Genome-Wide Association in Tomato Reveals 44 Candidate Loci for Fruit Metabolic Traits[1][W]

Christopher Sauvage*, Vincent Segura, Guillaume Bauchet, Rebecca Stevens, Phuc Thi Do, Zoran Nikoloski, Alisdair R. Fernie, and Mathilde Causse

Institut National de la Recherche Agronomique, UR1052, Génétique et Amélioration des Fruits et Légumes, 84143 Montfavet cedex, France (C.S., G.B., R.S., M.C.); Institut National de la Recherche Agronomique, UR0588, 45075 Orleans cedex 2, France (V.S.); Syngenta Seeds, 31790 Saint Sauveur, France (G.B.); Max-Planck-Institut für Molekulare Pflanzenphysiologie, 14476 Potsdam-Golm, Germany (P.T.D., Z.N., A.R.F.); and Faculty of Biology, University of Science, Vietnam National University, Thanh Xuan, Hanoi, Vietnam (P.T.D.)

Genome-wide association studies have been successful in identifying genes involved in polygenic traits and are valuable for crop improvement. Tomato (*Solanum lycopersicum*) is a major crop and is highly appreciated worldwide for its health value. We used a core collection of 163 tomato accessions composed of *S. lycopersicum*, *S. lycopersicum* var *cerasiforme*, and *Solanum pimpinellifolium* to map loci controlling variation in fruit metabolites. Fruits were phenotyped for a broad range of metabolites, including amino acids, sugars, and ascorbate. In parallel, the accessions were genotyped with 5,995 single-nucleotide polymorphism markers spread over the whole genome. Genome-wide association analysis was conducted on a large set of metabolic traits that were stable over 2 years using a multilocus mixed model as a general method for mapping complex traits in structured populations and applied to tomato. We detected a total of 44 loci that were significantly associated with a total of 19 traits, including sucrose, ascorbate, malate, and citrate levels. These results not only provide a list of candidate loci to be functionally validated but also a powerful analytical approach for finding genetic variants that can be directly used for crop improvement and deciphering the genetic architecture of complex traits.

In crops, linkage mapping has proved invaluable for detecting quantitative trait loci (QTLs) for traits of interest and to unravel their underlying genetic architecture. This approach is based on the analysis of the segregation of polymorphism between the parental lines and their progeny. However, one of the limitations of this approach is the reduced number of recombination events that occur per generation (for review, see Korte and Farlow, 2013). This leads to extended linkage blocks that reduce the accuracy of the linkage mapping. An alternative to linkage-based mapping studies is to perform linkage disequilibrium (LD) mapping in a population of theoretically unrelated individuals. The ancestral polymorphism segregating through this population (or panel) is far more informative compared with the polymorphism of the parental lines of the linkage mapping population (Mauricio, 2001). LD mapping, also known as genome-wide association (GWA), relies on the natural patterns of LD in the population investigated. The aim of GWA is to reveal trait-associated loci by taking advantage of the level of LD. Depending on the decay of LD, the mapping resolution can be narrowed from a large genomic portion where the level of LD is relatively high to a single marker when the LD level is very low.

Following domestication, crops are prone to (1) increased levels of LD, (2) population structure (remote common ancestry of large groups of individuals), and (3) cryptic relatedness (the presence of close relatives in a sample of unrelated individuals; Riedelsheimer et al., 2012). Population structure and cryptic relatedness may lead to false-positive association in GWA studies (Astle and Balding, 2009), but their effect is now relatively well accounted for in mixed linear models (for review, see Sillanpää, 2011; Listgarten et al., 2012). The problem of high LD in GWA scans also must be taken into account: Segura et al. (2012) investigated this difficulty by proposing a multilocus mixed model (MLMM) that handles the confounding effect of background loci that may be present throughout the genome due to LD. This approach revealed multiple loci in LD and associated with sodium concentration in leaves in Arabidopsis (*Arabidopsis thaliana*), while previous methods failed to identify this complex pattern (Segura et al., 2012).

In parallel, the development of cost-effective high-throughput sequencing technologies has identified increasingly dense variant loci necessary to conduct GWA scans, especially in model species such as rice (*Oryza sativa*) for agronomic traits (Huang et al., 2010)

[1] This work was supported by the European Union (European Solanaceae grant no. PL016214–2).

* Address correspondence to christopher.sauvage@avignon.inra.fr.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Christopher Sauvage (christopher.sauvage@avignon.inra.fr).

[W] The online version of this article contains Web-only data.

www.plantphysiol.org/cgi/doi/10.1104/pp.114.241521

 © 2014 American Society of Plant Biologists. All Rights Reserved.

or maize (*Zea mays*) for drought tolerance (Lu et al., 2010; for review, see Soto-Cerda and Cloutier, 2012). However, GWA is not restricted to model species and is becoming increasingly widespread in nonmodel ones such as sunflower (*Helianthus annuus*; Mandel et al., 2013) and tomato (*Solanum lycopersicum*; Xu et al., 2013), where numerous associations have been successfully identified for traits related to plant architecture (branching in the case of sunflower) and fruit quality (e.g. fresh weight in tomato).

Tomato is a crop of particular interest, as the fruit are an important source of fiber and nutrients in the human diet and a model for the study of fruit development (Giovannoni, 2001). Over the last two decades, numerous QTLs have been identified for traits such as fresh weight using linkage approaches (Frary et al., 2000; Zhang et al., 2012; Chakrabarti et al., 2013) but also for other fruit-related traits such as fruit ascorbic acid levels (Stevens et al., 2007), sensory and instrumental quality traits (Causse et al., 2002), sugar and organic acids (Fulton et al., 2002), and metabolic components (Schauer et al., 2008). Large tomato germplasm collections have been characterized at the molecular level using simple sequence repeat (Ranc et al., 2008) and single-nucleotide polymorphism (SNP) markers (Blanca et al., 2012; Shirasawa et al., 2013), giving insights into population structure, tomato evolutionary history, and the genetic architecture of traits of agronomic interest. These screens of nucleotide diversity were made possible (for review, see Bauchet and Causse, 2012) in the last couple of years due to the release of the tomato genome sequence (Tomato Genome Consortium, 2012) and derived genomic tools such as a high-density SNP genotyping array (Sim et al., 2012). The combination of large germplasm collections, high-throughput genomic tools, and traits of economic interest provide a framework to apply genome-wide association study (GWAS) in this species. In tomato, previous association studies have been limited to a targeted region (e.g. chromosome 2; Ranc et al., 2012), used low-density genome-wide-distributed SNP markers (Xu et al., 2013), or investigated a limited number of agronomic traits with low precision on the association panel (Shirasawa et al., 2013).

Using tomato as a model, we aimed to investigate the genetic architecture of traits related to fruit metabolic composition at high resolution. To reach this objective, we carried out an investigation into LD patterns at the genome-wide scale and a GWA scan using the MLMM approach. We present results on the genetic architecture of fruit metabolic composition for metabolites such as organic acids, amino acids, sugars, and ascorbate in tomato.

RESULTS

Phenotyping

We phenotyped a panel composed of 163 accessions for a total of 76 metabolic traits, including amino acids, organic acids, and sugars. The tomato diversity panel was composed of 28 *S. lycopersicum* (S.L), 119 *S. lycopersicum* var *cerasiforme* (S.C), and 16 *Solanum pimpinellifolium* (S.P) samples derived from the previously published core collection described by Xu et al. (2013). From the set of 76 phenotypes, 36 of these (47.3%) were highly correlated over the 2 years of sampling. Of these 36 phenotypes, significant differences between the three groups of tomato accessions were identified for 26 phenotypes (70.3%; Fig. 1). The post hoc Tukey's honestly significant difference test provided a more thorough investigation of the significant differences among the three groups for each trait. Comparisons including S.P (S.P-S.L and S.P-S.C) were more significantly different than the comparison S.L-S.C (Fig. 2).

The correlation pattern revealed clusters of highly correlated compounds in the metabolic profile that largely corresponded to a functional classification of the metabolites (Fig. 1). For example, the concentration of Fru, Suc, maltitol, erythritol, and maltose clustered together with soluble solid content (SSC), while amino acids (e.g. Ser, Thr, Met, and Asn) also clustered together. We conducted GWA on this set of 36 phenotypes, which were stable (correlated) over the 2 experimentation years using the MLMM approach (for the complete phenotype data set, see Supplemental Table S1).

Genotyping

From the initial 8,784 SNPs of the SOLCAP genotyping array, 7,720 (87.8%) passed the manufacturing quality control and constituted our raw data set (see "Materials and Methods"). From this raw data set, the quality filtering gave a total of 5,995 reliable SNPs (77.6%), thus constituting the analyzed data set for GWA. The overall average percentage of missing data per locus was estimated at 3.84% in the whole population while ranging from 2.25% in S.L to 4.07% in S.P. The missing data were imputed by the most common allele of the SNP, as no missing data are allowed in the MLMM.

The minor allele frequency (MAF) values were evenly distributed from 0.001% to 0.5% and showed differences in their distribution between groups. The S.L accessions showed an excess of rare variants with a skewed distribution of the MAF values (median MAF = 0.107), while the S.C and S.P accessions showed a broader distribution of the MAF values (median MAF = 0.161 and 0.214, respectively). Such a low median MAF in the S.L accessions may be attributed to (1) a higher proportion of nearly monomorphic SNPs and (2) the shared ancestry within this group. This observation is supported by a previous study that investigated the MAF pattern in subpopulations of tomato (Sim et al., 2012).

Population Structure

The pairwise-population genetic differentiation index was estimated to be approximately 1% (0.0102) between S.L and S.C, while between S.L and S.P and

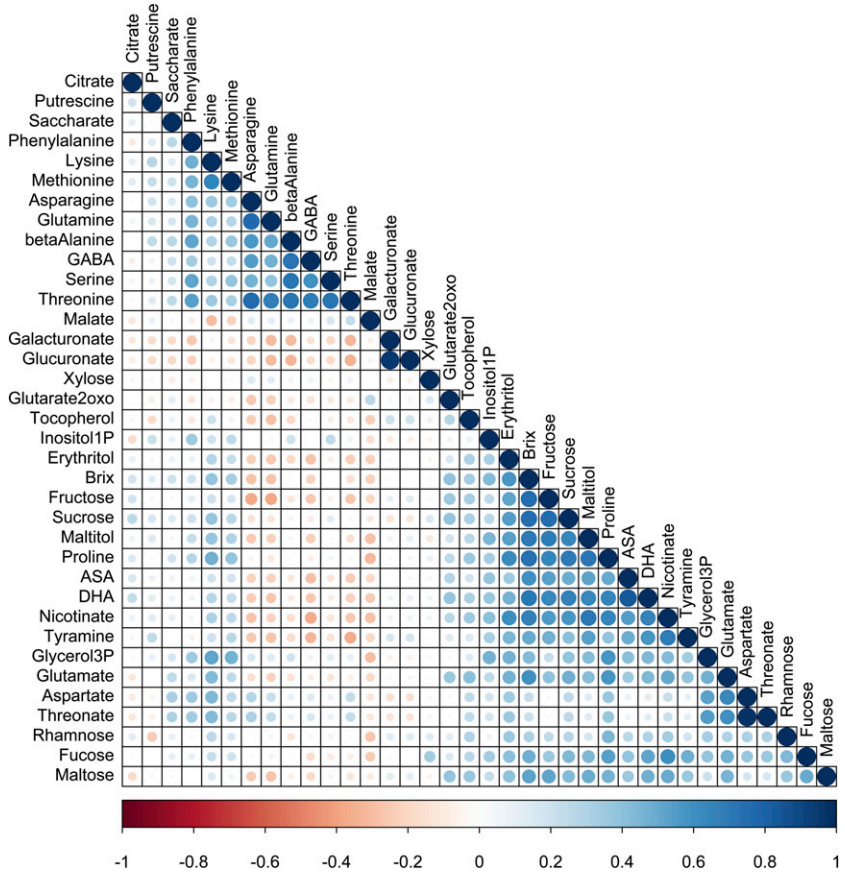

Figure 1. Lower matrix displaying the correlations between each analyzed phenotype adjusted for the year effect. The correlation coefficient (Spearman) ranges from −1 (red color) to +1 (blue color). GABA, γ-Aminobutyrate; ASA, ascorbic acid; DHA, dehydroascorbate.

between S.C and S.P, stronger population differentiation values, estimated to be 0.2132 and 0.1583, respectively, were detected. These results are supported by the estimation of the population structure using the Structure software. Following the ad hoc statistic ΔK, population structure was apparent with the number of ancestral populations estimated to be 2 (K = 2). Whereas the first group was composed of a cluster of the S.L. accessions and the S.C accessions (n = 147), the second group was composed of a cluster of the S.P accessions only (n = 16).

Estimates of Kinship and LD in the Collection

Within the 163 accessions, the pairwise kinship estimates revealed a low degree of relatedness between individuals, with a mean overall estimate of 0.0738. Pairwise LD estimates (r_s^2) within each group revealed different levels of LD along chromosomes. On average, LD was higher in S.L (r_s^2 = 0.57), medium in S.C (r_s^2 = 0.54), and lower in S.P (r_s^2 = 0.34). Within each group and for the 12 chromosomes, r_s^2 estimates ranged from 0.29 (K3) to 0.39 (K12) in S.P, from 0.5117 (K12) to 0.5619 (K11) in S.C, and from 0.52 (K9) to 0.62 (K6) in S.L. More details on LD estimates for each chromosome in the three groups by chromosome are given in Table I.

GWA

GWA was conducted trait by trait in order to dissect the optimal model obtained from the MLMM. After correcting for multiple testing, GWA scan identified a total of 44 loci that were significantly associated with 19 of the 36 traits (52.7%). These 44 loci were spread unevenly over the genome, as all chromosomes carried at least one association (chromosomes 1 and 12) but up to 10 associated loci were located on chromosome 2. Moreover, the number of associated loci per trait ranged from one (for eight traits in total) to nine (for SSC). Table II reports the detailed statistics of GWA (i.e. P value and genomic location) for the loci associated with these 19 traits. For each trait, the heritability (estimated at step 0 of the model, based on the variance component σ^2_g, computed for all markers and representing the estimated genetic variance of the trait) ranged from 0.168 (threonate level) to 0.773 (Pro level), with a median value of 0.553 (overall traits), while the missing heritability (not explained by the markers included in the model) ranged from 0.007 (Thr level) to 0.458 (nicotinate level), with a median value of 0.250. The percentage of variation explained for each trait was estimated from the optimal model obtained from the MLMM. The percentage of variation explained ranged from 16.2% to 74.3% for the Asp level and the dehydroascorbate level traits, respectively, while for

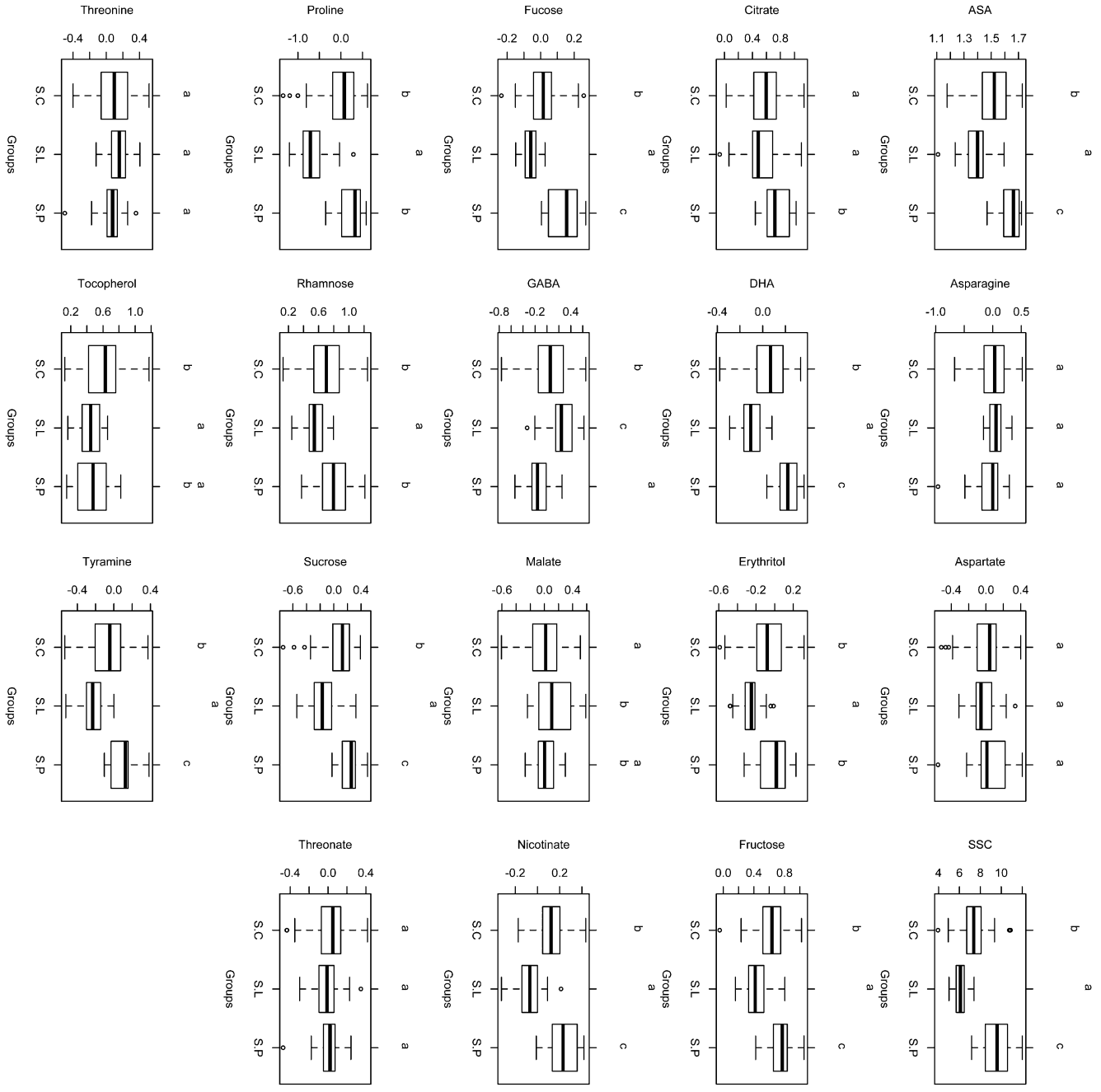

Figure 2. Box-plot representations of the distribution for the 19 traits that showed significant association. In all graphs, mean values labeled with different letters are significantly different, whereas those with the same letters are not (Tukey's test, $P < 0.05$).

the SSC, the percentage of variation was estimated as 0.611 (61%; for details, see Table III). For each trait, the Manhattan plot displaying *P* values for each locus in relation to its genomic location are shown in Supplemental Figure S1.

Finally, the peak SNP associated with SSC (SOLCAP_snp_sl_26678) that belongs to a candidate gene (Solyc09g010080.2 [*lin5*], a fruit-specific β-fructofuranosidase or invertase), which plays a role in sugar metabolism (Fridman et al., 2004), validates the methodological approach we employed by its mapping in our panel. We identified putative candidate genes in this study, especially in close proximity to four peak SNPs. For example, the peak SNP SOLCAP_snp_sl_26678 (chromosome 9, position: 2,411,368 bp) is associated with fruit ascorbate content and is located approximately 423 kb upstream of a monodehydroascorbate reductase (NADH)-like protein (MDHAR; Solyc09g009390.2,

Table I. *Intrachromosomal LD (r_s^2) in each tomato group*
This estimate takes into account the effect of population structure (Mangin et al., 2012).

Mean Pairwise r_s^2	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	K11	K12	All K
S.L	0.5508	0.5988	0.5895	0.5570	0.6029	0.6235	0.5416	0.5397	0.5231	0.5539	0.5938	0.5389	0.5678
S.C	0.5391	0.5318	0.5394	0.5191	0.5500	0.5530	0.5320	0.5337	0.5204	0.5315	0.5619	0.5117	0.5353
S.P	0.3323	0.3239	0.2884	0.3872	0.3557	0.2604	0.3478	0.3338	0.3431	0.3923	0.2917	0.3968	0.3378

position: 2,835,367 bp) shown previously to be linked to fruit ascorbate levels under stress conditions (Stevens et al., 2008). Similarly, the peak SNPs associated with nicotinate (SOLCAP_snp_sl_29349), malate (SOLCAP_snp_sl_19899), and Suc (SOLCAP_snp_sl_17956) levels are located at 680, 7.9, and 68 kb, respectively, from three putative candidate genes that play roles in the genetic architecture of the variations of these traits, which are described as a nicotinate phosphoribosyl transferase protein (Solyc02g093290.2, position: 48,771,224 bp), an aluminum-activated malate transporter-like (Solyc06g072910.2, position: 41,337,629 bp), and a sugar transporter (galactosylgalactosylxylosyl protein 3-β-glucuronosyltransferase; Solyc04g076920.2, position: 59,461,803 bp), respectively.

As a case study, we focused on the results associated with fruit malate content by compiling all the results obtained for this trait. Figures 3 and 4 illustrate these results. Malate levels were stable over the 2 years of the experiment (r^2 = 0.621; Fig. 3A), differences in malate levels were significant between groups (Figs. 2 and 3B), and the trait was normally distributed within the panel of accessions (Fig. 3C). GWA identified two significant SNPs associated with malate levels (Fig. 3D) without inflation in the distribution of P values at the optimal step of the model (Fig. 3E), suggesting that population structure was well controlled. These two SNPs explained a proportion of the trait variation estimated at 39% (Fig. 3F). For each peak SNP, located on chromosomes 2 and 6, the allelic effects of each genotypic class (homozygote and heterozygote) were estimated (Fig. 3, G and H). Finally, we used the pairwise LD estimates (r_s^2) for each genomic location to (1) narrow the genomic interval and (2) seek putative candidate genes in the vicinity of the two peak SNPs (Fig. 4), providing a local overview of the extent of LD and revealing an aluminum-activated malate transporter-like (Solyc06g072910.2, position: 41,337,629 bp) as a good candidate gene (see above).

DISCUSSION

The aims of this study were to (1) investigate LD patterns in a panel of 163 tomato accessions, including wild, admixed, and cultivated accessions, and (2) implement a stepwise GWA approach to reveal associations between SNP markers and traits related to fruit metabolites. We successfully achieved this objective with (1) the investigation of the LD patterns revealing different levels of LD along chromosomes and between the three groups constituting the panel, and (2) the detection of GWA for 19 fruit metabolic traits. Finally, we demonstrated that GWA is powerful enough to link the metabolic composition of fruits in tomato with genetic variation at a high resolution, despite a high level of LD and population structure.

Metabolite Profiling and Phenotyping of Traits

The phenotypic traits focused on in this study were measured for 2 years in a row (2007 and 2008) under similar growth conditions on an identical set of 163 tomato accessions. Only 36 traits (47.3%) were stable over the 2 years, suggesting that metabolite profiling is highly sensitive to the environmental conditions. Previous studies have reported developmental stage $\times$ genotype or environment $\times$ genotype interactions for metabolite profiles, supporting our results. For example, in tomato, metabolite profiling of 26 compounds revealed significant genotype $\times$ ripening stage interactions, whereas in durum wheat (*Triticum durum*), significant variations in metabolites were attributed to genotype $\times$ environment interactions (Beleggia et al., 2013).

Investigation of the correlations between the metabolites revealed significant relationships between traits (Fig. 1). For example, a first cluster composed of sugar-related traits (i.e. Suc and Fru) as well as ascorbate and dehydroascorbate levels were positively correlated. A second cluster of positively correlated metabolites composed of eight proteogenic amino acids could be distinguished. The traits of these two clusters were related and had significant and negative correlations. These relationships have already been shown, notably between several amino acids and sugar-related traits (i.e. Fru and Glc).

Within the set of stable traits, ANOVA revealed significant differences between the three groups of accessions (S.L, S.C, and S.P) for 25 of 37 (67.5%) of these traits (for a box-plot representation, see Fig. 2). For example, the most significant differences ($P < 1 \times 10^{-9}$) were observed for ascorbate and dehydroascorbate levels or SSC with higher levels in S.P compared with S.C and S.L. This was previously observed through the detection of QTLs related to ascorbate levels (Stevens et al., 2007) and related to SSC (Prudent et al., 2009) as well as through GWA for SSC (Xu et al., 2013).

Exploitation of the Patterns of LD

The clear population structure allowed us to estimate the patterns of intrachromosomal LD in the three

Table II. *Detailed information for the 44 significant associations detected within the 36 traits analyzed using the MLMM*

Phenotype	SNP[a]	Chromosome	SNP Position[b]	P[c]	UniRef Annotation	Locus Name (International Tomato Annotation Group 2.3)[d]
			bp			
ASA	SOLCAP_snp_sl_12749	6	36,931,366	1.42e-05	Peptide transporter, Transcription Growth Factor-β receptor, type I/II extracellular region	Solyc06g065020.2
ASA	SOLCAP_snp_sl_37057	7	63,886,939	2.94e-10	Conserved gene of unknown function	Solyc07g064580.2
ASA	SOLCAP_snp_sl_26678	9	2,411,418	1.09e-07	Repressor of silencing1	Solyc09g009080.2
ASA	SOLCAP_snp_sl_46662	9	61,773,785	1.07e-05	Gene of unknown function	Solyc09g074480.1
ASA	SOLCAP_snp_sl_62616	11	3,393,838	4.66e-08	ATP-dependent RNA helicase	Solyc11g010310.1
Asn	SOLCAP_snp_sl_32389	2	48,943,496	1.93e-07	Copine-like protein	Solyc02g093520.2
Asp	SOLCAP_snp_sl_11456	4	58,318,210	1.67e-07	Basic helix-loop-helix transcription factor	Solyc04g074810.2
SSC	SOLCAP_snp_sl_26136	2	29,851,816	7.79e-26	Man-6-P isomerase	Solyc02g063220.2
SSC	CT232_snp229	2	43,207,682	7.73e-10	UV excision DNA repair protein RAD23	Solyc02g085840.2
SSC	SOLCAP_snp_sl_63048	3	71,026	0.0006	CXE carboxylesterase	Solyc03g005100.2
SSC	SOLCAP_snp_sl_35206	6	1,748,271	2.92e-21	Auxin signaling F-box1 family protein	Solyc06g007830.1
SSC	SOLCAP_snp_sl_53288	7	60,078,938	1.22e-12	β-1,3-Galactosyl-*O*-glycosyl-glycoprotein β-1,6-*N*-acetylglucosaminyltransferase7	Solyc07g054440.2
SSC	SOLCAP_snp_sl_65072	8	59,477,446	5.57e-08	Agenet domain-containing protein	Solyc08g078530.2
SSC	SOLCAP_snp_sl_39725	9	3,477,979	1.34e-33	β-Fructofuranosidase (*lin5*)	Solyc09g010080.2
SSC	SOLCAP_snp_sl_10594	11	2,481,288	1.89e-13	Single-stranded nucleic acid-binding R3H domain protein	Solyc11g008250.1
SSC	SOLCAP_snp_sl_659	12	45,751,611	2.41e-06	Gene of unknown function	Nonavailable
Citrate	SOLCAP_snp_sl_19899	6	41,345,468	1.48e-07	Conserved gene of unknown function	Solyc06g072930.2
DHA	SOLCAP_snp_sl_69445	9	64,606,433	3.16e-39	Ubiquitin C-terminal hydrolase family protein	Solyc09g089560.2
DHA	SOLCAP_snp_sl_21770	11	3,063,738	8.49e-07	Pentatricopeptide repeat-containing protein	SGN-U564017
Erythritol	SOLCAP_snp_sl_13558	2	36,559,326	1.24e-07	Pollen allergen Chenopodium a1	Solyc02g076860.2
Erythritol	SOLCAP_snp_sl_60698	10	64,445,598	5.98e-16	Flavin oxidoreductase/NADH oxidase	Solyc10g086220.1
Fru	SOLCAP_snp_sl_16136	5	59,787,171	9.31e-07	Conserved gene of unknown function	Solyc05g050500.1
Fru	SOLCAP_snp_sl_27215	6	38,384,375	9.05e-07	Katanin p60 ATPase-containing subunit	Solyc06g066810.2
Fuc	SOLCAP_snp_sl_20802	3	60,860,146	2.70e-07	UV excision repair protein RAD23	Solyc03g117780.2
Fuc	SOLCAP_snp_sl_53149	4	53,628,534	1.63e-06	Structural constituent of ribosome	Solyc04g056530.1
GABA	SOLCAP_snp_sl_35255	6	1,330,594	5.53e-08	D-type of twin-arginine translocation DNase domain-containing DNase	Solyc06g007310.2
Malate	SOLCAP_snp_sl_6196	2	13,905,175	1.28e-06	Gene of unknown function	SGN-U565892
Malate	SOLCAP_snp_sl_19899	6	41,345,468	2.48e-08	Conserved gene of unknown function	Solyc06g072930.2
Nicotinate	SOLCAP_snp_sl_29349	2	49,451,582	3.83e-06	Uridylyltransferase PII	Solyc02g094300.2
Pro	SOLCAP_snp_sl_100675	2	28,798,838	3.71e-06	Conserved gene of unknown function	Nonavailable
Pro	SOLCAP_snp_sl_32499	6	21,807,134	3.91e-07	Membrane-associated progesterone receptor component1	Solyc06g035870.2
Rha	SOLCAP_snp_sl_40309	1	84,253,735	2.61e-08	Embryo-specific3	SGN-U565850
Rha	SOLCAP_snp_sl_34196	3	59,102,190	2.32e-09	Conserved gene of unknown function	Solyc03g115250.2
Rha	SOLCAP_snp_sl_56631	8	1,403,227	9.41e-06	Patatin1-Kuras2	Solyc08g006860.2
Rha	SOLCAP_snp_sl_39722	9	3,484,890	2.10e-10	Gene of unknown function	SGN-U565153
Suc	SOLCAP_snp_sl_13549	2	36,490,995	2.57e-06	Conserved gene of unknown function	Solyc02g076800.1
Suc	SOLCAP_snp_sl_17956	4	59,392,982	6.01e-05	Glutamyl-tRNA reductase	Solyc04g076870.2
Suc	SOLCAP_snp_sl_29483	5	4,037,126	9.51e-09	Glycosyltransferase family GT8 protein	Solyc05g009820.2
Threonate	SOLCAP_snp_sl_11456	4	58,318,160	5.73e-06	Basic helix-loop-helix transcription factor	Solyc04g074810.2
Thr	SOLCAP_snp_sl_32389	2	48,943,446	3.75e-07	Copine-like protein	Solyc02g093520.2
Tocopherol	SOLCAP_snp_sl_46445	10	2,199,297	4.35e-07	Conserved gene of unknown function	Solyc10g008030.2
Tyramine	SOLCAP_snp_sl_14531	8	2,587,919	1.12e-05	Conserved gene of unknown function	Solyc08g008120.2
Tyramine	SOLCAP_snp_sl_64706	8	57,571,484	1.18e-07	Lys-specific demethylase5A	Solyc08g076390.2
Tyramine	SOLCAP_snp_sl_36166	11	762,353	1.54e-06	Transcription regulator	SGN-U275742

[a]SNP names as given in the SOLCAP SNP array (http://solcap.msu.edu). [b]SNP genomic position on the tomato reference genome (version 2.40). [c]SNP P values. [d]Name of the locus to which the peak SNP belongs (according to the tomato genome annotation version 2.30).

groups of accessions (S.L, S.C, and S.P) using a total of 5,995 genome-wide markers. Our analysis of LD revealed considerable variation across the tomato genome in the populations investigated. The same observation was made in two previous studies that investigated LD patterns in tomato on a genome-wide

Table III. *Summary of trait associations showing the heritability of the trait (h^2; step 0 in the MLMM), the missing heritability (h^2 at the optimal model), the percentage of associated variation of the trait (PVE), and the number of significant loci associated with the trait variation*

Phenotype	Trait h^2	Missing h^2	PVE	No. of Associations
ASA	0.553	0.333	0.561	5
Asn	0.417	0.208	0.220	1
Asp	0.284	0.301	0.162	1
Brix	0.600	0.185	0.611	9
Citrate	0.423	0.299	0.181	1
DHA	0.595	0.192	0.743	2
Erythritol	0.534	0.286	0.358	2
Fru	0.565	0.250	0.386	2
Fuc	0.415	0.365	0.481	2
GABA	0.415	0.184	0.237	1
Malate	0.642	0.182	0.390	2
Nicotinate	0.595	0.458	0.279	1
Pro	0.773	0.282	0.461	2
Rha	0.579	0.195	0.504	4
Suc	0.585	0.220	0.439	3
Threonate	0.168	0.174	0.170	1
Thr	0.348	0.007	0.187	1
Tocopherol	0.306	0.261	0.224	1
Tyramine	0.612	0.347	0.472	3
Minimum	0.168	0.007	0.162	1
Maximum	0.773	0.458	0.743	9
Median	0.553	0.250	0.386	2

scale. Similar average r^2 estimates (r^2 = 0.464) were obtained in fresh market tomato populations, reflecting the effects of selection on genome variation and the breeding history of tomato toward market specialization (Robbins et al., 2011). However, a second study highlighted a biased decay of LD between euchromatic and heterochromatic regions (Shirasawa et al., 2013). These previous studies support the high level of LD identified in our study. The different levels of LD may be interpreted as a direct effect of the domestication that tomato (especially in S.L) has undergone during its history, through bottlenecks and selective breeding, that has led to a reduction in nucleotide diversity and an extended LD following the elimination of recombinant lineages (Hamblin et al., 2011).

In cultivated tomato, LD decays over large genomic regions (i.e. several hundreds of kb up to several Mb) and is advantageous for an association mapping approach, as it requires fewer markers to cover the entire genome. On the other hand, the difficulty in identifying the underlying causal polymorphism responsible for the phenotypic variation represents the main drawback of these large blocks of LD. Identifying the causal polymorphism from GWA signals remains challenging, especially in species where dense genome coverage is still not achieved. To overcome these limitations, the MLMM proposed by Segura et al. (2012) handles the confounding effect of background loci due to LD at the GWA scan step. This approach outperformed the existing mixed linear models, notably by reducing the number of significantly associated SNPs rather than the number of peaks. This reduced the number of candidate loci it was necessary to screen in order to identify the causal polymorphism.

GWA for Metabolic Traits

The GWA scan revealed a total of 44 loci (or peak SNPs) associated with the variation of 19 traits. These 44 loci accounted for various levels of estimated trait heritability (from 0.168 to 0.773), missing heritability (from 0.007 to 0.458), and percentage of trait variation (from 16.2% to 74.3%). These results suggest that different traits have different genetic architectures: in some cases, a few genes may explain a large proportion of the phenotypic variation (i.e. two loci explain 74.3% of the variation in fruit dehydroascorbate), while numerous genes may only explain a fraction of the phenotypic variation (i.e. five loci explain 33.2% of the ascorbate level variation). These results are supported by similar observations in rice, where various genetic architectures were revealed using a GWAS approach for different traits of agronomic interest (Zhao et al., 2011). However, it should be noted that in most GWAS, significantly associated loci might contribute to a larger proportion of phenotypic variation, as many other small- to medium-effect loci were not detected due to the stringent threshold used in GWAS (false discovery rate [FDR]) and the lack of statistical significance for the control of false negatives caused by small effect sizes (Visscher et al., 2012). Furthermore, the estimates of the missing heritability suggest that for some traits, most (or nearly all) loci underlying the variation in these traits have been identified through the genome scan we conducted. For example, for Thr, the missing heritability has been estimated to be 0.007, which means that (1) all the loci responsible for the variation of this trait may have been identified, and (2) the genetic architecture of this trait may rely on a small number of genes (only one associated locus in our study). Conversely, for nicotinate, the genetic architecture of the trait requires further investigation, since the missing heritability has been estimated at 0.458, which means that a large number of small-effect loci or a limited number of large-effect loci remain to be identified. Taken together, these observations suggest that the investigated genetic architecture is usually more complex than it appears.

A total of 35 of the 44 associated loci (79.5%) were associated with the metabolic traits (nine loci are associated with SSC) and accounted for between 16.2% (Asp) and 50.4% (Rha) of the variation of these traits. In a previous investigation of QTLs related to metabolic traits using a lower number of lines (n = 76), Schauer et al. (2008) detected 104 metabolite QTLs for 22 distinct amino acids in tomato. Our results obtained using a GWAS approach contrast with these results in terms of the number of QTLs. However, this difference may reflect the methodological principles underlying both approaches. The more stringent threshold used in GWAS (i.e. FDR) and the confounding effect of

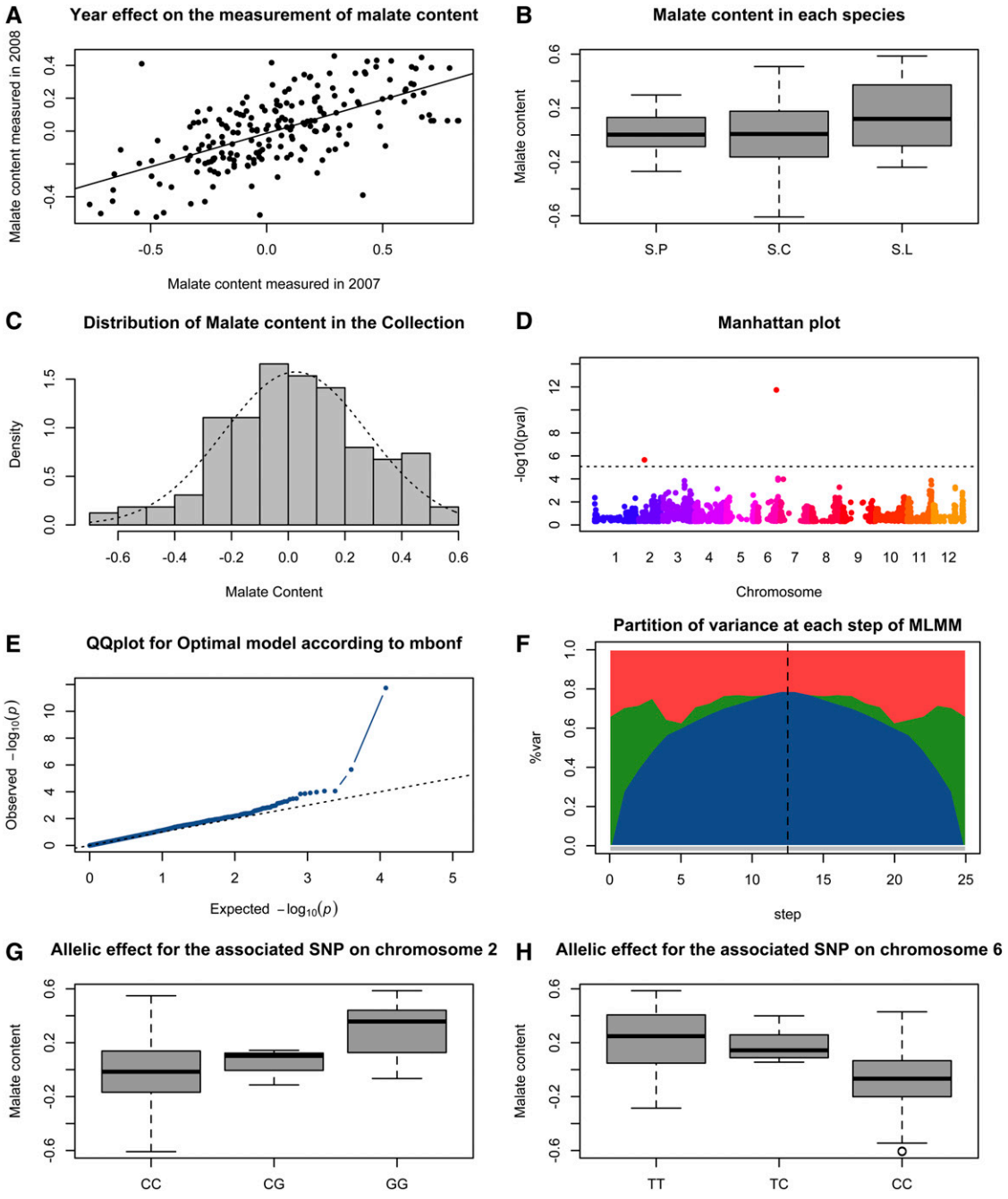

Figure 3. A focus on malate level results. A, Correlation for the malate level over the 2 years of sampling in the collection of 163 accessions. B, Variation of malate level adjusted for the year effect within the three groups. C, Distribution of the adjusted malate level in the collection. D, Manhattan plot for the 12 tomato chromosomes (*x* axis) and associated *P* values for each marker (*y* axis). E, QQplots of the observed *P* value distribution. F, Evolution of genetic variance at each step of the MLMM (blue, genetic variance explained; green, total genetic variance; red, error) for the optimal model (step indicates extended Bayesian information criterion). G and H, Allelic effect for the two associated markers on chromosomes 2 and 6.

population structure may explain this difference. This has been observed in a study of glucosinolate metabolites in Arabidopsis (Chan et al., 2010) and a study of leaf metabolic profiles in maize (Riedelsheimer et al., 2012). In the latter study, when comparing a linkage mapping experiment and a GWA scan, increased genetic variation was reported, suggesting that the genetic variability is greater in the GWAS, as it relies on a larger genetic pool (from several up to hundreds of individuals), whereas a linkage experiment relies on a much narrower genetic pool (i.e. a couple of parental lines; Riedelsheimer et al., 2012).

The 44 associated loci are spread over the tomato genome, as every chromosome carries at least one association (chromosomes 1 and 12), with up to 10 on chromosome 2. In tomato, chromosome 2 was suggested to be interesting, as it carries a lot of QTLs for traits of interest such as fresh weight (*fw2.2*; Frary et al., 2000), fruit morphology (Causse et al., 2002), and locule number (*lcn2.1*; Muños et al., 2011), to name a few (Ranc et al., 2012; Xu et al., 2013), supporting the identification of numerous associations on this chromosome in our study.

In several cases, peak SNPs of different traits colocalized or the same peak SNP was associated with different traits. For example, peak SNPs for Suc and erythritol levels (SOLCAP_snp_sl_12459 and SOLCAP_snp_sl_13558) colocalized in a region of 68 kb on chromosome 2. In a similar way, peak SNPs associated with Asn, Thr, and nicotinate levels were localized within an interval of 508.1 kb on chromosome 2 (SOLCAP_snp_sl_32389 and SOLCAP_snp_sl_29349, respectively). Such colocalization of peak SNPs has been observed in other GWAS experiments in tomato (Xu et al., 2013), Arabidopsis (Bergelson and Roux, 2010), and rice (Zhao et al., 2011), suggesting the presence of genes with pleiotropic effects or closely linked genes.

On chromosome 2, both Asn and Thr level traits are associated with the same peak SNP (SOLCAP_snp_sl_32389, annotated as a copine-like protein). These two α-amino acids belong to the class of polar uncharged side chain amino acids and are indirectly linked to the Krebs cycle, as their biosynthesis relies on oxaloacetate, which, as an acceptor compound of this cycle, is one of its major metabolic intermediates. The identification of an association between Asn and Thr levels at the same peak SNP means that this locus is located in close proximity to one or several crucial and pleiotropic effect gene(s) directly involved in the metabolic pathway of both Asn and Thr synthesis. This observation suggests that the genomic region around the peak SNP (SOLCAP_snp_sl_32389) has to be investigated further to seek for causal polymorphisms and candidate genes underlying the genetic architecture of the Asn and Thr level traits.

Malate and citrate levels were associated with one peak SNP located on chromosome 6 (SOLCAP_snp_sl_19899) and two peak SNPs located on chromosomes 2 and 6 (SOLCAP_snp_sl_6196 and SOLCAP_snp_sl_19899), respectively (for detailed Manhattan plots and LD patterns, see Fig. 4). Interestingly, malate and citrate were associated with the same peak SNP (SOLCAP_snp_sl_19899), located on chromosome 6 and annotated as a conserved gene of unknown function. This observation suggests either that, in this genomic region, the LD block that this peak SNP belongs to is particularly extended or, alternatively, this peak SNP was identified close to a gene involved in the citrate and malate metabolic pathways. Figure 4B, representing the local pattern of LD around the peak SNP, suggests that the LD level is relatively

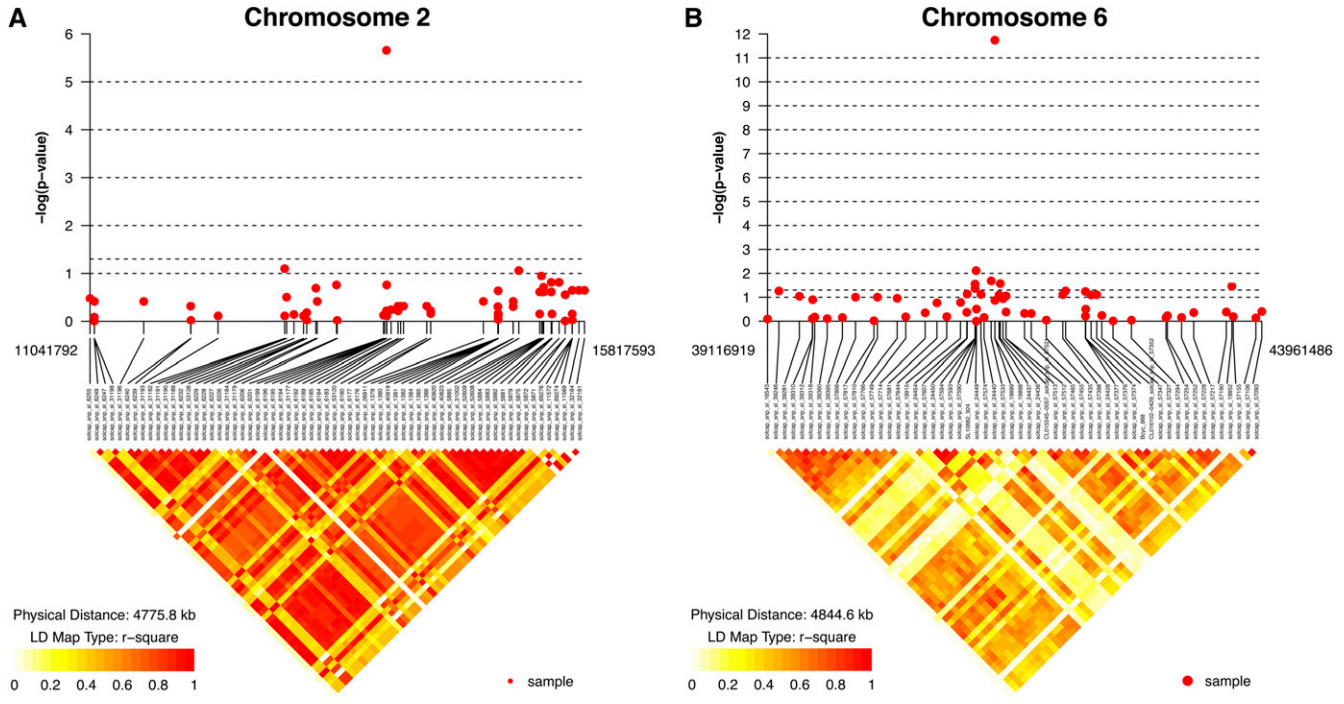

Figure 4. Manhattan plots displaying the $-\log_{10}$ *P* values (*y* axis) over genomic positions (*x* axis) in a window of 2.5 Mb upstream and downstream of the two loci associated with the malate level trait that are located on chromosome 2 (A) and chromosome 6 (B). Different colors are used to represent the pairwise LD estimates (r_s^2) for each genomic location.

low in this genomic region. From a functional point of view, citrate and malate are two organic compounds found in most ripe fruits (Etienne et al., 2013) and have been demonstrated to be highly correlated with many important regulators of ripening in studies that have investigated early fruit development (Mounet et al., 2009; Centeno et al., 2011). This result suggests that, in this study, we were able to identify a peak SNP that is located near one or several putative candidate gene(s) playing a crucial role in the citrate and malate metabolic pathways. Moreover, for the malate level trait, the phenotype of heterozygous individuals is intermediate to that of homozygotes (Fig. 3), suggesting an incomplete dominance effect.

Searching for Candidate Genes

In this study, we conducted GWA using an MLMM to identify more precisely putative candidate genes involved in the genetic architecture of fruit metabolic traits by taking advantage, notably, of the LD pattern. For numerous peak SNPs, their functional annotation is not directly linked to the trait they are associated with. However, for some of these peak SNPs, they directly target a previously characterized candidate gene or are located in close proximity to putative candidate genes. For example, the peak SNP associated with fruit SSC (SOLCAP_snp_sl_26678) belongs to a previously validated candidate gene (Solyc09g010080.2, *lin5*) that encodes a cell wall invertase and is a locus for a QTL that positively affects tomato fruit sugar content; hence, the important soluble solids commercial trait (Schauer et al., 2006). However, more putative candidate genes have been identified in this study. For example, the peak SNP (SOLCAP_snp_sl_26678) located on chromosome 9 (position: 2,411,368 bp) and associated with fruit ascorbate levels is located near (423 kb) a monodehydroascorbate reductase (NADH)-like protein (MDHAR; Solyc09g009390.2, position: 2,835,367 bp) that has been identified previously using a QTL fine-mapping approach (Stevens et al., 2008). Similarly, the peak SNPs associated with nicotinate, malate, and Suc levels (SOLCAP_snp_sl_29349, SOLCAP_snp_sl_19899, and SOLCAP_snp_sl_17956, respectively) are also located near (680, 7.9, and 68 kb, respectively) putative candidate genes that play roles in the genetic architecture of the variation of these traits. Indeed, these three putative candidate genes are described as a nicotinate phosphoribosyl transferase protein (Solyc02g093290.2, position: 48,771,224 bp), an aluminum-activated malate transporter-like (Solyc06g072910.2, position: 41,337,629 bp), and a sugar transporter (galactosylgalactosylxylosyl protein 3-β-glucuronosyltransferase; Solyc04g076920.2, position: 59,461,803 bp), respectively. Thus, these results open the door for subsequent analyses based on either fine localization of the putative candidate gene using a targeted resequencing approach combined with GWAS to identify and confirm the causal polymorphism or functional validation, for example by transgenic approaches to investigate the biological role of the putative candidate gene (e.g. fine-mapping of the *fw3.2* locus; Chakrabarti et al., 2013).

CONCLUSION

These results show that high-resolution GWA, by using an MLMM, has been successful in tomato in deciphering the genetic architecture of fruit composition traits. This led to the identification of promising candidate loci that underlie the genetic architecture of traits such as fruit malate and citrate levels, opening the door to further validation and functional investigation of this locus. The next analytical step will rely on the integration of recent methodological developments such as data imputation (Marchini and Howie, 2010; Howie et al., 2012; Porcu et al., 2013) and haplotype-based models (Powell et al., 2012) and should facilitate the identification of novel loci with a higher degree of accuracy.

MATERIALS AND METHODS

Plant Material

The tomato diversity panel consisted of 163 accessions composed of 28 *Solanum lycopersicum*, 119 *S. lycopersicum* var *cerasiforme*, and 16 *Solanum pimpinellifolium* samples derived from the previously published core collection described by Xu et al. (2013). Cherry-type tomato (S.C) is an admixture between tomato (S.L) and its closest wild relative (S.P), possibly resulting from the frequent hybridizations between them (Nesbitt and Tanksley, 2002; Ranc et al., 2008). In GWA experiments, the power to detect genetic effects is linked to MAF at genotyped loci. The mixing of different groups, populations, or subspecies within a panel will enhance the efficiency of the approach by capturing rare and common variants. Using a diversity panel composed of several subpopulations or species is a common practice in GWA experiments. For example, in rice (*Oryza sativa*), Zhao et al. (2011) used a worldwide diversity panel composed of five different species to unravel the complex genetics underlying the natural variation of 34 traits in rice. In this study, we mixed accessions of S.L, S.C, and S.P in order to (1) cover the broader range of phenotypic and genetic diversity, as we did not expect a uniform distribution of SNP MAF within each group, and (2) overcome the limited statistical power of GWA due to the skewed distribution of SNP MAF within the panel, especially in terms of the detection of false-positive associations (Tabangin et al., 2009).

Plants (four replicates) were grown in a tunnel in Avignon, France, during the summers of 2007 and 2008 (growth conditions are also described in Xu et al., 2013). Fruits were harvested at the ripe stage. Pericarp tissue from five fruits per accession was collected to be frozen in liquid nitrogen and stored at −80°C for metabolomic profiling. DNA was isolated from 100 mg of frozen leaves using to the DNeasy Plant Mini Kit (Qiagen) for the subsequent genotyping assay. Leaf samples corresponded to fully expanded but nonsenescent leaves. DNA was quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen) according to the manufacturer's protocol.

Phenotypic Variation

Metabolite Profiling

Tomato pericarp tissue was homogenized, and the exact amount used for metabolite extraction was defined. Three biological replicates were analyzed for each accession. Metabolite extraction, derivatization, gas chromatography-mass spectrometry, and data processing were performed as described by Schauer et al. (2006). Metabolites were identified in comparison with database entries of authentic standards (Kopka et al., 2005; Schauer et al., 2005). A total of 76 metabolites were measured in our experiment (for the complete list, see

Supplemental Table S1), including amino acids, sugars and sugar alcohols, and organic acids.

Ascorbic Acid Level

Ascorbate and dehydroascorbate were measured separately from the metabolite profiling using a microplate assay as described by Stevens et al. (2006) on the frozen pericarp material stored at −80°C. Extractions and assays were carried out in ice-cold 6% (w/v) TCA in triplicate. The assay used was a spectrophotometric assay based on the detection of dipyridyl-Fe^{2+} complexes following the reduction of Fe^{3+} to Fe^{2+} by the reduced form of ascorbate present in the samples and comparison with standards of known concentrations. Total ascorbate (reduced ascorbate + dehydroascorbate) was measured by mixing the sample with 5 mM dithiothreitol, to reduce dehydroascorbate, prior to the assay. Dehydroascorbate concentrations, therefore, were calculated to be the difference between the samples with and without dithiothreitol added.

SSC and Sugars

The concentrations of Fru, Glc, and Suc were determined within the 163 accessions for the 2 years of sampling using the micromethod developed by Gomez et al. (2007). This method is precise, linear, and accurate when compared with HPLC methods. Measurement of SSC in degrees Brix was performed as described by Xu et al. (2013) on fruit frozen powder derived from blending fruits with liquid N_2. SSC values primarily represent estimates of sugar content in fruits and vegetables.

Data Normalization and Statistical Analyses

All descriptive statistics and analyses were performed using R 2.15.1 except as otherwise specified. A nonparametric Kendall test was used to assess agreement among the biological replicates and to remove any outlier measurements. Thus, for each year, biological replicates were averaged and the normal distribution of the data was tested using a Shapiro-Wilk test. The normality test revealed that 29 of the 76 phenotypes (38.1%) were not normally distributed and were $\log_{10}$ transformed. For each phenotype, a linear regression revealed correlation between the 2 years. Only highly correlated phenotypes ($r^2 > 0.6$) between the 2 years of sampling were averaged and used in the GWAS. The pairwise correlation between phenotypes was evaluated using a Spearman test ($P < 0.05$). The graphic representation of the pairwise correlation between the phenotypes was produced using the R package Corrplot and the hclust clustering method (Friendly, 2002). Finally, for each phenotype, an ANOVA tested significant differences between the groups of accessions (S.L, S.C, and S.P). Then, a post-ANOVA Tukey's honestly significant difference test created a set of confidence intervals on the differences between the means of each trait, for which a significant association was detected, to test for significant differences between the pairwise means among the three groups of the panel (i.e. significant difference for the mean level of malate between S.L and S.P). Significance was declared at $P < 0.05$. Finally, the individual missing phenotype data, which ranged from 0.033% to 12.7% (median of 0.32%), were replaced by the mean value of the trait computed for all the accessions of the panel, as required by the MLMM.

Genotyping Array and SNP Selection

The SNP genotyping was performed using the Infinium assay (Illumina), developed by the Solanaceae Coordinated Agricultural Project (Hamilton et al., 2012; Sim et al., 2012) to genotype the collection of tomato accessions (according to the manufacturer's standard protocol). The probe sequences and SNP information are available from the Solanaceae Coordinated Agricultural Project (http://solcap.msu.edu). The SNP calling rate threshold per locus was set at 90%. A MAF ranging from 0.037 to 0.45 was used to filter the raw genotype data set. The minimal MAF was set according to the formula [number of chromosomes/(2 × number of individuals)], as proposed by Aulchenko et al. (2007). The minimal success rate of genotyping per accession was fixed to 90%. All SNPs and accessions that did not respect these criteria were removed using the –maf option implemented in Plink! (Purcell et al., 2007).

Estimation of Population Differentiation and Structure

Initially, the fixation index (Weir and Cockerham, 1984) estimation was performed between the three groups of tomato accessions to get an overview of the population structure. Then, the Structure software 2.3.3 (Pritchard et al., 2000; Falush et al., 2003) was used to infer the number of ancestral populations based on the filtered SNP data set (for the number of sites used, see "Results") and thus to assign the 163 individuals to populations (Q matrix). The most likely number of clusters K in all simulations was assumed to be in the range of K = 1 to K = 10. Ten replicates were conducted for each K with a burn-in period of 1×10^6, followed by 5×10^6 MCMC steps using the Bioportal computing resource (http://www.mn.uio.no/ibv/bioportal/index.html; Kumar et al., 2009). These parameters met the requirements for the use of the Structure software proposed by Gilbert et al. (2012) to ensure the reproducibility of the results of this study. The ad hoc statistic ΔK was used to determine the most probable K (Evanno et al., 2005). The ancestry estimation using Admixture software (Alexander et al., 2009), based on the maximum likelihood estimation of individual ancestries from multilocus SNP genotype data sets, was used to support the identification of the ancestral populations performed with the Structure software.

Kinship and LD Estimation

SPAGeDi software (Hardy and Vekemans, 2002) was used to estimate the Ritland (1995) matrix of pairwise kinship coefficient (K matrix) from the filtered SNP data set (see "Results") using a 10,000 bootstrap resampling procedure. Then, the intrachromosomal LD between all pairs of sites was estimated using an unbiased (as individuals are not independent) estimation (named r_s^2) that uses the population structure matrix and consisting of information about the origins of each individual and the admixture proportions of each individual genome. The method is implemented in the R package called LDcorSV (Mangin et al., 2012). Finally, the snp.plotter R package gave a graphic representation of the pairwise LD estimates at a local scale (http://cran.r-project.org/web/packages/snp.plotter/index.html).

GWA Mapping

GWA analyses were performed with correction for population structure (Q) and modeling phenotypic covariance with the kinship (K) matrix. QQplots was used to determine the most appropriate correcting method for each analyzed phenotype. Thus, these matrices were implemented into a modified version of the MLMM described by Segura et al. (2012) that takes into account the population structure as a cofactor (see the mlmm_cof.r R script at https://cynin.gmi.oeaw.ac.at/home/resources/mlmm). Briefly, the MLMM is based on the Emma library (Kang et al., 2008). The approach relies on a simple, stepwise mixed-model regression with forward inclusion and backward elimination while reestimating the variance components of the model at each step. This method increases the detection power and reduces the FDR when compared with traditional single-locus approaches. Two model selection criteria are implemented in MLMM for multitesting correction: the extended Bayesian information criterion (Chen and Chen, 2008) and the multiple-Bonferroni criterion, defined as the largest model in which all cofactors have P values below a Bonferroni-corrected threshold (we used a threshold of 0.05; for details, see Segura et al., 2012). From the optimal model obtained with MLMM (according to both the extended Bayesian information criterion and the multiple-Bonferroni criterion), the percentage of variation explained was obtained for each phenotype. Moreover, for each trait, the phenotypic heritability was obtained at step 0 in the MLMM, when no marker is included in the model, whereas the missing heritability (the percentage of the variance not explained by the markers) was obtained at the optimal step of the MLMM. Briefly, the MLMM partitions the phenotypic variance into genetic, random, and explained variance at each step, suggesting a natural stopping criterion (genetic variance of 0) for including cofactors. This estimates the explained and unexplained heritable variance of the analyzed trait.

Data Availability

To ensure their accessibility, the genotyping (5,995 SNPs) and phenotyping (n = 36) data, as well as the structure and kinship matrices for the 163 accessions, were deposited on the GNPis repository hosted at https://urgi.versailles.inra.fr/association (Steinbach et al., 2013). The complete phenotype data set is also available in Supplemental Table S1.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure S1. Manhattan plots for all the studied traits.

Supplemental Table S1. Phenotypic and genotypic data used to perform the GWA.

ACKNOWLEDGMENTS

We thank the associate editor, Jocelyn Rose, and two anonymous reviewers for critical input. We thank Hélène Burck and Yolande Carretero (Institut National de la Recherche Agronomique, UR1052 research unit) for invaluable help, characterizing and maintaining the Institut National de la Recherche Agronomique tomato Genetic Resources collection, and the Experimental Installation Unit of the Institut National de la Recherche Agronomique UR1052 research unit for maintaining the facilities. We also thank Sophie Rolland (Institut National de la Recherche Agronomique, Rennes), Nicolas Ranc (Syngenta Seeds), and Stéphane Munos (Institut National de la Recherche Agronomique, Toulouse) for early involvement in the experiment, Gisèle Riqueau and Renaud Duboscq (Institut National de la Recherche Agronomique, UR1052) for DNA and tissue sampling and ascorbate phenotyping, and Martin Ganal (TraitGenetics) for help with SNP array genotyping.

Received April 16, 2014; accepted May 31, 2014; published June 3, 2014.

LITERATURE CITED

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res **19:** 1655–1664

Astle W, Balding DJ (2009) Population structure and cryptic relatedness in genetic association studies. Stat Sci **24:** 451–471

Aulchenko YS, Ripke S, Isaacs A, van Duijn CM (2007) GenABEL: an R library for genome-wide association analysis. Bioinformatics **23:** 1294–1296

Bauchet G, Causse M (2012) Genetic diversity in tomato (*Solanum lycopersicum*) and its wild relatives. *In* C Mahmut, ed, Genetic Diversity in Plants. InTech http://www.intechopen.com/books/genetic-diversity-in-plants/genetic-diversity-in-tomato-solanum-lycopersicum-and-its-wild-relatives (March 23, 2013)

Beleggia R, Platani C, Nigro F, De Vita P, Cattivelli L, Papa R (2013) Effect of genotype, environment and genotype-by-environment interaction on metabolite profiling in durum wheat (*Triticum durum* Desf.) grain. J Cereal Sci **57:** 183–192

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

Blanca J, Cañizares J, Cordero L, Pascual L, Diez MJ, Nuez F (2012) Variation revealed by SNP genotyping and morphology provides insight into the origin of the tomato. PLoS ONE **7:** e48198

Causse M, Saliba-Colombani V, Lecomte L, Duffé 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

Centeno DC, Osorio S, Nunes-Nesi A, Bertolo AL, Carneiro RT, Araújo WL, Steinhauser MC, Michalska J, Rohrmann J, Geigenberger P, et al (2011) Malate plays a crucial role in starch metabolism, ripening, and soluble solid content of tomato fruit and affects postharvest softening. Plant Cell **23:** 162–184

Chakrabarti M, Zhang N, Sauvage C, Muños S, Blanca J, Cañizares J, Diez MJ, Schneider R, Mazourek M, McClead J, et al (2013) A cytochrome P450 regulates a domestication trait in cultivated tomato. Proc Natl Acad Sci USA **110:** 17125–17130

Chan EKF, Rowe HC, Kliebenstein DJ (2010) Understanding the evolution of defense metabolites in *Arabidopsis thaliana* using genome-wide association mapping. Genetics **185:** 991–1007

Chen J, Chen Z (2008) Extended Bayesian information criteria for model selection with large model spaces. Biometrika **95:** 759–771

Etienne A, Génard M, Lobit P, Mbeguié-A-Mbéguié D, Bugaud C (2013) What controls fleshy fruit acidity? A review of malate and citrate accumulation in fruit cells. J Exp Bot **64:** 1451–1469

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

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics **164:** 1567–1587

Frary A, Nesbitt TC, Grandillo S, Knaap E, Cong B, Liu J, 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, Carrari F, Liu YS, Fernie AR, Zamir D (2004) Zooming in on a quantitative trait for tomato yield using interspecific introgressions. Science **305:** 1786–1789

Friendly M (2002) Corrgrams: exploratory displays for correlation matrices. Am Stat **56:** 316–324

Fulton TM, Bucheli P, Voirol E, Lopez J, Petiard V, Tanksley SD (2002) Quantitative trait loci (QTL) affecting sugars, organic acids and other biochemical properties possibly contributing to flavor, identified in four advanced backcross populations of tomato. Euphytica **127:** 163–177

Gilbert KJ, Andrew RL, Bock DG, Franklin MT, Kane NC, Moore JS, Moyers BT, Renaut S, Rennison DJ, Veen T, et al (2012) Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program STRUCTURE. Mol Ecol **21:** 4925–4930

Giovannoni J (2001) Molecular biology of fruit maturation and ripening. Annu Rev Plant Physiol Plant Mol Biol **52:** 725–749

Gomez L, Bancel D, Rubio E, Vercambre G (2007) The microplate reader: an efficient tool for the separate enzymatic analysis of sugars in plant tissues. Validation of a micro-method. J Sci Food Agric **87:** 1893–1905

Hamblin MT, Buckler ES, Jannink JL (2011) Population genetics of genomics-based crop improvement methods. Trends Genet **27:** 98–106

Hamilton JP, Sim SC, Stoffel K, Van Deynze A, Buell CR, Francis DM (2012) Single nucleotide polymorphism discovery in cultivated tomato via sequencing by synthesis. Plant Genome **5:** 17–29

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:** 612–620

Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet **44:** 955–959

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, et al (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet **42:** 961–967

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E (2008) Efficient control of population structure in model organism association mapping. Genetics **178:** 1709–1723

Kopka J, Schauer N, Krueger S, Birkemeyer C, Usadel B, Bergmüller E, Dörmann P, Weckwerth W, Gibon Y, Stitt M, et al (2005) GMD@CSB. DB: the Golm Metabolome Database. Bioinformatics **21:** 1635–1638

Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. Plant Methods **9:** 29

Kumar S, Skjaeveland A, Orr RJ, Enger P, Ruden T, Mevik BH, Burki F, Botnen A, Shalchian-Tabrizi K (2009) AIR: a batch-oriented web program package for construction of supermatrices ready for phylogenomic analyses. BMC Bioinformatics **10:** 357

Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D (2012) Improved linear mixed models for genome-wide association studies. Nat Methods **9:** 525–526

Lu Y, Zhang S, Shah T, Xie C, Hao Z, Li X, Farkhari M, Ribaut JM, Cao M, Rong T, et al (2010) Joint linkage-linkage disequilibrium mapping is a powerful approach to detecting quantitative trait loci underlying drought tolerance in maize. Proc Natl Acad Sci USA **107:** 19585–19590

Mandel JR, Nambeesan S, Bowers JE, Marek LF, Ebert D, Rieseberg LH, Knapp SJ, Burke JM (2013) Association mapping and the genomic consequences of selection in sunflower. PLoS Genet **9:** e1003378

Mangin B, Siberchicot A, Nicolas S, Doligez A, This P, Cierco-Ayrolles C (2012) Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. Heredity (Edinb) **108:** 285–291

Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet **11:** 499–511

Mauricio R (2001) Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology. Nat Rev Genet **2:** 370–381

Mounet F, Moing A, Garcia V, Petit J, Maucourt M, Deborde C, Bernillon S, Le Gall G, Colquhoun I, Defernez M, et al (2009) Gene and metabolite regulatory network analysis of early developing fruit tissues highlights new candidate genes for the control of tomato fruit composition and development. Plant Physiol **149:** 1505–1528

Muños S, Ranc N, Botton E, Bérard A, Rolland S, Duffé P, Carretero Y, Le Paslier MC, Delalande C, Bouzayen M, et al (2011) Increase in tomato locule number is controlled by two single-nucleotide polymorphisms located near *WUSCHEL*. Plant Physiol **156:** 2244–2254

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

Porcu E, Sanna S, Fuchsberger C, Fritsche LG (2013) Genotype imputation in genome-wide association studies. Curr Protoc Hum Genet **Chapter 1:** Unit 1.25

Powell JE, Kranis A, Floyd J, Dekkers JCM, Knott S, Haley CS (2012) Optimal use of regression models in genome-wide association studies. Anim Genet **43:** 133–143

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

Prudent M, Causse M, Génard M, Tripodi P, Grandillo S, Bertin N (2009) Genetic and physiological analysis of tomato fruit weight and composition: influence of carbon availability on QTL detection. J Exp Bot **60:** 923–937

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet **81:** 559–575

Ranc N, Muños 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, Munos 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*. G3 **2:** 853–864

Riedelsheimer C, Lisec J, Czedik-Eysenberg A, Sulpice R, Flis A, Grieder C, Altmann T, Stitt M, Willmitzer L, Melchinger AE (2012) Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. Proc Natl Acad Sci USA **109:** 8872–8877

Ritland K (1996) Estimators for pairwise relatedness and individual inbreeding coefficients. Genet Res **67:** 175–185

Robbins MD, Sim SC, Yang W, 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

Schauer N, Semel Y, Balbo I, Steinfath M, Repsilber D, Selbig J, Pleban T, Zamir D, Fernie AR (2008) Mode of inheritance of primary metabolic traits in tomato. Plant Cell **20:** 509–523

Schauer N, Semel Y, Roessner U, Gur A, Balbo I, Carrari F, Pleban T, Perez-Melis A, Bruedigam C, Kopka J, et al (2006) Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. Nat Biotechnol **24:** 447–454

Schauer N, Steinhauser D, Strelkov S, Schomburg D, Allison G, Moritz T, Lundgren K, Roessner-Tunali U, Forbes MG, Willmitzer L, et al (2005) GC-MS libraries for the rapid identification of metabolites in complex biological samples. FEBS Lett **579:** 1332–1337

Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, Nordborg M (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet **44:** 825–830

Shirasawa K, Fukuoka H, Matsunaga H, Kobayashi Y, Kobayashi I, Hirakawa H, Isobe S, Tabata S (2013) Genome-wide association studies using single nucleotide polymorphism markers developed by re-sequencing of the genomes of cultivated tomato. DNA Res **20:** 593–603

Sillanpää MJ (2011) Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. Heredity (Edinb) **106:** 511–519

Sim SC, Durstewitz G, Plieske J, Wieseke R, Ganal MW, Van Deynze A, Hamilton JP, Buell CR, Causse M, Wijeratne S, et al (2012) Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. PLoS ONE **7:** e40563

Soto-Cerda BJ, Cloutier S (2012) Association mapping in plant genomes. *In* C Mahmut, ed, Genetic Diversity in Plants. InTech. http://www.intechopen.com/books/genetic-diversity-in-plants/association-mapping-in-plant-genomes (April 29, 2013)

Steinbach D, Alaux M, Amselem J, Choisne N, Durand S, Flores RL, Keliet AO, Kimmel E, Lapalu N, Luyten I,, et al (2013) GnpIS: an information system to integrate genetic and genomic data from plants and fungi. Database (Oxford) **2013:** bat058

Stevens R, Buret M, Duffé P, Garchery C, Baldet P, Rothan C, Causse M (2007) Candidate genes and quantitative trait loci affecting fruit ascorbic acid content in three tomato populations. Plant Physiol **143:** 1943–1953

Stevens R, Buret M, Garchery C, Carretero Y, Causse M (2006) Technique for rapid, small-scale analysis of vitamin C levels in fruit and application to a tomato mutant collection. J Agric Food Chem **54:** 6159–6165

Stevens R, Page D, Gouble B, Garchery C, Zamir D, Causse M (2008) Tomato fruit ascorbic acid content is linked with monodehydroascorbate reductase activity and tolerance to chilling stress. Plant Cell Environ **31:** 1086–1096

Tabangin ME, Woo JG, Martin LJ (2009) The effect of minor allele frequency on the likelihood of obtaining false positives. BMC Proc (Suppl 7) **3:** S41

Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485:** 635–641

Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet **90:** 7–24

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

Xu J, Ranc N, Muños S, Rolland S, Bouchet JP, Desplat N, Le Paslier MC, Liang Y, Brunel D, Causse M (2013) Phenotypic diversity and association mapping for fruit quality traits in cultivated tomato and related species. Theor Appl Genet **126:** 567–581

Zhang N, Brewer MT, van der Knaap E (2012) Fine mapping of *fw3.2* controlling fruit weight in tomato. Theor Appl Genet **125:** 273–284

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