Multiple haplotype-based analyses provide genetic and evolutionary insights into tomato fruit weight and composition

Jiantao Zhao1,2, Christopher Sauvage1,3, Frédérique Bitton1 and Mathilde Causse1,*

1 INRA, UR1052, Centre de Recherche PACA, Génétique et Amélioration des Fruits et Légumes, Domaine Saint Maurice, 67 Allée des Chênes CS 60094 – 84140, Montfavet Cedex, France
2 Boyce Thompson Institute for Plant Research, Cornell University, 533 Tower Road, Ithaca, NY 14853-1801, USA
3 Syngenta SAS France, 1228 Chemin de l’Hobit, Saint Sauveur 31790, France
*Corresponding author. E-mail: mathilde.causse@inrae.fr

Abstract

Improving fruit quality traits such as metabolic composition remains a challenge for tomato breeders. To better understand the genetic architecture of these traits and decipher the demographic history of the loci controlling tomato quality traits, we applied an innovative approach using multiple haplotype-based analyses, aiming to test the potentials of haplotype based study in association and genomic prediction studies. We performed and compared haplotype vs SNP-based associations (hapQTL) with multi-locus mixed model (MLMM), focusing on tomato fruit weight and metabolite contents (i.e. sugars, organic acids and amino acids). Using a panel of 163 tomato accessions genotyped with 5995 SNPs, we detected a total of 784 haplotype blocks, with an average size of haplotype blocks ∼58 kb. A total of 108 significant associations for 26 traits were detected thanks to Haplotype/SNP-based Bayes models. Haplotype-based Bayes model (97 associations) outperformed SNP-based Bayes model (50 associations) and MLMM (53 associations) in identifying marker-trait associations as well as in genomic prediction (especially for those traits with moderate to low heritability). To decipher the demographic history, we identified 24 positive selective sweeps using the integrated haplotype score (iHS). Most of the significant associations for tomato quality traits were located within selective sweeps (54.63% and 71.7% in hapQTL and MLMM models, respectively). Promising candidate genes were identified controlling tomato fruit weight and metabolite contents. We thus demonstrated the benefits of using haplotypes for evolutionary and genetic studies, providing novel insights into tomato quality improvement and breeding history.

Introduction

Deciphering the genetic architecture of agronomical traits is important to plant breeding. Up to now, most marker-trait associations were revealed using QTL and genome-wide association studies (GWAS) and demonstrated their success and limits. The later approach relies on the extent of linkage disequilibrium (LD) between the functional variant locus controlling the variation of the trait and molecular markers. While being successful for detecting loci of large effect, GWAS remains limited to decipher the additional medium and especially low effect loci to track down the missing heritability (the degree of variation for a given phenotypic traits due to the genetic variation) of traits of interest [1].

Haplotypes are the particular combinations of alleles/markers observed on a chromosomal segment in a given population [2]. Haplotype blocks are characterized by rare historical recombination events and are composed of few common haplotypes. Genotyping only a few carefully chosen tag-SNPs should provide enough information to identify the common haplotypes [2, 3], which provide the possibility to test multiple allelic interaction effects and could increase the GWAS statistical power [4] by reducing the number of tests compared to using all SNPs [5]. Alleles within the same haplotype block are more likely to be inherited together [6], while sharing a similar minor allele frequency (MAF). Haplotype-based analyses examine groups of SNPs simultaneously rather than individual SNPs and enhance the statistical detection power for many aspects of both population and quantitative genetics, including identifying signals of recent positive selection [7] and GWAS [8, 9], respectively. For example, a haplotype/SNP-based Bayes model (hapQTL), which integrated phase uncertainty and avoided arbitrariness in specifying haplotypes, discovered new associations in humans compared with other models [5]. This approach was also recently applied to detect marker-trait association in rice [10].
The nature and diversity of haplotypes also witnesses the forces that shaped the genetic variation in a genome across time and the underlying consequences for breeding, notably. Among the four major evolutionary forces (mutation, random genetic drift, gene flow and natural selection), selection refers to any non-random, differential propagation of an allele. Positive selection causes a beneficial allele and hitchhikes variants to sweep to high prevalence (soft-sweep) or even to fixation (hard sweep) within a population, thereby producing a population-wide reduction of genetic diversity [11]. Identifying genomic regions with unusually high local haplotype homozygosity represents a powerful strategy to track natural or artificial recent selective events [12] that occurred in the past recent. Among all available methods, the integrated haplotype score (iHS) compares the pattern of extended haplotype homozygosity (EHH) for both the ancestral and new alleles [13]. Thus, iHS provides helpful insights into the genome-wide distributions of very recent selective events in favor of alleles that have not yet reached fixation [13].

To sustain modern breeding and increase genetic gains, genomic selection (GS) was progressively adopted [14–16]. GS improves selection efficiency, may reduce phenotyping costs and orientate breeding schemes [17]. At first, single SNP analyses were implemented for genomic prediction but haplotypes were recently adopted to increase prediction accuracy by harnessing the similar genetic information with predictors (the number of haplotypes is usually smaller than that of SNPs). Simulations and analyses showed that haplotype-based genomic prediction further improves the prediction accuracy [18–20]. SNP- and haplotype-based regional associations have also been applied to narrow-down the candidate rice grain size/weight MetaQTLs [10], with a detailed step-by-step guideline [21]. However, the application of genomic selection in tomato is limited to simulations [22], or cross validation experiments [23, 24] without any public report of its application.

Tomato (Solanum lycopersicum L.) is the most consumed vegetable worldwide (http://www.fao.org/faostat). Cultivated tomato has experienced severe bottlenecks during its domestication and breeding history, resulting in a narrow genetic diversity [25–27]. Modern tomato flavor is often criticized, especially when compared with cherry tomatoes [28]. Many important metabolites in tomato are flavor-related, but often time-consuming and difficult to quantify [29]. High-throughput genomic and metabolomic approaches thus provide new opportunities for tomato quality improvement [30–32].

Here, we studied a core collection of tomato accessions that were genotyped with SNPs markers and phenotyped for 26 traits, including fruit weight, sugars, organic acids and amino acids [33]. We aimed at describing the haplotype architecture of the tomato genome and use it to (1) deepen our knowledge about the recent breeding history, (2) test the potential of haplotypes to detect new candidate regions for QTL and (3) to predict phenotypes in genomic selection context. We first assessed the haplotype blocks. We then identified the recent positive selective sweeps, which were compared with domestication and improvement sweeps previously identified. We then compared associations based on haplotype or SNP-based Bayes models with multi-locus mixed model. We finally compared the genomic prediction accuracies of multiple Bayes models using haplotypes and or SNPs. Our multiple haplotype-based analyses demonstrated the potentials of using haplotypes for several aspects.

Results
Haplotype block estimation and genome-wide distributions
To investigate the haplotype landscape in the tomato genome, we first determined the haplotype blocks within all accessions and within each of the three subgroups, composed of 116 S. l. cerasiforme accessions (cherry tomato, CER), 31 S. lycopersicum (large-fruit tomato, BIG) and 16 Solanum pimpinellifolium (the closest wild species, PIM). Within all accessions, we detected 784 haplotype blocks (Table S1). Within the three genetic subgroups, we detected 704, 259 and 134 haplotype blocks for CER, BIG and PIM groups, respectively (Table S2–4). We observed a significant positive correlation between the number of tomato accessions and the number of haplotype blocks (R² = 0.99, P = 0.0497). We also randomly selected 12 accessions 100 times for each subgroup and found that the number of haplotype blocks of CER and BIG was still significantly higher than that of PIM (Figure S1a), which could be mainly caused by different proportion of accessions. We tested this similarly by randomly selecting different number of accessions of CER and found that 1) with the increase of accessions proportion, the variation decreased greatly (Figure S1b) and 2) in order to achieve a comparable value of haplotype blocks from different subgroups, at least of 60 accessions for each subgroup should be included with an ideally similar total number of accessions.

The average size of haplotype blocks was 58.085 kbp, ranging from 8 bp to 199.95 kbp. The genome-wide distribution pattern of the haplotype block size was similar in three subgroups (Figure 1a), with a major peak for blocks smaller than 40 bps (Figure S2). On average, there were 5.2 SNPs per block, ranging from 2 to 46 SNPs (Figure 1b). Among the 5995 SNPs, 3841 were located within haplotype blocks. The overlapped genes were only significantly enriched in the process of amylase activity (Table S5). The chromosome ends had higher frequency of SNPs and haplotype blocks in this germplasm collection (Figure 1c and S3).

Identification of selective sweeps
We used both integrated haplotype score (iHS) to identify positive selective sweeps (PSS) and nucleotide diversity (π) in the three subgroups to identify domestication sweeps (DS, πPIM/CER > 3.43) and improvement sweeps
Zhao et al. | 3

Figure 1. Haplotype block distributions within the 163 tomato accessions that were genotyped with 5995 SNPs. (a) Distribution patterns of the size of haplotype blocks within all accessions and subgroups. All, 163 tomato accessions; BIG, 31 large-fruit tomato accessions; CER, 116 cherry tomato accessions; PIM, 16 wild tomato species. (b) Distribution patterns of the number of SNPs within all accessions and subgroups. (c) Genome-wide distribution of SNPs and haplotype blocks. SNP: distribution of 5995 SNPs; All: haplotype blocks within 163 tomato accessions; BIG: haplotype blocks within 31 large-fruit tomato accessions; CER: haplotype blocks within 116 cherry tomato accessions; PIM: haplotype blocks within 16 wild tomato species.

(IS, $\pi_{CER/BIG} > 6.16$). With iHS, we identified a total of 24 positive selective sweeps (PSS01 – PSS24) (Table S6). We then calculated the linkage disequilibrium of the SNPs with the highest iHS score for each positive selective sweep and used $R^2 = 0.5$ as the threshold to define the PSS size, which ranged from 12 kbp to 18.7 Mbp, with an average size of 3.43 Mb (Table S6). A total of 5965 genes were identified within these 24 PSS (Table S7), significantly narrowing down the number of genes that experienced strong selective pressures since domestication. In order to classify whether the PSS were caused during the domestication or the improvement stage, we calculated the total nucleotide diversity ($\pi$) in PIM, CER and BIG, and identified 132 domestication sweeps (DS001-DS132) (Table S8) and 93 improvement sweeps (IS001-IS093) (Table S9). All the genes within these sweeps were listed in Table S10 and Table S11. Among these, 13 domestication sweeps and 10 improvement sweeps overlapped with 8 and 6 PSS, respectively. Among the 24 PSS, 9 were not overlapping with either domestication or improvement sweeps. There were 3455 and 5700 genes located within DS and IS, respectively.
In addition, PSS, DS and IS covered a total genomic size of 43.959, 52.539 and 75.208 Mb, respectively, accounting for 3.73%, 4.46% and 6.38% of the tomato genome size.

**Haplotype-based association model outperformed multi/single-locus association model**

We then tested the efficiency of haplotypes in identifying associations between traits and markers. We performed the regional association mapping using the haplotype/SNP-based Bayes factor model implemented in hapQTL software [5]. A total of 108 significant associations were detected for 26 traits, among which 98 were identified in haplotype-based Bayes model (Bayes factorhap), 50 were identified in SNP-based Bayes model (Bayes factorSNP) and 40 were common to both models (Figure 2a; Table S12). Manhattan plots and regional marker local haplotype score (mLHS) for all the associations are provided in Figure S4–29.

In order to validate the benefits of haplotypes in identifying associations, we compared results of haplotype/SNP-based Bayes association model with multi-locus mixed model (MLMM) [34]. We divided the traits into four groups of acids (6), amino acids (9), sugars (10) and fruit weight (fw). In addition to previously analyzed metabolites [33], we detected 9 significant associations for fruit weight (fw), generating a total of 53 significant associations across the 26 traits (Table S13). The number of associations detected with MLMM (53 associations) was between Bayes factorhap (97 associations) and Bayes factorSNP (50 associations) (Figure 2a).

Notably, there were 13 significant associations that were co-detected across all three models. Among all the associations detected in hapQTL (combining Bayes factorhap and Bayes factorSNP), the largest number of associations was detected for sugars (42 associations) (Figure 2b). In contrast, only 9 associations were detected for fw in MLMM. Instead, MLMM was more efficient in identifying associations for sugars (Figure S30).

**Associations for fruit weight and diverse metabolites were located within selective sweeps**

In order to test whether fruit weight and metabolite associations experienced selective events, we cross-validated the overlap between hapQTL (combining Bayes factorhap and Bayes factorSNP) and MLMM results with selective sweep locations, including PSS, DS and IS (Figure S31). Among all the associations detected in hapQTL (108 associations) and MLMM (53 associations), 59 and 38 were located within a selective sweep, respectively (Figure S31). Though the total number of significant associations detected with hapQTL was larger than with MLMM, the proportion of associations within selective sweeps in hapQTL (54.63%) (Figure S31a) was smaller compared to MLMM (71.7%) (Figure S31b). The largest proportion of associations within selective sweeps was observed for sugars, with 20.37% (22 associations) and 28.3% (15 associations) in hapQTL and MLMM results, respectively (Figure S32).

**Fruit weight was notably improved due to positive selection events**

Fruit weight (fw) is one of the most selected traits during the long-term domestication and improvement processes. Though fruit weight was strongly linked to the population structure, MLMM and hapQTL identified 9 and 23 significant associations, respectively (Figure 3a, b). Among these, 7 (77.8%) and 14 (60.9%) were within selective sweeps (Table S12-S13). The two strongest positive selective sweeps (PSS12 and PSS16) overlapped with two fw associations (Figure 3c). We took these associations as examples for further illustration together with another on chromosome 6, located between DS066 and IS061, but not within any PSS. The strongest positive selective sweep PSS16 decayed on a large genomic region located from 1.34 Mb to 6.02 Mb. The peak SNP associated with fw detected both in MLMM and hapQTL was located between 1.82 Mb and 3.82 Mb, within PSS16. Within this region, two domestication sweeps (DS076 and DS077) and two
Improvement sweeps (ISO64 and ISO65) were identified. However, the domestication and improvement sweeps were intertwined, indicating that this region had experienced both domestication and improvement, which together generated a large positive selective sweep (PSS16) (Figure 3d, right panel). In contrast, though the second strongest positive selective sweep PSS12, overlapped with fw association, PSS12 only covered about 0.15 Mb. But, a larger improvement sweep ISO55 was located nearby (Figure 3d, left panel).

We then compared the extended haplotype homozygosity of the peak SNPs in these three regions and major differences between ancestral and derived allele (referred as Allele A and B) were observed for the associations on chr5 and chr7 (Figure 3e), indicating that these two regions have undergone positive selections (supported by PSS12 and PSS16) (Figure 3c). In contrast, the extended haplotype homozygosity profiles of the peak SNP on chr6 were similar between both alleles (Figure 3e).

**Most associations for metabolomic traits experienced selective pressures**

Among all the associations detected in hapQTL for the metabolic traits (sugars, organic acids and amino acids), 45 associations were located within selective sweeps (either in PSS, DS or IS). For all the metabolites, most associations (74%) detected in MLMM were within selective sweeps. Up to 85.7% of all associations detected for organic acids experienced selective footprints. In order to understand the impacts of positive selections on those important metabolites in tomato, we took one trait from each group as example (malate for organic acids, fructose for sugars and proline for amino acids), based on their physiological importance and also the significance of their associations. The strongest association for
malate on chr3 overlapped with positive selective sweep PSS10 (Figure 4a,b). The association for fructose on chr5 overlapped with PSS13 (Figure 4a,c) and the strongest association for proline on chr9 overlapped with PSS21 (Figure 4a,d).

We also compared the allelic diversity (\(\pi\)) between BIG, CER and PIM to see whether any domestication and improvement events occurred in the nearby region. For malate, the associated candidate region also experienced a major improvement sweep (IS043–043) (Figure 4e). The extended haplotype homozygosity of the peak SNP showed major differences between Allele A and B, which provided supporting evidence of positive selection (Figure 4h). In comparison, near the associated region for fructose, an improvement sweep (IS056) extended to a large region (Figure 4f). The extended haplotype homozygosity of the peak SNP showed that positive selective of this region had significantly reduced the number of haplotypes upstream of the peak SNP, for both alleles (Figure 4i). For proline, the associated region was located within a single positive selective sweep (PSS21). Though this region was not overlapped with any domestication or improvement sweeps (Figure 4g), significant difference of the extended haplotype homozygosity of the peak SNP was also observed between both alleles (Figure 4j). In addition, significant differences of relative contents of these three metabolites were observed between alleles A and B (Figure 4k-m).

Combining selection footprint and associations for the identification of candidate genes of tomato metabolic regulations

Promising candidate genes were identified associated with tomato metabolites. Among all the associations detected, we found that fructose, glucose and sucrose were all co-associated with the same candidate region on chr2:47218316 (chromosome: bps) (Figure 5a) and chr6:41994475. For both loci, significant difference was observed for all three sugars between sugar content and alleles as well as between different subgroups (Figure S33). The association on chr2:47218316 was located within positive selective sweep PSS07. The length of haplotype block in the nearby region was different in three groups and was longest in CER, followed by BIG (Figure 5b). Combining the overlapped haplotype block and PSS07, we narrowed down the candidate region to 28 kb, containing five genes (Figure 5c).

In the downstream region where the haplotype block was broken between BIG and CER was the gene _wuschel_ (Solyc02g083950). Together with functional annotations of the other four genes (oxidoreductase 2OG-Fe oxygenase, Solyc02g083960; chromosome 18 contig 1 DNA sequence, Solyc02g083970; endoglucanase 1, Solyc02g083980; calcium-dependent protein kinase CPK1 adapter protein 2-like, Solyc02g083990), _wuschel_ was likely the most promising candidate gene for this association. _Wuschel_ and its homologs is controlling plant embryo development and meristem maintenance, with different pleiotropic functions that might be conserved among species [35]. _Wuschel homeobox 7 (WOX7)_ in _Arabidopsis thaliana_ could inhibit lateral root development via a sugar-dependent approach [36]. Within this region, a total of 13 unique haplotypes were observed for the 163 accessions and five of them were rare (Hap4, Hap7, Hap10, Hap11 and Hap13). The total sugar content of the three sugars was different between different haplotypes (Figure 5d). We compared the total sugar contents for the SNP (chr2:47190360) located in the exon of _wuschel_ and the one (chr2:47204573) between _wuschel_ and the remaining genes and observed high significant differences between the ancestral and derived allele (Figure 5e).

The extended haplotype homozygosity carrying the derived allele (G) for chr2:47204573 extended much further compared to the ancestral allele (A). In contrast, more diversity was observed for the ancestral allele (Figure S34). Within the candidate region, there were 9 accessions carrying the ancestral allele, all of which were CER (8) and PIM (1). In comparison, there were only three accessions carrying the derived allele, which were all BIG, providing supporting evidence of positive selection in this region (Figure 5f). One of the main challenges of improving tomato sugar content was the general negative correlation between sugar content and fw. We found that the correlation for BIG and PIM were weak and not significant and was only significant for CER with a low negative value (\(R = -0.27, p = 2.1 \times 10^{-5}\)) (Figure 5g). Among the three BIG accessions carrying the derived haplotype within the candidate region, CR017 carried the core haplotype with high fw (142.56 g) and moderate sugar content (9.65 mg/g fw).

Haplotypes based models explain higher heritability and improved genomic prediction accuracy

We compared the heritabilities explained by SNPs and haplotypes (using pseudo SNPs) for 10 traits, which had variable heritabilities based on phenotypes. For all the traits, haplotype-based models explained a larger part of heritability compared to SNPs (except for fw, where the heritabilities were similar) (Figure 6a). Haplotype was especially more efficient compared to SNPs for those traits with moderate to low heritability, such as citrate and lysine (Figure 6a). Then, we tested the efficiency of using haplotypes in genomic prediction using several Bayes-based prediction models (Table S14).

Overall haplotype-based models outperformed SNP-based models in 39/50 cases (Figure 6b). Notably, for fw and SSC, the two most important agronomical traits, haplotype-based prediction models had always a higher prediction accuracy compared to SNP-based models. However, the best prediction models were different for fw and SSC, being BayesA and BL, respectively. Haplotype-based BRR had the highest prediction accuracy for asparagine (Asn, \(R^2 = 0.63\)), lysine (0.25) and sucrose (\(R^2 = 0.59\), Figure 6b). For those traits that are likely
controlled by a few major QTLs, such as malate and citrate, the prediction of haplotype-based BayesB outperformed all the other models (Figure 6b). Taken all these results together, we observed a highly significant correlation ($R = 0.79$, $p < 0.01$) between trait heritability and prediction accuracy (Figure 6c).

**Discussion**

**Benefits of haplotypes in identifying new associations**

Compared to whole genome sequencing (WGS), SNP array is still a reliable, highly accurate and relatively cheap technology for GWAS, especially for large samples...
Figure 5. Identification of associations significantly co-associated with fructose, glucose and sucrose using hapQTL. (a) Regional Manhattan plot of the significant association for fructose, glucose and sucrose. (b) Haplotype blocks near the associated candidate regions of big-fruit tomato (BIG, *Solanum lycopersicum*), cherry tomato (CER, *S. lycopersicum* var *cerasiforme*) and the closest wild species (PIM, *Solanum pimpinellifolium*). (c) Genes within the candidate region. The start and end position of the haplotype blocks and gene locations were consistent. (d) All the haplotypes detected within the candidate regions and their distribution of total sugars content. (e) Significant t-tests between total sugar content and the ancestral and derived allele for the first two loci. (f) Bifurcation diagram for the extended haplotypes starting from the allele A and G of the peak SNP within the candidate region. The "1" and "2" appeared at the end of the accession names were the two haplotype groups the accessions belonged to. (g) Correlations between total sugar content and fresh fruit weight of three subgroups. Core alternative haplotype-carrying accessions were indicated with arrows and the accession with high fruit weight and moderate total sugar content was highlighted.

Haplotype association mapping takes into account not only allelic heterogeneity, but also possible statistical interactions among close markers, which is more powerful than single marker and multiple marker analysis [5]. We identified more associations using haplotype-based Bayes model, compared to SNP-based Bayes model and MLMM, demonstrating the potential benefit of haplotypes in identifying new associations. In human genetics, the threshold of hapQTL is usually set at $10^{-6}$, which was comparable to the genome-wide threshold ($10^{-8}$) in a typical human GWAS study [5]. We thus thought this suggestive threshold based on the effective number of SNPs was appropriate for comparing the number of significant associations. The overlap between the most significant associations of MLMM and hapQTL provided additional support for this threshold.

**Metabolite composition and fruit weight were improved by selection events at various stages and positions**

During the tomato breeding history, flavor has not been the priority compared to yield, disease resistances and fruit shelf life [28, 29]. However, selection of the main breeding targets during tomato domestication and improvement might have left long-term direct or indirect effects on diverse flavor-related metabolites. In this study, 24 PSS were identified, which together accounted for 8.36% of the tomato genome size.
Figure 6. Comparison of haplotype and SNP based models in calculating narrow sense trait heritability and their performances in genomic prediction of different Bayes models. (a) Comparison of heritability using haplotypes and SNPs for the 10 selected traits. (b) Comparison of the genomic prediction accuracy of five Bayes models between haplotypes and SNPs based on 100 replicates. FW, fruit weight; Fru, fructose; Suc, sucrose; ASA, ascorbic acid; Asn, asparagine; Pro, proline; Lys, lysine. BL, Bayesian LASSO; BRR, Bayesian Ridge Regression. Models with "_hap" indicated that the prediction was based on haplotypes. (c) Correlation between trait heritability and prediction accuracy in the 100 replicates.

In contrast, though more domestication and improvement sweeps were identified (132 DS and 93 IS, respectively), they covered a smaller genomic size compared to PSS, which was lower than a previous study where 360 accessions were resequenced [27]. This could be explained by the limited genomic coverage of SNP arrays and the size and composition of our panel. Although several major fruit weight QTLs were identified in tomato via linkage mapping, we identified new significant associations using hapQTL and MLMM. Besides, the strongest selective sweeps identified via iHS overlapped with fruit weight associations. Among these, there was one strong association detected on chr12 both in hapQTL and MLMM, which was near fw12.1 [38]. The candidate gene for the strongest association on chr5 was annotated as a cell division cycle associated 7 (Cell division cycle associated 7), which could impact the fruit weight by regulating the cell division and growth [39].

In this study, the majority of associations for metabolites both in hapQTL and MLMM (54.63% and 71.70%, respectively) were located within selective sweeps (either in PSS, DS or IS). These results demonstrated that some genes with major effects might have undergone positive artificial selection, due to their major phenotypic effects. This is the case for instance for Al-Activated Malate Transporter 9 (Sl-ALMT9), the major QTL responsible for variation in malate accumulation in fruit [30, 31, 33, 40, 41]. This gene has been selected during domestication stage [41] by a gradual increase in the frequency of the haplotypes carrying the favorable allele. We also
found that it was located within a domestication sweep (DS069). In addition, we observed that citrate was also significantly associated with this region, indicating that Sl-ALMT9 may influence other closely related metabolites, such as citrate. Lin5, a major QTL regulating SSC in tomato [42], was located within both DS149 and PSS20, indicating strong selection experienced by this gene. The locus fu3.2 (Solyco3g114940), a major fruit weight QTL [43], was located within PSS10, supporting previous results that showed lower level of genetic diversity at this locus. In addition, we also found cases where associations with different traits overlapped with multiple sweeps. For instance, we found the association with brix and fructose on chr5 were quite close and were overlapped with both PSS13 and IS056. These results indicated that highly correlated traits were likely to be co-selected and also could be genetically controlled by some common loci [31].

The haplotype landscape revealed selective footprints that occurred during selective events, a typical footprint of human-driven selective process. For example, for the two associations that were co-associated with fructose, glucose and sucrose, the first association was located within one major haplotype. The length of the haplotype, which was located within PSS07 and a previously reported improvement sweep (IS030) [27], differed in BIG, CER and PIM, indicating that this haplotype has been strongly selected. In contrast, for these sugars, the peak SNP of association on chr6 were located between two major haplotypes and no selective sweeps were identified in the nearby region.

Promising candidate genes involved in tomato metabolic regulation

Identifying promising candidate genes controlling traits of interest is one of the major outputs in GWAS. However, choosing the candidate genes to focus on remains challenging, especially when the LD near the peak SNP is large. However, for most of the cases, the most promising candidate gene was quite close to the peak SNP. For example, we found the significant association for malate and citrate on chr6 was consistent with previous studies, corresponding to Sl-ALMT9 [31, 41]. The candidate gene for the association for SSC on chr9 that overlapped with PSS20 included Lin5 (Solyco9g010080), the first QTL cloned in tomato for sugar content [42].

Promising candidate genes involved in the content of primary metabolites and volatiles have already been identified [31, 33, 40]. In this study, we identified several additional candidate genes. Functional annotations and expression patterns in fruit of the candidate genes reinforced their possible involvement in regulating tomato metabolites. For example, phosphoenolpyruvate carboxylase (PEP) (Solyco4g006970) was within the associated regions on chr4:993580 (chromosome: position in bps) for SSC. This protein is involved in carbon fixation process as well as in regulating flux through the Krebs cycle and was highly expressed in fruits [44]. The candidate gene for the other association for SSC on chr11:49507731 was annotated as a sugar transporter (Solyco11g062360), which was functionally involved in the biological process of sugar transportation [45]. For proline, the candidate gene of the association on chr2:51094633 was annotated as proline dehydrogenase (Solyco04g074530), which was directly involved in the regulation of proline. The candidate gene of the association for sugars on chr6:41994475 was annotated as a glucose transporter (Solyco06g066600), thus directly involved in glucose transport. The candidate gene of the association for glutarate2oxo (2-oxoglutarate or α-ketoglutaric acid) on chr6:44818656 was annotated as a short-chain dehydrogenase/reductase (Solyco6g072670) and could play an important role in the Krebs cycle and regulate a diverse set of metabolites. Furthermore we demonstrated that wuschel, apart from controlling meristem in diverse higher plant species including tomato [46, 47], could also regulate tomato sugars following an unknown mechanism. The wuschel related homeobox WOX7, which regulates Arabidopsis lateral root development through a sugar-dependent manner, demonstrated one possible examples of wuschel gene family in regulating sugars in plants [36], and might be also conserved in other species [35]. Further functional validation is needed to definitively validate these candidate genes.

The genomic prediction accuracy can be influenced by many factors, including the population structure, the size of the training and testing populations, the number and density of markers, etc., which have been analyzed and discussed for our panel by Duangjit et al. (2016) [23]. Based on their results, we found that the overall prediction accuracy could be further improved by using pseudo-SNPs from haplotypes. However, this method did not always achieve the best performance compared to single SNPs, which could be caused by various factors above mentioned, but also because the number of pseudo-SNPs was different from SNPs.

In conclusion, we demonstrated the benefits of haplotypes versus SNPs in many aspects, including identifying marker-trait associations, revealing the haplotype bifurcation patterns near the most significantly associated SNPs, explaining more missing heritability and increasing the genomic prediction accuracy. These results will be helpful for breeders and researchers not only focusing on tomato, but also other agronomical crops.

Materials and methods

Materials

The studied panel consisted of 163 tomato accessions derived from a core collection previously described [33]. Briefly, there were 116 S. l. cerasifera accessions (CER, cherry tomato), 31 S. lycopersicum (BIG, large-fruit tomato) and 16 S. pimpinellifolium (PIM, the closest wild species). Plants were grown in greenhouse during summers of 2007 and 2008, in Avignon, France. Environmental conditions were detailed explained previously [33]. Three sets
of pericarp tissues from five ripe fruits were collected and stored at −80°C before metabolomic profiling. Genomic DNA was isolated from 100 mg frozen leaves.

Genotyping and quality control
All accessions were genotyped with the SOLCAP SNP array [49, 50]. SNPs with genotyping call rate lower than 90% were removed. The remaining SNPs were then filtered with minor allele frequency (MAF, 0.037 < MAF < 0.45), generating a total of 5995 high quality SNPs [33].

Haplotype block estimation
We first estimated the haplotype blocks for all accessions using PLINK v1.9 with “—blocks” [51], which was estimated following the default procedure of Haploview v4.2 [52]. We then estimated the haplotype blocks within each group (BIG, CER and PIM), following the same parameters.

Identification of positive selective sweeps using iHS
We used integrated haplotype score (iHS) as implemented in rehh v2.0 R package [12] to identify positive selective sweeps. The genotypic data were first phased using SHAPEIT v2 [53] with default settings (except the number of effective population size set at 2000) [31]. For those SNPs passing the significant threshold (−log10 [1−2|ΦiHS−0.5|] > 2), we calculated the linkage disequilibrium (r²) around the peak SNP with the following parameters: —ld-window-kb 100 000 —ld-window 1000 —r2 0.5 —maf 0.037 in PLINK 1.9 (https://www.cog-genomics.org/plink2). We used r² = 0.5 as the LD decay threshold to group those SNPs passing the significant iHS threshold as one selective sweep (sweeps closer 100 kb were also combined as one).

Identification of domestication and improvement sweeps
In order to check whether the positive selective sweeps occurred during domestication or improvement stages, two key stages during tomato evolution, we calculated the total nucleotide diversity (π) between three subgroups using a 100 kb window with a step size of 10 kb in PIM, CER and BIG, separately, using vcftools v 0.1.17 with —window-pi 100 000 —window-pi-step 10000 [54]. We then scanned the ratios of genetic diversity (π) between PIM and CER (πPIM/πCER) for domestication sweeps (DS) and between CER and BIG (πCER/πBIG) for improvement sweeps (IS). We selected windows with the top 5% of ratios as the domestication and improvement sweeps, respectively (3.43 and 6.16 for domestication and improvement, respectively). All sweeps within a window of 100 kb were merged into a single selected region. Compared with iHS results, we assessed which domestication and improvement sweeps were also positive selective sweeps.

Phenotyping
Briefly, ten fruits per accession were measured for fruit weight. The metabolite profiles were measured on three pools of five fruits as previously described. Only phenotypes with a high correlation over two years were retained for further analyses [33].

Regional haplotype/SNP-based association and significant threshold
We performed regional haplotype/SNP-based association using hapQTL v1.00 [5]. The number of EM runs was set at 10 to avoid uncertainty in LD inference. The number of upper clusters was set at 3. The genome-wide significant threshold of hapQTL should be lower than the -log10 transformed threshold of typical GWAS based on other SNP-based models, such as EMMAX [55], as the Bayes values of some SNPs were negative and were removed, which reduced the number of effective markers. The -log10 transformed suggestive threshold of typical GWAS was generally used as the significant threshold of hapQTL [5]. For the 5995 SNPs, the suggestive and significant p-value was 4.1 × 10^-4 and 2.05 × 10^-5, respectively, which was calculated in Genetic type 1 Error Calculator (GEC) [56]. In order to test the validity of this threshold set to 3.387 (−log10 [the suggestive p-value 4.10 × 10^{-4}] as the genome-wide significant threshold for hapQTL, we selected malate as an example and performed 100 permutations by randomly reordered malate content and also the corresponding DAPC cofactor. We then ran the hapQTL with the same parameters. For the 100 runs, the highest Bayes factor value was 3.576, slightly higher than 3.387 (among all associations, only two associations were higher than 3.387) (Table S14). However, all the three identified significant associated regions for malate were higher than 3.576. For all the associations passing 3.387 for all traits, the majority (80%) of them passed 3.576 (Table S11). So for all the traits analyzed, we used 3.387 (−log10 [the suggestive p-value 4.10 × 10^{-4}] as the genome-wide significant threshold for hapQTL [5]. Significant associations in the same strong LD block or haplotype regions were treated as a unique association and the peak SNP was retained.

Marker local haplotype sharing (mLHS) was calculated based on 10 independent EM runs with the same parameter, which could be used to define the LD block around the peak SNP [5]. We found that the threshold of 0.25 (2.5/10) was too stringent for most loci that they could not cover most of the top ancestral haplotypes. So, we adopted 0.20 as the LD block threshold. Gene annotations were retrieved from the tomato genome annotation version 2.40. For those SNPs passing the significant threshold, we then calculated the LD (r² = 0.5) for each marker to group those closely linked SNPs as one selective sweep. Those sweeps within 100 kb were also grouped as one single sweep.

Multi-locus mixed model for association study
In order to compare the efficiency of hapQTL, we compared our results with those obtained by multi-locus mixed-model (MLMM) [34], as previously described [33]. Population structure was estimated in Structure v2.3.3.
[57] and kinship matrix estimated in SPAGeDi v 1.5 [58]. We then compared the number of significantly associated regions obtained by the three GWAS methods. Same approach and parameters were also performed for fruit weight, including MLMM and hapQTL.

**Candidate gene expression patterns and visualization of local haplotype structure**

To provide supporting evidence whether candidate genes were functionally related to the analyzed phenotypic traits, we first screened the tomato genome annotation (version 2.40). For those candidate genes of related biological function, we checked their expression level in tomato fruit at different developing stages. Data were retrieved from the Tomato Expression Atlas database (http://tea.solgenomics.net/expression_viewer/input) [59]. Candidate gene expression levels at different tissues and developing stages were also checked using ePlant [60]. Visualization of local haplotype structure around a peak SNP near the candidate gene was performed using the bifurcation.diagram() function in rehh 2.0 R package [12]. We checked whether there was major difference between the ancestral and derived haplotypes (in order to avoid this uncertainty, we used allele A and B instead of ancestral and derived allele, as the true ancestral and derived alleles of the SNPs are unknown).

**Haplotype/SNP-based genomic prediction using Bayes models**

To test the potential benefits of using haplotypes compared to single SNP in improving the genomic prediction accuracy of traits of interest, we used both SNPs and haplotypes into a genomic prediction framework. To do so, we first converted the haplotypes within each haplotype block to pseudo-SNPs. Briefly, if the unique haplotype appears in one accession, we treated it as allele A; if not, we treated it as alternative allele B. Rare pseudo SNPs with low minor allele frequency were removed. After filtering minor allele frequency lower than 0.037 [33], a total of 2496 pseudo-SNPs remained. They were then combined with those SNPs located outside haplotype blocks, generating a total of 4650 SNPs (Table S15), which was used to compare with the 5995 SNPs.

In order to compare the prediction accuracy, we selected 10 traits as examples based on their biological significance and wide range of heritability, including fruit weight, SSC, fructose, sucrose, ascorbic acid, malate, citrate, asparagine, proline and lysine. We compared multiple Bayes genomic prediction models, in order to fit all expected genetic architecture of these traits and compare their differences and efficiencies for different traits, including BayesA, BayesB, BayesC, BL and BRR using BGLR package v1.0.8 [61]. Then, to test the predictive accuracy of each model, we implemented a cross validation approach: we randomly selected 75% of the accessions as the training population (TP), trained the model and predicted the phenotypic values within the remaining accessions or test population, with 100 replications. In order to test the marker effects on prediction accuracy, we also used malate as an example following the same approach with 100 replicates. We compared the prediction accuracy among all 5995 SNPs (SNP), 3481 SNPs that were within haplotype blocks and were converted to pseudo-SNPs (HAP) and also the final pseudo-SNPs + remaining SNPs (SNP + HAP). We found that HAP and SNP + HAP had a better performance than SNP alone for all the models, except for BayesA, where SNP + HAP had the lowest performance (Figure S35). The narrow sense heritability (additive) was calculated using “sommer” package [62]. Significance of correlations between predicted and observed values was tested using t-test in ggpubr package (https://cran.r-project.org/web/packages/ggpubr/index.html). All scripts are available as supplemental data DS1.

**Acknowledgements**

Jiantao Zhao was funded by Chinese Scholarship Council (CSC) scholarship (No. 201606300007). We thank Yongtao Guan from Baylor College of Medicine, USA for suggestions and his detailed explanations on the threshold of hapQTL and the example data and R codes he provided to calculate mLHS. We thank Anurag Daware and Akhilesh Tyagi from National Institute of Plant Genome Research, New Delhi, India for sharing the R codes and example data in rice to calculate mLHS. We are also grateful to Jacques Lagnel from UR1052 GAFL INRA, France for the bioinformatic assistance he provided to perform the analyses described in the present manuscript. We than Cao Xu from the IGBD, CAS, Beijing, China and Zachary B. Lippman from CSHL, USA for their discussions about wuschel.

**Author contributions**

Study design/conception: M.C., J-T.Z., C.S.; supervision: C.S., M.C.; data collection and analysis: J-T.Z.; data interpretation: J-T.Z., F.B., C.S., M.C.; first draft of the manuscript: J-T.Z.; critical revisions of the manuscript: all co-authors.

**Data Availability**

To ensure their accessibility, the genotyping (5995 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 ref. 33. The haplotype blocks are described in Table S1-S4.

**Conflict of interest**

The authors declare that they have no conflict of interests.
Supplementary data

Supplementary data is available at Horticulture Research Journal online.

References

1. Eichler EE, Flint J, Gibson G. et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11:446–50.
2. Belmont JW, Hardenbol P, Willis TD. et al. The international HapMap project. Nature. 2003;426:789–96.
3. Johnson GCL, Esposito L, Barratt B. et al. Haplotype tagging for the identification of common disease genes. Nat Genet. 2001;29:233–7.
4. Bader JS. The relative power of SNPs and haplotype as genetic markers for association tests. Pharmacogenomics. 2001;2:11–24.
5. Xu H, Guan Y. Detecting local haplotype sharing and haplotype association. Genetics. 2014;197:823–38.
6. Farashi S, Kryza T, Clements J. et al. A primary assembly of a bovine haplotype block map based on a 15,036-single-nucleotide polymorphism panel genotyped in Holstein–Friesian cattle. Genetics. 2007;176:763–72.
7. Maldonado C, Mora F, Scapim CA. et al. Genome-wide haplotype-based association analysis of key traits of plant lodging and architecture of maize identifies major determinants for leaf angle: HAPla4. PLoS One. 2019;14:e0212925. https://doi.org/10.1371/journal.pone.0212925.
8. Klee HJ, Tieman DM. The genetics of fruit flavour preferences. Nat Rev Genet. 2018;19:347–56.
9. Darrow AV, Srivastava R, Singh AK. et al. Regional association analysis of MetaQTLs delineates candidate grain size genes in Rice. Front Plant Sci. 2017;8:807.
10. Vitti JJ, Grossman SR, Sabeti PC. Detecting natural selection in genomic data. Annu Rev Genet. 2013;47:97–120.
11. Gautier M, Klassmann A, Vitalis, R. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype genomic data. Nat Biotechnol. 2019;37:744–58.
12. Crossa J, Perez-Rodriguez F, Cuevas J. et al. Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 2017;22:961–75.
13. Hickey LT, Haefez A, Robinson H. et al. Breeding crops to feed 10 billion. Nat Biotechnol. 2019;37:744–54.
14. Hickey JM, Chirurugi T, Mackay I. et al. Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. Nat Genet. 2017;49:1297–303.
15. Calus MPL, Meuwissen THE, Roos APW. et al. Accuracy of genomic selection using different methods to define haplotypes. Genetics. 2008;178:553–61.
16. Minamikawa MF, Kunihiisa M, Noshita K. et al. Tracing founder haplotypes of Japanese apple varieties: application in genomic prediction and genome-wide association study. Hortic Res. 2021;8:49.
17. Klee HJ, Tieman DM. The genetics of fruit flavour preferences. Nat Rev Genet. 2018;19:347–56.
18. Blanca J, Montero-Pau J, Sauvage C. et al. Genomic variation in tomato, from wild ancestors to contemporary breeding accessions. BMC Genomics. 2019;16:257.
19. Lin T, Zhu G, Zhang J. et al. Genomic analyses provide insights into the history of tomato breeding. Nat Genet. 2014;46:1220–6.
20. Živić A, Reic K, Reic M. et al. A chemical genetic roadmap to improved tomato flavor. Science. 2017;355:391–4.
21. Zhao J, Sauvage C, Zhao J. et al. Meta-analysis of genome-wide association studies provides insights into genetic control of tomato flavor. Nat Commun. 2019;10:1534.
22. Zhu G, Wang S, Huang Z. et al. Rewiring of the fruit metabolome in tomato breeding. Cell. 2018;172:249.e12–61.e12.
23. Sauvage C, Segura V, Bauchet G. et al. Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. Plant Physiol. 2014;165:1120–32.
24. Segura V, Vilhjalmsdottir BJ, Platt A. et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet. 2012;44:825–30.
25. Hendelman A, Zebbell S, Rodriguez-leal D. et al. Conserved pleiotropy of an ancient plant homeobox gene uncovered by cis-regulatory dissection. Cell. 2021;184:1724–1739.e16.
26. Kong D, Hao Y, Cui H. The WUSCHEL related Homeobox protein ACTIVATED MALATE TRANSPORTER9 is selected during tomato domestication and determines fruit malate contents and aluminum tolerance. Plant Cell. 2017;29:2249–68.
42. Fridman E, Pleban T, Zamir D. 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 U S A. 2000;97:4718–23.

43. Chakrabarti M, Zhang N, Sauvage C. et al. A cytochrome P450 regulates a domestication trait in cultivated tomato. Proc Natl Acad Sci U S A. 2013;110:17125–30.

44. Zhang C, Li X, He Y. et al. Physiological investigation of C4-phosphoenolpyruvate-carboxylase-introduced rice line shows that sucrose metabolism is involved in the improved drought tolerance. Plant Physiol Biochem. 2017;115:328–42.

45. Durán-Soria S, Pott DM, Osorio S. et al. Sugar signaling during fruit ripening. Front Plant Sci. 2020;11:564917, https://doi.org/10.3389/fpls.2020.564917.

46. Somssich M, Je BI, Simon R. et al. CLAVATA-WUSCHEL signaling in the shoot meristem. Development. 2016;143:3238–48.

47. Rodriguez-Leal D, Lemmon ZH, Man J. et al. Engineering quantitative trait variation for crop improvement by genome editing. Cell. 2017;171:470, e8–480.e8.

48. Gupta PK, Kulwal PL, Jaiswal V. Association mapping in plants in the post-GWAS genomics era. Adv Genet. 2019;104:75–154.

49. Hamilton JP, Stoffel K. et al. Single nucleotide polymorphism discovery in cultivated tomato via sequencing by synthesis. The Plant Genome. 2012;5:17–29.

50. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155:945–59.

51. Hardy OJ, Vekemans X. SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol Ecol Notes. 2002;2:618–20.

52. Perez P, de los Campos G, Goddard ME. Genome-assisted prediction of quantitative traits using the R package sommer. PLoS One. 2016;11:e0156744.