.3	12	TD008-95	112.5	5.77×10^{-05}	0.09	0.19	ns
Log (LCN) (0.24–0.30)	2.2	2	TD049-528	72.4	1.54×10^{-04}	0.09	0.36	ns
	2.3	2	TD133-395	83.8	9.27×10^{-05}	0.14	0.34	0.045
SSC (0.41–0.47)	1.2	1	TD011-260	97.3	0.003	0.06	0.13	ns
	1.3	1	Z2300-99	140.6	9.56×10^{-04}	0.10	0.12	ns
	2.3	2	TD280-108	88.6	8.94×10^{-05}	0.13	0.43	0.023
	2.4	2	TD116-707	122.2	0.001	0.09	0.34	ns

Table 2 continued

Trait (correlation with Q1[a])	SNP group[b]	Chromosome	Locus	Location[c]	Model A			Model B
					Corrected *p* value[d]	R^2[e]	MAF	Corrected *p* value[d]
SSC (0.41–0.47)	2.5	2	TD178-104	138.4	0.020 ns	0.12	0.11	0.037
	3.1	3	CON57-121	63.5	0.001	0.08	0.13	ns
	4.1	4	TD200-317	6.6	1.80×10^{-05}	0.12	0.19	ns
	6.1	6	TD025-87	20.8	1.71×10^{-04}	0.11	0.12	ns
	9.1	9	Z1707-100	38.0	3.54×10^{-07}	0.16	0.17	ns
	9.2	9	Z1475-87	51.0	0.003	0.07	0.17	ns
	9.3	9	TD237-253	70.0	0.003	0.07	0.14	ns
	9.4	9	TD168-241	99.5	1.02×10^{-04}	0.12	0.13	ns
	10.1	10	CON369-191	52.6	0.003	0.08	0.34	ns
	11.3	11	TD255-218	56.9	2.09×10^{-06}	0.05	0.13	ns
	11.4	11	CON50-294	62.3	0.001	0.07	0.13	ns
	12.1	12	Z2302-103	21.0	0.004	0.08	0.15	ns
SUG (0.39–0.52)	1.3	1	Z2300-99	140.6	1.06×10^{-05}	0.14	0.12	ns
	2.1	2	TD091-657	49.5	0.004	0.07	0.22	ns
	2.2	2	TD139-547	72.2	1.37×10^{-04}	0.08	0.20	ns
	2.3	2	TD133-395	83.8	7.66×10^{-07}	0.15	0.34	0.194 ns
	2.4	2	TD114-259	121.7	6.72×10^{-04}	0.09	0.14	ns
	2.5	2	TD178-104	138.4	1.34×10^{-06}	0.14	0.11	0.206 ns
	3.1	3	CON57-121	63.5	0.001	0.08	0.13	ns
	4.1	4	TD200-317	6.6	2.15×10^{-06}	0.14	0.19	ns
	4.5	4	CON105-290	131.7	5.55×10^{-04}	0.08	0.12	ns
	6.1	6	TD025-87	20.8	1.94×10^{-04}	0.11	0.12	ns
	9.1	9	Z1707-100	38.0	3.68×10^{-05}	0.12	0.17	ns
	9.2	9	Z1475-87	51.0	5.37×10^{-04}	0.09	0.17	ns
	9.4	9	TD168-241	99.5	3.57×10^{-04}	0.11	0.13	ns
	10.1	10	CON369-191	52.6	6.59×10^{-04}	0.10	0.34	0.011
	11.3	11	TD255-218	56.9	4.42×10^{-04}	0.07	0.13	ns
	11.4	11	CON50-294	62.3	1.23×10^{-04}	0.09	0.13	ns
	12.1	12	Z2302-103	21.0	3.81×10^{-04}	0.11	0.15	ns
TA (0.26–0.22)	2.3	2	TD275-101	88.5	0.110 ns	0.04	0.13	0.048
	4.3	4	Z1370-98	66.8	ns	0.08	0.14	0.003

Model A: MLM model, with structure and kinship based on 20 SSR (*p* values lower than 0.005 are shown with indication on allele effect), Model B: MLM model with structure and kinship based on 121 SNP (*p* value lower than 0.05 are shown). Only the most significant association from each group is shown. See Supplemental Table S5 for detail

MAF minimum allele frequencies

[a] Q1 is the probability that an individual belongs to the "cultivated" subpopulation generated from STRUCTURE2.1 software (Pritchard et al. 2000). Correlations with the *Q* value defined by 20 SSR markers and 121 SNPs

[b] Associated SNPs in less than 10 cM on each chromosome were grouped together. SNP which is 10 cM apart from the other SNPs was assigned as an independent group. Groups are detailed in Supplemental Table S5

[c] Genetic distance of the marker on EXPEN2000 reference map (http://www.solgenomics.net)

[d] *p* values were corrected following the standard Bonferroni procedure; *ns* non-significant

[e] R^2 were calculated using a Q model

24 and 17 % of the phenotypic variation, respectively. The association with TD212-247 was also significant with Model B.

For FW, 15 groups of association (involving 23 SNPs) were detected with Model A on chromosomes 1, 2 (eight SNPs), 3 (two SNPs), 4 (two SNPs), 9 (six SNPs), 10 (two SNPs) and 12. Marker TD116-707 in group 2.4 and marker TD243-38 in group 9.4 showed the most significant associations and explained 16 and 13 % of the FW variation. Using Model B, TD116-707 and CON300-472 showed significant associations, on chromosomes 2 and 4, respectively.

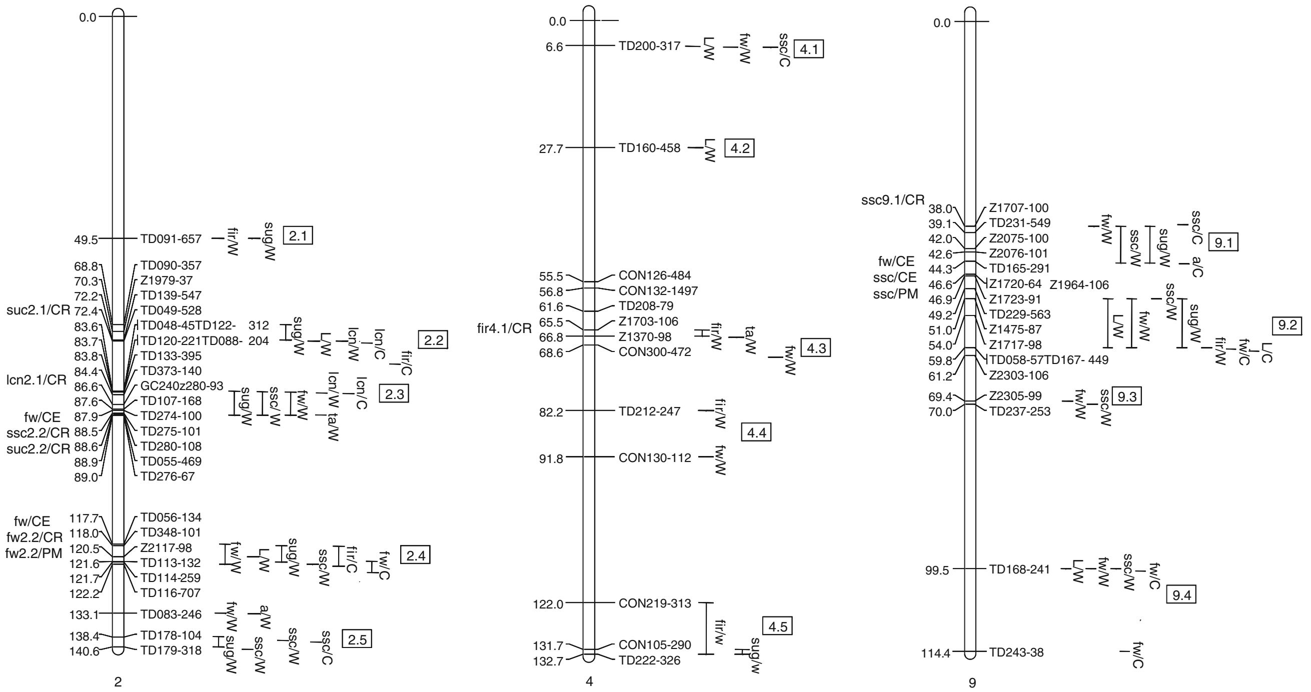

Fig. 4 Comparison of associations and QTLs identified by linkage mapping on chromosomes 2, 4 and 9. SNPs were mapped on tomato EXPEN 2000 reference map (http://www.solgenomics.net). Associations detected in the 188 accessions (W) and in 127 *S. l. cerasiforme* accessions (C) are indicated to the *right* of the chromosomes. Associations were estimated with K + Q model, Model A with structure and kinship based on 20 SSR marker, Model B with structure and kinship based on 121 SNP (*common font* associations detected with Model A, in *italic* associations detected with Model B, in *bold* associations detected with both models). *Horizontal line* correspond to the genetic location of associated marker, associations are linked together by a *vertical line* when linked markers in less than 10 cM are associated to the same trait. Associated SNPs in less than 10 cM on each chromosome were grouped together. SNP which is 10 cM apart from the others were assigned as independent groups. Groups are named as consecutive number according to their genetic location on each chromosome. Traits are *FIR* firmness, *FW* fruit weight, *SSC* soluble solids content, *SUG* total sugar content, *LCN* locule number, *a**, *L* color, *TA* titratable acidity. Only SNPs significantly associated with one trait are represented on chromosome 2, where markers are too dense. QTLs identified by linkage mapping in the populations from crosses of *S. lycopersicum* × *S. l. cerasiforme* (Saliba-Colombani et al. 2001), *S. lycopersicum* × *S. pimpinellifolium* (Grandillo and Tanksley 1996) and *S. lycopersicum* × *S. l. cheesmanii* (Goldman et al. 1995) are shown to the *left* of the chromosomes (CR = QTL from *S. l. cerasiforme*, CE = QTL from *S. l. cheesmanii*, PM = QTL from *S. pimpinellifolium*). Only QTL co-localizing with an association are shown

For LCN, two groups of association (involving two SNPs) were identified on chromosome 2. The two associations involved marker TD133-395 in group 2.3 (also significant with Model B) and TD049-528 in group 2.2, explained 14 and 9 % of the phenotypic variation.

For SSC, 16 groups of association (involving 28 SNPs) were detected with Model A on chromosomes 1 (two SNPs), 2 (ten SNPs), 3, 4, 9 (eight SNPs), 10 (two SNPs), 11 (two SNPs) and 12. The most significant associations involved markers Z1707-100 in group 9.1 and TD255-218 in group 11.3, explained 16 and 5 % of the soluble solid variation. Two associations on chromosome 2 were significant with Model B.

For SUG, 17 groups of association were identified on chromosomes 1, 2 (22 SNPs), 3, 4 (three SNPs), 6, 9 (eight SNPs), 10 (two SNPs), 11 (two SNPs) and 12 with Model A. The strongest association involved marker TD133-395 in group 2.3 and TD178-104 in group 2.5 on chromosome 2, which explained 15 and 14 % of the sugar content variation. The only significant association with Model B involved CON369-191 on chromosome 10.

We then performed association analysis using Model A and B on the subset of 127 *S. l. cerasiforme* accessions (Table 3; Fig. 4; Supplemental Fig. S1). Population structure accounted for much less variation of all the traits, with the highest correlation between *Q* values and FIR (Table 3). The comparison of the probabilities associated to the tests using simple linear model, Model A and Model B showed that Model A was still intermediate between simple linear model and Model B for L, FW, FIR and was similar to the simple linear model for LCN. The three models were very close for a*, b*, SUG, SSC, pH and TA (Fig. 5; Supplemental Fig. S3). For Model A, ten significant associations were found for FIR, FW, LCN and SSC. Population structure accounted for 25, 11, 13 and 2 % of the phenotypic variation for these traits, respectively. Eight significant associations were common with associations found with 188 accessions. Two new associations were

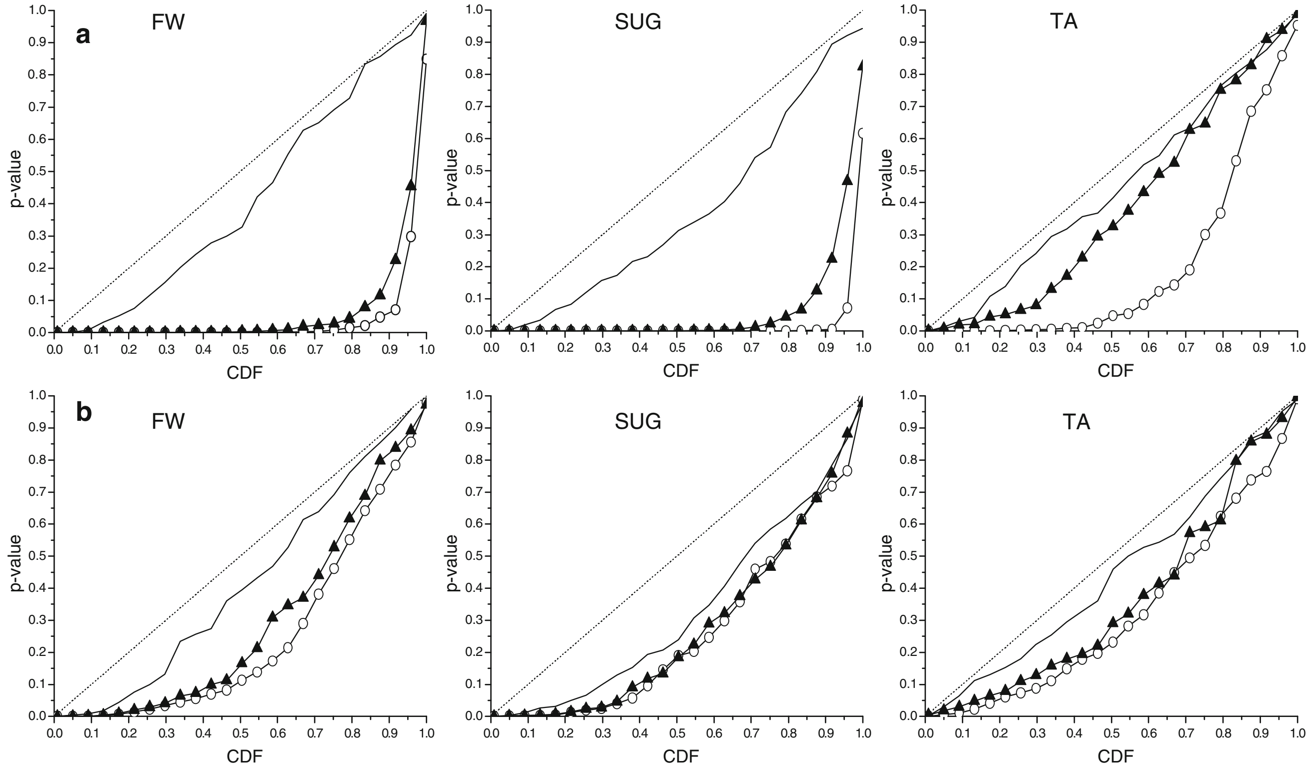

Fig. 5 Cumulative density functions (CDF) using three alternative models of association for fruit weight (FW), sugar content (SUG) and titratable acidity (TA). Associations are tested for 121 polymorphic sites on 188 accessions (**a**) and 127 *S. l. cerasiforme* (**b**). Simple linear model (*empty circle*) and K + Q models, with structure and kinship based on SSR markers (*black triangle*), and on 121 SNP markers (*black line*) were tested. The *diagonal* indicates uniform distribution of *p* values under the expectation that random SNPs are unlinked to the polymorphisms controlling these traits (H0, no SNP effect)

observed between marker Z1117-98 and FW, TD032-112 and SSC, responsible for 18 and 17 % variation, respectively. No significant associations were found for a*, b* and L, pH and TA. With Model B, 18 significant associations were detected for seven traits. Three associations were common with associations found in 188 accessions. Five associations were detected with both models, for FIR and for LCN on chromosome 2, for FW on chromosome 2 and 9 and for SSC on chromosome 5. Only two associations were common to the two models and the two samples, one for FW with TD116-707 and one for LCN with TD133-395.

Discussion

We herein present the phenotypic and genetic diversity of a large collection of tomato accessions representing wild relatives, intermediate and cultivated types characterized using a SNPlex^{TM} genotyping assay. The percentage of SNPs successfully scored (73 %) is consistent with the success rate reported by Pindo et al. (2008) and Berard et al. (2009). The results suggest that this assay is reliable, flexible and cost-effective for medium-throughput SNP detection. This pioneering technology opened the way to new technologies more flexible or with higher throughput like the one proposed by Fluidigm (Moonsamy et al. 2011) or Illumina GoldenGate^{TM}. Although SNPlex^{TM} assay is no more used, the SNPs used in this assay may be adapted to other genotyping platforms to be used by tomato breeders.

A source of phenotypic variability

The sample consisted of 127 cherry type tomato accessions *S. l. cerasiforme*, 44 *S. lycopersicum* large fruited accessions and 17 *S. pimpinellifolium* accessions. The genome structure of *S. l. cerasiforme* accessions was previously described as a mosaic of *S. lycopersicum* and *S. pimpinellifolium* genomes (Ranc et al. 2008). Compared to *S. lycopersicum*, *S. l. cerasiforme* and *S. pimpinellifolium* fruits are much smaller (less than 20 g for cherry types, less than 5 g for wild types) with only two or three locules and higher sugar content, soluble solid content and titratable acidity. The same trend of variation was already observed in smaller samples (Davies and Hobson 1981; Causse et al. 2003). Correlations among traits were quite homogenous in the whole collection and among the three groups.

Table 3 Significant associations for color (a*), color (L), firmness (FIR), fruit weight (FW), locule number (LCN), sugar content (SUG), and soluble solid content (SSC) estimated with K + Q models on 127 *S. l. cerasiforme* accessions

Trait (correlation with Q1[a])	Chromosome	Locus	Location[b]	Model A			Model B
				p value[c]	R^{2d}	MAF	*p* value[c]
a* (0.12–0.20)	5	CON310-990	71.9	ns	0.01	0.21	0.002
	9	Z1723-91	46.9	ns	0.03	0.18	0.008
L (0.17–0.25)	9	TD167-449	59.8	ns	0.03	0.11	0.011
FIR (0.25–0.35)	2	TD018-103	75.3	ns	0.01	0.23	0.026
	2	TG454z273-252	75.5	ns	0.05	0.13	0.040
	2	TD348-101	118.0	0.001	0.12	0.16	ns
	2	TD113-132	121.6	8.13×10^{-4}	0.16	0.10	0.001
	2	TD114-259	121.7	0.002	0.15	0.15	0.018
Log (FW) (0.11–0.20)	2	Z2117-98	120.5	0.001	0.18	0.19	ns
	2	TD116-707	122.2	6.75×10^{-7}	0.26	0.36	0.002
	9	TD058-57	59.8	ns	0.05	0.15	3.88×10^{-4}
	9	TD167-449	59.8	ns	0.04	0.11	0.025
	9	TD168-241	99.5	0.004	0.13	0.08	ns
	9	TD243-38	114.4	8.22×10^{-4}	0.15	0.30	ns
Log (LCN) (0.13–0.18)	2	TD049-528	72.4	4.00×10^{-4}	0.16	0.36	ns
	2	TD120-221	83.7	0.009 ns	0.12	0.19	0.007
	2	TD133-395	83.8	6.67×10^{-5}	0.20	0.33	0.009
SSC (0.02–0.03)	2	TD178-104	138.4	0.028 ns	0.10	0.08	0.045
	4	TD200-317	6.6	ns	0.08	0.15	0.015
	5	TD032-112	72.3	0.002	0.17	0.19	0.009
	9	Z1707-100	38.0	0.051 ns	0.10	0.12	0.016
SUG (0.02–0.14)	5	TD032-112	72.3	ns	0.10	0.19	0.033
	11	TD247-57	6.4	ns	0.08	0.12	0.015

Model A: MLM model, with structure and kinship based on 20 SSR (only *p* values lower than 0.005 are shown with indication on allele effect), Model B MLM model with structure and kinship based on 121 SNP (*p* value lower than 0.05 are shown)

MAF minimum allele frequencies

[a] Q1 is the probability that an individual belongs to the "cultivated" subpopulation generated from STRUCTURE2.1 software (Pritchard et al. 2000). Correlations with the *Q* value defined by 20 SSR markers and 121 SNPs

[b] Genetic distance of the marker on EXPEN2000 reference map (http://www.solgenomics.net)

[c] *p* values were corrected following the standard Bonferroni procedure; *ns* non-significant

[d] R^2 were calculated using a Q model

S. pimpinellifolium and cherry type accessions may be useful sources of alleles for tomato fruit quality improvement, particularly to improve firmness or the content in sugars and acids, but the strong negative correlation between fruit size and soluble solids or sugar content may hamper the simultaneous improvement of both traits. A better knowledge of the loci controlling these traits may thus help breeders to use these resources.

LD decay and population structure

Association mapping requires a thorough understanding of LD and population structure in the collection. In tomato, LD remains high over genetic distance. Robbins et al. (2011) observed that LD decayed at 6–8 cM in a collection of 102 tomato varieties, 6–14 cM within 39 processing varieties, and 3–16 cM within 24 fresh market varieties. In our study, slightly higher level of LD was observed in the whole collection, although it was lower in *S. l. cerasiforme.* This result is consistent with van Berloo et al. (2008) who found LD extent to 15–20 cM using AFLP markers in a sample of 18 cherry tomato accessions. Such extent of LD will allow the identification of regions carrying QTL rather than direct associations with candidate genes. Nevertheless, at the physical scale, some SNPs may appear in complete equilibrium with their neighbors (Munos et al. 2011). At the physical scale, recombination hotspots may be detected with very low LD in short distances (Ranc et al. 2012). When considering LD among chromosomes, only a few pairs of markers (less than 10 loci) exhibited high LD (data not shown). SSR and SNP markers revealed similar structure patterns with two main groups and many intermediates. This result is consistent with Hamblin et al. (2007) who compared the structure based on 89 SSR to the structure based on 847 SNPs in a set of 259 maize lines. The SSRs performed better to cluster the germplasm into populations, but the population structure assessed by both marker types was similar. Laval et al. (2002) stated that $k - 1$ times more bi-allelic markers are needed to obtain the same genetic distance accuracy as a set of microsatellites with k alleles. In the present study, the average number of alleles per SSR locus was about 7, thus 20 SSR markers should correspond to 120 biallelic SNP markers and should thus provide the same accuracy. Nevertheless, several individuals were not classified in the same groups and both structures did not correct for structure the same way (Fig. 5). The SSR were less efficient than SNP markers. This result may be due to the fact that the SNPs revealed more loci than SSR and were for a large extent chosen to be located in regions, where QTLs were previously detected in crosses between one wild or cherry accession and a cultivated one, and thus may be linked to the polymorphisms responsible of the structure.

Associations confirmed previously identified QTLs and detected new candidate polymorphisms

Compared with previous association analysis (Nesbitt and Tanksley 2002; Mazzucato et al. 2008; Munos et al. 2011), it is the first time that associations are analyzed between more than 100 SNP markers and ten tomato fruit quality traits in a large collection. Ranc et al. (2012), in a pilot study on chromosome 2 and 90 accessions, showed that association mapping permitted to map QTLs that were already cloned. They showed that to get just a location of major QTLs, a few thousand SNPs will be sufficient, while their precise characterisation and the identification of mutations which have evolved under balancing selection and introgressed into many accessions (like *Lcn2.1)* may require a much larger number of SNPs (more than 50,000). The way we take into account the population structure influences the results, as population structure may cause false association results (Mezmouk et al. 2011). Statistical methods have been developed to deal with the effect of population structure (Pritchard et al. 2000; Price et al. 2006; Yu et al. 2006). The MLM model has been shown to efficiently correct for the effects of population structure by including the structure and a matrix of genetic similarity among the accessions (Yu et al. 2006; Atwell et al. 2010). Population structure accounted for a large part of the phenotypic variation for several traits in the 188 tomato accessions and accounted for much less phenotypic variation in the 127 *S. l. cerasiforme* accessions. Although a structure with two subgroups was detected with both SSR and SNP markers, the classification of some individuals changed. It thus appeared that Model A did not correct well for the structure, leading to a large number of associations, particularly for FW, which is strongly correlated with structure. To reduce the false positive associations, a more stringent p value threshold was used in Model A. In the collection of 127 *S. l. cerasiforme* accessions, the structure is less significantly correlated with the trait values and the number of associations detected with both models was less different. Associations were found for most of the traits. With Model A, 132 and ten SNPs were associated with the traits, for all accessions and *S. l. cerasiforme* accessions, respectively, while with Model B these numbers were 11 and 18. Only eight and three associations were detected in both sets of accessions when using Model A and B, respectively. Associations between markers and fruit quality traits were mostly localized on chromosome 2, 4 and 9, but this is partly due to the higher number of markers representing these chromosomes (50, 12 and 18 markers, respectively) than the other chromosomes. Almost all marker groups, except group 4.2, were associated with two or more traits. For example, group 2.3 was associated with sugar, soluble solid content, fruit weight, locule number and titratable acidity. Such co-localization of associations for several traits was found in several studies (Zhao et al. 2011; Bergelson and Roux 2010). Co-localized associations for soluble solid content and sugar content, fruit weight and locule number were also frequent. Such co-localization might be related to the pleiotropic effects of the same genes or due to genetic linkage, as already shown for QTL (Lecomte et al. 2004).

In tomato, QTLs for fruit size, shape and quality traits have been mapped in several bi-parental populations involving one wild species (Paterson et al. 1991; Goldman et al. 1995; Eshed and Zamir 1995; Grandillo and Tanksley 1996; Frary et al. 2000; van der Knaap and Tanksley 2001, 2003; Causse et al. 2002; Barrero and Tanksley 2004; Causse et al. 2004; Lecomte et al. 2004). A few genes controlling fruit trait QTL have been cloned, like *FW2.2* which controls fruit weight (Frary et al. 2000), *Lin5* which is responsible for fruit sugar content (Fridman et al. 2000) or *Lcn2.1* which controls locule number (Munos et al. 2011).

The localization of associations for eight quality traits (a*, L, FIR, FW, LCN, SUG, SSC and TA) were compared with those of QTLs previously detected from populations derived from crosses of *S. lycopersicum* × *S. l. cerasiforme* (hereafter named EC × CR) (Saliba-Colombani et al. 2001), *S. lycopersicum* × *S. pimpinellifolium* (hereafter named EC × PM) (Grandillo and Tanksley 1996) and *S. lycopersicum* × *S. lycopersicum cheesmanii* (hereafter named EC × CE), another species closely related to the cultivated tomato (Goldman et al. 1995). An association was considered to be in the same region as a QTL when it mapped within a 20 cM region of the tomato EXPEN 2000 map (http://www.solgenomics.net) around the QTL. On average, 30 % of the associations were localised in a region, where a QTL for the same trait has been mapped. With Model A, 40 associations (two for FIR, ten for FW, two for LCN, 12 for SUG and 14 for SSC) were co-localized with previously identified QTL or known genes, and many other associations for these traits and associations for a*, L and TA were detected in regions, where no known QTL have been located to date (Fig. 4; Supplemental Fig. S1). More associations were found to be co-localized with previously identified QTL with Model A than with Model B, because on one hand, Model B revealed less associations and on another hand, the structure being better taken into account, the QTL responsible for this structure may be more difficult to detect.

For FW, the markers of group 2.4 associated with FW shared the same position as the major QTL, fw2.2. Actually, TD056-134 corresponds to a polymorphism in fw2.2 promoter. Nesbitt and Tanksley (2002) failed to detect any association in the region around fw2.2 in a small set of *S. l. cerasiforme* accessions, but using a larger sample, Ranc et al. (2012) could detect one and identify the accessions that carry the large fruit allele in this region. In the present study, the association was detected with both Models with TD116, a marker less than 2 cM far from fw2.2. Association of group 2.3, group 3.1 and group 9.1 were also co-localized with QTL for FW detected in EC × CE population (Goldman et al. 1995). For LCN, association of group 2.3 co-localized with lcn2.1, a QTL controlling LCN identified in the EC × CR population (Saliba-Colombani et al. 2001) and recently cloned by Munos et al. (2011). Sequencing 1,800 bp around the lcn2.1 locus in 90 accessions allowed the identification of two SNPs strongly associated with the variation of locule number of tomato fruit. These SNPs were also associated with the 188 accessions. TD133-395 was in less than 2 cM from these two SNP. These results show that our resolution does not allow the precise localisation of the responsible genes, but we may detect regions carrying relevant QTLs. The combination of QTL fine mapping and association study is much more efficient for this purpose.

For SUG, associations of group 2.2 and group 2.3 co-localized with two QTLs detected in the EC × CR population (Saliba-Colombani et al. 2001). Marker TD274 (group 2.3) was also strongly associated with FW and SSC. It was defined in the 5′ region of the gene *Solyc02g085170.2* coding for a glucose transporter protein. Marker TD055-469 (group 2.3) was also found to be associated with FW and SSC. It was designed in the 5′ region of a gene *Solyc02g085500.2* coding for the *ovate* protein. The Ovate locus is responsible for pear fruit shape and no effect of this locus on FW, SSC or SUG has been reported before. This polymorphism could thus only be linked to a causative polymorphism. Association of group 10.1 with sugar content was co-localized with a QTL detected for sugar content in EC × CR population (Saliba-Colombani et al. 2001). For SSC, association of group 2.3 was located in the same region as QTL for SSC detected in the EC × CR population (Saliba-Colombani et al. 2001). Association of group 3.1 was found to be co-localized with a QTL for SSC in the EC × PM population (Grandillo and Tanksley 1996). Association of group 9.1 was co-localized with a QTL detected in the three studies. Marker Z1707-100 in this group was closely linked to the previously cloned QTL *lin5* (Fridman et al. 2000) and found in association with soluble solid content. For FIR, the main association of group 4.3 was located in the same region as a QTL detected in the EC × CR population (Saliba-Colombani et al. 2001).

In conclusion, we identified several associations between SNP markers and fruit traits in a large sample of tomato accessions. The large LD and frequently low MAF in the cultivated group may hamper association discovery in this group. The *S. l. cerasiforme* accessions represent intermediate type between cultivated and wild species with various degrees of introgression as shown by the admixture structure of these accessions. This group exhibited higher MAF on average than cultivated group, lower LD and a less structured pattern. Association mapping should thus be easier with this group. About half of the associations detected with Model B were also detected with Model A in both sets of accessions. Around 30 % of the associations detected with Model A were localized in regions, where QTLs were previously mapped. We thus presented these results, although we are aware that several of these associations may be false positives. This approach will thus have to be combined with QTL fine mapping to identify the relevant polymorphisms as suggested by Nemri et al. (2010). Model B allowed the detection of 25 associations. Nevertheless, the density of SNP is too low to identify SNPs in candidate genes. The availability of large panels of SNPs (Sim et al. 2012) will soon allow whole genome scan for association.

Acknowledgments The research was financed by INRA AIP Bio-resources project. Jiaxin Xu was supported by China Scholarship Council. We thank Hélène Burck for characterizing and maintaining the INRA tomato Genetic Resources collection. We are grateful to Yolande Carretero, Esther Pelpoir and Laure David for their help with growing and phenotyping cherry tomato accessions.

References

Agrama HA, Eizenga GC, Yan W (2007) Association mapping of yield and its components in rice cultivars. Mol Breed 19:341–356

Aranzana MJ, Kim S, Zhao KY, Bakker E, Horton M, Jakob K, Lister C, Molitor J, Shindo C, Tang CL, Toomajian C, Traw B, Zheng HG, Bergelson J, Dean C, Marjoram P, Nordborg M (2005) Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. PLoS Genet 1:531–539

Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Li Y, Meng DZ, Platt A, Tarone AM, Hu TT, Jiang R, Muliyati NW, Zhang X, Amer MA, Baxter I, Brachi B, Chory J, Dean C, Debieu M, de Meaux J, Ecker JR, Faure N, Kniskern JM, Jones JDG, Michael T, Nemri A, Roux F, Salt DE, Tang CL, Todesco M, Traw MB, Weigel D, Marjoram P, Borevitz JO, Bergelson J, Nordborg M (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465:627–631

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

Berard A, Le Paslier MC, Dardevet M, Exbrayat-Vinson F, Bonnin I, Cenci A, Haudry A, Brunel D, Ravel C (2009) High-throughput single nucleotide polymorphism genotyping in wheat (*Triticum* spp.). Plant Biotechnol J 7:364–374

Bergelson J, Roux F (2010) Towards identifying genes underlying ecologically relevant traits in *Arabidopsis thaliana*. Nat Rev Genet 11:867–879

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635

Causse M, Saliba-Colombani V, Lecomte L, Duffe P, Rousselle P, Buret M (2002) QTL analysis of fruit quality in fresh market tomato: a few chromosome regions control the variation of sensory and instrumental traits. J Exp Bot 53:2089–2098

Causse M, Buret M, Robini K, Verschave P (2003) Inheritance of nutritional and sensory quality traits in fresh market tomato and relation to consumer preferences. J Food Sci 68:2342–2350

Causse M, Duffe P, Gomez MC, Buret M, Damidaux R, Zamir D, Gur A, Chevalier C, Lemaire-Chamley M, Rothan C (2004) A genetic map of candidate genes and QTLs involved in tomato fruit size and composition. J Exp Bot 55:1671–1685

Davies JN, Hobson GE (1981) The constituents of tomato fruit—the influence of environment, nutrition, and genotype. Crit Rev Food Sci Nutr 15:205–280

De La Vega FA, Lazaruk KD, Rhodes MD, Wenz MH (2005) Assessment of two flexible and compatible SNP genotyping platforms: TaqMan (R) SNP genotyping assays and the SNPlex (TM) genotyping system. Mutat Res Fundam Mol Mech Mutagen 573:111–135

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

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

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357–374

Frary A, Nesbitt TC, Grandillo S, van der Knaap E, Cong B, Liu JP, Meller J, Elber R, Alpert KB, Tanksley SD (2000) fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. Science 289:85–88

Fridman E, Pleban T, Zamir D (2000) A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene. Proc Natl Acad Sci USA 97:4718–4723

Goldman IL, Paran I, Zamir D (1995) Quantitative trait locus analysis of a recombinant inbred line population derived from a *Lycopersicon esculentum x Lycopersicon cheesmanii* cross. Theor Appl Genet 90:925–932

Grandillo S, Tanksley SD (1996) QTL analysis of horticultural traits differentiating the cultivated tomato from the closely related species *Lycopersicon pimpinellifolium*. Theor Appl Genet 92:935–951

Hall D, Tegstrom C, Ingvarsson PK (2010) Using association mapping to dissect the genetic basis of complex traits in plants. Brief Funct Genomics 9:157–165

Hamblin MT, Warburton ML, Buckler ES (2007) Empirical comparison of simple sequence repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. Plos One 2(12):e1367

Hardy OJ, Vekemans X (2002) SPAGEDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol Ecol Notes 2:618–620

Laval G, SanCristobal M, Chevalet C (2002) Measuring genetic distances between breeds: use of some distances in various short term evolution models. Genet Select Evol 34:481–507

Lecomte L, Saliba-Colombani V, Gautier A, Gomez-Jimenez MC, Duffe P, Buret M, Causse M (2004) Fine mapping of QTLs of chromosome 2 affecting the fruit architecture and composition of tomato. Mol Breed 13:1–14

Mamidi S, Chikara S, Goos RJ, Hyten DL, Annam D, Moghaddam SM, Lee RK, Cregan PB, McClean PE (2011) Genome-wide association analysis identifies candidate genes associated with iron deficiency chlorosis in soybean. Plant Genome 4:154–164

Mazzucato A, Papa R, Bitocchi E, Mosconi P, Nanni L, Negri V, Picarella ME, Siligato F, Soressi GP, Tiranti B, Veronesi F (2008) Genetic diversity, structure and marker-trait associations in a collection of Italian tomato (*Solanum lycopersicum* L.) landraces. Theor Appl Genet 116:657–669

Mezmouk S, Dubreuil P, Bosio M, Decousset L, Charcosset A, Praud S, Mangin B (2011) Effect of population structure corrections on the results of association mapping tests in complex maize diversity panels. Theor Appl Genet 122:1149–1160

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

Moonsamy PV, Bonella PL, Williams TC, Holcomb CL, Turenchalk GS, Blake LA, Hoglund BN, Rastrou M, Daigle DA, Simen BB, Goodridge D, Hillman G, Hamilton A, May AP, Erlich HA (2011) Use of the fluidigm (R) access array (TM) system provides simplified amplicon library preparation in next generation sequencing for high throughput hla genotyping. Hum Immunol 72:S142–S142

Munos S, Ranc N, Botton E, Berard A, Rolland S, Duffe P, Carretero Y, Le Paslier MC, Delalande C, Bouzayen M, Brunel D, Causse M (2011) Increase in tomato locule number is controlled by two single-nucleotide polymorphisms located near Wuschel. Plant Physiol 156:2244–2254

Nemri A, Atwell S, Tarone AM, Huang YS, Zhao K, Studholme DJ, Nordborg M, Jones JD (2010) Genome-wide survey of Arabidopsis natural variation in downy mildew resistance using combined association and linkage mapping. Proc Natl Acad Sci USA 107:10302–10307

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

Orsini E, Krchov LM, Uphaus J, Melchinger AE (2012) Mapping of QTL for resistance to first and second generation of European corn borer using an integrated SNP and SSR linkage map. Euphytica 183:197–206

Paterson AH, Damon S, Hewitt JD, Zamir D, Rabinowitch HD, Lincoln SE, Lander ES, Tanksley SD (1991) Mendelian factors underlying quantitative traits in tomato—comparison across species, generations, and environments. Genetics 127:181–197

Pindo M, Vezzulli S, Coppola G, Cartwright DA, Zharkikh A, Velasco R, Troggio M (2008) SNP high-throughput screening in grapevine using the SNPlex (TM) genotyping system. BMC Plant Biol 8:128

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909

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

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

Ranc N, Munos S, Santoni S, Causse M (2008) A clarified position for *Solanum lycopersicum* var. *cerasiforme* in the evolutionary history of tomatoes (solanaceae). BMC Plant Biol 8:130

Ranc N, Muños S, Xu J, Le Paslier MC, Chauveau A, Bounon R, Rolland S, Bouchet JP, Brunel D, Causse M (2012) Genome-wide association mapping in tomato (*Solanum lycopersicum*) is possible using genome admixture of *Solanum lycopersicum* var. *cerasiforme*. Genes Genomes Genet 2:853–864

Robbins MD, Sim SC, Yang WC, Van Deynze A, van der Knaap E, Joobeur T, Francis DM (2011) Mapping and linkage disequilibrium analysis with a genome-wide collection of SNPs that detect polymorphism in cultivated tomato. J Exp Bot 62:1831–1845

Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. Mol Ecol Notes 4:137–138

Saliba-Colombani V, Causse M, Langlois D, Philouze J, Buret M (2001) Genetic analysis of organoleptic quality in fresh market tomato 1. Mapping QTLs for physical and chemical traits. Theor Appl Genet 102:259–272

Sim S-C, Durstewitz G, Plieske J, Wieseke R, Ganal MW, Van Deynze A, Hamilton JP, Robin Buell C, Causse M, Wijeratne S, Francis DM (2012) Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. PLoS One 7(7):e40563. doi:10.1371/journal.pone.0040563

Szalma SJ, Hostert BM, LeDeaux JR, Stuber CW, Holland JB (2007) QTL mapping with near-isogenic lines in maize. Theor Appl Genet 114:1211–1228

Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, Flint-Garcia S, Rocheford TR, McMullen MD, Holland JB, Buckler ES (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat Genet 43:U113–U159

Tobler AR, Short S, Andersen MR, Paner TM, Briggs JC, Lambert SM, Wu PP, Wang Y, Spoonde AY, Koehler RT, Peyret N, Chen C, Broomer AJ, Ridzon DA, Zhou H, Hoo BS, Hayashibara KC, Leong LN, Ma CN, Rosenblum BB, Day JP, Ziegle JS, De La Vega FM, Rhodes MD, Hennessy KM, Wenz HM (2005) The SNPlex genotyping system: a flexible and scalable platform for SNP genotyping. J Biomol Tech (JBT) 16:398–406

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 der Knaap E, Tanksley SD (2001) Identification and characterization of a novel locus controlling early fruit development in tomato. Theor Appl Genet 103:353–358

van der Knaap E, Tanksley SD (2003) The making of a bell pepper-shaped tomato fruit: identification of loci controlling fruit morphology in Yellow Stuffer tomato. Theor Appl Genet 107:139–147

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

Voorrips RE (2002) MapChart: software for the graphical presentation of linkage maps and QTLs. J Hered 93:77–78

Wang JK, Wan XY, Crossa J, Crouch J, Weng JF, Zhai HQ, Wan JM (2006) QTL mapping of grain length in rice (*Oryza sativa L.*) using chromosome segment substitution lines. Genet Res 88:93–104

Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208

Zhao H, Nettleton D, Soller M, Dekkers JCM (2005) Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. Genet Res 86:77–87

Zhao K, Tung CW, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J, McClung AM, Bustamante CD, McCouch SR (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. Nat Commun 2:467