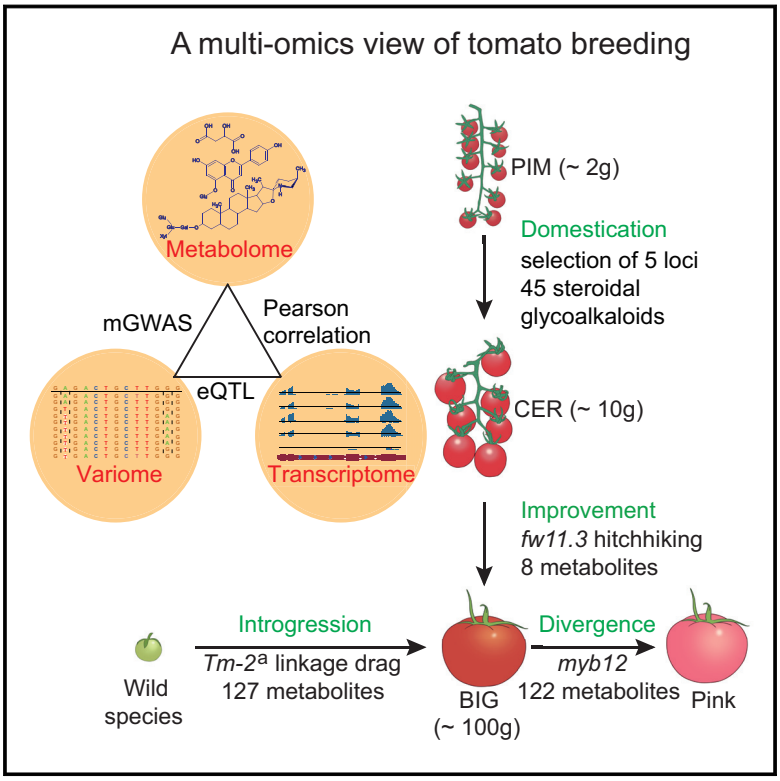
Article

Cell

Rewiring of the Fruit Metabolome in Tomato Breeding

Graphical Abstract



Highlights

- Multi-omic analysis of tomato fruits revealed new metabolic genes and pathways
- Selection of fruit mass gene-altered metabolites altered due to nearby hitchhiking genes
- Domestication acted on five major loci that reduced anti-nutritional compounds
- Pink tomato breeding modified hundreds of metabolites, leading to unexpected changes

Authors

Guangtao Zhu, Shouchuang Wang, Zejun Huang, ..., Alisdair R. Fernie, Jie Luo, Sanwen Huang

Correspondence

jie.luo@mail.hzau.edu.cn (J.L.), huangsanwen@caas.cn (S.H.)

In Brief

Multi-omic analysis reveals how the appearance/taste-oriented breeding process modulates the metabolic makeup of tomato.

Zhu et al., 2018, Cell *172*, 249–261
January 11, 2018
https://doi.org/10.1016/j.cell.2017.12.019

CellPress

Article

Cell

Rewiring of the Fruit Metabolome in Tomato Breeding

Guangtao Zhu,[1,9] Shouchuang Wang,[2,9] Zejun Huang,[3] Shuaibin Zhang,[3] Qinggang Liao,[1] Chunzhi Zhang,[1] Tao Lin,[1] Mao Qin,[1] Meng Peng,[2] Chenkun Yang,[2] Xue Cao,[3] Xu Han,[1] Xiaoxuan Wang,[3] Esther van der Knaap,[4] Zhonghua Zhang,[3] Xia Cui,[3] Harry Klee,[5] Alisdair R. Fernie,[6,7] Jie Luo,[2,8,*] and Sanwen Huang[1,3,10,*]

[1]Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong 518124, China

[2]National Key Laboratory of Crop Genetic Improvement and National Center of Plant Gene Research (Wuhan), Huazhong Agricultural University, Wuhan, Hubei 430070, China

[3]Key Laboratory of Biology and Genetic Improvement of Horticultural Crops of the Ministry of Agriculture, Sino-Dutch Joint Laboratory of Horticultural Genomics, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China

[4]Department of Horticulture, University of Georgia, Athens, GA 30602, USA

[5]Horticultural Sciences, Plant Innovation Center, University of Florida, Gainesville, FL 32611, USA

[6]Max-Planck-Institute of Molecular Plant Physiology, Potsdam-Golm 144776, Germany

[7]Center of Plant System Biology and Biotechnology, Plovdiv 4000, Bulgaria

[8]Institute of Tropical Agriculture and Forestry of Hainan University, Haikou, Hainan 572208, China

[9]These authors contributed equally

[10]Lead Contact

*Correspondence: jie.luo@mail.hzau.edu.cn (J.L.), huangsanwen@caas.cn (S.H.)

https://doi.org/10.1016/j.cell.2017.12.019

SUMMARY

Humans heavily rely on dozens of domesticated plant species that have been further improved through intensive breeding. To evaluate how breeding changed the tomato fruit metabolome, we have generated and analyzed a dataset encompassing genomes, transcriptomes, and metabolomes from hundreds of tomato genotypes. The combined results illustrate how breeding globally altered fruit metabolite content. Selection for alleles of genes associated with larger fruits altered metabolite profiles as a consequence of linkage with nearby genes. Selection of five major loci reduced the accumulation of anti-nutritional steroidal glycoalkaloids in ripened fruits, rendering the fruit more edible. Breeding for pink tomatoes modified the content of over 100 metabolites. The introgression of resistance genes from wild relatives in cultivars also resulted in major and unexpected metabolic changes. The study reveals a multi-omics view of the metabolic breeding history of tomato, as well as provides insights into metabolome-assisted breeding and plant biology.

INTRODUCTION

Metabolomics is defined, by analogy to transcriptomics and proteomics, as the analysis of the metabolic complement of an organism (Wishart et al., 2007). While metabolome coverage is not as comprehensive (Fernie et al., 2004), advances in high-throughput metabolic profiling have rendered metabolomics an important tool for both fundamental and applied research (Saito and Matsuda, 2010). The plant kingdom is exceptionally rich in metabolic diversity, harboring in excess of 200,000 structurally distinct metabolites (Wurtzel and Kutchan, 2016). These metabolites not only play important roles in plant growth, development, and adaptation to environmental changes but also are important sources of human food, medicine, and energy (Butelli et al., 2008; Chen et al., 2016). Over the past decade, the integration of metabolic profiling with other omics tools has proven to be highly effective for functional gene identification and pathway elucidation in plant primary and secondary metabolism (Kusano et al., 2011; Matsuda et al., 2010).

Tomato is the highest value fruit and vegetable crop worldwide and makes a substantial nutritional contribution to the human diet. Tomato fruit quality at harvest is an integrative embodiment of multiple metabolites. Large datasets document the dynamic changes of both tomato fruit metabolites and the structural and regulatory genes that control their abundance throughout development and ripening (Carrari et al., 2006). During fruit development there are large changes in the levels of primary metabolites, including carbohydrates and acids, while at the onset of ripening flavonoids and carotenoids begin to accumulate (Muir et al., 2001), and the content of the bitter glycoalkaloid, α-tomatine, markedly decreases (Iijima et al., 2009).

The combination of metabolomics, linkage mapping studies, and metabolome-based genome-wide association studies (mGWAS) has provided considerable insight into the extent of natural variation in metabolism and its genetic and biochemical control in tomato. GWAS have been recently conducted to map the genetic loci for important metabolic traits (Sauvage et al., 2014; Tieman et al., 2017). Metabolic profiling combined with transcriptome analysis also has been used to dissect secondary metabolic pathways such as steroidal glycoalkaloid (SGA), phenylpropanoid, and flavonoid biosynthesis, revealing

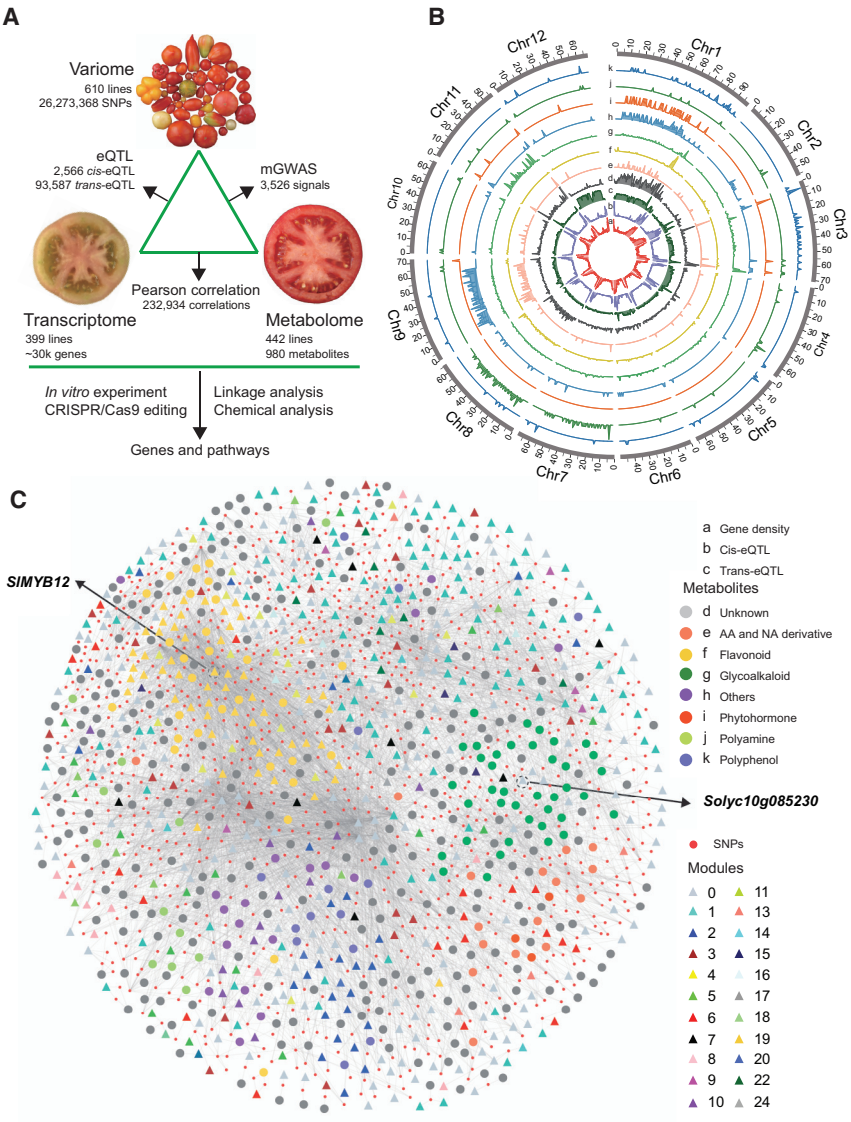


Figure 1. Generation and Integration of Multi-omics Data

(A) Layout of the study.

(B) Genomic distribution of mGWAS and eQTL. The eQTL include *cis*-eQTL and *trans*-eQTL. Metabolites are grouped into eight categories and marked with different colors and tags (d–k).

(C) Network built on correlation among metabolites, genes, and SNPs. Metabolites are shown as large dots colored like in Figure 1B. Genes are shown as reverse arrows with a distinct color per co-expression module. The significant SNPs for mGWAS or eQTL are shown as small red dots. The *SlMYB12* and *Solyc10 g085230* genes are denoted. An interactive version is at http://218.17.88.51:8080/network/index.html.

See also Figures S1 and S2, Tables S1, S2, and S3, and Data S1.

important regulatory roles of novel gene clusters as well as of the *GAME9* transcription factor (Alseekh et al., 2015; Cárdenas et al., 2016; Itkin et al., 2013). Combining metabolomic and genomic data has allowed a comprehensive refinement of SGA biosynthesis (Schwahn et al., 2014), which is predominantly abundant in *Solanaceous* species. Such studies provide insight into the processes underlying the evolution of metabolism. It is important to note, however, that despite the great in-roads made in the targeted studies described above, global insights into metabolic regulation remain rare.

Tomato (*Solanum lycopersicum* var. *cerasiforme* and *S. lycopersicum*) was domesticated from its wild ancestor, selected for adaptive traits to the environment in which they were grown, and, lately, differentiated into different types. Modern breeding specifically focuses on developing varieties that incorporate multiple disease-resistant loci that are introgressed from wild relatives. In a previous study, we reconstructed the history of tomato breeding, including domestication, improvement, divergence, and introgression (Lin et al., 2014). The metabolic changes that have accompanied these human-guided evolutionary processes are essentially unknown. Here, we generated a large dataset spanning the genome, transcriptome, and metabolome on a population of between 399 and 610 diverse tomato accessions (Figure 1A). Integration of the resultant data identified 3,526 mGWAS signals, 2,566 *cis*-eQTL (expression quantitative trait locus), 93,587 *trans*-eQTL, and 232,934 expression-metabolite correlations. The clues gained from these multi-omics datasets were subsequently experimentally tested by linkage mapping, molecular biology, biochemical assays, genetic complementation, and CRISPR-Cas9 knockout. This study provides major insights into how breeding changed the tomato metabolome, a knowledge base for fruit quality improvement, and a rich resource for plant metabolic biology.

RESULTS

Generation and Characterization of a Multi-omics Dataset

We collected a total of 610 tomato accessions (Table S1), including 42 accessions of wild species and 568 accessions from the red-fruited clade (*S. pimpinellifolium*, *S. lycopersicum* var. *cerasiforme* and *S. lycopersicum*), representing various geographical origins, consumption type, and improvement status. Resequencing of these accessions generated a total of 6.6 trillion nucleotides, with a median depth of 6.6× and coverage of 97.2% of the assembled genome (SL2.50) (Consortium, 2012). We generated a final set of 26,273,368

SNPs and identified 500,919 nonsynonymous SNPs in 33,088 genes. This variation map adds 14.7 million SNPs on the basis of our previous one (11,620,517 SNPs of 360 accessions) (Lin et al., 2014).

To understand the natural variation of the metabolome in the red-fruited tomato population, we selected 442 accessions for metabolite quantification. These accessions represent a cross-section of the set selected based on their passport information, morphological traits, and phylogenetic relationships (Figure S1A). We quantified fruit metabolites using a broadly targeted liquid chromatography-tandem mass spectrometry (LC-MS/MS)-based metabolic profiling method. A total of 980 distinct analytes were identified in pericarp tissue of ripe fruits (Data S1), including 362 annotated metabolites. A principal component analysis (PCA) of all metabolite data revealed that the three subgroups largely formed independent clusters, consistent with the evolutionary relationship of the red-fruited tomato clade built by genetic makers (Figure S1B).

Two independent estimates of broad-sense heritability (H^2) revealed that 65.9% (656 of 980) of the metabolites displayed values greater than 0.5 (Figure S1C; Table S2). We additionally observed that 96.1% (942 of 980) of the coefficients of variation (CVs) were greater than 0.5, and the distribution of the CVs in the three subgroups was similar to that of the entire dataset (Figure S1D). However, we found that the phenotypic variation of the PIM group was substantially lower than *S lycopersicum* var. *cerasiforme* group (CER) ($p < 0.0045$) and *S. lycopersicum* group (BIG) ($p < 0.0039$), displaying mean CVs of 1.13, 1.22, and 1.24, respectively (Figure S1E), a result inconsistent with the level of genetic diversity in the sub-populations.

The transcriptomes of orange stage (about 75% ripe) fruit pericarp of 399 accessions were next analyzed in order to explore the relationship between gene expression and metabolites. Expression of a total of 30,326 genes was detected in the RNA sequencing (RNA-seq) dataset, accounting for 88.4% of the annotated genes. On average, 20,226 genes were detected in all of the samples (Figure S1F), whereas a total of 18,675 shared genes could be detected in 80% of the samples (Figure S1G). We found a total of 5,563 genes with a significant difference between the subgroups ($p < 2.2 \times 10^{-6}$, multiple test), including 2,964 PIM-CER, 3,655 CER-BIG, and 4,427 PIM-BIG differentially expressed genes. We applied a weighted correlation network analysis to find modules of highly correlated genes and identified 31 such modules (Figures S1H and S1I; Data S1). To find the potential biological and molecular connections, we imported modules into the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. While multiple metabolic pathways co-existed within single modules, we were able to identify some clear patterns (Data S1). For example, metabolism of multiple amino acids was enriched in module 2; phenylalanine biosynthesis was enriched in module 19.

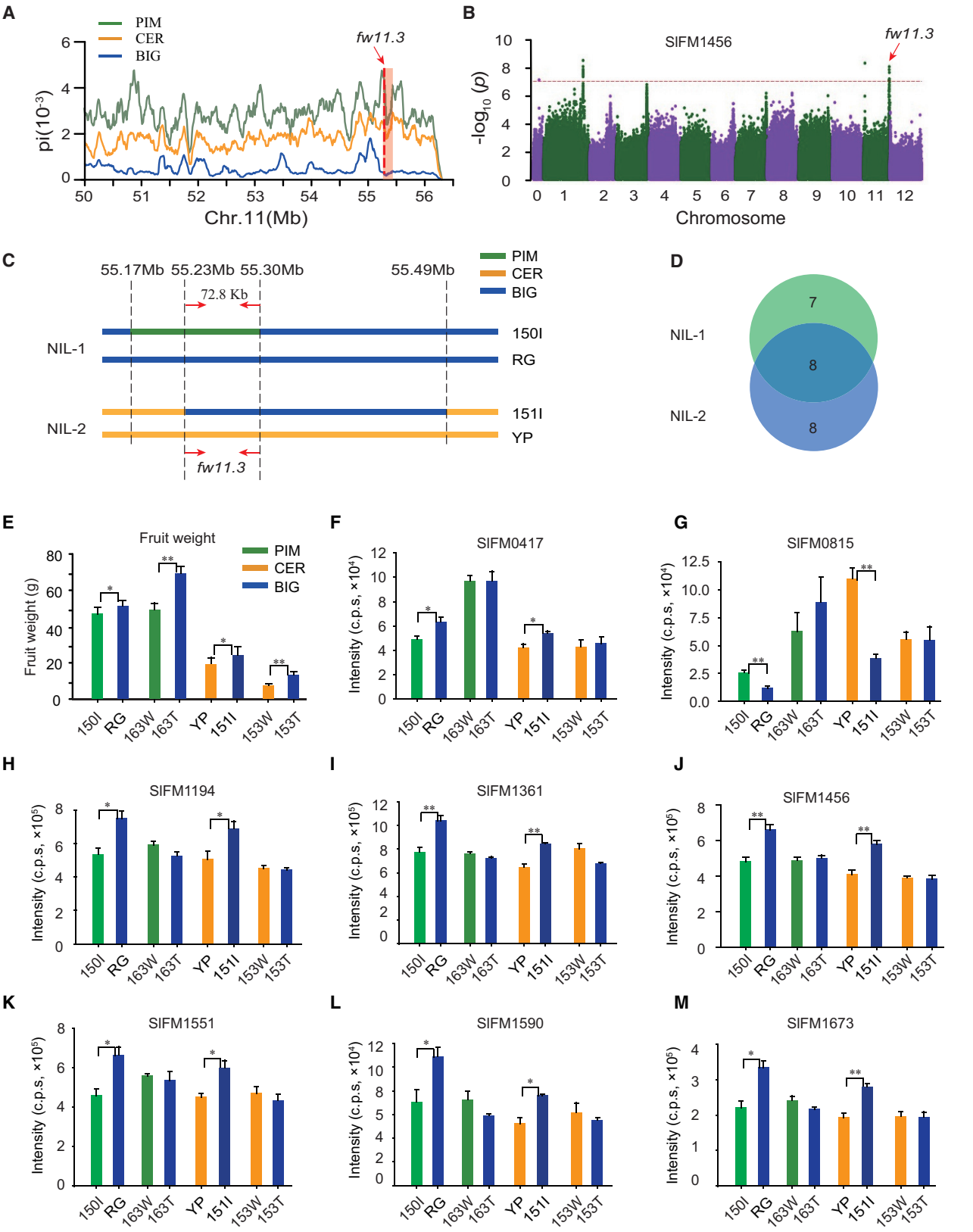
Correlations between the Variome, Transcriptome, and Metabolome

To uncover genetic control of the fruit metabolic traits, we selected 2,678,533 common SNPs (minor allele frequency [MAF] > 0.05 and missing ratio < 10%) in a 442 member population to perform mGWAS. A Bonferroni correction of $p = 9.05 \times 10^{-8}$ was employed as the genome-wide threshold for all trait associations, and a total of 3,526 signals for 514 metabolites was detected including 1,291 and 2,235 signals corresponding to 163 annotated and 351 unknown metabolites, respectively (Figure S2A; Data S1). The identified signals showed a significantly uneven distribution ($\chi^2 = 2{,}206.6$, $p < 4.9 \times 10^{-4}$) deviating from a random distribution across the genome and we found 102 potential hotspots (signals number > 9, permutation test, $p < 0.01$), predominantly located on chromosome 1 (chr 1) (Figure S2B). Some metabolites detected here were associated with previously identified flavor volatiles in mGWAS (Tieman et al., 2017). For example, 18 metabolites and 2 volatiles, 6-methyl-5-hepten-2-one and geranylacetone, were mapped to the same signal (ch03:4024959), which could contribute to the study of flavor-related pathways. In summary, we identified a large set of genetic loci controlling tomato fruit metabolites and the results will facilitate both functional verification of genes and elucidation of the metabolic networks facilitating tomato quality improvement.

Many genetic variants could influence phenotype by means of modulating gene expression. Therefore, to bridge the gap between the variome and transcriptome, we next tested correlations between genomic polymorphism and gene expression. To reduce false positives, the hidden confounding factors (Figure S2C) of expression variation and population co-variates were taken into account (Stegle et al., 2012). A total of 2,588,483 common SNPs (MAF > 0.05 and missing < 10%) and 22,480 genes (missing < 80%) was selected for further analyses. We detected a total of 434,809 SNPs significantly ($p < 8.6 \times 10^{-13}$, multiple test) associated with 3,465 genes using the linear module. eQTL were further subdivided into *cis*-eQTL and *trans*-eQTL according their distance (Michaelson et al., 2009). To reduce the repeatability of eQTL for certain genes, the leading SNP within a 30-kb interval was selected and defined as an eQTL (Figure S2D). A total of 2,566 *cis*-eQTL and 93,587 *trans*-eQTL was identified for 2,566 and 2,461 genes, respectively (Figure 1B; Data S1).

To find patterns linking the transcriptome and metabolome, correlations were calculated between the abundance of each metabolite and transcript. A rigorous multiple test correction, $p = 4.5 \times 10^{-8}$, was used to filter the genes that significantly correlated with each metabolite. A total of 232,934 expression-metabolite correlations involving 820 chemicals and 9,150 genes were identified.

Next, we integrated the above data by building a multi-omics network. The overlap of mGWAS and eQTL results generated 13,361 triple relationships (metabolite-SNP-gene) (Figure 1C; Table S3), which includes 371 metabolites, 970 SNPs, and 535 genes. This dataset thus facilitates both candidate gene identification and metabolic pathway elucidation. For example, one mGWAS signal (03:67080052) of the SGA hydroxytomatidenol (SlFM0964) was also supported by the eQTL of *Solyc03g118100*, an oxidoreductase gene that was previously reported to play an important role in SGAs biosynthesis (Umemoto et al., 2016). Further examples of insight derived from the data integration are provided in the examples below describing both SGA biosynthesis and regulation by *SlMYB12*.



(legend on next page)

Metabolome Alteration Subjected to Fruit-Mass-Targeted Selection

We previously showed that domestication (PIM-CER) and improvement (CER-BIG) targeted two independent sets of QTL and together made the modern tomato fruit ~100 times larger than its ancestor. Using the new variome map, we identified 168 domestication sweeps and 151 improvement sweeps covering 7.85% and 8.19% of the assembled genome and harboring 4,095 and 4,547 genes, respectively (Data S1). Human selection of such large portions of the genome would be expected to have a major influence on the chemical contents of fruits. Between PIM (fruit weight, 2.0 g) and CER (13.3 g), 389 metabolites were significantly different, whereas 614 were significantly different between CER and BIG (111.4 g) ($p < 0.05$) (Data S1), suggesting that selections during the improvement stage have had a larger impact on metabolite content than during the domestication stage.

To identify the altered fruit metabolites specific to fruit-mass targeted selection, we created two 500 member F_2 populations, one derived from a cross between a PIM and a CER accession and a second derived from a cross between a CER and a BIG accession (Figures S3A and S3B). QTL-seq and metabolite analyses were performed on pooled fruit from the 10% of plants producing the smallest and largest fruit in each F_2 population. Two regions corresponding to the defined fruit weight QTL *fw1.2* and *fw3.2* were identified in the segregating population derived from the PIM by CER cross (Figure S3C) and 343 metabolites were significantly ($p < 0.05$) different in the two corresponding bulked pools. Among these metabolites, 116 overlapped with those that were different between the PIM and CER groups (Figure S3D; Data S1). For the other bulked pools, four genomic regions that include *fw2.2*, *fw9.1*, *fw9.3*, and *fw11.1-fw11.2-fw11.3*, were identified as contributing to fruit weight (Figure S3E). In this instance, a total of 375 metabolites were significantly ($p < 0.05$) different based on fruit weight. Among these metabolites, 172 overlap with those that were different between the CER and BIG groups (Figure S3F; Data S1). During the improvement phase, 17 primary metabolites (five amino acid and derivatives and three vitamins and nine nucleotides derivatives) increased.

To recapitulate, ~30% (116/389) of the metabolites differentiating PIM and CER and ~28.0% (172/614) of the metabolites differentiating CER and BIG are likely associated with breeding for larger fruit. This result indicates that fruit mass-targeted selection led to considerable changes in the chemical composition of the fruit but does not account for the majority of the chemical differences associated with domestication and improvement.

Linkage to Fruit Weight Genes Contributes to Metabolite Alteration

The previously identified fruit weight genes enabled us to further explore which metabolites are changed in fruit mass targeted selection. To date, the three fruit weight genes *fw2.2*, *fw3.2*, and *fw11.3* have been cloned, and the corresponding larger fruit alleles are almost entirely fixed in the BIG group (Chakrabarti et al., 2013; Frary et al., 2000; Huang and van der Knaap, 2011). All three genes have undergone selection and are situated in sweep regions SW53, SW75, and SW255, respectively. A total of 53 signals for 51 metabolites were identified in the three sweeps and 47 signals for 47 metabolites were mapped in SW255 harboring *fw11.3* (Figures 2A and 2B).

To ascertain whether these metabolites were directly affected by the fruit weight genes, we focused on *fw11.3* since many more mGWAS signals were located within the sweep region and we anticipated it to be responsible for the largest number of metabolic changes. Two sets of near-isogenic lines (NILs) harboring different alleles for *fw11.3* were generated. One NIL (150I) harbors a PIM (LA1589) introgression (from 55.17 to 55.30 Mb) in a BIG (Rio Grande) background, while the other NIL (151I) harbors a BIG introgression from the variety Gold Ball Livingston (from 55.23 to 55.49 Mb on chr 11) in a CER (yellow pear) background (Figure 2C). The overlapping region was 72.8 kb from 55,228,328 to 55,301,104, where signals for 38 metabolites were mapped. The two NILs with different alleles displayed significantly different fruit weight compared to their near-isogenic parents (Figure 2E). Comparing the fruit metabolites in the lines and their respective near-isogenic backgrounds, we found 15 of the 38 metabolites were significantly ($p < 0.05$) different in NIL-1 and 16 in NIL-2 (Data S1). The 15 and 16 metabolites of the two NILs display an overlap of eight metabolites (Figure 2D), hypothesized to be driven by *fw11.3* selection.

To directly test whether *fw11.3* is the causative locus for the metabolite differences, we evaluated tomato lines that carried the derived allele of *fw11.3* (Mu et al., 2017). As expected, fruit weight was significantly altered in the transgenic plants, but none of the eight metabolites were significantly different. The same result was obtained when the derived allele was transformed into a PIM background (Figures 2F–2M; Data S1), indicating that *fw11.3* is not directly responsible for the altered metabolite levels. Correlation analysis between metabolite

Figure 2. Hitchhiking of Fruit Weight Locus *fw11.3* Altered Metabolites

(A) The *fw*11.3 locus is located within a sweep region with much lower nucleotide diversity in BIG than in either CER or PIM.

(B) Manhattan plot for the metabolite of SlFM1456. Metabolite content was genetically associated with the sweep harboring *fw*11.3.

(C) A 72.8-kb genomic region overlapped by two *fw11.3* NILs. Line 150I carries a PIM fragment in the RG (Rio Grande) background, and line 151I carries a BIG fragment in the YP (yellow pear) background.

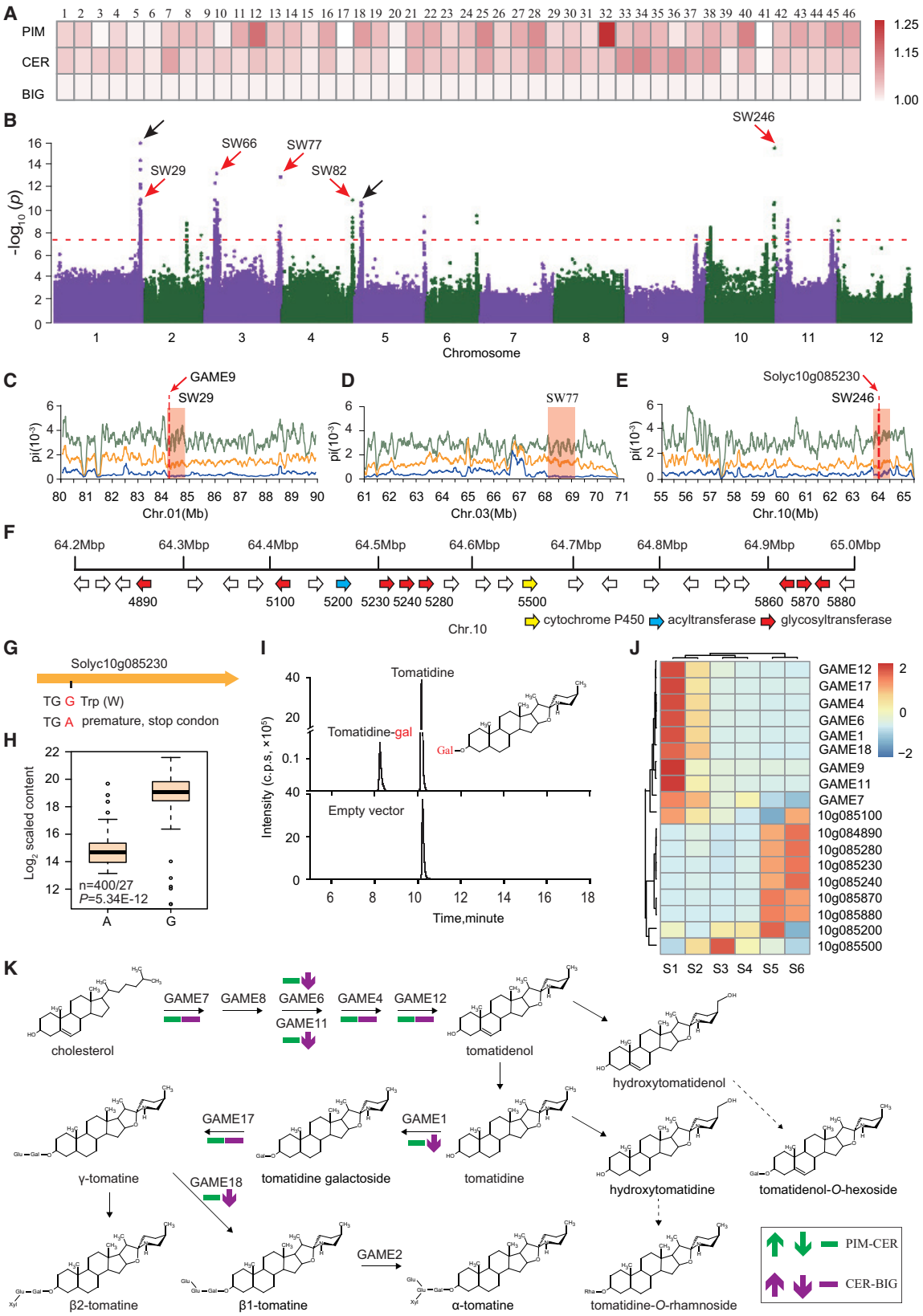
(D) NIL-1 lines differed by 15 metabolites, and the NIL-2 lines differed by 16 metabolites. Eight significantly different metabolites were present in the two *fw11.3* NILs.

(E) Fruit weight in *fw11.3* NILs and transgenic complementation lines. The difference was measured by Student's t test. * and ** indicate significant levels at the 0.05 and the 0.01 levels, respectively. W, wild-type; T, transgenic materials.

(F–M) Relative content of eight metabolites, SlFM0417 (F), SlFM0815 (G), SlFM1194 (H), SlFM1361 (I), SlFM1456 (J), SlFM1551 (K), SlFM1590 (L), and SlFM1673 (M) in *fw11.3* NILs and in transgenic complementation lines.

(E–M) Data represent mean ± SEM and was analyzed by Student's t test.

See also Data S1.



(legend on next page)

content and gene expression of the sweep regions denoted two nearest flanking gene of *fw11.3* that might be the causative genes, as the correlation coefficients for the seven metabolites were relatively high ($r = 0.14–0.18$, $p < 10^{-4}$) (Data S1). Whether that gene is indeed causative awaits further experimental validation. However, this analysis indicates that while many metabolites were changed during fruit mass based selection in domestication, as *fw11.3* illustrates, they are not necessarily caused by the fruit weight gene itself; rather, linked genes within the corresponding sweeps may be responsible.

Domestication of Steroidal Glycoalkaloids and Underlying Biosynthetic Network

SGAs have potent resistance properties, protecting the plant from predators, but are considered to be anti-nutritional in plants; however, this is only true for a subset of SGAs, while some exhibit positive nutritional properties (Friedman, 2002). In general, they are bitter and humans have selected for fruits with reduced SGA contents (Rick et al., 1994). Of the metabolites we annotated, a total of 46 are SGAs. Through intergroup comparison, only one SGA, SlFM0959 ($p < 0.05$), increased from PIM to CER, and all the others declined from PIM to BIG, consistent with strong negative selection for SGAs during domestication (Figure 3A). The combination of mGWAS and genomic selection analysis enabled us to determine how SGAs were domesticated. mGWAS identified seven major signals (two each on chr 1 and 3, one each on chr 4, 5, and 10) for 44 SGAs. Five out of the seven major signals were located within domestication/improvement sweeps (Figure 3B), including SW29, SW66, SW77, SW82, and SW246. *GAME9* is located within the sweep region of SW29 on chr 1 (Figure 3C). We discovered a SNP (ch01:84029382) within the CDS of *GAME9* that leads to a non-synonymous mutant (V-A) and associates with content alteration of eight SGAs. In addition, we found the frequency of the allele for lower SGA content increased from 0% in PIM, to 26.3% in CER, and to 57.3% in BIG (Figure S4A). These data indicate that *GAME9* is highly likely to have had an important role in SGA domestication. The sweep SW77 on chr 3 harbors 5 candidate genes including 2 cytochrome *c* oxidoreductases, 1 glycosyltransferase, and 2 ethylene-responsive transcription factors. (Figure 3D; Data S1).

In the SW246 sweep on chr 10, we discovered a new co-expression gene cluster that consists of 1 acetyltransferase, 1 cytochrome P450, 1 acyl-CoA dehydrogenase, and 7 UDP-glucosyltransferase genes, all of which could potentially be involved in SGA biosynthesis (Figures 3E and 3F). A SNP (ch01:64501127) that introduces a premature stop codon was found in the exon of a UDP-glycosyltransferase (*Solyc10 g085230*), and significant metabolic changes are associated with this SNP (Figures 3G and 3H). These findings thus complement the identification of this gene as being causal for tomatidine-hexose in a *S. pennellii* introgression line population and subsequent validation of its function via the virus induced gene-silencing approach (Alseekh et al., 2015). Furthermore, following the expression of *Solyc10 g085230* in *E. coli* BL-21, we could demonstrate that it converted tomatidine into tomatidine-galactose, providing a further confirmation of its function (Figure 3I). Interesting, this newly uncovered gene cluster has a markedly different expression pattern compared to the GAME1 cluster. The GAME1 cluster is predominantly expressed in immature fruits, whereas that reported here is predominantly expressed during ripening (Figure 3J). The mutations described above substantially decrease levels of SGAs beyond this stage, indicating that they play an important role in detoxifying the ripe fruits in readiness for consumption and hence seed dispersal. On investigating the transcriptome data, we found most of the cloned genes were downregulated in BIG rather than CER, a fact that is consistent with the reduction of SGA contents during improvement (Figure 3K).

The identified causative variants and effective SNPs exist within the natural tomato germplasm and could readily be adopted in marker-assisted breeding strategies. We examined three major SGAs (SlFM1785, SlFM1885 and SlFM1985) and two loci (SW29 and SW246) as examples. Selection for low SGA alleles at both SW29 and SW246 can reduce the three SGAs to 20.9% (Figure S4B). Pyramiding of these loci with other high value loci would, therefore, likely mitigate any deleterious or anti-nutritional effects of using wild germplasm in breeding.

Metabolome Divergence between Pink and Red Tomatoes

Following domestication and improvement, breeders developed different types of tomatoes based on human preference, usage and local climates. In general, cultivated tomatoes can be separated into fresh market and processing types. Within the fresh

Figure 3. Domesticated SGAs and Related Metabolic Pathway

(A) Heatmap of all SGAs detected in this study. The relative values of SGAs content were scaled to the BIG group for each chemical.

(B) Genomic distribution of major signals for all SGAs detected. Red arrows denote signals located within sweeps regions.

(C–E) Nucleotide diversity of three groups and the major signals on chr 1 (C), chr 3 (D), and chr 10 (E). The sweeps and genes are denoted by pink shading and red bars, respectively.

(F) Schematic map of genes identified in the duplicated genomic regions in chr 10. Specific gene families are indicated by colored arrows, and the other gene families are shown by white arrows. The numbers under each arrow are the last four digits of the gene ID.

(G) The nonsense mutant produces a stop codon and prematurely truncated protein.

(H) The effect of different alleles on the content of one SGA. A truncated protein resulted in significantly reduced content of SlFM1885 compared to wild-type.

(I) Validation of gene function *in vitro*.

(J) Expression pattern of clustered genes in Chr10 and previously identified GAME genes. The FPKM of each gene was scaled to the maximum value of all stages. Gene expressions of Heinz 1706 fruit are shown at different stages: S1, 1 cm; S2, 2 cm; S3, 3 cm; S4, mature green; S5, break; S6, break +10 days. (Heinz 1706 expression data is from the database Tomato Functional Genomics Database [TFGD].)

(K) Expression of structural genes in the SGA biosynthesis pathway from cholesterol to dehydrolycoperoside. The green and the purple arrows represent the gene expression shift from PIM to CER and CER to BIG, respectively.

See also Figure S4 and Data S1.

market type, pink tomatoes are preferred by some consumers and have become especially popular in Asian countries. Genetic and biochemical evidence has demonstrated that *SlMYB12*, a transcription factor that is a key regulator of the flavonoid biosynthesis pathway, is responsible for the red phenotype (Adato et al., 2009; Ballester et al., 2010). Our previous studies demonstrated that deletion of a *cis*-acting element in the transcriptional promoter and several nonsense mutations within the coding sequence together inactivate gene function resulting in a colorless peel, due to the absence of flavonoids (Lin et al., 2014).

To uncover the full range of metabolites associated with the phenotype, we compared all pericarp metabolites in 44 pink BIG lines and 191 red BIG lines, identifying 122 metabolites that were significantly different ($p < 10^{-3}$, multiple test correction) (Figure 4A; Data S1). Mining the 13,361 triple relationships, we found that the signature SNP of *SlMYB12* (SNP_y) was associated with 56 metabolites via mGWAS and 69 genes via eQTL (Figure 4B). All 56 metabolites are included within the 122 metabolites identified by pink-red comparison.

Next, we investigated which genes are directly or indirectly regulated by *SlMYB12* and are therefore responsible for the observed metabolic profile of the pink tomato. Two CRISPR/Cas9 constructs with different single-guide RNAs (sgRNAs) were created and transformed into the red cultivar Moneymaker. Four independent homozygous transformed lines with different mutations were obtained (Figure 4C). In these plants, the pink fruit phenotype indicates that knockout of *SlMYB12* was achieved (Figure 4D). Since *SlMYB12* is predominantly expressed in peel at the turning stage, we also isolated peel tissue and quantified the transcripts of orange stage fruits. Within the peel tissue, we found 869 significant alterations in gene expression ($p < 0.01$, at least 2-fold change in expression) including 658 up- and 211 downregulated genes in common to all mutant lines. As expected, most (49 out of 69), of the genes that were subject to regulation by *SlMYB12* using eQTL analysis are included within the 869 genes. Interestingly, 18 of the 44 genes that were reported to be directly regulated by *AtMYB12* in ChIP (chromatin immunoprecipitation assay) experiments (Zhang et al., 2015) were eQTL of *SlMYB12* (Figure 4E; Data S1), supporting the notion that they are direct downstream targets of *SlMYB12*. Integration of these data provides new insights into the regulation network of *SlMYB12*.

Relative to the wild-type fruit, 152 metabolites were altered in the peel of all four knockout mutants ($p < 0.05$; Table S4). The decrease in content of the major flavonoids is consistent with the transparent peel of pink tomatoes. Some other chemicals, including 16 glycoalkaloids, 12 polyamines, 5 polyphenols, and 11 primary metabolites were also changed (Figure 4F), indicating that *SlMYB12* directly or indirectly influences more than just flavonoid metabolism. Among the 12 changed polyamines, our mGWAS analysis of SlFM0516 and SlFM0756 identified two polyamine biosynthesis genes, *N*-acetyltransferase (*Solyc05g041860*) and caffeoyl-CoA *O*-methyltransferase (*Solyc04g063210*), that were key enzymes of polyamines biosynthesis pathway (Grienenberger et al., 2009; Peng et al.,2016). In addition, their expression was also significantly altered in the knockout mutants (Figure 4G). A causal relationship between *SlMYB12* and the content of these other metabolites awaits further experimentation. All the above data pinpoint *SlMYB12* as a major hub gene of tomato fruit metabolism and an excellent example of how selection for one trait can have a major impact on seemingly unrelated traits that can potentially impact fruit quality.

Influence of Wild Introgressions on Fruit Metabolome

In recent decades, wild relatives of tomato have been used to introduce alleles into elite cultivars, particularly those relating to biotic stress resistance, including, among others, *Tm-2*^{a} (tomato mosaic virus resistance gene) from *S. peruvianum* and *Ty-1*/*Ty-3* from *S. chilense* (Verlaan et al., 2013). Many of these resistance genes are derived from inedible green-fruited species and linkage drag might be expected to negatively impact fruit quality especially in wide crosses. In a previous study (Lin et al., 2014), we delimited the regions of introgressed fragments involving some of these genes. However, how exotic introgressions change the fruit metabolome has not yet been determined. Here, we used the *S. peruvianum*-derived *Tm-2*^{a} as a case study.

We identified 11 accessions within the BIG group harboring the exotic introgression on chr 9 with a shared genomic fragment from 7.45 to 62.70 Mb. To identify which metabolites were changed by the *Tm-2*^{a} introgression in big-fruited cultivars, a segregating population derived from a cross between a *Tm-2*^{a}-harboring inbred line and a susceptible line was analyzed. After genotypic confirmation of the progeny, those individuals displaying homozygous resistant (R) and susceptible (S) genotypes were pooled (about 30 lines each). We found that 346 metabolites were significantly ($p < 0.05$) altered between the S and R pools. By comparing the metabolite profiles of the PIM accessions ($n = 31$) and the BIG accessions without the *Tm-2*^{a} introgression ($n = 276$), we identified 589 metabolites that were significantly changed during the domestication process ($p < 0.05$). Among these metabolites, we identified 52 metabolites that were increased in the R pool and were decreased from PIM to BIG, as well as 75 metabolites that were decreased in the R pool and were increased from PIM to BIG. The fate of these 127 metabolites is, therefore, reversed by the resistance breeding using wild introgression (Data S1).

DISCUSSION

In the broadest sense, the metabolome is what we eat from a tomato fruit that determines the nutritional and consumption value of this important crop. Metabolomics analyses have been applied to *Arabidopsis* and some crops to reveal the natural variations in chemical composition (Luo, 2015). However, the changes in metabolism during the domestication process have been rarely studied. To our knowledge, this question has been only addressed at the metabolome level in a single study, but the genetic variations underlying the metabolite changes were not explored (Beleggia et al., 2016). In the current study, we provide the first multi-omics data to understand the impact of human intervention on chemical composition of crop.

A Multi-omics Dataset for Plant Metabolic Biology

The multi-omics data generated in this study provides a valuable resource for further studies on biosynthetic pathways and regulatory circuits of plant metabolites. Owing to the limitations of single-data-type approaches, combining multiple datasets can

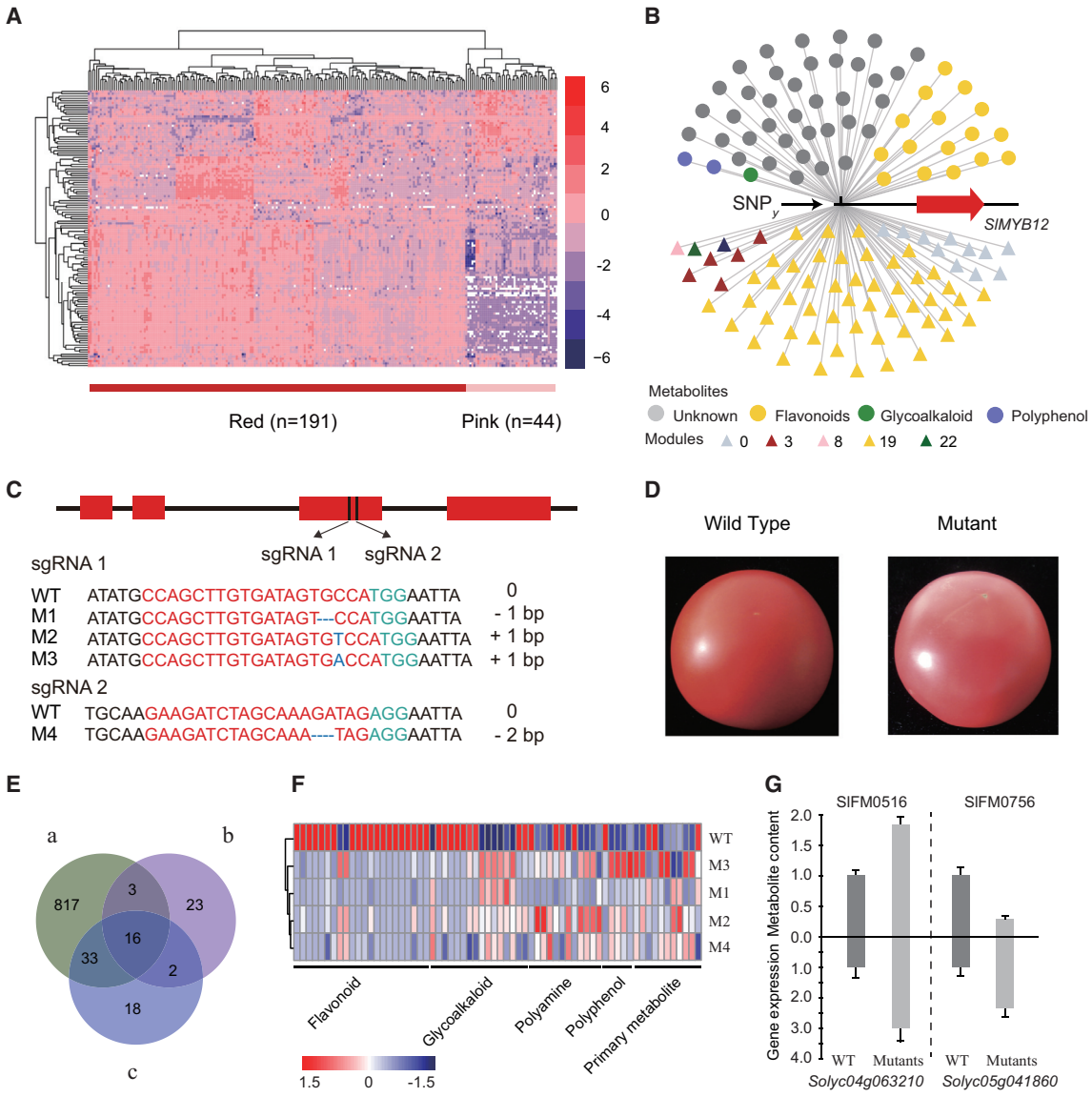


Figure 4. Metabolites and Pathway Affected by *SlMYB12*

(A) Heatmap of 122 significant differentially accumulating metabolites in the red and the pink populations. Red (n = 191) and pink (n = 40) accessions could be distinctly separated into two clusters.

(B) Triple relationships centered on SNP$_y$. A total of 56 metabolites and 69 genes were identified by mGWAS and eQTL analyses.

(C) Generation of *SlMYB12* mutations by CRISPR/Cas9 using two independent single-guide RNAs (sgRNA1 and sgRNA2). Sequences of *SlMYB12* mutant M1-M4 are shown. sgRNA targets and a protospacer-adjacent motif (PAM) are indicated in red and in green, respectively. Deletions and insertions are indicated by dashes and in blue, respectively.

(D) Fruit color of wild-type and *SlMYB12* mutant.

(E) Cross-validation between three independent datasets. a, 869 genes with at least 2-fold differential expression between *SlMYB12* mutant and wild-type in peel tissue. b, a total of 44 genes that were reported to be directly regulated by *AtMYB12* in ChIP. c, a total of 69 genes potentially subject to regulation by *SlMYB12* using eQTL.

(F) Altered metabolites in fruit peel of *SlMYB12* mutants. 152 annotated chemicals were significantly changed in all four mutant lines, including 16 steroidal glycoalkaloids, 18 flavonoids, 12 polyamines, 5 polyphenols, and 11 primary metabolites.

(G) Two changed polyamines and corresponding differentially expressed genes in the peel of mutant lines. The metabolite contents and gene expression values of mutant lines were scaled to the wild-type values. Data represent mean ± SEM and was analyzed by Student's t test.

See also Tables S3 and S4 and Data S1.

compensate for missing or unreliable information in any single data type (Ritchie et al., 2015). Multi-dimensional analysis and multi-staged analysis have been increasingly used to provide clues for understanding biological mechanisms (Holzinger and Ritchie, 2012). Here, we have applied multi-dimensional analysis of genomes, transcriptomes, and metabolomes to provide leads for discovery of candidate genes contributing to metabolic pathways, including those of flavonoids and SGA metabolism. These examples provide clear highlights how the study of extensive natural genetic variance will greatly aid attempts to improve the coverage of metabolomics.

These relationships resulted from this multi-omics analyses will greatly facilitate large-scale interactive gene-metabolites annotation and pathway elucidation. The associations between 278 known metabolites and 7,478 genes will enrich the annotation of genes by metabolic functions. These data could also be used to annotate unknown chemicals. For example, both SlFM1290 and SlFM1379 were associated with a diacylglycerol kinase (*Solyc12g005380*), which participates in the fatty acid biochemical process. Among them, SlFM1290 was annotated as lyso phosphatidylcholines (1-acyl 18:2), but SlFM1379 was an unknown chemical. By comparing their retention time and fragmentation patterns, we tentatively annotated SlFM1379 as a fatty acid, phosphatidylcholines (O-18:1(9Z)/2:0). Further investigations shall lead to the annotation of more unknown metabolites.

The Metabolome Is Shaped by Direct and Indirect Selection

Genetically, the PIM group is ancestral to and more diverse than the CER and BIG groups. However, we found that the metabolic diversity of the CER and BIG groups was higher than that of PIM, inconsistent with the corresponding genetic diversity. Many more metabolites were changed during the improvement stage than in the domestication stage, possibly the consequence of stronger selection (measured by genetic diversity ratio) or larger phenotypic variation (van der Knaap and Tanksley, 2003). One explanation could be that constraints on the phenotype in the wild population might to be released or "hidden" gene effects manifested in an agricultural context (Kalisz and Kramer, 2008). Alternatively, phenotypic diversification might be caused by targeted selection. For example, the distinct traits selected for processing tomatoes are very different from those of round shaped fresh market tomatoes (Tanksley, 2004).

The impact of tomato breeding on the fruit metabolome can be summarized in two contexts. The first is associated with direct selection, as with domestication of bitter chemicals such as the SGAs. Without direct knowledge of SGAs, humans instinctively selected for less bitter fruit, progressively selecting plants with tastier fruits and consequently, reducing the levels of the more bitter SGAs. A similar story has emerged in cucumber when ancient native collectors selected non-bitter lines that carried a lowly expressed allele of the fruit bitterness gene *Bt* (Shang et al., 2014). The second context for selection is indirect and includes the hitchhiking of metabolic genes with fruit weight genes as ever larger fruits were selected as well as the linkage drag associated with *R* genes that have been introgressed from wild relatives. During fruit-mass targeted selection, hundreds of metabolites (measured by dry weight) were changed. Among the changed metabolites, the increased primary metabolite content between BIG (big-pool) and PIM (small-pool) might be the consequence of a larger metabolic sink in domesticated fruits. Evidence provided here following the generation of two sets of NILs strongly suggests that many, if not most of these metabolic changes may not be caused by the fruit weight genes themselves but rather, as the results with *fw11.3* indicate, be the consequence of linked genes. The same phenomenon was also observed in maize and rice (Olsen et al., 2006; Palaisa et al., 2004), albeit for different phenotypic traits. These results point to the possibility that significant changes affecting flavor and nutritional alterations linked to selection for larger fruit may be able to be corrected with precision molecular breeding.

Metabolome-Assisted Breeding in Tomato

Modern tomato breeding has focused considerable effort on yield, shelf life, and resistance to disease, while flavor has been relatively neglected. The contributions of sugars, acids, and volatiles to flavor have been well characterized, while the contribution of other metabolites detected by LC/MS to tomato flavor have not been as extensively evaluated. In Asian countries, there is a public perception that pink-fruited tomatoes are tastier. It is generally recognized that the difference between red and pink fruits was due to the lack of the yellow-hued naringenin chalcone in the peel of pink fruit (Adato et al., 2009; Ballester et al., 2010). Previously, using consumer taste panels and targeted metabolomics (Tieman et al., 2017), we identified 34 metabolites that are significantly correlated with consumer preference ($p < 0.05$). Here, using a widely targeted approach, we identified a large number of metabolites with significant differences between red and pink varieties, many of which are present in the pericarp tissue. It is reasonable to assume that at least some of these chemicals also influence taste preferences. The molecular markers linked with these genomic variations may thus be useful for metabolome-assisted breeding.

Throughout the history of plant breeding, phenotype-targeted selection has been the foundation of modern agriculture. Modern molecular tools offer the opportunity to improve crops with great precision. High throughput molecular breeding combined with precision genome editing has huge potential to accelerate crop improvement (Soyk et al., 2017; Townsend et al., 2009), and reduce or eliminate linkage drag. The work presented here illustrates examples in which linked genes, not the target introgressed gene, clearly have major effects on the fruit metabolome. This phenomenon of undesired alterations as a consequence of movement of desirable alleles, particularly from wild relatives of tomato, at the least illustrates the need to incorporate the latest genome information into molecular breeding strategies but also should focus attention on the advantages of genome editing to precisely alter specific traits.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING

- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Plant material and growth conditions
- METHOD DETAILS
  - Samples preparation
  - Metabolite profiling
  - SNP identification and annotation
  - Phylogenetic analysis and population structure
  - Read mapping and Expression profiling
  - Co-expression modules identification and KEGG enrichment for transcriptome
  - Genome-wide association analysis and linkage disequilibrium
  - Detection of mGWAS signals hotspots
  - Normal quantile transformation of expression and eQTL analysis
  - Network building for metabolome, transcriptome and variome
  - Detection of domestication and improvement sweeps
  - Bulked segregant analysis of the F_2 population by whole-genome resequencing
  - Glycosyltransferase assay
  - CRISPR/Cas9 constructs design
- QUANTIFICATION AND STATISTICAL ANALYSES
- DATA AND SOFTWARE AVAILABILITY
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures, four tables, and one dataset and can be found with this article online at https://doi.org/10.1016/j.cell.2017.12.019.

ACKNOWLEDGMENTS

We thank Solab lab members for preparing fruit samples and Dr. Jiabing Ji from the Agricultural Genomics Institute at Shenzhen for his suggestions on the manuscript. We thank Ruowang Li from the University of Pennsylvania for data analysis. The work was supported by China National Key Research and Development Program for Crop Breeding (2016YFD0100307 to S.H.), National Science Fund for Distinguished Young Scholars (31625021 to J.L.; 31225025 to S.H.), National Natural Science Foundation of China (31530066 to S.H.; 31530052 to J.L.; 31601756 to G.Z.), the "973" Program (2012CB113900 to S.H.), the funding of the PlantaSYST Project by the European Union's Horizon 2020 Research and Innovation Programme (SGA-CSA 664621; 739582 under FPA no. 664620), and National Science Foundation (IOS-1539831 to H.K.). This work was also supported by the Chinese Academy of Agricultural Science (ASTIP-CAAS and CAAS-XTCX2016001), Leading Talents of Guangdong Province Program (00201515 to S.H.), and the Shenzhen municipal (The Peacock Plan KQTD2016113010482651 to S.H.) and Dapeng district governments.

AUTHOR CONTRIBUTIONS

S.H. and J.L. conceived and designed the research. G.Z., X.H., X. Cao, and C.Y. participated in the material preparation. S.W. conducted metabolic profiling. Z. Huang and E.v.d.K. created NILs and transgenic lines of *fw11.3*. S. Zhang, C.Z., and X. Cui performed plasmid construction and genetic transformation. M.P. carried out validation *in vitro*. Q.L., M.Q., and Z.Z. performed the RNA-seq analysis. X.W. and G.Z. created the genetic population. T.L. and X.H. performed the SNP calling. G.Z., S.W., S.H., J.L., Z.Z., H.K., and A.R.F. analyzed the data. G.Z., S.W., S.H., and J.L. wrote the manuscript. A.R.F., H.K., and X. Cui revised the manuscript.

DECLARATION OF INTERESTS

All authors declare no competing interests.

Received: July 31, 2017
Revised: October 3, 2017
Accepted: December 15, 2017
Published: January 11, 2018

REFERENCES

Adato, A., Mandel, T., Mintz-Oron, S., Venger, I., Levy, D., Yativ, M., Domínguez, E., Wang, Z., De Vos, R.C., Jetter, R., et al. (2009). Fruit-surface flavonoid accumulation in tomato is controlled by a SlMYB12-regulated transcriptional network. PLoS Genet. *5*, e1000777.

Aflitos, S., Schijlen, E., de Jong, H., de Ridder, D., Smit, S., Finkers, R., Wang, J., Zhang, G., Li, N., Mao, L., et al.; 100 Tomato Genome Sequencing Consortium (2014). Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. Plant J. *80*, 136–148.

Alseekh, S., Tohge, T., Wendenberg, R., Scossa, F., Omranian, N., Li, J., Kleessen, S., Giavalisco, P., Pleban, T., Mueller-Roeber, B., et al. (2015). Identification and mode of inheritance of quantitative trait loci for secondary metabolite abundance in tomato. Plant Cell *27*, 485–512.

Ballester, A.R., Molthoff, J., de Vos, R., Hekkert, Bt., Orzaez, D., Fernández-Moreno, J.P., Tripodi, P., Grandillo, S., Martin, C., Heldens, J., et al. (2010). Biochemical and molecular analysis of pink tomatoes: deregulated expression of the gene encoding transcription factor SlMYB12 leads to pink tomato fruit color. Plant Physiol. *152*, 71–84.

Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. (2005). Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics *21*, 263–265.

Beleggia, R., Rau, D., Laidò, G., Platani, C., Nigro, F., Fragasso, M., De Vita, P., Scossa, F., Fernie, A.R., Nikoloski, Z., and Papa, R. (2016). Evolutionary metabolomics reveals domestication-associated changes in tetraploid wheat kernels. Mol. Biol. Evol. *33*, 1740–1753.

Butelli, E., Titta, L., Giorgio, M., Mock, H.P., Matros, A., Peterek, S., Schijlen, E.G., Hall, R.D., Bovy, A.G., Luo, J., and Martin, C. (2008). Enrichment of tomato fruit with health-promoting anthocyanins by expression of select transcription factors. Nat. Biotechnol. *26*, 1301–1308.

Cárdenas, P.D., Sonawane, P.D., Pollier, J., Vanden Bossche, R., Dewangan, V., Weithorn, E., Tal, L., Meir, S., Rogachev, I., Malitsky, S., et al. (2016). GAME9 regulates the biosynthesis of steroidal alkaloids and upstream isoprenoids in the plant mevalonate pathway. Nat. Commun. *7*, 10654.

Carrari, F., Baxter, C., Usadel, B., Urbanczyk-Wochniak, E., Zanor, M.I., Nunes-Nesi, A., Nikiforova, V., Centero, D., Ratzka, A., Pauly, M., et al. (2006). Integrated analysis of metabolite and transcript levels reveals the metabolic shifts that underlie tomato fruit development and highlight regulatory aspects of metabolic network behavior. Plant Physiol. *142*, 1380–1396.

Chakrabarti, M., Zhang, N., Sauvage, C., Muños, S., Blanca, J., Cañizares, J., Diez, M.J., 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.

Chen, W., Gong, L., Guo, Z., Wang, W., Zhang, H., Liu, X., Yu, S., Xiong, L., and Luo, J. (2013). A novel integrated method for large-scale detection, identification, and quantification of widely targeted metabolites: application in the study of rice metabolomics. Mol. Plant *6*, 1769–1780.

Chen, W., Wang, W., Peng, M., Gong, L., Gao, Y., Wan, J., Wang, S., Shi, L., Zhou, B., Li, Z., et al. (2016). Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals. Nat. Commun. *7*, 12767.

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

DeLuca, D.S., Levin, J.Z., Sivachenko, A., Fennell, T., Nazaire, M.D., Williams, C., Reich, M., Winckler, W., and Getz, G. (2012). RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics *28*, 1530–1532.

Eschenfeldt, W.H., Lucy, S., Millard, C.S., Joachimiak, A., and Mark, I.D. (2009). A family of LIC vectors for high-throughput cloning and purification of proteins. Methods Mol. Biol. *498*, 105–115.

Fernie, A.R., Trethewey, R.N., Krotzky, A.J., and Willmitzer, L. (2004). Metabolite profiling: from diagnostics to systems biology. Nat. Rev. Mol. Cell Biol. *5*, 763–769.

Frary, A., Nesbitt, T.C., Grandillo, S., Knaap, E., Cong, B., Liu, J., Meller, J., Elber, R., Alpert, K.B., and Tanksley, S.D. (2000). *fw2.2*: a quantitative trait locus key to the evolution of tomato fruit size. Science *289*, 85–88.

Friedman, M. (2002). Tomato glycoalkaloids: role in the plant and in the diet. J. Agric. Food Chem. *50*, 5751–5780.

Grienenberger, E., Besseau, S., Geoffroy, P., Debayle, D., Heintz, D., Lapierre, C., Pollet, B., Heitz, T., and Legrand, M. (2009). A BAHD acyltransferase is expressed in the tapetum of Arabidopsis anthers and is involved in the synthesis of hydroxycinnamoyl spermidines. Plant J. *58*, 246–259.

Holzinger, E.R., and Ritchie, M.D. (2012). Integrating heterogeneous high-throughput data for meta-dimensional pharmacogenomics and disease-related studies. Pharmacogenomics *13*, 213–222.

Huang, Z., and van der Knaap, E. (2011). Tomato fruit weight 11.3 maps close to fasciated on the bottom of chromosome 11. Theor. Appl. Genet. *123*, 465–474.

Iijima, Y., Fujiwara, Y., Tokita, T., Ikeda, T., Nohara, T., Aoki, K., and Shibata, D. (2009). Involvement of ethylene in the accumulation of esculeoside A during fruit ripening of tomato (Solanum lycopersicum). J. Agric. Food Chem. *57*, 3247–3252.

Itkin, M., Heinig, U., Tzfadia, O., Bhide, A.J., Shinde, B., Cardenas, P.D., Bocobza, S.E., Unger, T., Malitsky, S., Finkers, R., et al. (2013). Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. Science *341*, 175–179.

Kalisz, S., and Kramer, E.M. (2008). Variation and constraint in plant evolution and development. Heredity (Edinb) *100*, 171–177.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360.

Kusano, M., Tabuchi, M., Fukushima, A., Funayama, K., Diaz, C., Kobayashi, M., Hayashi, N., Tsuchiya, Y.N., Takahashi, H., Kamata, A., et al. (2011). Metabolomics data reveal a crucial role of cytosolic glutamine synthetase 1;1 in coordinating metabolic balance in rice. Plant J. *66*, 456–466.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009a). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K., and Wang, J. (2009b). SNP detection for massively parallel whole-genome resequencing. Genome Res. *19*, 1124–1132.

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., and Wang, J. (2009c). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics *25*, 1966–1967.

Li, M.X., Yeung, J.M., Cherny, S.S., and Sham, P.C. (2012). Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. Hum. Genet. *131*, 747–756.

Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., Zhang, Z., Lun, Y., Li, S., Wang, X., et al. (2014). Genomic analyses provide insights into the history of tomato breeding. Nat. Genet. *46*, 1220–1226.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. Nat. Methods *8*, 833–835.

Luo, J. (2015). Metabolite-based genome-wide association studies in plants. Curr. Opin. Plant Biol. *24*, 31–38.

Matsuda, F., Hirai, M.Y., Sasaki, E., Akiyama, K., Yonekura-Sakakibara, K., Provart, N.J., Sakurai, T., Shimada, Y., and Saito, K. (2010). AtMetExpress development: a phytochemical atlas of Arabidopsis development. Plant Physiol. *152*, 566–578.

Michaelson, J.J., Loguercio, S., and Beyer, A. (2009). Detection and interpretation of expression quantitative trait loci (eQTL). Methods *48*, 265–276.

Mu, Q., Huang, Z., Chakrabarti, M., Illa-Berenguer, E., Liu, X., Wang, Y., Ramos, A., and van der Knaap, E. (2017). Fruit weight is controlled by cell size regulator encoding a novel protein that is expressed in maturing tomato fruits. PLoS Genet. *13*, e1006930.

Muir, S.R., Collins, G.J., Robinson, S., Hughes, S., Bovy, A., Ric De Vos, C.H., van Tunen, A.J., and Verhoeyen, M.E. (2001). Overexpression of petunia chalcone isomerase in tomato results in fruit containing increased levels of flavonols. Nat. Biotechnol. *19*, 470–474.

Olsen, K.M., Caicedo, A.L., Polato, N., McClung, A., McCouch, S., and Purugganan, M.D. (2006). Selection under domestication: evidence for a sweep in the rice waxy genomic region. Genetics *173*, 975–983.

Palaisa, K., Morgante, M., Tingey, S., and Rafalski, A. (2004). Long-range patterns of diversity and linkage disequilibrium surrounding the maize *Y1* gene are indicative of an asymmetric selective sweep. Proc. Natl. Acad. Sci. USA *101*, 9885–9890.

Peng, M., Gao, Y., Chen, W., Wang, W., Shen, S., Shi, J., Wang, C., Zhang, Y., Zou, L., Wang, S., et al. (2016). Evolutionarily distinct BAHD *N*-acyltransferases are responsible for natural variation of aromatic amine conjugates in rice. Plant Cell *28*, 1533–1550.

Retief, J.D. (2000). Phylogenetic analysis using PHYLIP. Methods Mol. Biol. *132*, 243–258.

Rick, C.M., Uhlig, J.W., and Jones, A.D. (1994). High alpha-tomatine content in ripe fruit of Andean Lycopersicon esculentum var. cerasiforme: developmental and genetic aspects. Proc. Natl. Acad. Sci. USA *91*, 12877–12881.

Rio, D.C., Ares, M., Jr., Hannon, G.J., and Nilsen, T.W. (2010). Purification of RNA using TRIzol (TRI reagent). Cold Spring Harb. Protoc. *2010*, pdb.prot5439.

Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A., and Kim, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. Nat. Rev. Genet. *16*, 85–97.

Ron, M., Kajala, K., Pauluzzi, G., Wang, D., Reynoso, M.A., Zumstein, K., Garcha, J., Winte, S., Masson, H., Inagaki, S., et al. (2014). Hairy root transformation using Agrobacterium rhizogenes as a tool for exploring cell type-specific gene expression and function using tomato as a model. Plant Physiol. *166*, 455–469.

Saito, K., and Matsuda, F. (2010). Metabolomics for functional genomics, systems biology, and biotechnology. Annu. Rev. Plant Biol. *61*, 463–489.

Sauvage, C., Segura, V., Bauchet, G., Stevens, R., Do, P.T., Nikoloski, Z., Fernie, A.R., and Causse, M. (2014). Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. Plant Physiol. *165*, 1120–1132.

Schwahn, K., de Souza, L.P., Fernie, A.R., and Tohge, T. (2014). Metabolomics-assisted refinement of the pathways of steroidal glycoalkaloid biosynthesis in the tomato clade. J. Integr. Plant Biol. *56*, 864–875.

Shabalin, A.A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics *28*, 1353–1358.

Shang, Y., Ma, Y., Zhou, Y., Zhang, H., Duan, L., Chen, H., Zeng, J., Zhou, Q., Wang, S., Gu, W., et al. (2014). Plant science. Biosynthesis, regulation, and domestication of bitterness in cucumber. Science *346*, 1084–1088.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment

for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.

Soyk, S., Müller, N.A., Park, S.J., Schmalenbach, I., Jiang, K., Hayama, R., Zhang, L., Van Eck, J., Jiménez-Gómez, J.M., and Lippman, Z.B. (2017). Variation in the flowering gene SELF PRUNING 5G promotes day-neutrality and early yield in tomato. Nat. Genet. *49*, 162–168.

Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat. Protoc. *7*, 500–507.

Takagi, H., Abe, A., Yoshida, K., Kosugi, S., Natsume, S., Mitsuoka, C., Uemura, A., Utsushi, H., Tamiru, M., Takuno, S., et al. (2013). QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. Plant J. *74*, 174–183.

Tanksley, S.D. (2004). The genetic, developmental, and molecular bases of fruit size and shape variation in tomato. Plant Cell *16* (*Suppl*), S181–S189.

Tieman, D., Zhu, G., Resende, M.F., Jr., Lin, T., Nguyen, C., Bies, D., Rambla, J.L., Beltran, K.S., Taylor, M., Zhang, B., et al. (2017). A chemical genetic roadmap to improved tomato flavor. Science *355*, 391–394.

Townsend, J.A., Wright, D.A., Winfrey, R.J., Fu, F., Maeder, M.L., Joung, J.K., and Voytas, D.F. (2009). High-frequency modification of plant genes using engineered zinc-finger nucleases. Nature *459*, 442–445.

Umemoto, N., Nakayasu, M., Ohyama, K., Yotsu-Yamashita, M., Mizutani, M., Seki, H., Saito, K., and Muranaka, T. (2016). Two cytochrome P450 monooxygenases catalyze early hydroxylation steps in the potato steroid glycoalkaloid biosynthetic pathway. Plant Physiol. *171*, 2458–2467.

van der Knaap, E., and Tanksley, S.D. (2003). The making of a bell pepper-shaped tomato fruit: identification of loci controlling fruit morphology in yellow stuffer tomato. Theor. Appl. Genet. *107*, 139–147.

Verlaan, M.G., Hutton, S.F., Ibrahem, R.M., Kormelink, R., Visser, R.G., Scott, J.W., Edwards, J.D., and Bai, Y. (2013). The tomato yellow leaf curl virus resistance genes *Ty-1* and *Ty-3* are allelic and code for DFDGD-class RNA-dependent RNA polymerases. PLoS Genet. *9*, e1003399.

Wishart, D.S., Tzur, D., Knox, C., Eisner, R., Guo, A.C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., et al. (2007). HMDB: the human metabolome database. Nucleic Acids Res. *35*, D521–D526.

Wurtzel, E.T., and Kutchan, T.M. (2016). Plant metabolism, the diverse chemistry set of the future. Science *353*, 1232–1236.

Zhang, Y., Butelli, E., Alseekh, S., Tohge, T., Rallapalli, G., Luo, J., Kawar, P.G., Hill, L., Santino, A., Fernie, A.R., and Martin, C. (2015). Multi-level engineering facilitates the production of phenylpropanoid compounds in tomato. Nat. Commun. *6*, 8635.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and Virus Strains		
BL21(DE3)	Transgen Biotech	CD601-01
GV3101	This study	N/A
Biological Samples		
DNA of leaf tissue from all accessions	This study	N/A
RNA of fruit pericarp on orange stage	This study	N/A
Pericarp tissue of ripen fruit for metabolic profiling	This study	N/A
Chemicals, Peptides, and Recombinant Proteins		
UDP-galactose	Sigma Aldrich	Cat#94333
Tomatidine	ChemFaces	Cat#CFN90632
Tris	BBI life Sciences	Cat#TB0195-500 g
KOD-Plus-High Fidelity DNA polymerase	TOYOBO	Cat#KOD-201
SspI enzyme	NEB	Cat# R0132
T4 DNA Polymerase	Promega	Cat# M4211
T4 PNK	NEB	Cat# M0201
BglII	Thermo Fisher	Cat#FD0084
dCTP	Promega	Cat# U122D
dGTP	Promega	Cat# U121D
Not I	Thermo Fisher	Cat#FD0595
The authentic reference compound	Sigma Aldrich	N/A
Methanol	Merck	CAS 67-56-1
Acetonitrile	Merck	CAS 75-05-08
Acetic acid	Fisher Scientific	CAS64-19-7
Critical Commercial Assays		
E.N.Z.A Cycle Pure Kit	OMEGA bio-tek	Cat#D6492-02
E.N.Z.A Plasmid Mini Kit I	OMEGA bio-tek	Cat#D6943-02
In-Fusion cloning kit	Clontech	Cat#638909
EasyPure Quick Gel Extraction Kit	Transgen	Cat# EG101-01
pLB vector	Tiangen	Cat# VT205-01
Deposited Data		
Whole-genome DNA-seq data	Lin et al., 2014	SRP045767
Whole-genome DNA-seq data	Tieman et al., 2017	PRJNA353161
Whole-genome DNA-seq data	Aflitos et al., 2014	PRJEB5226–PRJEB5228, and PRJEB5253
RNA-seq data for Heinz 1706	Consortium, 2012	http://ted.bti.cornell.edu
RNA-seq data	This study	PRJNA396272
The original data and results	This study	Mendeley https://dx.doi.org/10.17632/gbz22vb344.1
Experimental Models: Organisms/Strains		
Tomato accessions including wild species and cultivars	See Table S1	N/A
Near isogenic lines and transgenic lines for *fw11.3*	This study	N/A
SlMYB12 mutant lines	See Table S4	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Oligonucleotides		
sgRNA for *SlMYB12* knock-out	See Figure 4C	N/A
pMR093-specific primers AtU6	Ron et al., 2014	N/A
Primers for *Solyc10 g085230* cloning	This study	N/A
Recombinant DNA		
pMAL-C2-GST::*Solyc10 g085230*	This study	N/A
pMR093:: *SlMYB12*	This study	N/A
Software and Algorithms		
Analyst Software	AB Sciex	https://sciex.com/
Cytoscape	Shannon et al., 2003	http://www.cytoscape.org/
MultiQuant Software	AB Sciex	https://sciex.com/
SOAP2	Li et al., 2009c	http://soap.genomics.org.cn
SOAPsnp	Li et al., 2009b	http://soap.genomics.org.cn
BWA	Li and Durbin, 2009	http://bio-bwa.sourceforge.net
Samtools	Li et al., 2009a	http://samtools.sourceforge.net
PHYLIP (version 3.695)	Retief, 2000	http://evolution.genetics.washington.edu/phylip.html
SIMCA-P	Umetrics	https://umetrics.com/downloads/simca-q
hisat2	Kim et al., 2015	https://github.com/infphilo/hisat2
FaST-LMM	Lippert et al., 2011	https://github.com/MicrosoftGenomics/FaST-LMM
GEC	Li et al., 2012	http://grass.cgs.hku.hk/gec/download.php
Haploview	Barrett et al., 2005	https://www.broadinstitute.org/haploview/haploview
RNA-QcSe	DeLuca et al., 2012	http://archive.broadinstitute.org/cancer/cga
WGACNA	Langfelder and Horvath, 2008	https://labs.genetics.ucla.edu/horvath/htdocs/CoexpressionNetwork/Rpackages/WGCNA
Matrix eQTL	Shabalin, 2012	http://bios.unc.edu/research/genomic_software/Matrix_eQTL
PEER	Stegle et al., 2012	https://github.com/PMBio/peer

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Sanwen Huang (huangsanwen@caas.cn).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Plant material and growth conditions

A total of 610 tomato accessions were collected from TGRC (Tomato Genetics Resource Center), USDA (United State Department of Agriculture), University of Florida, EU-SOL (The European Union-Solanaceae project), INRA (The National Institute for Agricultural Research) and IVF-CAAS (The Institute of Vegetables and Flowers, Chinese Academy of Agricultural Science). These accessions include 42 wild tomato accessions (four *S. pennellii*, nine *S. habrochaites*, three *S. chilense*, one *S. corneliomuelleri*, seven *S. peruvianum*, three *S. huaylasense*, three *S. neorickii*, three *S. arcanum*, two *S. chmielewskii*, five *S. cheesmaniae* and two *S. galapagense*), 56 *S. pimpinellifolium*, 142 *Sl*. var *cerasiforme*, and 370 *S. lycopersicum* accessions (Table S1). Tomato plants were grown in greenhouses of AGIS-CAAS (Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences). For metabolite profiling, 442 accessions were selected to represent a corss-section of the tomato germplasm sets, consisting of PIM (31 accessions of *S. pimpinellifolium*), CER (124 accessions of *S. lycopersicum* var. *cerasiforme*) and BIG (287 big-fruited varieties of *S. lycopersicum*). Independent biological samples of each accession were grown in two different locations. For each accession, eight plants were grown. At ripening, five plants with representative phenotype were randomly selected. At least one fruit from each plant was pooled together into one sample. Two independent biological samples were metabolically profiled.

METHOD DETAILS

Samples preparation

Each fruit sample contained five or more fruits. Genomic DNA was extracted from young leaves using the CTAB method. The fruit pericarp at orange stage was detached and froze using liquid nitrogen for RNA extraction. Fruit pericarp tissue at the fully ripe stage were harvested and freeze-dried for metabolites profiling. The lyophilized tissues were ground using a mixer mill (MM400, Retsch) with a zirconia bead for 1.5 min at 30 Hz. 100 mg powder was weighed and extracted overnight at 4°C with 1.0 mL 70% aqueous methanol and pure methanol for water and lipid-solubility metabolites, respectively. Following centrifugation at 10,000 g for 10 min, all the supernatants were pooled and filtered with a membrane (SCAA-104, 0.22 μm pore size; ANPEL, Shanghai, China, http://www.anpel.com.cn/) before LC-MS analysis.

Metabolite profiling

A liquid chromatography-electrospray ionization-tandem mass spectrometry system was used for the relative quantification of widely targeted metabolites in dried tomato fruit samples. The MS^2 spectral tag (MS2T) library including the MS and MS/MS spectra was established by the untargeted method based on the total scan ESI (ESI–QqTOF–MS/MS) and multiple ion monitoring-enhanced product ions (MIM-EPI)-based target metabolic methods. Full time-of-flight (TOF) scans were acquired in the mass range of m/z 50–1500 using an AB SCIEX TripleTOF 5600 system. A modified MIM–EPI strategy, called stepwise scan MIM–EPI, was adopted on a triple quadrupole-linear ion trap mass spectrometer (Q TRAP), API 4000 Q TRAP LC/MS/MS System, in which Q1 (Q3) was set from 50.1 to 1000.0 Da in positive scan mode, and the mass step was 1.0 Da, such as from 50/50, 51/51, 52/52, to 1500/1500. Subsequently, the mass signals with MS^2 spectral obtained by these two methods using the MRM mode to filter the peak type and remove the redundant signals. Quantification of metabolites was carried out using a scheduled multiple reaction monitoring method. To produce maximal signal, collision energy (CE) and de-clustering potential (DP) were optimized for each precursor–product ion (Q1–Q3) transition (Chen et al., 2013). We made a mixture of 200 randomly chosen extracts from the association panel that contains all the 980 metabolic features as the reference control. All of the data were normalized by those reference control that could reflect the change of every metabolite feature, and then log2 transformed for further normalization. A data matrix containing the 980 relative intensities of metabolites from 884 runs (442 accessions × two sample sets) was produced for the tomato population (Data S1). The Metabolite (m-trait) data of the association panel are the mean of the two biological sample sets for the LC-MS/MS as shown below: $P_{m,i} = 1/2(P_{m,i,1}+P_{m,i,2})$, where $P_{m,i}$ represents the m-trait data for metabolite m (m = 1, 2, 3, ..., 980 in tomato) in accession i (i = 1, 2, 3, ..., 442), and $P_{m,i,1}$ and $P_{m,i,2}$ are the normalized metabolite levels determined in the two biological sample sets, respectively.

SNP identification and annotation

The 610 accessions used in this study were characterized by whole genome re-sequencing. The raw data had been previously genotyped and deposited in the NCBI Sequenced Archive (SRA) under accession SRP045767, PRJNA353161 and the European Nucleotide Archive under accession PRJEB5235. DNA was isolated from young leaves and sequencing libraries with insert sizes of approximately 500 bp were constructed following manufacturer's instructions (Illumina). The samples were sequenced on an Illumina HiSeq 2000 platform with paired-end 100-bp and 125-bp reads. We used SOAP2 (Li et al., 2009c) to map all the sequencing reads from each accession to the tomato reference genome with the following parameters: -m 100, -x 888, -s 35, -l 32, -v 3. Mapped reads were filtered to remove PCR duplicates. Both paired-end and single-end mapped reads were then used for SNP calling throughout the entire collection of tomato accessions using SOAPsnp with the following parameters: -L 100 -u -F 1 (Li et al., 2009b). We generated the genotype likelihood across the population for each SNP with quality > = 40 and base quality > = 40. The identified SNPs were further categorized as variations in intergenic regions, UTRs, coding sequences and introns according to the tomato genome annotation (release ITAG2.4). SNPs in coding sequences were further classified into synonymous SNPs (not causing amino acid changes) and nonsynonymous SNPs (causing amino acid changes) using Python scripts.

Phylogenetic analysis and population structure

A subset of 18,286 SNPs at four-fold degenerate sites (MAF > 5% and missing data < 10%) were filtered to build a neighbor-joining tree for 448 accession using PHYLIP (version 3.695) with 100 bootstrap replicates (Retief, 2000). These accessions include four wild, one *S. galapagense*, one *S. cheesmaniae* and 442 red-fruited accessions. Four green-fruited wild accessions (including one *S. habrochaites* and three *S. peruvianum*) were used as an outgroup. The principal component analyses were performed by the log-scaled chemical content for 442 population to demonstrate the structure of the tomato population, and the software SIMCA-P (https://umetrics.com/) with default settings was used to cluster all the metabolites.

Read mapping and Expression profiling

Total RNA of fruit pericarp tissues at orange stage was extracted by a Trizol method (Rio et al., 2010). The transcriptome libraries of were sequenced using 150-bp paired-end Illumina sequencing with libraries of 350-bp insert sizes. After filtering out reads with low sequencing quality, an average of 33.87 million reads were obtained and 22.68 million reads were uniquely mapped to the tomato reference genome for each sample. Those reads were used to calculate genome-wide gene expression patterns. To quantify

gene expression, reads were mapped to the genome by hisat 2 (Kim et al., 2015). The reads that uniquely mapped to the reference were conserved. Some of the reads in the remaining paired reads containing repetitive sequences or error of the sequencing instrument were removed. In total, 97.56% of the clean reads were mapped to the genome. On average, 30.48 million reads for each sample were uniquely mapped to the tomato reference genome, and these reads were used to calculate the whole genome expression pattern. About 80% of the reads for each sample mapped to the exons, which is expression profiling efficient for calculating RPKM per transcript. RNA-SeQC was used to calculate the RPKM for each gene among samples (DeLuca et al., 2012).

Co-expression modules identification and KEGG enrichment for transcriptome

We applied Weighted Correlation Network Analysis (WGCNA) to gene modules with distinct expression patterns. A total of 18,675 genes (80% samples shared) were used in module constructions. Construction of a weighted co-expression network needs the soft-thresholding powers β, which were calculated by the pick Soft Threshold function of the R Package (Langfelder and Horvath, 2008). We chose the power 10, which is the lowest power for which the scale-free topology fit index curve flattens out reaching a high value (above 0.9) (Figure S1H). A total of 13,752 genes were assigned to 31 modules. We evaluated each module in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database for enrichment (Kanehisa and Goto, 2000). We first obtained the corresponding KO number for all the genes and conducted pathway enrichment analysis. All the pathways with significant threshold ($p < 0.05$) were presented in Data S1.

Genome-wide association analysis and linkage disequilibrium

A total of 2,678,533 SNPs (MAF > 5% and Missing rate < 10%) for 442 accessions were used to perform the genome-wide association analysis. Factored Spectrally Transformed Linear module (software FaST-LMM) was used to conduct all associations (Lippert et al., 2011). The matrix of pairwise genetic distances was used as the variance-covariance matrix for random effect and the first ten principal components produced by GCTA were included as fixed effects. The genome-wide significance thresholds of all the traits were set with a uniform threshold ($p = 1/n$, n is the effective number of independent SNPs). The effective number of independent SNPs was calculated using Genetic type 1 Error Calculator (GEC) software (Li et al., 2012). The unified threshold ($p = 9.05 \times 10^{-8}$) was used to filter the SNPs for all the metabolites. LD (Linkage disequilibrium) analyses were performed based on all the SNPs (MAF > 0.05) using Haploview software (Barrett et al., 2005). The parameters were as follows: -n -pedfile -info -log -minMAF 0.05 -hwcutoff 0 -dprime -memory 2096 -maxdistance 2 Mb. LD decay was calculated based on the R^2 value and corresponding distance between two SNPs. The average R^2 values for all pairwise SNPs within 200 bp distance were calculated and plotted against the average distance. To reduce the redundancy of mGWAS signals, the lead SNP within one Mb window for each metabolite was extracted as one signal.

Detection of mGWAS signals hotspots

A permutation test was used to assess the statistical significance of the deviation of the observed signal distribution from uniform distribution. In permutation, all the signals were randomly assigned to the genomic regions for each 1 Mb interval, and the number of signal for each interval were recorded. After 10,000-permutation test, the value of significant ($p < 0.01$) number per Mb would be 10 for 3,526 signals.

Normal quantile transformation of expression and eQTL analysis

One of the assumptions of detecting eQTL through linear mixed model is that the expression values follow a normal distribution in each genotype class, which is violated by outliers or non-normality in gene expression estimated from the sequencing reads. For each gene, the expression was normalized by QQ-normal of R package. Finally, a dataset including 22,480 genes (missing rate < 80% and a median expression level > 0) were obtained to conduct downstream analyses. To find hidden batch effects and other confounders in the expression data, we employed the Probabilistic Estimation of Expression Residuals (PEER) method to detect factors (Stegle et al., 2012). We included 20 PEER factors that maximized our sensitivity in the eQTL discovery process, capturing ~56.3% of the total variance in gene expression (Figure S2C). The linear regression mode of Matrix eQTL Package was used to detect associations for SNP-gene pairs (Shabalin 2012). We corrected for the following covariates: the first five genotyping principal components (PC's), the first 20 expression PEER factors and quantile normalized expression matrices for population. To deal with the false positive of association between 558,650 SNPs and genes expression, Multiple hypothesis adjustment produced a rigorous threshold ($p < 8.6 \times 10^{-13}$) by controlling genome-wide error at level a 0.05 using Bonferroni method. The eQTL could be subdivided into *cis*-eQTL and *trans*-eQTL. The intergenic distance of pairwise neighboring genes for the whole genome was calculated and we found a sharp drop of distance at 30 kb with 85.4% pairwise genes (Figure S2D). If the SNPs resides within the corresponding genes or less than 30 kb from the transcriptional start site or the end of a gene, it was classified as *cis*-acting, otherwise as *trans*-acting.

Network building for metabolome, transcriptome and variome

The mGWAS signals were used to connect the metabolites and genetic loci, and eQTL was used to link the genetic loci and genes. A total of 3,526 significant signals for 514 metabolites were detected. Within 3,526 mGWAS signals, we identified 1,626 SNPs (eQTL) for 535 genes, and the shared loci between mGWAS and QTL as linker for metabolites and genes. The network involved a total of

13,661 triple connections (metabolite-locus-genes) including 371 metabolites, 970 loci and 535 genes. All the associations among mGWAS (metabolite-loci), shared eQTL (shared loci-gene) and corresponding gene modules (among transcriptomes) were shown by Cytoscape software.

Detection of domestication and improvement sweeps

To identify genomic regions affected by domestication and improvement, two key stages in tomato evolution, we first measured the level of genetic diversity (π) using a 100 kb window with a step size of 10 kb in PIM, CER and BIG. By scanning the ratios of genetic diversity between PIM and CER (PIM/CER) as well as between CER and BIG (CER/BIG), we selected windows with the top 5% of ratios (2.98 and 7.81 for domestication and improvement, respectively) as candidate regions for further analysis. The regions defined by domestication or improvement sweeps were regarded as sweep regions (SW), and windows that were 200 kb apart were merged into a single selected region. These sweep regions should have undergone selection during domestication or improvement stage.

Bulked segregant analysis of the F_2 population by whole-genome resequencing

We planted 500 F_2 progeny derived from the cross between TS-19 (a PIM line, 1.7g) and TS-450 (a CER line, 13.6g). For each individual, the average weight of ten representative fruits was recorded. We ranked plants by average mass of their individual fruit. Plants producing fruit that were in the top or bottom 10% in weight were pooled together. The two pools were referred as *large* and *small*. Bulk DNA samples for large-pools and small-pools were constructed by mixing equal amounts of DNA from selected lines, respectively. Roughly 50 x genome coverage short reads data for each bulk was aligned against the reference genome using the Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009). SNPs between the two parental genomes were identified using SAMtools software (Li et al., 2009a). The SNP index was calculated for both bulk samples expressing the proportion of reads harboring SNPs that were identical to those in the big-fruit parent. A ΔSNP index was obtained by subtracting the SNP index for the small-fruit bulk sample from that of the big-fruit bulk sample (Takagi et al., 2013). An average SNP index for big-fruit and small-fruit bulk samples was calculated using a 1,000-kb sliding window with a step size of 10 kb. For each position, a 95% confidence intervals of ΔSNP index under the null hypothesis of no QTL were calculated to filter out the significant genomic region. The segregated genomic regions controlling fruit weight derived from the cross between TS-450 and TS-400 (a BIG line, 260.1g) were detected using the same pipeline.

Glycosyltransferase assay

The full length CDS of *Solyc10 g085230* was amplified with KOD-Plus-Neo high fidelity DNA polymerase using primers (Forward: TACTTCCAATCCAATGCGATGAAAATAGAAAGAAAACAGAGTG, and Reverse: TTATCCACTTCCAATGCGCTAAGACTCTATGAT ACACTTGCTTGC) and the following program: 95°C for 3 min, 40 cycles of 95°C for 30 s, 60°C for 30 s and 68°C for 30 s. The amplified sequence was inserted into the expression vector pMAL-C2-GST by a ligation-independent cloning method as previously described (Eschenfeldt et al., 2009). In brief, the vector was digested with SspI and treated with T4 DNA polymerase and dGTP to create overhang. The amplified CDS was treated with T4 PNK and then with T4 DNA polymerase and dCTP to create overhang. The overhung CDS and vector were annealed at room temperature and then transformed into transformation-competent *E. coli* cells.

Recombinant proteins were expressed in BL21 (DE3) cells (Novagen) following induction by addition of 0.1 mM isopropy-β-D-thiogalactoside (IPTG) and grown continually for 16 h at 20°C. Cells were harvested and pellets were resuspended in lysis buffer (50 mM Tris-HCl, pH 8.0, 400 mM NaCl). The cells were disrupted by high pressure and cell debris was removed by centrifugation (14000 g, 1 h). Glutathione Sepharose 4B agarose (GE Healthcare) was added to the supernatant containing the target proteins. After incubation for 1 h, the mixture was transferred to a disposable column and washed extensively with lysis buffer (5 column volumes). Target proteins were confirmed by SDS-PAGE and purified recombinant proteins were selected for enzyme assays and kinetic determination.

The *in vitro* glucosyltransferase assay for was performed at 37°C in a total volume of 100 μL containing 200 μM tomatidine substrate, 1.5 mM UDP-galactose, 5 mM $MgCl_2$ and 500 ng purified protein in Tris-HCl buffer (100 mM, pH 7.4). After incubating for 1 h, the reaction was stopped by adding 300 μL of ice-cold methanol. The reaction mixture was then filtered through a 0.2 μm filter (Millipore) before being used for LC-MS analysis. HPLC conditions for the analysis of SGAs were described in Metabolite profiling section. Peak identification of each component was confirmed using authentic samples and post-run by LC-MS/MS analysis.

CRISPR/Cas9 constructs design

Two *SlMYB12* target sites (sgRNA1 and sgRNA2) of 19 nucleotide (nt) were manually selected (Figure 4C). To confirm the specificity, the sgRNA sequences were aligned to the tomato whole genome, and no potential sites with mismatches over six bases were found. The CRISPR/Cas9 constructs were generated following the description (Ron et al., 2014). The sgRNA sequences were incorporated into two 60 nt oligonucleotides (sgRNA1-Forward: GAAGCTGAGTTTATATACAGCTAGAGTCGAAGTAGT GATTGCCAGCTTGTGATAGTGCCA, and Reverse: GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAACTGGCACTATC ACAAGCTGG; sgRNA2-Forward, GAAGCTGAGTTTATATACAGCTAGAGTCGAAGTAGTGATTGGAAGATCTAGCAAAGATAG, and Reverse, GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAACCTATCTTTGCTAGATCTTCC). The two primers were annealed and extended to make 100-bp double-strand DNA fragment, and then cloned into NcoI-linearized pMR093 vector with the In-Fusion cloning kit (Clontech). The plasmid with the correct insertion was introduced into *Agrobacterium tumefaciens* strain GV3101 for tomato transformation.

Transformation of *S. lycopersicum* cv. Moneymaker was performed. Regenerated plantlets capable of growing on Basta-containing (1 $mg \cdot L^{-1}$) medium were detected with the pMR093-specific primers AtU6 (Forward: CCGGGGATCCTCTAGAAGCTTCGTTG AACAACGG and Reverse: CGTCGGGCCCTCTAGAAAAAAAAAGCACCGACTCGG). For the positive T_0 plants, the gene fragment of *SlMYB12* was amplified with primers that flanked both sgRNA targets. The PCR products were purified and cloned into the pLB vector for sequence determination. The T_0 plants carrying the targeted mutation were transplanted into the greenhouse. To test the stable heritability of CRISPR/Cas9-induced mutations, 96 plants were grown for each family. All T_1 plants were genotyped with pMR093-specific primers. For the T_1 plants not containing the CRISPR/Cas9 cassette, the PCR products of *SlMYB12* were sequenced.

QUANTIFICATION AND STATISTICAL ANALYSES

The values of the coefficient of variation were calculated for each metabolite in PIM, CER, BIG and whole population. The formula is as follows: σ/u, where σ and u are the standard deviation and mean of each metabolite in the population, respectively. Broad-sense heritability (H^2) was estimated by treating accessions as a random effect and the biological replication as a replication effect using the following formula: $H^2 = = \mathrm{var}(G)/(\mathrm{var}(G) + \mathrm{var}(E))$, where var(G) and var(E) are the variances derived from genetic and environmental effects, respectively. The quantification for the other parts can be found in the relevant sections of the Method Details.

DATA AND SOFTWARE AVAILABILITY

The deposited number for the RNA-seq reads reported in this paper is PRJNA396272.

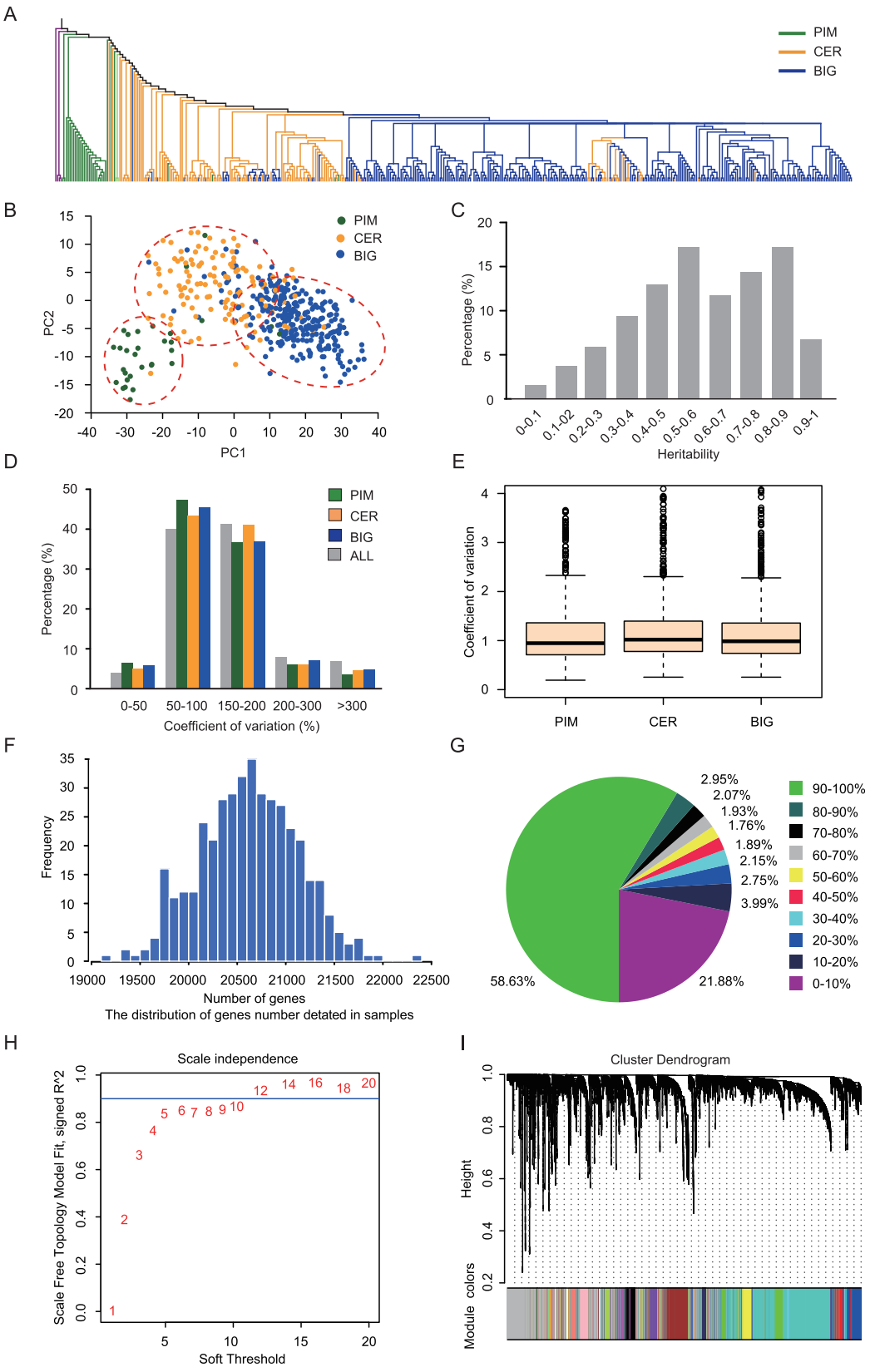
The original data and results were presented in Tables M1-M9 at Mendeley Data and could be found at https://data.mendeley.com/datasets/gbz22vb344/1.

ADDITIONAL RESOURCES

Tomato core collection: http://tgrc.ucdavis.edu

Tomato Functional Genomics Database (TFGD): http://ted.bti.cornell.edu

Data deposition: https://ncbi.nlm.nih.gov/sra and https://www.ebi.ac.uk/ena



(legend on next page)

Figure S1. Characters of Variome, Metabolome, and Transcriptome in Three Populations, Related to Figure 1

(A) Neighbor-joining of 448 varieties calculated from whole genome fourfold degenerate SNPs. 448 varieties included four green-fruited wild accessions (one *S. habrochaites* and three *S. peruvianum*), one *S. galapagense*, one *S.cheesmaniae* and 442 red-fruited accessions.

(B) Principle component analyses of 980 metabolites in red-fruited population.

(C) Distribution of broad-sense heritability (H^2) of metabolic traits (n = 980) detected in the metabolite panel across two biological replicates.

(D) Coefficient of variation for PIM, CER, BIG and whole group, respectively.

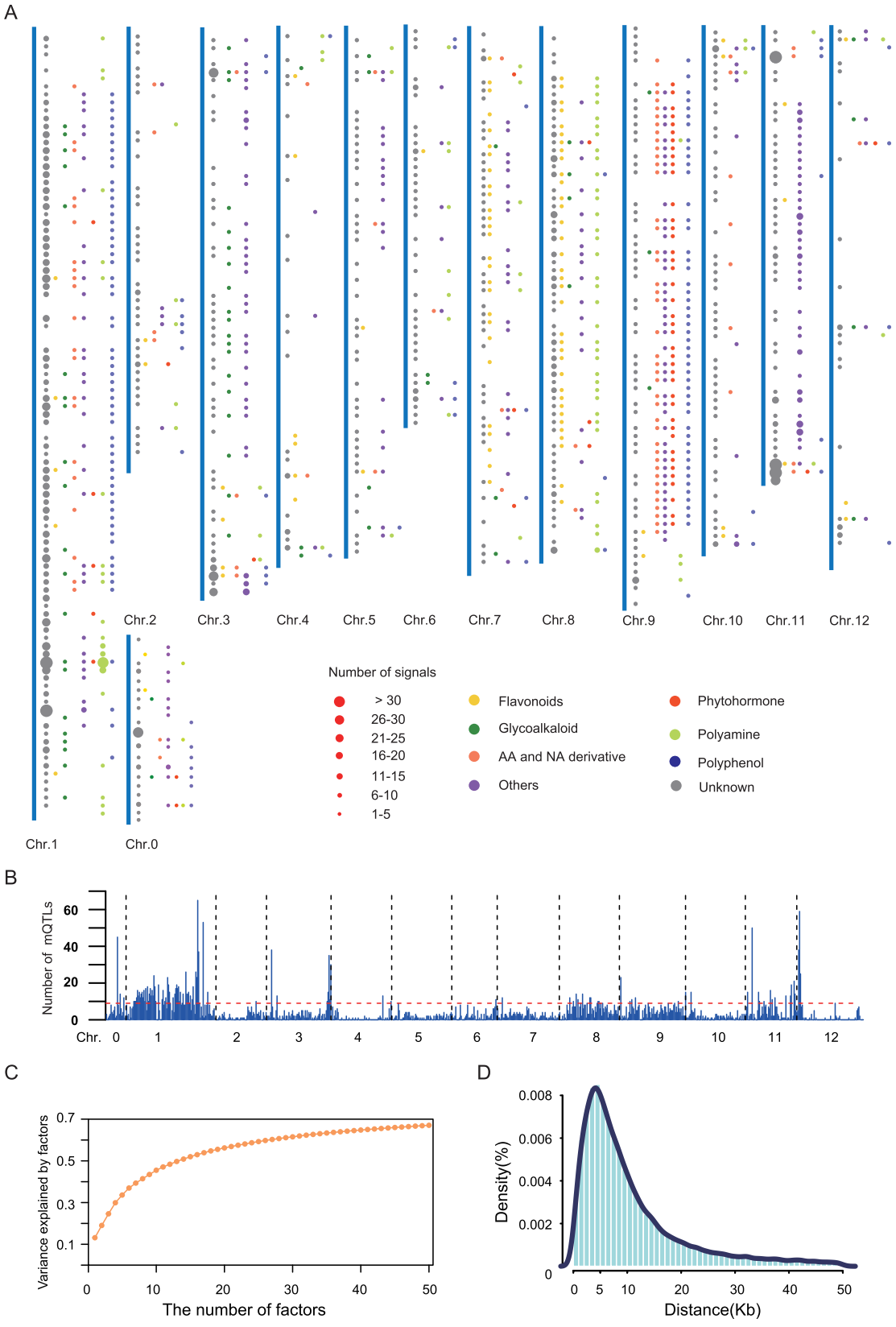
(E) Boxplot of coefficient of metabolite for three population. The mean value of 980 metabolites is 1.13, 1.22 and 1.24 for PIM, CER and BIG, respectively.

(F) Distribution of detected genes in each tomato samples. The mean gene number for all the samples is 20,226.

(G) Percentage of genes detected in the tomato population. A total of 18,675 (61%) expressed genes could be detected in more than 80% of the samples.

(H) The parameter, soft threshold, determination for module construction. The best value is 10 for this dataset.

(I) Genes cluster dendrogram of fruit transcriptome. 31 modules were built based on gene expression value. Each color indicates a different gene module.



(legend on next page)

Figure S2. Distribution of mGWAS and Determination of eQTL Signals, Related to Figure 1

(A) Distribution of mGWAS signals for different metabolic classes. A total of 3,526 signals were identified for 514 metabolites. Circles next to the genome segment indicate the positions and were proportional to the number of signals for each compound class. Different metabolic classes are marked with different colored circles.

(B) Distribution of mGWAS signal number per 1 Mb window across the genome. The horizontal dashed line indicates the threshold (permutation test < 0.01) for signal hotspots.

(C) Diagnostic analysis of hidden confounding factor by PEER. The x axis is the factor number and the y axis is the variance explained by corresponding factor number.

(D) Distribution of pairwise genes distance. 85.35% of pairwise genes were less than 30 kb, the point of separation for *cis*-eQTL and *trans*-eQTL.

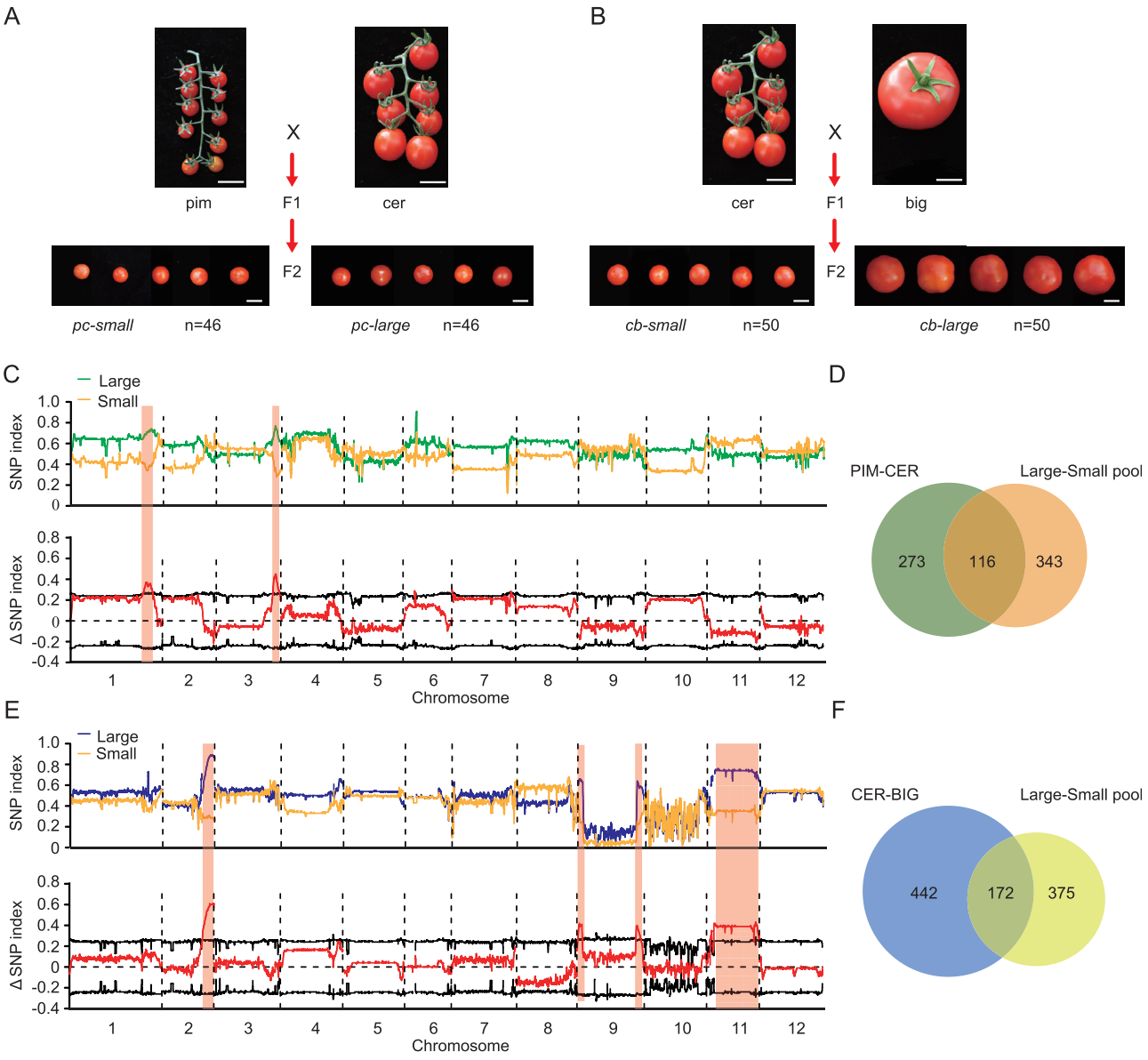


Figure S3. Effect of Domestication and Improvement Related to Fruit Weight on the Genome and Metabolome, Related to Figure 3

(A) Fruit phenotypes of the parental lines pim (*S. pimpinellifolium*) and cer (*S. lycopersicum* var. *cerasiforme*), and the two bulked pools with extreme fruit weight from the F_2 population, each containing 46 individuals. Scale bars in (A) and (B), 1 cm.

(B) Fruit phenotypes of the parental lines cer and big (*S. lycopersicum*), and the two bulked pools with extreme fruit weight from the F_2 population, each containing 50 individuals.

(C) Genomic regions of the fruit weight genes differing between pim and cer. Two regions in Chr.1 and Chr.3 were identified. The SNP indices (ratio of the SNPs that are identical to those in the big-fruited parent) of the small and large pools are shown with green and orange lines, respectively. The ΔSNP index (subtracting the SNP index of the small-pool from that of the large-pool) and its 95% confidence interval are shown with red and black lines, respectively. Regions with a ΔSNP index above the confidence line are highlighted with pink shadow.

(D) Overlap between bulked pool (pim-cer) shifted metabolites and PIM-CER group shifted metabolites.

(E) Genomic regions of the fruit weight genes differing between cer and big. Four regions in Chr.1, Chr.9 and Chr.11 were identified.

(F) Overlap between bulked pool (cer-big) shifted metabolites and CER-BIG group shifted metabolites.

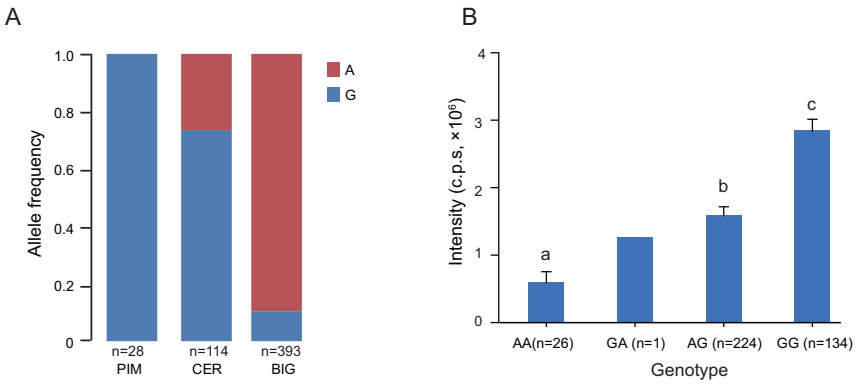


Figure S4. Frequency of One SGA Locus and Genetic Effect of Two SGAs loci, Related to Figure 2

(A) Allele frequencies of locus on ch01:84029382 (CDS of *GAME9*) in three groups.

(B) SGAs content in varieties containing the two allele combinations. Total content of three major SGA chemicals (SlFM1785, SlFM1885 and SlFM1985) was calculated for each combination of two alleles (01:84029382 and 10:64501127). Letters above the bars indicate significant differences as determined by Student' s pairwise t test ($p < 0.05$). Data represent mean ± SEM and was analyzed by Student's t test.