ARTICLES
https://doi.org/10.1038/s41588-019-0410-2
nature genetics

# The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor

Lei Gao[1,10], Itay Gonda[1,2,10], Honghe Sun[1], Qiyue Ma[1], Kan Bao[1], Denise M. Tieman[3], Elizabeth A. Burzynski-Chang[4], Tara L. Fish[5], Kaitlin A. Stromberg[1], Gavin L. Sacks[4], Theodore W. Thannhauser[5], Majid R. Foolad[6], Maria Jose Diez[7], Jose Blanca[7], Joaquin Canizares[7], Yimin Xu[1], Esther van der Knaap[8], Sanwen Huang[9], Harry J. Klee[3], James J. Giovannoni[1,5*] and Zhangjun Fei[1,5*]

**Modern tomatoes have narrow genetic diversity limiting their improvement potential. We present a tomato pan-genome constructed using genome sequences of 725 phylogenetically and geographically representative accessions, revealing 4,873 genes absent from the reference genome. Presence/absence variation analyses reveal substantial gene loss and intense negative selection of genes and promoters during tomato domestication and improvement. Lost or negatively selected genes are enriched for important traits, especially disease resistance. We identify a rare allele in the *TomLoxC* promoter selected against during domestication. Quantitative trait locus mapping and analysis of transgenic plants reveal a role for TomLoxC in apocarotenoid production, which contributes to desirable tomato flavor. In orange-stage fruit, accessions harboring both the rare and common *TomLoxC* alleles (heterozygotes) have higher *TomLoxC* expression than those homozygous for either and are resurgent in modern tomatoes. The tomato pan-genome adds depth and completeness to the reference genome, and is useful for future biological discovery and breeding.**

Tomato is one of the most consumed vegetables worldwide with a total production of 182 million tons worth more than US$60 billion in 2017 (http://www.fao.org/faostat). A reference genome sequence was released[1] and has greatly facilitated scientific discoveries and molecular breeding of this important crop. Cultivated tomato (*Solanum lycopersicum* L.) has experienced severe bottlenecks during its breeding history, resulting in a narrow genetic base[2]. However, modern cultivated tomatoes exhibit a wide range of phenotypic variation[3] and metabolic diversity[4], mainly because of natural and human breeding-mediated introgressions from wild relatives[5], in addition to spontaneous mutations that have also contributed to this seeming paradox[3]. Consequently, individual cultivars are expected to contain alleles or loci that are absent in the reference genome[6].

*S. lycopersicum* L. can be further divided into two botanical types: large-fruited tomatoes *S. lycopersicum* var. *lycopersicum* (SLL) and cherry-sized early domesticates *S. lycopersicum* var. *cerasiforme* (SLC). Following the release of the tomato reference genome, hundreds of diverse cultivated and wild tomato accessions have been resequenced, and the resulting data have been analyzed to reveal genomic changes through the history of tomato breeding. This has led to identifying specific genome regions targeted by human selection[7–10]. Notably, in these studies, reported genomic variation was revealed through mapping of short reads to the reference genome, an activity whose very nature ignores sequence information that is absent from the reference genome, precluding the discovery of previously unknown loci and highly divergent alleles.

A pan-genome comprising all genetic elements from cultivated tomatoes and their wild progenitors is crucial for comprehensive exploration of domestication, assessment of breeding histories, optimal utilization of breeding resources and a more complete characterization of tomato gene function and potential. We constructed a tomato pan-genome using the 'map-to-pan' strategy[11], based on resequencing data of 725 accessions belonging to the Lycopersicon clade, which consists of *S. lycopersicum* L. and its close wild relatives, *Solanum pimpinellifolium* (SP), and *S. cheesmaniae* and *S. galapagense* (SCG). The pan-genome captured 4,873 additional genes not in the reference genome. Comparative analyses using the constructed pan-genome revealed abundant presence/absence variations (PAVs) of functionally important genes under selection and identified a rare allele defined by promoter variation in the tomato lipoxygenase gene, *TomLoxC*. TomLoxC is known to influence fruit flavor by catalyzing the synthesis of lipid-derived C5 and C6 volatiles. Further characterization reveals a role of TomLoxC in apocarotenoid production. The rare allele of *TomLoxC* may have undergone negative selection in the early domesticates, followed by more recent reintroduction. The PAV dynamics presented here provide a case model of the profound impact of human selection on the gene repertoire of an important modern crop, in addition to a more complete picture

[1]Boyce Thompson Institute for Plant Research, Cornell University, Ithaca, NY, USA. [2]Unit of Aromatic and Medicinal Plants, Newe Ya'ar Research Center, Agricultural Research Organization, Ramat Yishay, Israel. [3]Horticultural Sciences, Plant Innovation Center, University of Florida, Gainesville, FL, USA. [4]Department of Food Science, Cornell University, Ithaca, NY, USA. [5]US Department of Agriculture–Agricultural Research Service, Robert W. Holley Center for Agriculture and Health, Ithaca, NY, USA. [6]Department of Plant Science, The Pennsylvania State University, University Park, PA, USA. [7]Institute for the Conservation and Improvement of Agricultural Biodiversity, Polytechnic University of Valencia, Valencia, Spain. [8]Department of Horticulture, University of Georgia, Athens, GA, USA. [9]Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China. [10]These authors contributed equally: Lei Gao, Itay Gonda. *e-mail: jjg33@cornell.edu; zf25@cornell.edu

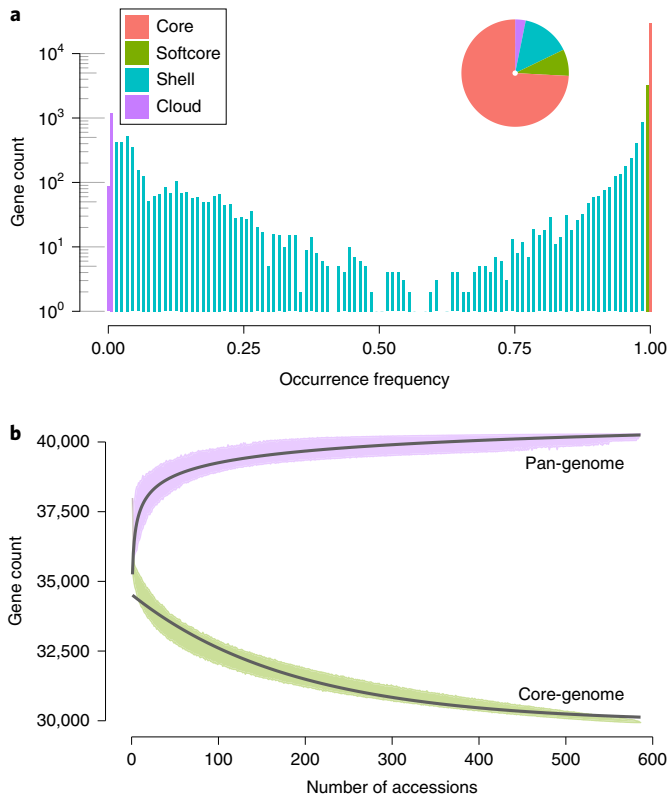


**Fig. 1 | Pan-genome of tomato. a**, Composition of the tomato pan-genome. **b**, Simulations of the increase of the pan-genome size and the decrease of core-genome size. Accessions were sampled as 10,000 random combinations of each given number of accessions. Upper and lower edges of the purple and green areas correspond to the maximum and minimum numbers of genes, respectively. Solid black lines indicate the pan- and core-genome curves fitted using points from all random combinations according to the models proposed by Tettelin et al.[41].

of the genome potential of tomato that will guide breeding for targeted traits.

## Results

**Pan-genome of cultivated tomato and close wild relatives.** Genome sequences were collected/generated for a total of 725 tomato accessions in the Lycopersicon clade, including 372 SLL, 267 SLC, 78 SP and 8 SCG (3 *S. cheesmaniae* and 5 *S. galapagense*) (Supplementary Tables 1 and 2). Among these accessions, genome sequences of 561 were available from previous reports[1,7–9,12–14], whereas genomes of 166 accessions (of which 2 were also sequenced previously), including 121 SLC, 26 SP and 19 SLL, were sequenced in this study to obtain broader regional and global representation. Among the 725 accessions, 98 and 242 had sequence coverage of more than 20× and 10×, respectively.

The genome for each accession was de novo assembled, producing a total of 306 Gb of contigs longer than 500 base pairs (bp) with an N50 value (the minimum contig length needed to cover 50% of the assembly) of 3,180 bp (Supplementary Table 2, Supplementary Fig. 1 and Supplementary Note). All assembled contigs were compared with the reference genome to identify previously unknown sequences. A total of 4.87 Gb of nonreference sequence with identity <90% to the reference genome was obtained (Supplementary Table 2, Supplementary Fig. 2 and Supplementary Note). After removing redundancies, 449,614 sequences with a total length of 351 Mb comprising the nonreference genome remained. Approximately 78.2% of the nonreference genome comprised repetitive elements, which was higher than that of the reference genome (63.5%)[1].

A total of 4,873 protein-coding genes were predicted in the nonreference genome (Supplementary Table 3). The reference 'Heinz 1706' genome contains 35,768 protein-coding genes (version ITAG3.2), of which 272 were potential contaminations and thus were removed (Supplementary Table 4 and Supplementary Note). The tomato pan-genome, including reference and nonreference genome sequences, had a total size of 1,179 Mb and contained 40,369 protein-coding genes. Among the nonreference genes, 2,933 could be annotated with gene ontology (GO) terms or Pfam domains. A total of 332 nonreference genes were covered by 'Heinz 1706' reads with a coverage fraction greater than 95%, and 170 were fully covered, suggesting that they were not assembled in the reference genome (Supplementary Table 3). Among them were two well-characterized genes, *GAME8* (*TomatoPan006500*), which encodes a CYP72 family P450 protein involved in regulation of steroidal glycoalkaloid biosynthesis[15], and *PINII* (*TomatoPan007410*), which encodes a wound-inducible proteinase inhibitor[16]. In addition, several other well-characterized genes, including *Hcr9-OR2A* (*TomatoPan017870*, a homolog of *Cf-9* involved in *Cladosporium fulvum* resistance[17]), *I2C-1* (*TomatoPan019380*, a disease resistance gene[18]) and *Pto* (*TomatoPan028750*, a protein kinase gene conferring disease resistance[19]), were not covered by any 'Heinz 1706' reads, suggesting their absence in the reference accession. Moreover, we found that 69.6% of the reference and 22.4% of the nonreference genes were expressed at >1 reads per kilobase (kb) of exon per million mapped reads (RPKM) in fruit pericarp tissues at the orange stage (about 75% ripe) in at least 1 of 397 accessions for which RNA-sequencing (RNA-Seq) data were available[4]. Gene expression analysis indicated generally lower expression levels of nonreference genes than reference genes (Supplementary Fig. 3a), similarly to pan-genome analysis of rice[20]. Given that the tomato RNA-Seq data used emanated from a single tissue at one developmental stage, these expression frequencies represent a conservative estimate with many additional nonreference genes likely expressed in other tissues.

**PAVs in protein-coding genes.** PAVs in genes among the wild, early domesticates and modern tomato accessions can reveal genetic changes through breeding history. High-depth sequencing data are preferable for robust PAV calling and have been deployed in several previous plant pan-genome studies examining relatively small numbers of accessions[20–26]. However, if sequencing data are uniformly distributed across the genome, low-depth data can still effectively cover a large proportion of the genome and provide sufficient evidence for PAV calling. Based on our analysis, we limited our investigation to a total of 586 accessions (294 SLL, 225 SLC, 60 SP and 7 SCG) for PAV calling (Supplementary Note and Supplementary Fig. 4).

The total number of detected genes from the 586 accessions was 40,283, accounting for 99.97% of genes in the tomato pan-genome (40,369). Similarly to Gordon et al.[24], we categorized genes in the tomato pan-genome according to their presence frequencies: 29,938 (74.2%) core genes shared by all the 586 accessions, and 3,232 softcore, 5,912 shell and 1,287 cloud genes defined as present in more than 99%, 1–99% and less than 1% of the accessions, respectively (Fig. 1a and Supplementary Table 5). The core and softcore groups contained highly conserved genes, whereas the shell and cloud groups contained the so-called flexible genes. Modeling of the pan-genome size by iteratively randomly sampling accessions suggested a closed pan-genome with a finite number of both pan and core genes (Fig. 1b). The most striking feature of the tomato pan-genome was its high core gene content (74.2%), as compared with those of *Arabidopsis thaliana*[23] (70%), *Brassica napus*[25] (62%), bread wheat[26] (64%), rice[11] (54%), wild soybean[22] (49%) and *Brachypodium distachyon*[24] (35%).

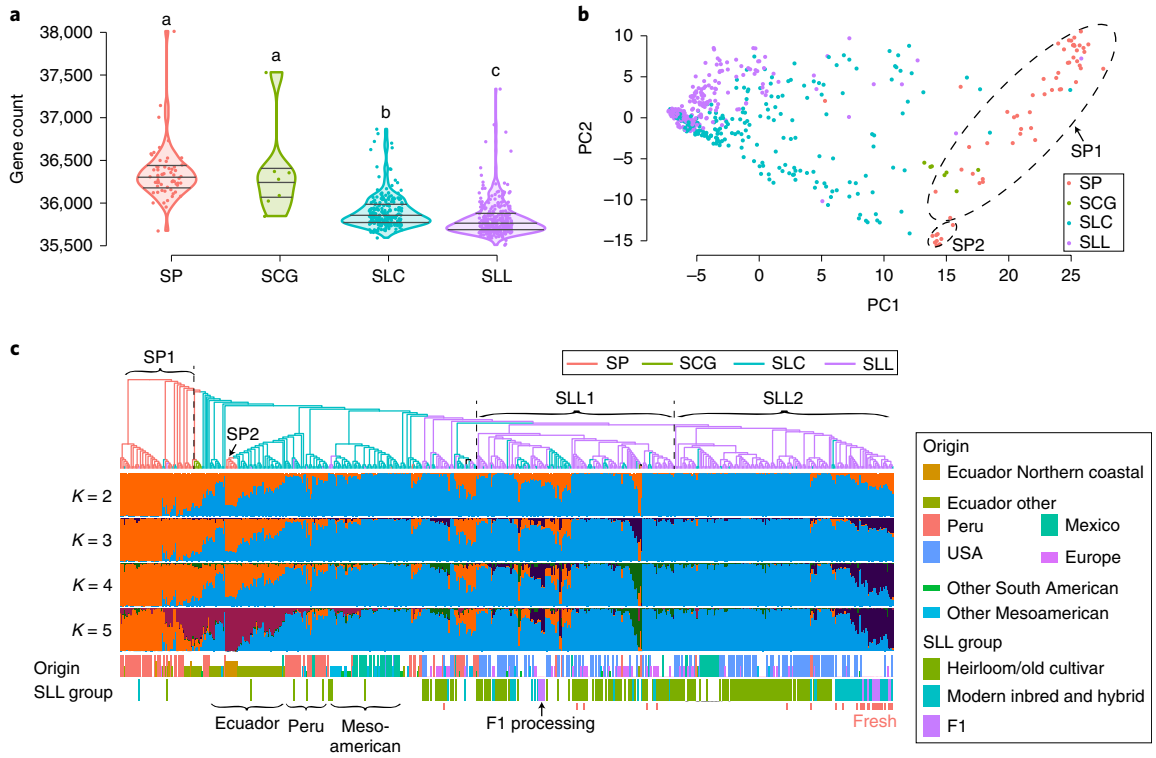


**Fig. 2 | PAVs of genes in wild and cultivated tomatoes. a**, Violin plots showing the number of detected genes in each group. Groups labeled with different letters indicate significant difference in gene contents at $P < 0.01$ (Tukey's HSD test). Three lines (from the bottom to the top) in each violin plot show the location of the lower quartile, the median and the upper quartile, respectively. **b**, Principal component analysis based on PAVs. **c**, Maximum-likelihood tree and model-based clustering of the 586 accessions with different numbers of ancestral kinships ($K = 2$, 3, 4 and 5) using the 10,345 identified PAVs.

Only *Brassica oleracea*[21] was higher (81%), although it is noteworthy that this pan-genome was based on only eight cultivated and one wild accession, and would likely shrink in core gene representation if additional accessions were sequenced.

The reference genome contained the majority of highly conserved genes (99.6%) but only around one-third of the flexible genes. We also observed lower expression levels of the flexible genes compared with conserved genes (Supplementary Fig. 3b), in line with reports in *A. thaliana*[23] and *B. distachyon*[24]. Moreover, conserved reference and nonreference genes displayed similar expression levels, whereas the flexible reference genes generally had higher expression levels than flexible nonreference genes (Supplementary Fig. 3c). Within the flexible genome, the occurrence of reference and nonreference genes displayed distinct distribution patterns (Supplementary Fig. 5): most of the former were sporadically absent in a small number of accessions, whereas the majority of the latter could be found in only a few accessions. The largest groups of genes in the flexible genome included those involved in the oxidation-reduction process, regulation of transcription and defense response (Supplementary Fig. 6a). Compared with the entire pan-genome, genes in the flexible genome were significantly enriched with those involved in biological processes, such as defense response, photosynthesis and biosynthetic processes (Supplementary Fig. 6b). It thus could be anticipated that divergence within the flexible genes among different tomato accessions would be related to corresponding phenotypic and metabolic variations.

**Selection of gene PAVs during tomato breeding.** Genomes of wild accessions (SP and SCG) encoded significantly more genes than SLC, whereas SLC contained significantly more genes than SLL (Fig. 2a), suggesting a general trend of gene loss during tomato domestication and subsequent improvement. Furthermore, more genes were lost during domestication than improvement. Phylogenetic and principal component analyses using the PAVs suggested that wild accessions clearly separated from domesticated accessions with only a few exceptions, and the two domesticated groups (SLC and SLL) separated but with clear overlaps (Fig. 2b,c).

Clustering of tomato accessions based on gene PAVs could be explained by geographic origin and domestication stage (Fig. 2c, Supplementary Fig. 7 and Supplementary Note). A small SP clade (SP2), nested in SLC, including nine accessions from the coastal region of northern Ecuador, possessed significantly fewer genes than the phylogenetically separated main SP clade (SP1), implying that environmental adaptation within SP may have taken place in this region. The continuing decrease of gene content and wild ancestral proportions of SLC accessions from Ecuador and Peru to Mesoamerica suggests that tomato domestication followed this trajectory. Similar gene content and homogeneous genetic structures were found in Mexican SLC and SLL, and older cultivars found in Europe and the rest of the world, supporting the completion of tomato domestication in Mexico with minimal gene loss during subsequent improvement. Modern breeding has left a conspicuous genetic signature on contemporary tomato genomes, because modern elite inbred lines and hybrid cultivars possess significantly higher gene content than SLL heirlooms. This could be at least partially attributed to the intense introgression of disease resistance and abiotic stress tolerance alleles from wild species into modern cultivars[5,27].

To identify gene PAVs under selection during the history of tomato breeding, we conducted two sets of comparisons of flexible gene frequencies, between SLC and SP for 'domestication' (Fig. 3a) and between SLL heirlooms and SLC for 'improvement' (Fig. 3b). Ten accessions that were positioned into an unexpected species group (Fig. 2c) were excluded from the downstream analyses.

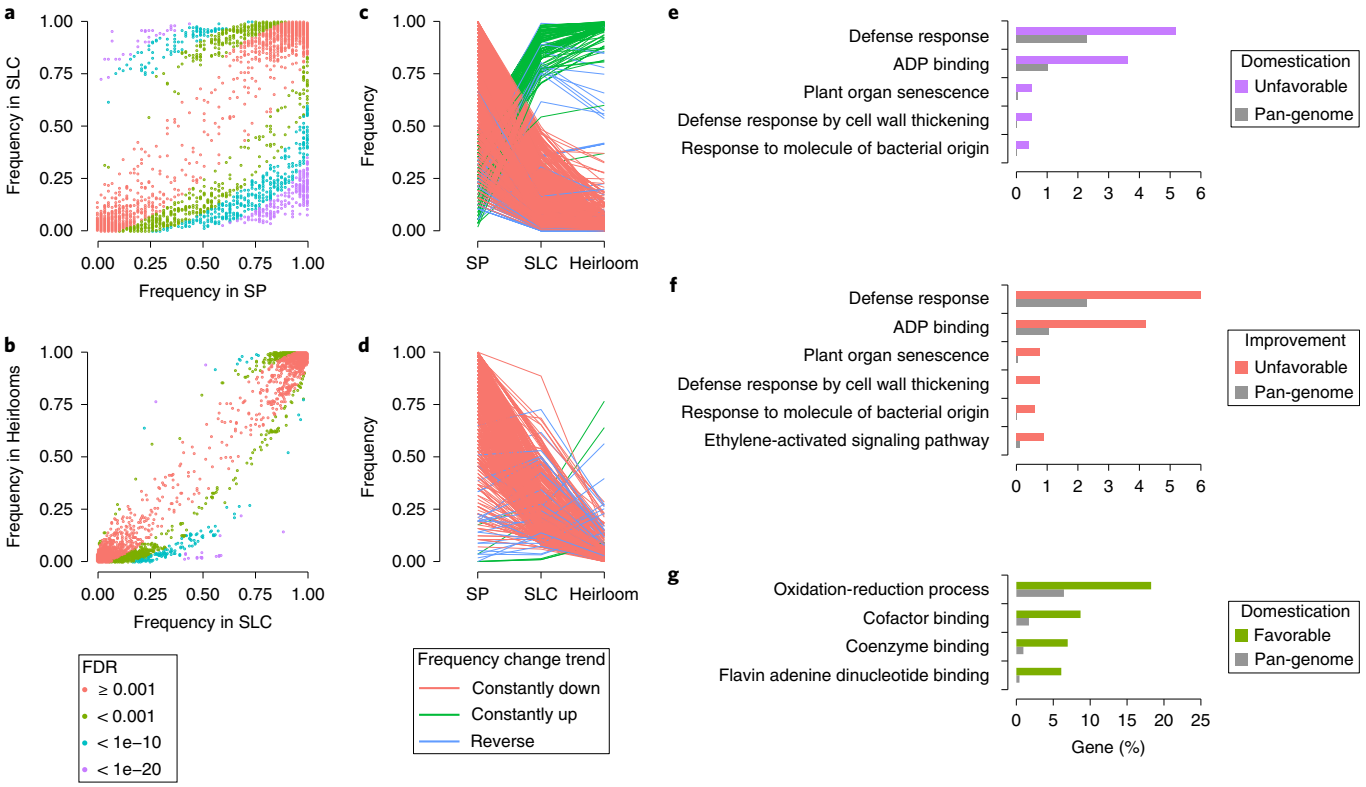


**Fig. 3 | Gene selection preference during tomato domestication and improvement. a**,**b**, Scatter plots showing gene occurrence frequencies in SP and SLC (**a**) and in SLL heirlooms and SLC (**b**). **c**,**d**, Occurrence frequency patterns of putative selected genes during domestication (**c**) and improvement (**d**). **e**–**g**, Enriched GO terms in unfavorable genes during domestication (**e**) and improvement (**f**), and favorable genes during domestication (**g**).

For each comparison, genes with significantly different frequencies between the two groups were identified as selected genes. We treated genes with higher frequencies in SLC than SP, or in SLL heirlooms than SLC as possible favorable genes, and those with lower frequencies as possible unfavorable genes. We note that the selection or loss of any particular gene could be random or due to respective positive or negative selection. In total, we identified 120 favorable and 1,213 unfavorable genes during domestication, and 12 favorable and 665 unfavorable genes during improvement (Supplementary Table 5). These results suggest that more genes were selected against than selected for during both domestication and improvement of tomato. For genes favorable or unfavorable in one stage, most (94.9%) showed the same trend in the other stage (Fig. 3c,d), suggesting the possibility of common and continued selection preferences from domestication to improvement.

Enrichment analysis indicated that defense response was the most enriched group of unfavorable genes during both domestication and improvement, and especially for genes related to cell wall thickening (Fig. 3e,f), which influences abiotic and biotic stress responses through fortification of the physical and mechanical strength of the cell wall. Cell wall modifications also can contribute to fruit firmness and flavor[28,29]. Aging and plant organ senescence were additional enriched classes of unfavorable genes, possibly reflecting selection for increased storability and shelf-life. Of the 120 favorable genes selected during domestication, 21 were related to oxidation-reduction processes (Fig. 3g). The unfavorable and favorable genes selected during domestication also showed distinct molecular functions, with the former enriched for ADP binding and the latter for cofactor, coenzyme and flavin adenine dinucleotide binding (Fig. 3e–g). No significantly enriched gene families were found in favorable genes during improvement.

It is worth noting that among the unfavorable genes, seven were not full length (Supplementary Table 6). These included *TomatoPan028690*, which corresponded to the truncated part of a fruit weight gene *Cell Size Regulator* (*CSR*) as previously reported[30]. *TomatoPan028690* was detected in all SP, 88.6% of SLC and 14.4% of SLL heirlooms, supporting that the deletion allele arose during domestication and has been largely fixed in cultivated tomatoes. Another nonreference gene, *TomatoPan005770*, corresponded to the 5′ part of a full-length gene encoding a UDP-glycosyltransferase, and the reference gene *Solyc05g006140* corresponded to the 3′ portion (Supplementary Table 6 and Supplementary Fig. 8). UDP-glycosyltransferases have been reported to catalyze the glycosylation of plant secondary metabolites and play an important role in plant defense responses[31]. *TomatoPan005770* has experienced strong negative selection during both domestication and improvement (present in all SP, 13.2% of SLC and 1.4% of SLL heirlooms), consistent with the loss of disease resistance in SLL heirlooms. Notably, for three of the seven genes, both truncated and full-length transcripts were expressed in orange-stage fruit (Supplementary Table 6), implying that these truncated genes might be functional, such as the gain-of-function truncation of *CSR* as reported in Mu et al.[30].

**Selection of promoter PAVs during tomato breeding.** A total of 90,929 nonreference contigs could be localized to defined regions (with both ends aligned) or linked sites (one end aligned) on the 'Heinz 1706' genome (Supplementary Table 7). The majority of these sequences were found in intergenic regions, whereas only 8.7% (7,912) overlapped with reference genes, much lower than the genic content of the reference genome (18.0%), implying a functional constraint against these structure variations. There were 3,741 nonreference sequences localized in putative promoter regions (<1 kb to

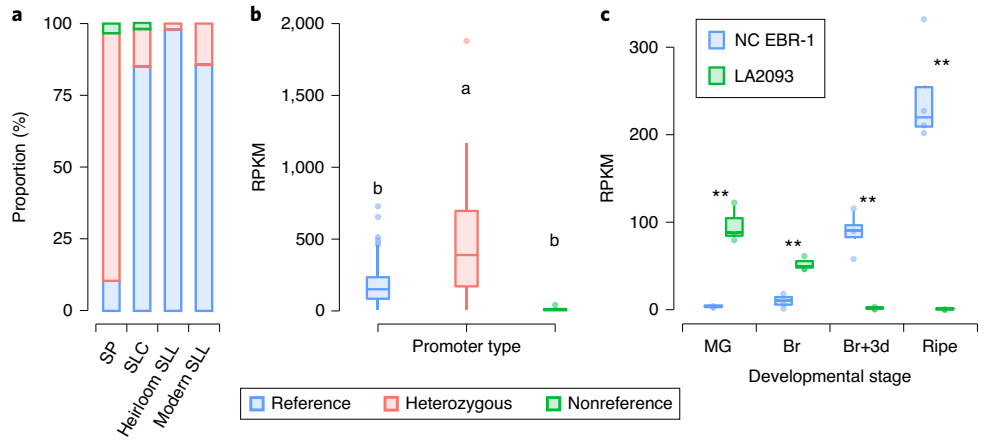


**Fig. 4 | Variation of *TomLoxC* expression under different promoter alleles. a**, Proportion of accessions within each group that have different *TomLoxC* promoter alleles. The numbers of accessions used for SP, SLC, heirloom and modern SLL are 57, 219, 222 and 69, respectively. **b**, Expression levels of *TomLoxC* in orange-stage fruit of accessions with different promoter alleles. RNA-Seq data for four accessions with homozygous nonreference allele were generated under this study, and for the remaining accessions were obtained from Zhu et al.[4]. The numbers of accessions with homozygous reference, nonreference and heterozygous *TomLoxC* promoter alleles are 295, 5 and 43, respectively. Groups labeled with different letters indicate significant difference in gene contents at $P < 0.01$ (Tukey's HSD test). **c**, Expression of *TomLoxC* during fruit development of LA2093 and NC EBR-1. $n = 3$ independent experiments for LA2093 and 4 for NC EBR-1. Br, breaker; Br + 3d, breaker plus 3 d; heterozygous, containing both alleles; MG, mature green; nonreference, homozygous nonreference allele; reference, homozygous reference allele; ripe, red ripe fruit. Asterisks (**) indicate significant difference (two-tailed Student's *t*-test; $\alpha < 0.01$) of *TomLoxC* expression between LA2093 and NC EBR-1. For each boxplot, the lower and upper bounds of the box indicate the first and third quartiles, respectively, and the center line indicates the median.

gene start positions) of 2,823 reference genes. To identify promoter sequences possibly under selection during tomato domestication and improvement, we checked PAV patterns of these promoters, as well as those in the reference genome (Supplementary Fig. 9a,b).

A total of 856 and 388 sequences were under selection during domestication and improvement, respectively (Supplementary Table 8). Similar to the selection pattern of protein-coding genes, domestication exerted greater influence on the promoter sequences than did improvement. Among these promoter sequences, 717 (83.8%) and 385 (99.2%) were unfavorable during domestication and improvement, respectively. A conserved selection preference from domestication to improvement was also observed for most unfavorable promoters, with 89.9% of them displaying a similar trend in frequency changes from SP to SLC and from SLC to SLL (Supplementary Fig. 9c,d).

For the 980 promoter sequences that were under selection in at least one of the two stages, we checked the expression of their downstream genes in the 397 accessions for which RNA-Seq data were available for orange-stage fruit[4]. Of these promoters, 240 had downstream genes with significantly different expression (adjusted $P$ value < 0.01, two-tailed Student's *t*-test) associated with their presence and absence (Supplementary Table 8), suggesting that human selection influenced fruit quality or additional phenotypes in some instances by targeting regulatory sequences.

**A rare promoter allele that modifies fruit flavor.** Aroma volatiles have long been known to provide some of the unique flavor components of tomato fruit[32,33]. Recent studies revealed the importance of specific volatiles to the overall liking of tomato fruit, as well as for aroma intensity and specific flavor characteristics[9,34]. In particular, short-chain alcohols and aldehydes derived from fatty acids, amino acids and carotenoids play crucial roles in determining consumer acceptance of tomato fruit[9,34]. Many of the favorable alleles at multiple loci have been lost in recent years as a result of breeding emphasizing production over quality traits[9].

Our pan-genome analysis identified an ~4-kb substitution in the promoter region of *TomLoxC* (*Solyc01g006540*) (Supplementary Note, Supplementary Table 9 and Supplementary Fig. 10), which encodes a 13-lipoxygenase previously shown to be essential for C5 and C6 green-leaf volatile production in tomato fruit[35,36]. The two identified alleles were 149 bp upstream of the transcriptional start site: a 4,724-bp allele present in the reference 'Heinz 1706' genome (reference allele) and a 4,151-bp nonreference allele captured in our pan-genome. The nonreference allele was present in 91.2% of SP, 15.1% of SLC, and 2.2% of SLL heirlooms, indicating strong negative selection during both domestication and improvement. Further analysis indicated that only six accessions (two SP and four SLC) contain the homozygous nonreference allele, whereas 95 (50 SP, 29 SLC, 5 heirloom SLL, 10 modern SLL and 1 SCG) contain both alleles and the remaining 473 possess the homozygous reference allele (Fig. 4a and Supplementary Table 9). The frequency of the nonreference allele was highest in SP (47.4%) and declined dramatically in SLC (8.4%) and SLL heirlooms (1.1%), but interestingly recovered in modern SLL cultivars (7.2%), most likely because of recent introgressions from wild into cultivated tomatoes. Gene expression analysis based on RNA-Seq data from orange-stage fruit revealed that accessions containing both alleles displayed significantly higher expression levels of *TomLoxC* than those homozygous for either the reference or nonreference allele (Fig. 4b and Supplementary Table 10).

Given the association of TomLoxC with fruit flavor, we performed quantitative trait locus (QTL) mapping for 65 volatiles, including those derived from nutritionally important molecules such as carotenoids, essential fatty acids and amino acids, using a recombinant inbred line (RIL) population (Supplementary Table 11). The RIL population was derived from a cross between LA2093, an SP accession containing the homozygous nonreference *TomLoxC* promoter allele, and NC EBR-1, an advanced breeding line harboring the homozygous reference allele[37]. LA2093 and NC EBR-1 displayed contrasting expression patterns of *TomLoxC* during fruit development (Fig. 4c). We identified 116 QTLs for 56 volatiles across the 12 chromosomes (Supplementary Note, Supplementary Figs. 11 and 12 and Supplementary Tables 12–17). Interestingly, 28 volatiles, including 19 fatty-acid-derived volatiles

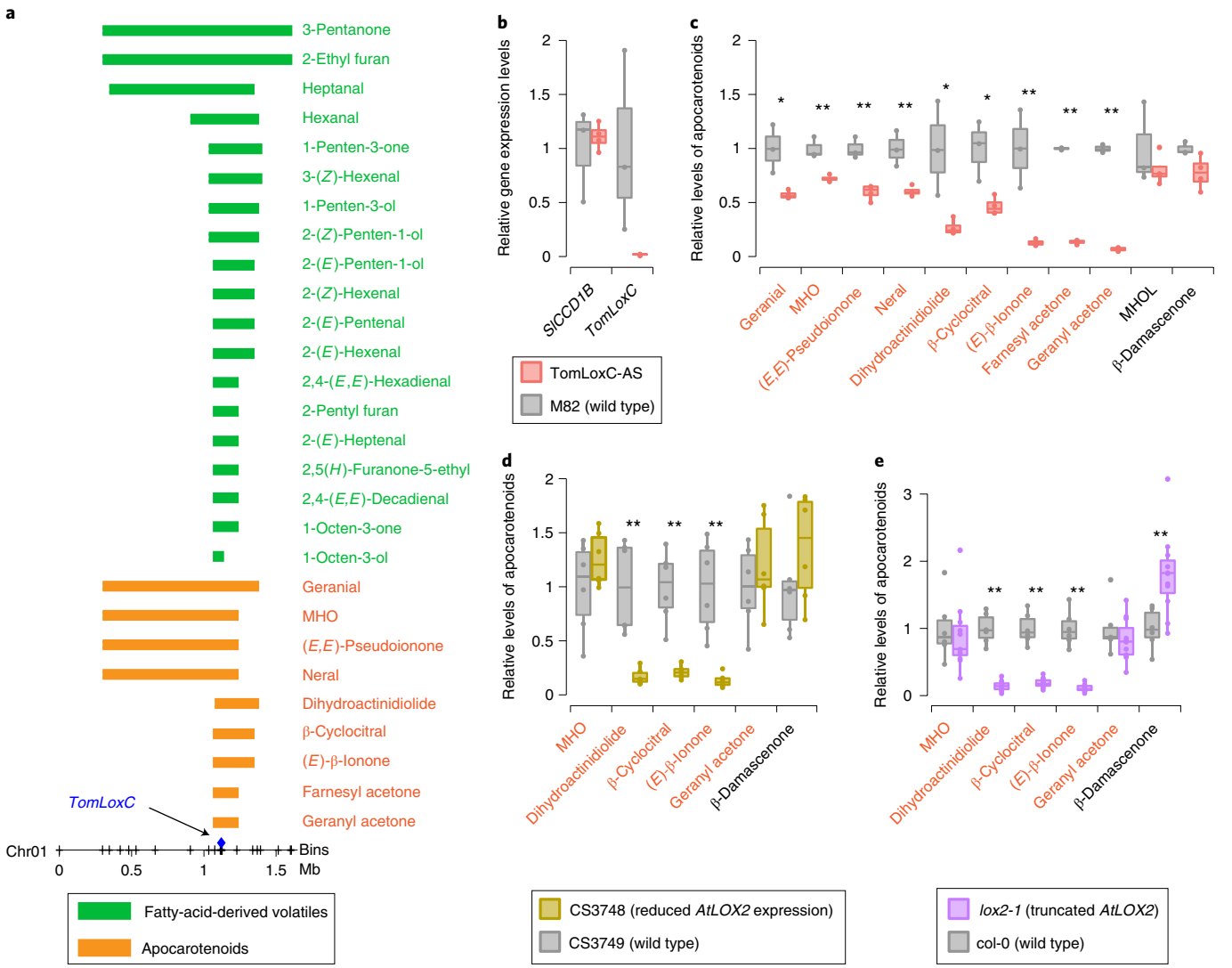


**Fig. 5 | Involvement of TomLoxC in apocarotenoid biosynthesis. a**, QTL interval for apocarotenoids and fatty-acid-derived volatiles on chromosome 1. **b**, Expression levels of *TomLoxC* and *SlCCD1B* in ripe fruits of TomLoxC-AS (*TomLoxC* antisense) and M82 plants. $n = 3$ independent experiments for M82 and 4 for TomLoxC-AS. **c**, Relative levels of apocarotenoids in ripe fruits of TomLoxC-AS and M82 plants. $n = 3$ independent experiments for M82 and 4 for TomLoxC-AS. **d**,**e**, Relative levels of apocarotenoids in Arabidopsis leaves of *AtLOX2* mutants and the corresponding controls. $n = 6$ independent experiments for CS3748, CS3749 and col-0, and 11 for *lox2-1*. Volatiles accumulated in significantly different levels (two-tailed Student's *t*-test) in target plants compared with the controls are marked with asterisks ($^*\alpha < 0.05$ or $^{**}\alpha < 0.01$). Apocarotenoids with QTL at the *TomLoxC* position are in red text and those without QTL are in black text. For each boxplot, the lower and upper bounds of the box indicate the first and third quartiles, respectively, and the center line indicates the median.

and 9 apocarotenoids, shared a QTL at the same location on chromosome 1 spanning a 153-kb interval (Fig. 5a) containing 19 genes including *TomLoxC*, which had the highest expression levels in RILs and largest expression difference between the two parents (Supplementary Table 18). The NC EBR-1 allele was associated with high levels of all 28 volatiles in concert with elevated expression of *TomLoxC* (Supplementary Table 12). These results strongly suggest that *TomLoxC* is the candidate gene underlying this QTL and might additionally play a role in apocarotenoid biosynthesis.

To verify the involvement of TomLoxC in apocarotenoid biosynthesis, we determined levels of 11 apocarotenoids and fatty-acid-derived volatiles in ripe fruits of transgenic tomatoes in which *TomLoxC* expression was repressed[36], and the expression of a previously known apocarotenoid biosynthesis gene, *SlCCD1B* (*Solyc01g087260*), remained unchanged (Fig. 5b). As expected, the majority of fatty-acid-derived volatiles showed significantly reduced levels in transgenic fruits (Supplementary Table 19). The levels of the nine apocarotenoids having a QTL at the *TomLoxC* position were also significantly reduced in transgenic fruits, whereas the levels of two other apocarotenoids without a QTL at this region, as well as their corresponding carotenoid substrates, were not affected (Fig. 5c and Supplementary Table 19). We further investigated apocarotenoid levels in two Arabidopsis mutants of the *AtLOX2* gene, the closest homolog of *TomLoxC*. Both mutants showed significantly reduced levels of specific apocarotenoids (Fig. 5d,e), further supporting the contribution of 13-lipoxygenases (for example, TomLoxC and AtLOX2) to apocarotenoid biosynthesis. Even though the involvement of LOX enzymes in volatile and nonvolatile apocarotenoid production was demonstrated in vitro in a co-oxidation mechanism coupled to fatty acid catabolism[38] (Supplementary Note), it is demonstrated here to be active in vivo. Furthermore, transgenic tomato fruits with decreased expression of *SlHPL*[35],

which follows LOX in C6 volatile biosynthesis, accumulated higher levels of C5 volatiles and cyclic apocarotenoids (Supplementary Note and Supplementary Figs. 13 and 14). Because the C5, not the C6, pathway has been proposed to additionally involve a LOX activity, this further supports the co-oxidation hypothesis. Finally, transgenic tomato with reduced *SlCCD1B* expression showed only up to 60% reduction in apocarotenoid levels[36]. The existence of a non-carotenoid cleavage dioxygenase pathway to apocarotenoids might explain this residual accumulation of these compounds (Supplementary Note).

## Discussion

We have constructed a pan-genome of cultivated tomato and its close relatives, which includes a 351-Mb sequence and 4,873 protein-coding genes not captured by the reference genome. The observation that 25.8% of genes in the pan-genome exhibit varying degrees of PAVs highlights the diverse genetic makeup of tomato with potential utility for future improvement. It is well known that cultivated tomatoes contain a narrow genetic base compared with their wild progenitors, although the specific lineages of SP contributing to domestication remain unknown. Here we show that at least part of this genetic diversity reduction could be attributed to substantial gene losses during domestication and improvement. Our PAV analysis suggests the loss of ~200 genes within SP took place in northern Ecuador, with gene losses continuing through subsequent domestication of SLC in South America and on to Mesoamerica. These findings point to northern Ecuador as a region for assessment of further accessions that may encompass additional genetic diversity useful for tomato breeding and in identifying more precisely the origins of domesticated tomatoes. Examination of the pan-genome further revealed that substantial gene content recovery has been achieved in modern commercial cultivars possibly because of intense introgression from diverse wild donors. Comparative analyses of the tomato pan-genome revealed extensive domestication- and improvement-associated loci and genes, with an evident bias toward those involved in defense response. It is unclear why these genes may have been disproportionally lost, although we speculate it could reflect a fitness cost of nonutilized defense genes (negative selection) or random loss caused by the absence of any positive selection force for their retention. Furthermore, it seems that selection against promoter regions that affect downstream gene expression had also shaped tomato domestication and improvement of genetic outcomes.

Modern tomato breeding has primarily focused on yield, shelf-life and resistance to biotic and abiotic stresses[39], often ignoring organoleptic/aroma quality traits that are difficult to select, resulting in decline of flavor-associated volatiles[9]. Because the reference genome is a modern processing tomato cultivar, at least some flavor-associated alleles may be absent in this accession. A nonreference allele of the *TomLoxC* promoter captured in the pan-genome represents a rare allele in cultivated tomatoes that reflects strong negative selection during domestication. Heterozygous *TomLoxC* promoter genotypes have the strongest expression in orange-stage fruit. Interestingly, the *TomLoxC* rare allele experienced a recovery in modern elite breeding lines (7.25% versus 1.13% in SLL heirlooms, all heterozygotes), consistent with its selection during modern breeding, possibly the consequence of selecting lines with superior stress tolerance in agricultural settings. In addition, QTL mapping pointed to *TomLoxC* as the cause of changed levels of flavor-associated lipid- and carotenoid-derived volatiles. Analysis of transgenic tomato fruit reduced in *TomLoxC* expression revealed a previously unknown alternative apocarotenoid production route, likely to be nonenzymatic, in addition to that initiated by carotenoid cleavage dioxygenases. Apocarotenoids are positively associated with flavor and overall liking of tomato fruit[9], and are components of the tomato fruit aroma[40]. Because of their very low perception threshold[33], apocarotenoids present an attractive target for improving tomato flavor at minimal metabolic expense.

The tomato pan-genome harbors useful genetic variation that has not been available to researchers and breeders relying on the 'Heinz 1706' reference genome alone. We demonstrate here that such variation may have important phenotypic outcomes that could contribute to crop improvement. The constructed tomato pan-genome represents a comprehensive and important resource to facilitate mining of natural variation for future functional studies and molecular breeding.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41588-019-0410-2.

Received: 18 November 2018; Accepted: 3 April 2019;
Published online: 13 May 2019

## References

1. The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635–641 (2012).
2. Bauchet, G. & Causse, M. in *Genetic Diversity in Plants* (Intech, 2012).
3. Tanksley, S. D. The genetic, developmental, and molecular bases of fruit size and shape variation in tomato. *Plant Cell* **16** (Suppl.), S181–S189 (2004).
4. Zhu, G. et al. Rewiring of the fruit metabolome in tomato breeding. *Cell* **172**, 249–261 (2018).
5. Labate, J. A. & Robertson, L. D. Evidence of cryptic introgression in tomato (*Solanum lycopersicum* L.) based on wild tomato species alleles. *BMC Plant Biol.* **12**, 133 (2012).
6. Kim, J. et al. Analysis of natural and induced variation in tomato glandular trichome flavonoids identifies a gene not present in the reference genome. *Plant Cell* **26**, 3272–3285 (2014).
7. Aflitos, S. et al. Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. *Plant J.* **80**, 136–148 (2014).
8. Lin, T. et al. Genomic analyses provide insights into the history of tomato breeding. *Nat. Genet.* **46**, 1220–1226 (2014).
9. Tieman, D. et al. A chemical genetic roadmap to improved tomato flavor. *Science* **355**, 391–394 (2017).
10. Blanca, J. et al. Genomic variation in tomato, from wild ancestors to contemporary breeding accessions. *BMC Genom.* **16**, 257 (2015).
11. Wang, W. et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**, 43–49 (2018).
12. Causse, M. et al. Whole genome resequencing in tomato reveals variation associated with introgression and breeding events. *BMC Genom.* **14**, 791 (2013).
13. Bolger, A. et al. The genome of the stress-tolerant wild tomato species *Solanum pennellii*. *Nat. Genet.* **46**, 1034–1038 (2014).
14. Strickler, S. R. et al. Comparative genomics and phylogenetic discordance of cultivated tomato and close wild relatives. *PeerJ* **3**, e793 (2015).
15. Itkin, M. et al. Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science* **341**, 175–179 (2013).
16. Graham, J. S. et al. Wound-induced proteinase inhibitors from tomato leaves. II. The cDNA-deduced primary structure of pre-inhibitor II. *J. Biol. Chem.* **260**, 6561–6564 (1985).
17. de Kock, M. J. D., Brandwagt, B. F., Bonnema, G., de Wit, P. J. G. M. & Lindhout, P. The tomato Orion locus comprises a unique class of *Hcr9* genes. *Mol. Breed.* **15**, 409–422 (2005).
18. Ori, N. et al. The *I2C* family from the wilt disease resistance locus *I2* belongs to the nucleotide binding, leucine-rich repeat superfamily of plant resistance genes. *Plant Cell* **9**, 521–532 (1997).
19. Martin, G. B. et al. Map-based cloning of a protein kinase gene conferring disease resistance in tomato. *Science* **262**, 1432–1436 (1993).
20. Zhao, Q. et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* **50**, 278–284 (2018).
21. Golicz, A. A. et al. The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nat. Commun.* **7**, 13390 (2016).
22. Li, Y. H. et al. *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* **32**, 1045–1052 (2014).
23. Contreras-Moreira, B. et al. Analysis of plant pan-genomes and transcriptomes with GET_HOMOLOGUES-EST, a clustering solution for sequences of the same species. *Front. Plant Sci.* **8**, 184 (2017).

24. Gordon, S. P. et al. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nat. Commun.* **8**, 2184 (2017).
25. Hurgobin, B. et al. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. *Plant Biotechnol. J.* **16**, 1265–1274 (2018).
26. Montenegro, J. D. et al. The pangenome of hexaploid bread wheat. *Plant J.* **90**, 1007–1013 (2017).
27. Menda, N. et al. Analysis of wild-species introgressions in tomato inbreds uncovers ancestral origins. *BMC Plant Biol.* **14**, 287 (2014).
28. Shinozaki, Y. et al. High-resolution spatiotemporal transcriptome mapping of tomato fruit development and ripening. *Nat. Commun.* **9**, 364 (2018).
29. Saladié, M. et al. A reevaluation of the key factors that influence tomato fruit softening and integrity. *Plant Physiol.* **144**, 1012–1028 (2007).
30. Mu, Q. et al. Fruit weight is controlled by *Cell Size Regulator* encoding a novel protein that is expressed in maturing tomato fruits. *PLoS Genet.* **13**, e1006930 (2017).
31. Tiwari, P., Sangwan, R. S. & Sangwan, N. S. Plant secondary metabolism linked glycosyltransferases: An update on expanding knowledge and scopes. *Biotechnol. Adv.* **34**, 714–739 (2016).
32. Buttery, R. G., Teranishi, R., Flath, R. A. & Ling, L. C. in *Flavor Chemistry: Trends and Developments*, Vol. 388 (eds Teranishi, R., Buttery, R. G. & Shahidi, F.) 213–222 (American Chemical Society, 1989).
33. Buttery, R. G., Seifert, R. M., Guadagni, D. G. & Ling, L. C. Characterization of additional volatile components of tomato. *J. Agr. Food Chem.* **19**, 524–529 (1971).
34. Tieman, D. et al. The chemical interactions underlying tomato flavor preferences. *Curr. Biol.* **22**, 1035–1039 (2012).
35. Shen, J. et al. A 13-lipoxygenase, TomloxC, is essential for synthesis of C5 flavour volatiles in tomato. *J. Exp. Bot.* **65**, 419–428 (2014).
36. Chen, G. et al. Identification of a specific isoform of tomato lipoxygenase (TomloxC) involved in the generation of fatty acid-derived flavor compounds. *Plant Physiol.* **136**, 2641–2651 (2004).
37. Ashrafi, H., Kinkade, M. & Foolad, M. R. A new genetic linkage map of tomato based on a *Solanum lycopersicum*×*S. pimpinellifolium* RIL population displaying locations of candidate pathogen response genes. *Genome* **52**, 935–956 (2009).
38. Hayward, S., Cilliers, T. & Swart, P. Lipoxygenases: From isolation to application. *Compr. Rev. Food Sci. Food Saf.* **16**, 199–211 (2017).
39. Klee, H. J. & Tieman, D. M. The genetics of fruit flavour preferences. *Nat. Rev. Genet.* **19**, 347–356 (2018).
40. Baldwin, E. A., Scott, J. W., Shewmaker, C. K. & Schuch, W. Flavor trivia and tomato aroma: Biochemistry and possible mechanisms for control of important aroma components. *HortScience* **35**, 1013–1022 (2000).
41. Tettelin, H., Riley, D., Cattuto, C. & Medini, D. Comparative genomics: The bacterial pan-genome. *Curr. Opin. Microbiol.* **11**, 472–477 (2008).


## Acknowledgements

This research was supported by grants from the US National Science Foundation (IOS-1339287 to Z.F. and J.J.G.; IOS-1539831 to Z.F., J.J.G. and H.J.K.; and IOS-1564366 to E.v.d.K., J.C. and D.M.T.), BARD, the US–Israel Binational Agricultural Research and Development Fund, a Vaadia-BARD Postdoctoral Fellowship Award (FI-508-14 to I.G.) and the USDA Agricultural Research Service.

## Author contributions

Z.F., J.J.G., H.J.K., S.H. and E.v.d.K. designed and managed the project. I.G., E.A.B.-C., K.A.S., T.L.F., G.L.S., T.W.T., D.M.T., Y.X., M.J.D., J.B., J.C., M.R.F. and E.v.d.K. collected samples and performed experiments. L.G., I.G., H.S., Q.M. and K.B. performed data analyses. L.G. and I.G. wrote the manuscript. Z.F. and J.J.G. revised the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41588-019-0410-2.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Correspondence and requests for materials** should be addressed to J.J.G. or Z.F.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Methods

**Genome sequences of tomatoes in the Lycopersicon clade.** Genome sequencing data of 561 tomato accessions in the Lycopersicon clade published previously[1,7–9,12–14], including species SLL, SLC, SP and SCG, were downloaded from the National Center for Biotechnology Information Sequence Read Archive database (Supplementary Table 1). Genome sequences of a total of additional 166 accessions were generated here, with two shared among the previously sequenced 561 accessions. Genomic DNA was extracted from a single seedling from each of these 166 accessions using Qiagen's DNeasy 96 Plant Kit. Paired-end libraries with insert sizes of ~500 bp were constructed using the NEBNext Ultra DNA Library Prep kit (Illumina Inc.) according to the manufacturer's instructions and sequenced on an Illumina NextSeq platform using the paired-end 2×150 bp mode. For quality evaluation, we also generated Illumina genome data of 45× coverage for the reference cultivar 'Heinz 1706'.

**Pan-genome construction.** Raw Illumina reads were processed to consolidate duplicated read pairs into unique read pairs. The resulting reads were then processed to trim adapters and low-quality sequences using Trimmomatic[42] with parameters 'SLIDINGWINDOW:4:20 MINLEN:50'. The final high-quality cleaned Illumina reads from each sample were de novo assembled using Megahit[43] with default parameters. The assembled contigs with lengths >500 bp were kept and then aligned to the tomato reference genomes, including the nuclear genome[1] (version SL3.0), chloroplast genome[44] (GenBank accession no.: NC_007898.3) and mitochondrial genome (SOLYC_MT_v1.50, http://www.mitochondrialgenome.org), using the nucmer tool in the Mummer package[45]. A reliable alignment was defined as a continuous alignment longer than 300 bp with sequence identity higher than 90%. Contigs with no reliable alignments were kept as unaligned contigs. For contigs containing the reliable alignments, if they also contained continuous unaligned regions longer than 500 bp, the unaligned regions were extracted as unaligned sequences. The unaligned contigs and unaligned sequences (>500 bp) were then searched against the GenBank nucleotide database using blastn[46]. Sequences with best hits from outside the green plants, or covered by known plant mitochondrial or chloroplast genomes, were possible contaminations and removed.

The cleaned nonreference sequences from all accessions were combined. The redundant sequences were consolidated into unique contigs using CD-HIT[47]. To further remove redundancies, we performed all-versus-all alignments with nucmer and blastn, respectively. The resulting nonredundant sequences were subsequently aligned against the reference genome using blastn to ensure no sequences were redundant with the reference genome. In all of the above filtering steps, the sequence identity threshold was set to 90%. The final nonredundant nonreference sequences and the reference tomato genome[1] (version SL3.0) were merged as the pan-genome.

The assembled contigs from the newly sequenced reads of the 'Heinz 1706' cultivar were aligned against the 'Heinz 1706' reference genome[1], using the nucmer tool[45], and sequences from the one-to-one alignment blocks were extracted and aligned with MUSCLE[48], to validate the quality of the de novo assemblies. Putative assembly errors were identified based on sequence variants between the assembled contigs and the reference genome.

**Annotation of the tomato pan-genome.** A custom repeat library was constructed by screening the pan-genome using MITE-Hunter[49] and RepeatModeler (http://www.repeatmasker.org/RepeatModeler/), and used to screen the nonreference genome to identify repeat sequences using RepeatMasker (http://www.repeatmasker.org/). Protein-coding genes were predicted from the repeat-masked nonreference genome using MAKER2 (ref. [50]). Ab initio gene prediction was performed using Augustus[51] and SNAP[52]. The 'tomato' model was selected for Augustus prediction, and SNAP was trained for two rounds based on RNA-Seq evidence according to MAKER2 instruction. RNA-Seq data of fruit pericarp tissues at the orange stage of 397 accessions reported in Zhu et al.[4] were used as transcript evidence. The raw RNA-Seq reads were processed to trim adapter and low-quality sequences using Trimmomatic[42]. Potential ribosomal RNA (rRNA) reads were filtered using SortMeRNA[53]. The final cleaned RNA-Seq reads were then mapped to the pan-genome using Hisat2 (ref. [54]), and the resulting alignments were used to construct gene models using StringTie[55]. Furthermore, reads mapped to the nonreference genome were extracted and then de novo assembled for each individual accession using Trinity[56]. The assembled transcripts from all accessions were combined, and the redundant sequences were removed using CD-HIT[47]. The resulting nonredundant sequences were aligned to the nonreference genome using Spaln[57]. In addition, protein sequences of Arabidopsis, rice and all asterid species were downloaded from RefSeq and aligned to the nonreference genome using Spaln[57]. Finally, gene predictions based on ab initio approaches, and transcript and protein evidence were integrated using the MAKER2 pipeline[50]. A set of high-confidence gene models supported by transcript and/or protein evidence were generated by MAKER2. The remaining ab initio predicted gene models were checked against the InterPro domain database using InterProScan[58]. Gene models containing InterPro domains were recovered and added to the final predicted gene set. Predicted genes with deduced protein length shorter than 50 amino acids, or overlapping with repeat sequences for more than 50% of their transcript length were removed.

Genes were functionally annotated by comparing their protein sequences against the GenBank nonredundant database and InterPro domain database. GO annotation and enrichment analysis were performed using the Blast2GO suite[59].

**PAV analysis.** Genome reads from each accession were aligned to the pan-genome using BWA-MEM[60] with default parameters. The presence or absence of each gene in each accession was determined using SGSGeneLoss[61]. In brief, for a given gene in a given accession, if less than 20% of its exon regions were covered by at least two reads (minCov = 2, lostCutoff = 0.2), this gene was treated as absent in that accession, otherwise it was considered present.

A maximum-likelihood phylogenetic tree was constructed based on the binary PAV data with 1,000 bootstraps using IQ-TREE[62]. Population structure based on the same PAV data was investigated using STRUCTURE[63]. Fifty independent runs for each *K* from 1 to 10 were performed with an admixture model at 50,000 Markov chain Monte Carlo (MCMC) iterations and a 10,000 burn-in period. The best *K* value was determined by the 'Evanno' method implemented in STRUCTURE HARVESTER[64]. Principal component analysis using the PAV data was performed with TASSEL5 (ref. [65]). To identify genes under selection during domestication or improvement, their presence frequencies in each of the three groups (SLL heirlooms, SLC and SP) were derived. The significance of the difference of the presence frequencies for each gene between the two compared groups (SP versus SLC for domestication and SLC versus SLL for improvement) was determined using the Fisher's exact test. The resulting raw *P* values of all genes in each of the two comparisons were then corrected via false discovery rate (FDR). Genes with significantly different frequencies (FDR < 0.001 and fold change >2) were identified as those under selection. GO enrichment analysis was performed for the favorable or unfavorable gene sets using the FatiGO package integrated in the Blast2GO suite[59] with a cutoff of FDR < 0.05.

**Anchoring of nonreference sequences and selection of promoter sequences.** For the nonreference sequences, if the ends of their source contigs had reliable and unique alignments to the reference genome (described earlier in this article), their defined genome positions could be assigned based on these alignments. For the remaining nonreference sequences, if they contained uniquely mapped hanging read pairs, that is, one read of the read pairs was uniquely mapped to the reference genome, their genomic positions on the reference genome could be deduced based on the alignments of these hanging read pairs. Because both of the earlier strategies were based on unique alignments, they might fail to localize sequences with extensive repeats on their ends.

PAV patterns of promoters (<1 kb to gene start positions) in both reference and nonreference sequences were derived. For promoters in the nonreference sequences, only those connected to the downstream genes supported by three or more hanging read pairs were included in the analysis. A promoter sequence in a given accession was considered 'present' if at least 50% of its length was covered by two or more reads, whereas a promoter sequence was considered 'absent' if no more than 20% of its length was covered. For each promoter sequence, accessions not assigned with presence or absence were excluded from subsequent analyses. Based on their PAV patterns, the promoter sequences were analyzed to identify those under selection during domestication and improvement, using the same method for protein-coding genes.

**RNA sequencing, SNP calling and expression analysis.** A total of 146 $F_{10}$ RILs and their two parents, *S. lycopersicum* breeding line NC EBR-1 and SP accession LA2093, were grown in triplicates in an open field in Live Oak, Florida. From each plant, at least four fruits were harvested at the red ripe stage, and pericarp tissues were flash-frozen in liquid $N_2$ and then pooled. Total RNA was extracted using the QIAGEN RNeasy Plant Mini Kit following the manufacturer's instructions (QIAGEN). RNA quality was evaluated via agarose gel electrophoresis, and the quantity was determined on a NanoDrop (Thermo Fisher Scientific).

Strand-specific RNA-Seq libraries were constructed from the total RNA using the protocol described in Zhong et al.[66], and sequenced on an Illumina HiSeq2500 platform with single-end 100-bp read length. At least three independent biological replicates were prepared for each sample. In addition, besides LA2093, RNA-Seq data were also generated from orange-ripe fruits of four additional accessions (BGV006231, BGV006859, BGV006904 and BGV006906) with the homozygous nonreference allele of *TomLoxC* promoter (Supplementary Table 10). Raw RNA-Seq reads were processed to remove adapter, low-quality and poly A/T tails using Trimmomatic[42]. Trimmed reads longer than 40 bp were kept and aligned to the SILVA rRNA database (https://www.arb-silva.de/) to filter out rRNA reads. The resulting high-quality cleaned reads were aligned to the reference 'Heinz 1706' genome (version SL3.0) using HISAT2 (ref. [54]) allowing two mismatches. Following alignments, raw counts for each gene were derived and normalized to RPKM.

To identify SNPs across the RILs and the two parents, we aligned the cleaned RNA-Seq reads to the reference 'Heinz 1706' genome using STAR[67] with the two-pass method and default parameters. Duplicated reads in each RNA-Seq library were marked using Picard (http://broadinstitute.github.io/picard/), and read alignments from biological replicates of the same samples were combined. SNPs were called using GATK (Genome Analysis Toolkit)[68] following the online

Best Practices protocol with recommended parameters for RNA-Seq data (https://software.broadinstitute.org/gatk/best-practices/). Other than high-quality SNPs assigned as 'PASS' by GATK, SNPs were further filtered to retain only those with different homozygous genotypes in the two parents, missing rate <0.2 and minor allele frequency >0.05.

**Volatile and carotenoid analyses.** Volatiles were analyzed via solid-phase microextraction (SPME) coupled to gas chromatography mass spectrometry according to Tikunov et al.[69] with minor modifications. In brief, 1.5 g frozen tissue powder was incubated for 2 min at 30 °C, and 1.5 ml of 100 mM EDTA (pH 7.5) was added to each sample and then thoroughly vortexed. Subsequently, 2 ml of the resultant slurry was transferred to a 10-ml glass vial containing 2.4 g $CaCl_2$, and 20 µl of 10 p.p.m. 2-octanone (Sigma-Aldrich) was added as the internal standard. Samples were sealed and stored at 4 °C for no more than 1 d before analysis. Samples were preheated to 50 °C for 5 min, and volatiles were sampled with a 1 cm long and 30/50 µm film thickness of divinylbenzene/Carboxen/polydimethylsiloxane SPME fiber (Supelco) at 50 °C for 30 min with 10 s agitation every 5 min.

Volatiles were analyzed by gas chromatography–time of flight (TOF)–mass spectrometry (Pegasus 4D; LECO Corp.), using a CP-Sil 8 CB (30 m × 0.25 mm × 0.25 µm) fused-silica capillary column (Agilent). The SPME fiber was introduced to the gas chromatography inlet, which was set to 250 °C in splitless injection, and 10 min was allowed for thermal desorption. Helium was used as a carrier gas at a constant flow rate of 1 ml × $min^{-1}$ in gas saver mode. The initial oven temperature was set to 45 °C for 5 min, then raised to 180 °C at a rate of 5 °C per minute, and then to 280 °C at 25 °C increase per minute and held for an additional 5 min. The TOF–mass spectrometry was operated in electron ionization (EI) mode with an ionization energy of 70 eV, and the electron multiplier voltage was set to 1,700 V. Mass spectrometry data from 41 to 250 *m/z* were stored at an acquisition rate of 8 spectra per second. Data processing was performed using LECO ChromaTOF software. To resolve retention indices, we injected a mixture of straight-chain alkanes (C6–C25) into the column under the same conditions. Calculated retention indices and mass spectra were compared with the NIST mass spectral database for compound identification. Relative quantification was done based on single ion area normalized to the internal standard.

Carotenoids were extracted according to Alba et al.[70] and analyzed using super-critical fluid chromatography equipped with a diode array detector according to Gonda et al.[71].

**Map construction and QTL mapping.** To generate a map of genomic bins composed of the genotype of every individual in the RIL population, we used SNPbinner[71] with default parameters except that emission probability was set to 0.99. QTL analysis was performed using R/qtl (ref. [72]) with a script developed by Spindel et al.[73]. In brief, interval mapping was used for initial QTL detection, followed by multiple-QTL-model analysis in additive-only mode. Traits that were not normally distributed (as determined by the Shapiro–Wilk W test) were transformed by $\log_{10}$ or square root, and outliers were removed to reach normal distribution. Traits that did not reach normal distribution after transformations were analyzed considering nonparametric models.

**Functional characterization of *TomLoxC* and *AtLOX2*.** Antisense transgenic tomato plants with decreased *TomLoxC* expression described in Chen et al.[36] and the corresponding wild-type plants (M82) were grown in triplicate in a greenhouse in Ithaca, New York, with a 16-h light period at 20 °C (night) to 25 °C (day). The Arabidopsis *lox2-1* mutant[74] carrying a point mutation causing a premature stop of *AtLOX2* was obtained from Prof. Edward E. Farmer (University of Lausanne, Switzerland). Seeds of the *AtLOX2* reduced expression line (CS3748) and the corresponding control (CS3749) were obtained from the Arabidopsis Biological Resource Center. Arabidopsis plants were grown in soil with a 16-h light period at 22 °C with 60% humidity and were harvested after 6 weeks. Each sample was composed of two plants from the same genotype to achieve sufficient plant material needed for the SPME analysis.

**Real-time PCR.** Total RNA was treated with DNase (Invitrogen), and complementary DNA was synthesized using ProtoScript II reverse transcriptase (New England Biolabs). Real-time PCR was carried out on an Applied Biosystems QuantStudio 6 Flex Real-Time PCR System using the SYBR Green master mix (Life Technologies). Primer sequences used for *SlCCD1B*, *TomLoxC* and *SlRPL2* (*Solyc10g006580*; the internal control) are listed in Supplementary Table 20. Relative expression values were determined as $2^{-(\Delta\Delta Ct)}$ (ref. [75]).

**Statistical analysis.** The statistical tests used are described throughout the article and in the figure legends. Specifically, Fisher's exact test with FDR corrected for multiple comparisons was used to identify genes selected during domestication or improvement, and to identify enriched GO terms, we used Tukey's honest significant difference (HSD) test to determine the significance of difference of detected gene counts among different tomato groups, *TomLoxC* expression levels among accessions with different promoter types and expression levels of genes belonging to different groups. The two-tailed Student's *t*-test was performed to compare *TomLoxC* expression levels between NC EBR-1 and LA2093 at each fruit developmental stage, expression levels of *TomLoxC* and *SlCCD1B* between *TomLoxC-AS* and M82 fruits, relative levels of each volatile between mutants and corresponding wild-type controls, expression levels of genes between presence and absence of the promoters, and expression levels between reference and nonreference and between conserved and flexible genes.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

Raw genome and RNA-Seq reads have been deposited into the National Center for Biotechnology Information Sequence Read Archive under accession codes SRP150040, SRP186721 and SRP172989, respectively. The nonreference genome sequences and annotated genes of the tomato pan-genome and SNPs called from the RIL population are available via the Dryad Digital Repository (https://doi.org/10.5061/dryad.m463f7k).

## References


42. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
43. Li, D. et al. MEGAHITv1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**, 3–11 (2016).
44. Daniell, H. et al. Complete chloroplast genome sequences of *Solanum bulbocastanum*, *Solanum lycopersicum* and comparative analyses with other Solanaceae genomes. *Theor. Appl. Genet.* **112**, 1503–1518 (2006).
45. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
46. Camacho, C. et al. BLAST+: Architecture and applications. *BMC Bioinform.* **10**, 421 (2009).
47. Li, W. & Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
48. Edgar, R. C. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* **5**, 113 (2004).
49. Han, Y. & Wessler, S. R. MITE-Hunter: A program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucl. Acids Res.* **38**, e199 (2010).
50. Holt, C. & Yandell, M. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinform.* **12**, 491 (2011).
51. Stanke, M. & Morgenstern, B. AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucl. Acids Res.* **33**, W465–W467 (2005).
52. Korf, I. Gene finding in novel genomes. *BMC Bioinform.* **5**, 59 (2004).
53. Kopylova, E., Noe, L. & Touzet, H. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
54. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
55. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
56. Haas, B. J. et al. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
57. Iwata, H. & Gotoh, O. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucl. Acids Res.* **40**, e161 (2012).
58. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
59. Gotz, S. et al. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucl. Acids Res.* **36**, 3420–3435 (2008).
60. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
61. Golicz, A. A. et al. Gene loss in the fungal canola pathogen *Leptosphaeria maculans*. *Funct. Integr. Genom.* **15**, 189–196 (2015).
62. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
63. Hubisz, M. J., Falush, D., Stephens, M. & Pritchard, J. K. Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.* **9**, 1322–1332 (2009).
64. Earl, D. A. & vonHoldt, B. M. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* **4**, 359–361 (2012).
65. Bradbury, P. J. et al. TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).

66. Zhong, S. et al. High-throughput Illumina strand-specific RNA sequencing library preparation. *Cold Spring Harb. Protoc.* **2011**, 940–949 (2011).
67. Dobin, A. et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
68. McKenna, A. et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
69. Tikunov, Y. et al. A novel approach for nontargeted data analysis for metabolomics: Large-scale profiling of tomato fruit volatiles. *Plant Physiol.* **139**, 1125–1137 (2005).
70. Alba, R. et al. Transcriptome and selected metabolite analyses reveal multiple points of ethylene control during tomato fruit development. *Plant Cell* **17**, 2954–2965 (2005).
71. Gonda, I. et al. Sequencing-based bin map construction of a tomato mapping population, facilitating high-resolution quantitative trait loci detection. *Plant Genome* **12**, 180010 (2019).
72. Broman, K. W., Wu, H., Sen, S. & Churchill, G. A. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889–890 (2003).
73. Spindel, J. et al. Bridging the genotyping gap: Using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. *Theor. Appl. Genet.* **126**, 2699–2716 (2013).
74. Glauser, G. et al. Velocity estimates for signal propagation leading to systemic jasmonic acid accumulation in wounded *Arabidopsis*. *J. Biol. Chem.* **284**, 34506–34513 (2009).
75. Pfaffl, M. W. A new mathematical model for relative quantification in real-time RT-PCR. *Nucl. Acids Res.* **29**, e45 (2001).

nature research

Corresponding author(s): Zhangjun Fei & James Giovannoni

Last updated by author(s): Mar 18, 2019

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☒ | ☐ | A description of all covariates tested |
| ☐ | ☒ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | No software was used in data collection. |
| Data analysis | No commercial and custom code was used in this study. We only used freely available bioinformatics software our data analysis. |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw genome and RNA-Seq reads have been deposited into the NCBI Sequence Read Archive (SRA) under accessions SRP150040, SRP186721 and SRP172989, respectively. The non-reference genome sequences and annotated genes of the tomato pan-genome, and SNPs called from the RIL population are available via the Dryad Digital Repository (https://doi.org/dryad.m463f7k).

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | No sample size calculation was required for this study. |
| Data exclusions | For genome and RNA-Seq data, we only excluded sequences that were of low quality and potential contaminants from the analysis. This is standard for these types of analyses. For PAV analysis, we excluded samples with low sequencing depth and coverage of the genome based on our analysis (please see Supplementary Fig. 4 of the manuscript). |
| Replication | For RNA-Seq experiment of the RIL fruits and qRT-PCR experiment of RILs and transgenic plants, we used at least three biological replicates. |
| Randomization | This is not relevant to our study. |
| Blinding | Blinding was not relevant to our study. |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology |
| ☒ | ☐ Animals and other organisms |
| ☒ | ☐ Human research participants |
| ☒ | ☐ Clinical data |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |