nature genetics

Genomic analyses provide insights into the history of tomato breeding

Tao Lin[1,2,16], Guangtao Zhu[1,16], Junhong Zhang[3,16], Xiangyang Xu[4,16], Qinghui Yu[5,16], Zheng Zheng[1,16], Zhonghua Zhang[1], Yaoyao Lun[1], Shuai Li[1], Xiaoxuan Wang[1], Zejun Huang[1], Junming Li[1], Chunzhi Zhang[1], Taotao Wang[3], Yuyang Zhang[3], Aoxue Wang[4], Yancong Zhang[6], Kui Lin[6], Chuanyou Li[7], Guosheng Xiong[2,7], Yongbiao Xue[8,9], Andrea Mazzucato[10], Mathilde Causse[11], Zhangjun Fei[12], James J Giovannoni[12], Roger T Chetelat[13], Dani Zamir[14], Thomas Städler[15], Jingfu Li[4], Zhibiao Ye[3], Yongchen Du[1] & Sanwen Huang[1,2]

The histories of crop domestication and breeding are recorded in genomes. Although tomato is a model species for plant biology and breeding, the nature of human selection that altered its genome remains largely unknown. Here we report a comprehensive analysis of tomato evolution based on the genome sequences of 360 accessions. We provide evidence that domestication and improvement focused on two independent sets of quantitative trait loci (QTLs), resulting in modern tomato fruit ~100 times larger than its ancestor. Furthermore, we discovered a major genomic signature for modern processing tomatoes, identified the causative variants that confer pink fruit color and precisely visualized the linkage drag associated with wild introgressions. This study outlines the accomplishments as well as the costs of historical selection and provides molecular insights toward further improvement.

Global food production relies heavily on the capacity and effectiveness of plant breeding[1]. More than 10,000 years ago, our Neolithic ancestors successfully domesticated hundreds of wild plant species into cultivated crops that remain principal human food sources[2,3]. Domestication can be regarded as the first stage of plant breeding and is often followed by species distribution along corridors of human migration. Migration and subsequent differential selection by local farmers likely contributed to geographical differences in preferences for cultivated species and traits[4]. Until recently, crop breeding has relied heavily on accumulated experience and careful observation. The current genomic era, enabled by next-generation DNA sequencing technologies, offers new and powerful tools for targeted and precise selection. Scientists can now 'read' entire genomes during selection and track the history of plant breeding via population genomics, as recently demonstrated in rice[5], maize[6], soybean[7] and cucumber[8]. These studies illustrate how human-involved evolutionary processes have shaped modern crop genomes and provide insights for further crop improvement.

Tomato (*Solanum lycopersicum*) has a worldwide distribution and is considered the leading vegetable crop, with a global yield of 162 million tons in 2012 (United Nations Food and Agriculture Organization (FAO) statistics; see URLs) and a net value of over $55 billion[9]. Tomato is also an important model system for plants and especially for fleshy fruit biology[10]. It represents the cornerstone for biological research on and genetic improvement of all solanaceous crops, including potato, pepper and eggplant. Tomato and its wild relatives originated from the Andean region of South America. Cherry tomato (*S. lycopersicum* var. *cerasiforme*) is considered the probable ancestor of the big-fruited tomato and was likely domesticated from the red-fruited wild species *Solanum pimpinellifolium*[11]. Tomatoes were brought to Europe by the conquistadors in the sixteenth century[12], and subsequent migration and continued selection reduced the genetic diversity of this crop. To further boost the performance of modern tomato cultivars, wild tomato genomes were deliberately introgressed into elite cultivars[13]. However, how human selection has changed the tomato genome remains largely unknown.

[1]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, China. [2]Agricultural Genome Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China. [3]Key Laboratory of Horticultural Plant Biology, Huazhong Agricultural University, Wuhan, China. [4]College of Horticulture, Northeast Agricultural University, Harbin, China. [5]Institute of Horticulture, Xinjiang Academy of Agricultural Sciences, Urumqi, China. [6]College of Life Sciences, Beijing Normal University, Beijing, China. [7]State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences and National Plant Gene Research Centre, Beijing, China. [8]State Key Laboratory of Molecular Developmental Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences and National Plant Gene Research Centre, Beijing, China. [9]Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China. [10]Department of Agriculture, Forests, Nature and Energy (DAFNE), University of Tuscia, Viterbo, Italy. [11]Institut National de la Recherche Agronomique (INRA), Unité de Génétique et Amélioration des Fruits et Légumes, Domaine Saint-Maurice, Montfavet, France. [12]Boyce Thompson Institute for Plant Research, US Department of Agriculture (USDA) Robert W. Holley Center for Agriculture and Health, Cornell University, Ithaca, New York, USA. [13]C.M. Rick Tomato Genetics Resource Center, Department of Plant Sciences, University of California, Davis, Davis, California, USA. [14]Robert H. Smith Institute of Plant Sciences and Genetics, Faculty of Agriculture, Hebrew University of Jerusalem, Rehovot, Israel. [15]Plant Ecological Genetics, Institute of Integrative Biology, Eidgenössische Technische Hochschule (ETH) Zurich, Zurich, Switzerland. [16]These authors contributed equally to this work. Correspondence should be addressed to S.H. (huangsanwen@caas.cn), Y.D. (duyongchen@caas.cn), Z.Y. (zbye@mail.hzau.edu.cn) or Jingfu Li (lijf_2005@126.com).

Received 20 July; accepted 22 September; published online 12 October 2014; doi:10.1038/ng.3117

RESULTS

A map of tomato genome variation

We constructed a genomic history of tomato breeding by analyzing the genomes of 360 diverse accessions collected from around the world. These included 333 accessions from the red-fruited clade (*S. pimpinellifolium*, *S. lycopersicum* var. *cerasiforme* and *S. lycopersicum*) representing various geographical origins, consumption types and improvement statuses, 10 accessions of wild tomato species including some known donors of disease resistance genes (*R* genes) and 17 modern commercial hybrids (F_1) (**Supplementary Table 1**). Resequencing of the 360 accessions generated a total of 2.6 trillion base pairs of sequence, with a median depth of 5.7× and coverage of 93.1% of the assembled genome[14] (release SL2.40) (**Supplementary Table 1**). After aligning the reads against the tomato reference genome, we generated a final set of 11,620,517 SNPs and 1,303,213 small indels (shorter than 5 bp) (**Supplementary Tables 2** and **3**). The accuracy of the identified SNPs was estimated to be 98.4% and 97.6% using Sanger sequencing (349 SNPs in 3 accessions) and existing SNP array data[15] (48 accessions with an average of 5,800 SNPs), respectively (**Supplementary Tables 4** and **5**, and **Supplementary Note**). We identified 207,306 nonsynonymous SNPs in 30,945 genes, including 12,035 nonsense SNPs in 7,678 genes that caused start codon changes, the introduction of premature stop codons or the production of elongated transcripts. This SNP data set represents a new resource for tomato biology and breeding.

We explored the phylogenetic relationships among the accessions using 20,111 SNPs (minor allele frequency (MAF) > 0.05) at fourfold-degenerate sites that represent neutral or near-neutral variants. The resulting neighbor-joining tree (**Fig. 1a**) supports the clustering of the red-fruited clade. Interestingly, three *Solanum cheesmaniae* accessions and one *Solanum galapagense* accession were situated within this clade, consistent with previous studies[16]. These two species bear small mature fruits (2–3 g) of yellow or red color. Endemic to the Galapagos Islands, they can be experimentally crossed to *S. lycopersicum* without difficulty. On the basis of passport information, fruit weight and other morphological traits, we assigned the 331 red-fruited accessions to 3 groups (2 accessions could not be assigned as fruit weight was highly segregated within each accession): PIM including 53 *S. pimpinellifolium* accessions (fruit weight = 2.04 ± 0.85 g), CER including 112 *S. lycopersicum* var. *cerasiforme* accessions (fruit weight = 13.29 ± 9.54 g) and BIG including 166 big-fruited *S. lycopersicum* accessions (fruit weight = 111.33 ± 68.19 g) (**Supplementary Fig. 1**). The neighbor-joining tree largely supported this division, although some discrepancies between phenotypic characterization and phylogenetic clustering can be anticipated due to shared ancestral variation and historical gene flow among these very closely related groups (**Fig. 1a**). For instance, some cherry tomatoes residing in the BIG group could be the feral descendants of cultivated tomatoes or the products of introgression between the groups, as previously hypothesized[17].

Model-based clustering analyses (**Fig. 1b** and **Supplementary Figs. 2** and **3**) enabled the division of the CER group into two main clusters. One cluster showed obvious admixture in genetic composition and consisted of 49 accessions mainly distributed in South America, where CER accessions might experience occasional gene flows with the local wild relative *S. pimpinellifolium*. Another cluster was homogeneous with the BIG group in genetic composition and contained 38 accessions mainly of non–South American origin (from Mesoamerica, Europe, North America and Asia). Within the BIG group, we identified a cluster of processing tomatoes, highlighting their genetic similarity and distinction from other accessions (**Fig. 1b**). Nucleotide diversity measured by the π value[18] for the PIM group (3.23×10^{-3}) was substantially higher than that for the CER (1.74×10^{-3}) and BIG (0.73×10^{-3}) groups. In addition, the PIM group had more private SNPs (582,954) than the CER (207,892) and BIG (194,919) groups. We aligned the *Solanum pennellii* genome[19] to the reference Heinz 1706 genome. Of the ~11.6 million SNP sites, ~3.5 million could be reliably recovered in the *S. pennellii* genome. Among the ~3.5 million SNPs, on average, 30.4% of the sites in PIM accessions, 6.6% of the sites in CER accessions and 2.8% of the sites in BIG accessions were identical to the corresponding *S. pennellii* sites that are presumably ancestral alleles. The decay of linkage disequilibrium (LD) with physical distance between SNPs occurred at 8.8 kb in PIM ($r^2 = 0.2$), 256.8 kb in CER ($r^2 = 0.35$) and 865.7 kb in BIG

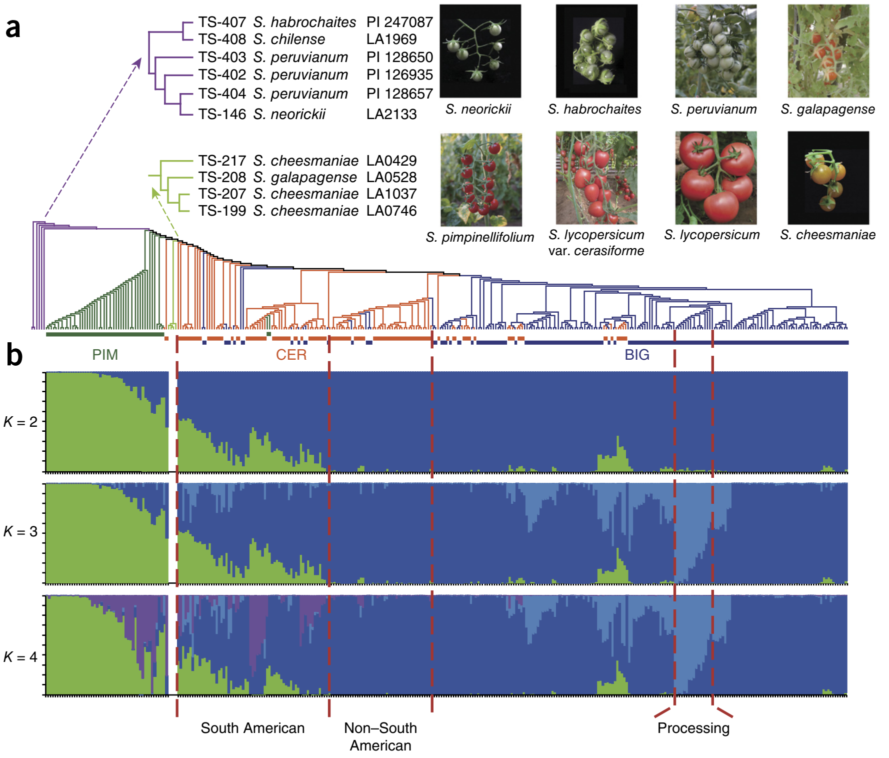

Figure 1 Genome-wide relationship and fruit morphology in cultivated tomato and its wild relatives. (**a**) The neighbor-joining tree of the population (331 accessions from the red-fruited clade and 10 wild accessions) was generated using 20,111 SNPs at fourfold-degenerate sites. The bars indicate the PIM (green), CER (orange) and BIG (blue) lines. The two branches containing wild accessions are enlarged for visualization. Typical fruits of the species studied are shown. (**b**) Model-based clustering analysis with different numbers of clusters (*K* = 2, 3 and 4). The *y* axis quantifies cluster membership, and the *x* axis lists the different accessions. The orders and positions of these accessions on the *x* axis are consistent with those for the neighbor-joining tree. South American CER, non–South American CER and processing tomato clusters are separated by dashed red lines.

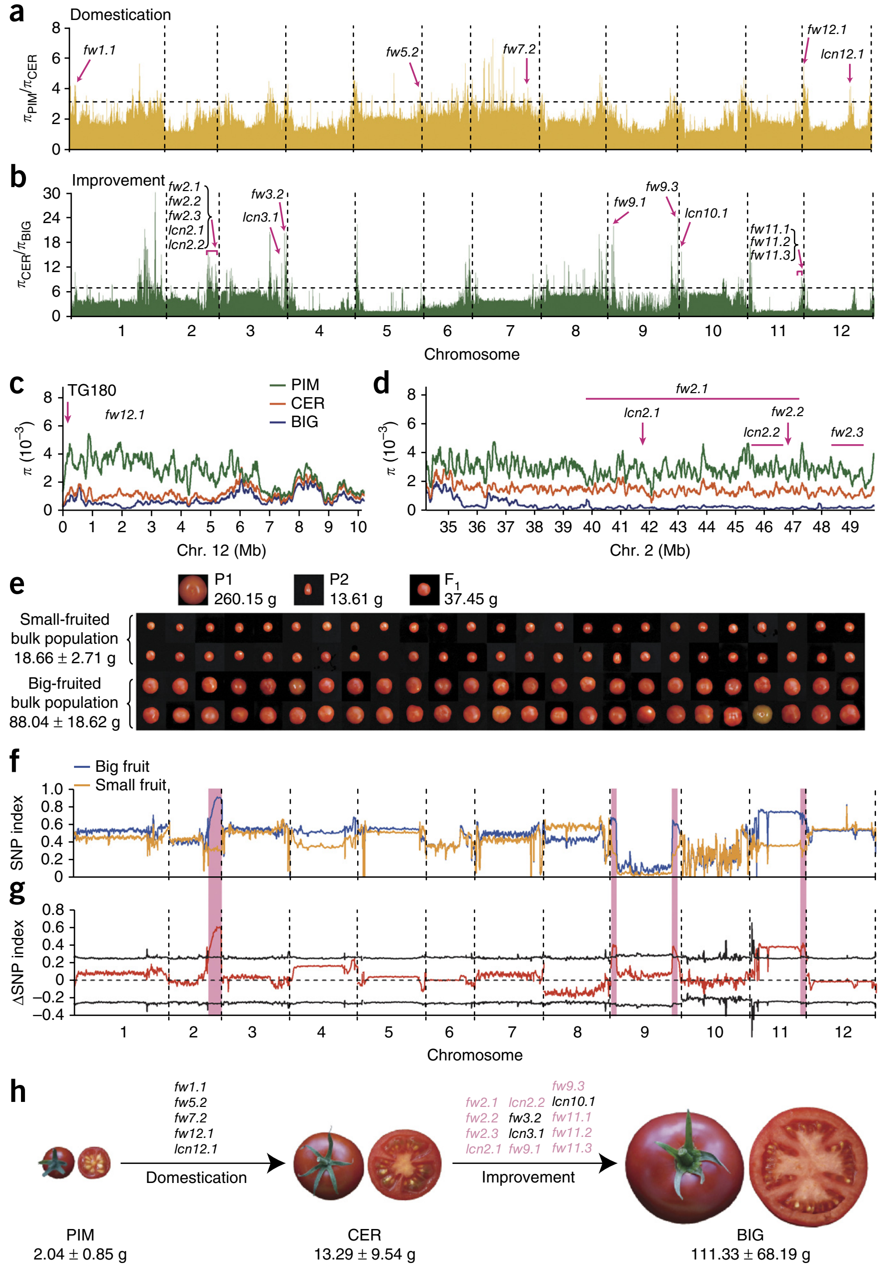

Figure 2 Evolution of fruit mass during domestication and improvement. (**a**,**b**) A total of 186 (top 5%; $\pi_{PIM}/\pi_{CER} \geq 3.0$) and 133 (top 5%; $\pi_{CER}/\pi_{BIG} \geq 6.9$) regions are considered to be candidate domestication sweeps (orange bars above the dashed horizontal threshold line) (**a**) and improvement sweeps (green bars above the dashed horizontal threshold line) (**b**), respectively. Pink arrows indicate the 5 and 13 QTLs related to fruit mass located within the domestication and improvement sweeps, respectively. (**c**,**d**) Distribution of nucleotide diversity (π) of the PIM (green), CER (orange) and BIG (blue) lines within the domestication sweep harboring *fw12.1* (with TG180 as the signature marker) (**c**) and within the improvement sweep harboring five fruit mass QTLs on chromosome 2 (**d**). (**e**–**g**) Verification of the improvement sweeps related to fruit mass. (**e**) Fruit phenotype and mass of the parental lines, the F_1 and the two bulk populations with extreme fruit size from the F_2 population each containing 50 individuals. (**f**) The SNP indices (ratio of the SNPs that are identical to those in the big-fruited parent) in the big-fruited and small-fruited bulk populations are shown using blue and orange lines, respectively. (**g**) The ΔSNP index (subtracting the SNP index of the small-fruited bulk population from that of the big-fruited bulk population) and its 95% confidence interval are shown using red and black lines, respectively. Regions with a ΔSNP index above the confidence line are highlighted with pink bars. (**h**) Schematic of the two-step evolution of tomato fruit size. QTLs that were putatively selected during domestication and improvement are listed, and those in pink were verified in this study. Note that the size of the photos is different for **e** and **h**.

($r^2 = 0.35$) accessions (**Supplementary Figs. 4** and **5**, and **Supplementary Note**). Overall, these morphological and genomic data support the PIM group as the ancestor of the red-fruited clade and the CER group as the evolutionary intermediate between PIM and BIG, consistent with a previous report[20]. We therefore define the evolutionary processes yielding the CER group from PIM as 'domestication' and yielding the BIG group from CER as 'improvement'. Demographic modeling using δaδi[21] suggested an effective population size of ~300 for tomato at domestication, an estimate similar to that for cucumber[8] (**Supplementary Table 6**).

Two-step evolution of fruit mass

It is well known that the indigenous people of the Andes domesticated quinoa, lima bean, peanut, potato, sweet potato and squash[2]. They likely also kept and propagated seeds from wild tomato plants with bigger and tastier fruits. Fruit mass is the key trait of human selection in tomato, as it affects both yield and quality. Typical fruits from PIM lines are tiny and have a thick skin, thin pericarp and high seed content (**Fig. 2**). Previous reports on segregating populations from crosses between PIM and BIG accessions identified multiple QTLs for fruit mass, including several genes that were cloned[22–25]. However, whether these QTLs and genes were selected during domestication or improvement remains elusive. To address this question, we scanned genomic regions with a drastic reduction in nucleotide diversity in the comparison of PIM and CER lines (π_{PIM}/π_{CER}; domestication sweeps) as well as in the comparison of CER and BIG lines (π_{CER}/π_{BIG}; improvement sweeps). In total, we identified 186 domestication sweeps and 133 improvement sweeps covering 8.3% (64.6 Mb) and 7.0% (54.5 Mb) of the assembled genome, harboring 5,605 and 4,807 genes, respectively (**Fig. 2a**,**b** and **Supplementary Tables 7**–**10**). We note that 21% of the domestication sweeps overlapped with improvement sweeps (8.0 Mb; 1.0% of the genome), indicating that some of the domestication loci might have undergone a second round of selection for further increase in fruit size and improvement of other agronomic traits. Jointly, the domestication and improvement sweeps occupied 111.0 Mb (14.2% of the assembled genome).

Five QTLs (**Supplementary Table 11**) related to fruit mass (*fw1.1*, *fw5.2*, *fw7.2*, *fw12.1* and *lcn12.1*) are located within the domestication sweeps and likely contributed to the enlargement of tomato fruits during the evolutionary transition from PIM to CER lines (**Fig. 2a**).

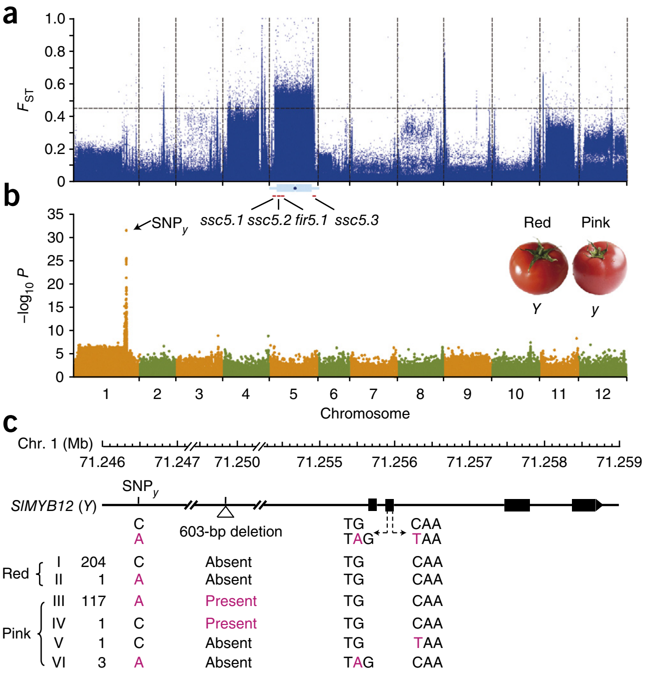

Figure 3 A major genomic signature of modern processing tomatoes and three causative variants for pink fruit. (**a**) F_{ST} values for all SNP sites between tomatoes for fresh consumption and modern processing tomatoes. Blue dots above the horizontal dashed line indicate highly divergent SNPs (top 1%; $F_{ST} = 0.4464$). Three SSC QTLs and a firmness QTL on chromosome 5 are indicated. (**b**) Manhattan plot of the GWAS f or fruit color using the compressed mixed linear model (MLM). A significantly associated SNP$_y$ is identified upstream of the *SlMYB12* (*y*) gene. (**c**) The structure of *SlMYB12* and the positions of SNP$_y$, the 603-bp deletion and the two other deleterious mutations. Exons are depicted as black blocks. The numbers of tomato accessions with the six genotypes (I–VI) are given. Note that genotypes II and IV are recombinants between SNP$_y$ and the 603-bp deletion.

For instance, the *fw12.1* QTL resides in a domestication sweep spanning the telomeric region of the short arm of chromosome 12 (**Fig. 2c**). Its signature marker, TG180, is physically close to the *Solyc12g005310* gene that encodes a putative auxin-responsive GH3-like protein with predominant expression in flower buds, making it a logical candidate for the gene corresponding to *fw12.1* (ref. 26). We detected 13 QTLs (**Supplementary Table 11**) located within improvement sweeps that might have contributed to the second round of fruit enlargement during the CER-to-BIG transition (**Fig. 2b**). Remarkably, a major improvement sweep spanned a 10.3-Mb region at the distal end of the long arm of chromosome 2, where two fruit mass QTLs (*fw2.2* and *lcn2.1*) were cloned and three others (*fw2.1*, *fw2.3* and *lcn2.2*) were mapped (**Fig. 2d**). Selection for these QTLs in the BIG group might be causative of the low genetic diversity in this region ($\pi = 0.22 \times 10^{-3}$ in the 10-Mb region versus 0.73×10^{-3} in the whole genome).

The gene *fw2.2* controls carpel cell number and contributes substantially to the evolution of tomato fruit mass and size[24]. However, its causative variation remains undetermined. We exploited the SNP data set to perform a local association study around *fw2.2* for allelic variation conferring the phenotypic change. A SNP in the promoter region (−912 bp relative to the start codon) showed a signal ($P < 1 \times 10^{-3}$) that was almost fixed in BIG accessions (97.3%) but not in CER accessions (66.7%). In PIM accessions, it was a minor allele (2.6%). Taking this finding together with the fact that *fw2.2* is not located within a domestication sweep, we infer that *fw2.2* is more likely an improvement rather than a domestication gene. The same held true for two other cloned QTLs: *lcn2.1*, which contributes to increased locule number[25], and *fw3.2*, which corresponds to a cytochrome P450 gene controlling fruit cell number[22]. The SNP for *fw2.2* could be developed as a marker for selecting recombinants to break apart the improvement sweep and to introduce new variations into this region for modern tomato breeding.

To further verify improvement sweeps related to fruit mass, we sequenced 2 bulk populations with extreme fruit size, each consisting of 50 progenies from an F_2 population of 500 individuals from a cross between the CER and BIG lines, to a depth of 50× (**Fig. 2e**, **Supplementary Fig. 6** and **Supplementary Note**). We called SNPs between two parental genomes, and we computed the SNP indices for the big-fruited and small-fruited bulk populations as well as their differences (ΔSNP index) using a 1,000-kb sliding window with a step size of 10 kb. This analysis led to the identification of four genomic regions contributing to fruit mass, all of which overlapped with previously identified improvement sweeps, i.e., the chromosome 2 distal end carrying *fw2.1*, *fw2.2*, *fw2.3*, *lcn2.1* and *lcn2.2*, both distal ends of chromosome 9 carrying *fw9.1* and *fw9.3*, and the distal end of the long arm of chromosome 11 carrying *fw11.1*, *fw11.2* and *fw11.3* (**Fig. 2f,g**). To summarize, we propose a two-step evolution of fruit mass that involved two different sets of loci (**Fig. 2h**), which jointly gave rise to modern tomatoes about 100 times larger in fruit size than their wild ancestor.

Divergence in big-fruited tomatoes

After domestication in South America, tomatoes were dispersed to other parts of the world and selected by local farmers and breeders. In general, big-fruited tomatoes were bred for fresh consumption or for processing into tomato paste. Modern processing tomatoes have several characteristic traits, including determinate growth for homogeneous fruit setting and harvest[27], jointless pedicel[28], increased firmness for mechanical harvest, and higher soluble solid content (SSC) and lycopene content for processing quality. However, the genome-wide genetic basis underlying the divergence between tomatoes for fresh consumption and processing tomatoes was not previously studied. To search for SNPs underlying this divergence, we computed the population differentiation statistic (F_{ST}) of each SNP site for 22 modern processing accessions and the remaining 144 BIG accessions. We observed a non-random distribution of highly divergent sites (the top 1% had $F_{ST} \geq 0.4464$; the genome average was 0.07). Intriguingly, 90.53% (63,009 of 69,603) of these sites resided on chromosome 5 (**Fig. 3a**), spanning the majority of the chromosome (from 3.5 to 62.8 Mb). We note that a previous study identified three SSC QTLs (*ssc5.1*, *ssc5.2* and *ssc5.3*) located on the short arm, in the centromeric region and on the long arm of chromosome 5, respectively[29]. A major fruit firmness QTL, *fir5.1*, also resides in the centromeric region of chromosome 5 (ref. 30). In addition, the chromosome has a large centromere with a length of ~50 Mb, extending from 10 to 60 Mb on the assembled chromosome[14]. Therefore, selection of the QTLs for higher SSC and better fruit firmness likely resulted in the hitchhiking of almost the entire chromosome 5, representing a genomic signature of modern processing tomatoes.

Red-fruited tomatoes are widely consumed, but pink-fruited tomatoes are especially popular in China and Japan[31]. Two independent studies[31,32] demonstrated that the pink gene *y* on chromosome 1 corresponds to *SlMYB12*, which controls the accumulation of yellow-colored flavonoid (naringenin chalcone) in the tomato fruit epidermis (peels from pink fruit are colorless owing to the absence

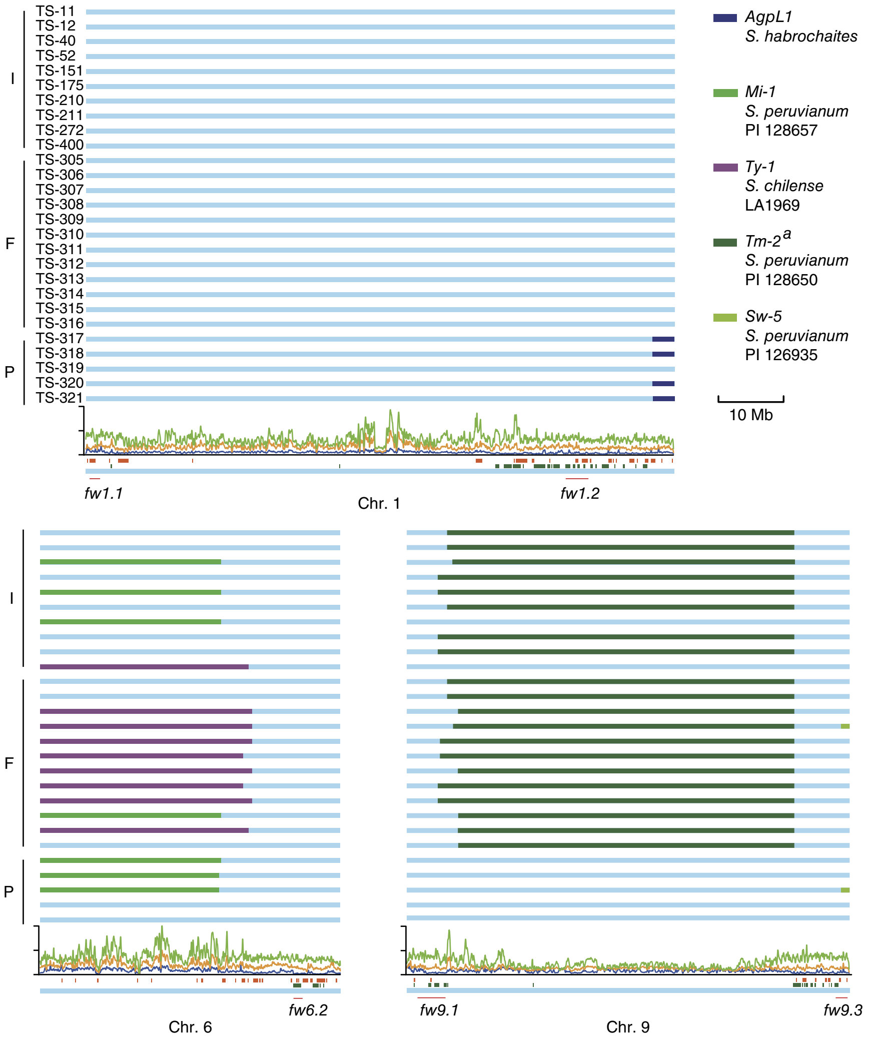

Figure 4 Introgressions and sweeps. The introgression fragments from different wild relatives of the 27 accessions (10 inbreeding lines (I), 12 fresh market hybrids (F) and 5 processing hybrids (P)) are displayed with colored bars. Light-blue bars depict the chromosomes. The nucleotide diversities of the PIM, CER and BIG groups are depicted as green, orange and blue lines, respectively. The orange and green bars above the chromosomes denote the identified domestication and improvement sweeps, respectively. The locations of fruit mass QTLs are marked by red lines. *AgpL1*, *Mi-1*, *Ty-1*, *Tm-2^a* and *Sw-5* represent genes encoding the ADP-Glc pyrophosphorylase large subunit, the root knot nematode resistance gene, the tomato yellow leaf curl virus resistance gene, the tomato mosaic virus resistance gene and the tomato spotted wilt virus resistance gene, respectively.

of flavonoid); however, the causative variant remains unknown. To identify the allelic variation underlying this phenotype, we performed a genome-wide association study (GWAS) using 231 tomato accessions with known phenotypes. The strongest association signal (SNP$_y$, $P < 1 \times 10^{-32}$) resided 8,616 bp upstream of the start codon of *SlMYB12* (**Fig. 3b**). We further analyzed the upstream and downstream sequences of the target gene for structural variants and discovered a 603-bp deletion in the upstream region of *SlMYB12* (−4,865 bp relative to the start codon) that was present in most pink-fruited accessions (**Fig. 3c**). Among 205 red-fruited accessions and 122 pink-fruited ones (96 additional pink-fruited accessions were added for this analysis), the 603-bp deletion coincided with the phenotype in all but 4 accessions. We sequenced the genic regions of *SlMYB12* in these four accessions and identified two nonsense mutations (a nucleotide substitution (C>T) and a 1-bp insertion (TG>TAG)), both resulting in the introduction of premature stop codons. It is noteworthy that the 603-bp deletion was more diagnostic than SNP$_y$, as two recombinants were found (genotype II and IV in **Fig. 3c**). We hypothesize that the deletion might impair the transcription of *SlMYB12*, whose expression is silenced in pink fruits[32]. The silencing of *SlMYB12* likely relaxed purifying selection in the coding sequence, which could accumulate more deleterious mutations as observed. The three recessive alleles of the *y* gene represent useful markers for pink tomato breeding.

Wild introgressions

Domestication and improvement have increased tomato productivity as well as narrowed its genetic basis. In recent decades, wild germplasm has increasingly been used as a source of new alleles for tomato breeding. These efforts rely largely on the pioneering research and comprehensive germplasm collections of Charles Rick at the University of California, Davis[33]. As an example, resistance genes (*R* genes) introgressed from wild species are necessary for the success of modern commercial cultivars. To determine how introgression changed the tomato genome, we scanned all accessions in the CER and BIG lines and the 17 F$_1$ hybrids for genome regions similar to those in wild accessions (**Fig. 4**). We detected a large exotic fragment on chromosome 9 (51.7–54.7 Mb in length) carrying the tomato mosaic virus resistance gene *Tm-2^a* derived from *Solanum peruvianum*[34] (PI 128650). In addition, there were two major introgressions on chromosome 6: one (26.6–27.7 Mb in length) carrying the root knot nematode resistance gene *Mi-1* from *S. peruvianum*[35] (PI 128657) and the other (30.9–32.5 Mb in lengths) carrying the tomato yellow leaf curl virus resistance gene *Ty-1* from *Solanum chilense*[36] (LA1969). Both introgressions occupied nearly the same genomic region, making it difficult to recombine both genes into a single cultivar. Even after multiple generations of backcrossing, these introgressed fragments remain intact, possibly owing to chromosomal inversions or a centromeric location that would inhibit recombination, as shown in the case of *Ty-1* and *Mi-1* (refs. 36,37). An introgression from *Solanum habrochaites* carrying the *AgpL1* gene on chromosome 1, which enhances SSC in mature fruits, was observed in four modern processing hybrids (**Fig. 4**). Understanding the precise position and size of these large wild introgressions will enable the deployment of molecular markers to minimize the limitation from linkage drag and maximize the potential of wild germplasm.

Intriguingly, introgressions carrying resistance genes showed relatively low overlap with the genomic locations of domestication and improvement sweeps (**Fig. 4**). Only one 4.1-Mb region out of the 92.2 Mb of introgressions (*Mi-1*, *Ty-1*, *Tm-2^a*, *Sw-5* and *AgpL1*) overlapped with the domestication and improvement sweeps (4.5% in comparison to the genome level of 14.2%; $P = 0.01$), indicating that introgressions are less likely to have occurred within swept regions. For the three cloned fruit mass genes (*fw2.2*, *fw3.2* and *lcn2.1*), gene action was either recessive or additive[22,24,25], requiring both 'big' alleles for full phenotypic penetration. This could also be true for other domestication and improvement genes[38]. Therefore, it may prove difficult to introgress new alleles into swept regions, implying that domestication and improvement bear a cost in terms of potential future improvement.

DISCUSSION

The genomic foundation for modern tomato breeding was shaped by human-involved selection, as illustrated in this study. Despite their historical contribution to desirable phenotypic traits, these human-induced processes also resulted in the near fixation of a large proportion of the tomato genome. As shown, the domestication and improvement sweeps and linkage drags associated with introgression jointly occupy nearly 200 Mb (25.6% of the assembled genome), limiting further improvement via conventional breeding. The genome sequence[14] and the variation map generated here will facilitate the separation of genes for favorable traits from their embedded sweeps and linkage drags by variome-guided selection for rare recombination or possibly by genome editing. These efforts should enable a redesign of the genomic foundation for future tomato breeding.

URLs. The SNPs from this study can be viewed in a genome browser at http://solgenomics.net/jbrowse/JBrowse-1.11.4/?data=data/json/tomato_variants. Food and Agricultural Organization of the United Nations (FAO) statistics, http://faostat.fao.org/; Tomato Heinz 1706 genome, ftp://ftp.sgn.cornell.edu/genomes/Solanum_lycopersicum/; SOAP software, http://soap.genomics.org.cn; PHYLIP software, http://evolution.genetics.washington.edu/phylip.html; STRUCTURE software, http://pritchardlab.stanford.edu/structure.html.

METHODS

Methods and any associated references are available in the online version of the paper.

Accessions codes. The sequence data have been deposited in the NCBI Sequence Read Archive (SRA) under accession SRP045767.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank J. Maloof (University of California, Davis) for providing tomato RNA sequencing data and L.A. Mueller and N. Menda (Cornell University) for setting up a genome browser of SNPs. This work was supported by funding from the National Program on Key Basic Research Projects in China (973 program; 2012CB113900 and 2011CB100600), the National Science Fund for Distinguished Young Scholars (31225025 to S.H.), the National HighTech Research Development Program in China (863 Program; 2012AA100101 and 2012AA100105), the National Natural Science Foundation of China (31272160, 31230064, 31272171 and 31171962), the Chinese Ministry of Finance (1251610601001), CAAS (an Agricultural Science and Technology Innovation Program grant to S.H.), the China Agriculture Research System (CARS-25-A-09 and CARS-25-A-15), the Special Fund for Agro-Scientific Research in the Public Interest of China (201303115), the Major Special Science and Technology Project during the Twelfth Five-Year Plan Period of Xinjiang (201230116-3) and the US National Science Foundation Plant Genome Program (IOS-0923312). This work was also supported by the Shenzhen municipal and Dapeng district governments.

AUTHOR CONTRIBUTIONS

S.H., Y.D., Z.Y. and Jingfu Li conceived and designed the research. T.L., G.Z., J.Z., X.X., Q.Y., Z. Zheng, Y.L., S.L., T.W. and Yuyang Zhang performed DNA sequencing and biological experiments. T.L., G.Z., Z. Zhang, K.L., Yancong Zhang, C.L., Y.X., X.W., Z.H., D.Z., Junming Li, G.X., C.Z., A.M., M.C., Z.F., J.J.G., R.T.C., A.W. and T.S. performed the data analysis. S.H., G.Z., T.L., J.Z., X.X., Q.Y. and Z. Zhang wrote the manuscript. Y.D., Z.Y., Jingfu Li, Z. Zhang, C.L., Y.X., A.M., M.C., Z.F., J.J.G., R.T.C., D.Z. and T.S. revised the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Borlaug, N.E. Contributions of conventional plant breeding to food production. *Science* **219**, 689–693 (1983).
2. Diamond, J.M. *Guns, Germs, and Steel* (W.W. Norton & Company, New York, 1997).
3. Doebley, J.F., Gaut, B.S. & Smith, B.D. The molecular genetics of crop domestication. *Cell* **127**, 1309–1321 (2006).
4. Gross, B.L. & Olsen, K.M. Genetic perspectives on crop domestication. *Trends Plant Sci.* **15**, 529–537 (2010).
5. Huang, X. *et al.* A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497–501 (2012).
6. Hufford, M.B. *et al.* Comparative population genomics of maize domestication and improvement. *Nat. Genet.* **44**, 808–811 (2012).
7. Lam, H.M. *et al.* Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* **42**, 1053–1059 (2010).
8. Qi, J. *et al.* A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nat. Genet.* **45**, 1510–1515 (2013).
9. Vincent, H. *et al.* A prioritized crop wild relative inventory to help underpin global food security. *Biol. Conserv.* **167**, 265–275 (2013).
10. Meissner, R. *et al.* A new model system for tomato genetics. *Plant J.* **12**, 1465–1472 (1997).
11. Ranc, N., Munos, S., Santoni, S. & Causse, M. A clarified position for *Solanum lycopersicum* var. *cerasiforme* in the evolutionary history of tomatoes (solanaceae). *BMC Plant Biol.* **8**, 130 (2008).
12. Jenkins, J. The origin of the cultivated tomato. *Econ. Bot.* **2**, 379–392 (1948).
13. Rick, C.M. Hybridization between *Lycopersicon esculentum* and *Solanum pennellii*: phylogenetic and cytogenetic significance. *Proc. Natl. Acad. Sci. USA* **46**, 78–82 (1960).
14. Tomatod Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635–641 (2012).
15. Sim, S.C. *et al.* High-density SNP genotyping of tomato (*Solanum lycopersicum* L.) reveals patterns of genetic variation due to breeding. *PLoS ONE* **7**, e45520 (2012).
16. Spooner, D.M., Peralta, I.E. & Knapp, S. Comparison of AFLPs with other markers for phylogenetic inference in wild tomatoes (*Solanum* L. section *Lycopersicon* (Mill.) Wettst.). *Taxon* **54**, 43–61 (2005).
17. Rick, C. & Holle, M. Andean Lycopersicon esculentum var. cerasiforme: genetic variation and its evolutionary significance. *Econ. Bot.* **44**, 69–78 (1990).
18. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
19. Bolger, A. *et al.* The genome of the stress-tolerant wild tomato species *Solanum pennellii*. *Nat. Genet.* **46**, 1034–1038 (2014).
20. Blanca, J. *et al.* Variation revealed by SNP genotyping and morphology provides insight into the origin of the tomato. *PLoS ONE* **7**, e48198 (2012).
21. Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H. & Bustamante, C.D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, e1000695 (2009).
22. Chakrabarti, M. *et al.* A cytochrome P450 regulates a domestication trait in cultivated tomato. *Proc. Natl. Acad. Sci. USA* **110**, 17125–17130 (2013).
23. Grandillo, S., Ku, H. & Tanksley, S. Identifying the loci responsible for natural variation in fruit size and shape in tomato. *Theor. Appl. Genet.* **99**, 978–987 (1999).
24. Frary, A. *et al.* *fw2.2*: a quantitative trait locus key to the evolution of tomato fruit size. *Science* **289**, 85–88 (2000).
25. Muños, S. *et al.* Increase in tomato locule number is controlled by two single-nucleotide polymorphisms located near *WUSCHEL*. *Plant Physiol.* **156**, 2244–2254 (2011).
26. Liu, K. *et al.* A *GH3*-like gene, *CcGH3*, isolated from *Capsicum chinense* L. fruit is regulated by auxin and ethylene. *Plant Mol. Biol.* **58**, 447–464 (2005).
27. Pnueli, L. *et al.* The *SELF-PRUNING* gene of tomato regulates vegetative to reproductive switching of sympodial meristems and is the ortholog of *CEN* and *TFL1*. *Development* **125**, 1979–1989 (1998).
28. Mao, L. *et al.* *JOINTLESS* is a MADS-box gene controlling tomato flower abscission zone development. *Nature* **406**, 910–913 (2000).
29. Tanksley, S.D. *et al.* Advanced backcross QTL analysis in a cross between an elite processing line of tomato and its wild relative *L. pimpinellifolium*. *Theor. Appl. Genet.* **92**, 213–224 (1996).

30. Xu, J. *et al.* Phenotypic diversity and association mapping for fruit quality traits in cultivated tomato and related species. *Theor. Appl. Genet.* **126**, 567–581 (2013).
31. Ballester, A.R. *et al.* Biochemical and molecular analysis of pink tomatoes: deregulated expression of the gene encoding transcription factor SlMYB$_{12}$ leads to pink tomato fruit color. *Plant Physiol.* **152**, 71–84 (2010).
32. Adato, A. *et al.* Fruit-surface flavonoid accumulation in tomato is controlled by a *SlMYB12*-regulated transcriptional network. *PLoS Genet.* **5**, e1000777 (2009).
33. Rick, C.M. The tomato. *Sci. Am.* **239**, 76–87 (1978).
34. Tanksley, S.D. *et al.* Yield and quality evaluations on a pair of processing tomato lines nearly isogenic for the *Tm2^{a}* gene for resistance to the tobacco mosaic virus. *Euphytica* **99**, 77–83 (1998).
35. Kaloshian, I. *et al.* Genetic and physical localization of the root-knot nematode resistance locus *Mi* in tomato. *Mol. Gen. Genet.* **257**, 376–385 (1998).
36. Verlaan, M.G. *et al.* 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 (2013).
37. Seah, S., Yaghoobi, J., Rossi, M., Gleason, C.A. & Williamson, V.M. The nematode-resistance gene, *Mi-1*, is associated with an inverted chromosomal segment in susceptible compared to resistant tomato. *Theor. Appl. Genet.* **108**, 1635–1642 (2004).
38. Schauer, N. *et al.* Mode of inheritance of primary metabolic traits in tomato. *Plant Cell* **20**, 509–523 (2008).

ONLINE METHODS

Plant materials and sequencing. A total of 360 tomato accessions were collected from TGRC (Tomato Genetics Resource Center), USDA (US Department of Agriculture), EU-SOL (European Union Solanaceae project), INRA (National Institute for Agricultural Research) and IVF-CAAS (Institute of Vegetables and Flowers, Chinese Academy of Agricultural Science). These accessions included 10 wild tomato accessions (1 *S. habrochaites*, 3 *S. cheesmaniae*, 1 *S. galapagense*, 3 *S. peruvianum*, 1 *Solanum neorickii* and 1 *S. chilense*), 53 PIM accessions, 112 CER accessions, 166 BIG accessions (2 accessions were excluded for extreme phenotype segregation) and 17 modern commercial hybrids (F_1) (**Supplementary Table 1**). Tomato plants were grown in the greenhouses of IVF-CAAS, Beijing, and the Huazhong Agricultural University, Wuhan, China. Genomic DNA was extracted from young leaves using the cetyltriethylammnonium bromide (CTAB) method[39]. At least 5 μg of genomic DNA was used for each accession to construct paired-end sequencing libraries with insert sizes of approximately 500 bp according to the manufacturer's instructions (Illumina). We generated more than 5 Gb of sequence data for each accession with 100-bp paired-end reads using the Illumina HiSeq 2000 platform.

Mapping and variation calling. To call SNPs, we used SOAP2 (ref. 40) to map all the sequence reads from each accession to the tomato reference genome[14] (release SL2.40) with the following parameters: -m 100, -x 888, -s 35, -l 32, -v 3. Mapped reads were filtered to remove PCR duplicates, assigned to the chromosomes and sorted according to mapping coordinates. Both paired-end and single-end mapped reads were then used for SNP detection throughout the entire collection of tomato accessions via the following procedures. Scripts and analysis pipelines for the SNP data set are available in **Supplementary Data Set**.

To identify SNPs for each genotype, we identified possible SNPs for each accession relative to the reference using SOAPsnp[41] with the following parameters: -L 100 -u -F 1. The likelihood of each individual's genotype in glf format was then generated for each chromosome with SNP quality ≥ 40 and base quality ≥40.

To integrate SNPs across the entire collection, we called each SNP using GLFmulti on the basis of the maximum-likelihood estimation of site frequency. The core set of SNPs was then obtained by filtering according to site frequency and the quality score given by GLFmulti. These SNPs were further filtered using the following criteria: (i) one position with more than two alleles was considered to be a polymorphic site in the population; (ii) the total sequencing depth had to be >150× and <3,500× and the SNP quality value had to be greater than 40; (iii) positions with an average mapping rate of reads of less than 1.5 were retained to rule out the effect of duplications; and (iv) the nearest SNPs had to be more than 1 bp away.

To obtain the final set of SNPs, we performed filtering on the basis of segregation tests and the proportion of homozygosity. Segregation tests can distinguish any segregation pattern from random sequencing errors on the basis of the sequencing depth of the two putative alleles in different individuals. We thus performed segregation tests on the contingency table of read depth for SNP alleles from the 360 tomato accessions. Permutations were used to determine the significance of allele depth in the population, and only sites with $P < 0.01$ were retained. In addition, we filtered out sites at which fewer than 85% of the lines appeared to be homozygous and sites with a proportion of heterozygous genotypes greater than three times that of the homozygous genotypes with the minor allele.

Finally, to detect small indels (≤5 bp in length), we mapped all the sequence reads from each accession with a gap of less than 5 bp allowed (parameter -g 5) using SOAP2 (ref. 40). Indels (1–5 bp) were called by the SOAPindel pipeline (see URLs).

Annotation of SNPs. The identified SNPs were further categorized as variations in intergenic regions, UTRs, coding sequences and introns according to the tomato genome annotation (release ITAG2.3). SNPs in coding sequences were further grouped into synonymous SNPs (not causing amino acid changes) and nonsynonymous SNPs (causing amino acid changes) (**Supplementary Table 2**).

Evaluation of SNPs. We used Sanger sequencing and previously released SNP array data to evaluate the accuracy of SNPs. First, we randomly selected 349 SNPs from 3 tomato accessions (PIM, TS-244; CER, TS-252; BIG, TS-137) for validation by PCR and Sanger sequencing. Second, we compared 285,508 SNPs for 48 tomato accessions identified in this study to previously published tomato SNP array data[15] (**Supplementary Tables 4** and **5**, and **Supplementary Note**).

Phylogenetic analysis and population structure of tomato. To build a neighbor-joining tree, we screened a subset of 20,111 SNPs at fourfold-degenerate sites (MAF > 5% and missing data < 10%) in the 341 tomato accessions (excluding the F_1 individuals) from the entire SNP data set. These SNPs should be under less selective pressure, thus more reliably reflecting population structure and demography. We constructed a phylogenetic tree using PHYLIP[42] (version 3.695) with 100 bootstrap replicates. Using the same data set, we also investigated the population structure using STRUCTURE[43] (version 2.3.1) on the basis of allele frequencies. To determine the most likely group number, STRUCTURE was run 20 times on 1,000 randomly selected SNPs at fourfold-degenerate sites for each K value from 2 to 19 (**Supplementary Fig. 2**). After determining ΔK, we used 20,111 SNPs at fourfold-degenerate sites to determine the group membership of each accession by 10,000 iterations with K values from 2 to 4. In addition, we performed principal-component analysis (PCA)[44] using 2,340,973 SNPs across the genome (MAF > 10%, missing < 5%). Two-dimensional coordinates were plotted for the 331 tomato accessions (excluding 10 wild accessions) (**Supplementary Fig. 3**).

Demographic analysis of tomato evolution history. The best parameters for fitting were estimated using δ*a*δ*i*[21]. We fitted the three-population model with PIM and BIG mixed together (**Supplementary Table 6**) for all three groups. The simulation was carried out 20 times, and each time we randomly selected 500,000 SNPs and estimated 95% confidence intervals on the basis of the best fitting parameters. The parameters inferred by δ*a*δ*i* were scaled by $2N_e$, with N_e being the ancestral population size. We estimated the ancestral population size using the formula $4N_e \times \mu \times L = \theta$, where μ is the mutation rate, L is the generation time and θ is the genetic diversity. We used θ_{PIM} to estimate θ (set to ~3.23×10^{-3}). Here the neutral mutation rate of 1×10^{-8} (ref. 45) was used for μ (the mutation rate per generation). Thus, $2N_e$ was estimated to be 1.615×10^5. All the parameters were then scaled by $2N_e$ to estimate time in years and the population size in number of individuals.

Detection of domestication and improvement sweeps. Nucleotide diversity (π) is often applied to measure the degree of variability in a group[18]. 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. The regions affected by domestication should have substantially lower diversity in CER than in PIM. Improvement sweeps should show a much stronger reduction of diversity in BIG in comparison to CER. If π_{PIM} for a window was lower than 0.002 in domestication analysis and π_{CER} was lower than 0.001 in improvement analysis, then the window was excluded. By scanning the ratios of genetic diversity between PIM and CER (π_{PIM}/π_{CER}) as well as between CER and BIG (π_{CER}/π_{BIG}), we selected windows with the top 5% of ratios (3.0 and 6.9 for domestication and improvement, respectively) as candidate regions for further analysis. Finally, windows that were ≤100 kb apart were merged into a single selected region (**Supplementary Tables 7–10**). To verify the empirical thresholds with low false discovery rate, we performed whole-genome permutation tests to ascertain the thresholds for identifying domestication and improvement sweeps (**Supplementary Note**). The regions shared by domestication and improvement sweeps were defined as overlapping regions, and these regions should have undergone further selection during improvement. In our study, we analyzed almost all the fruit mass QTLs and genes that segregated between PIM parents and cultivated parents. If closely linked markers or the mapped intervals were located in domestication (improvement) sweeps, we considered them to be candidate domestication (improvement) QTLs or genes (**Supplementary Table 11**).

Linkage disequilibrium analysis. LD values for PIM, CER and BIG were calculated on the basis of SNPs (MAF > 0.05) using Haploview software[46]. The parameters were as follows: -n -pedfile -info -log -minMAF 0.05 -hwcutoff 0 -dprime -memory 2096. LD decay was calculated on the basis of the r^2 value and corresponding distance between two SNPs (**Supplementary Fig. 5**).

Bulked segregant analysis of the F_2 population by whole-genome resequencing. We planted an F_2 population of 500 individuals derived from the cross between TS-400 (a big-fruit accession) and TS-my (a small-fruit accession) in the fall of 2013 in IVF-CAAS, China. For each individual, the average weight of approximately ten representative fruits was recorded (**Fig. 2e**, **Supplementary Fig. 6** and **Supplementary Note**) and genomic DNA was isolated from fresh leaves using the CTAB method. For bulked segregant analysis, bulk DNA samples for big- and small-fruit accessions were constructed by mixing equal amounts of DNA from 50 individuals showing extremely big and small fruits, respectively. Roughly 20× genome sequences for each parent (TS-400 and TS-my) and 50× data for each bulk sample (big fruit and small fruit) were generated. Short reads were aligned against the reference genome (release SL2.40) using the Burrows-Wheeler Aligner (BWA)[47], and SNPs were identified using SAMtools[48]. SNPs between two parental genomes were identified for further analysis when the base quality value was ≥20 and the SNP quality value was ≥20. On the basis of these criteria and the number of SNPs with read depth from 4 to 200, a 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 (TS-400). A ΔSNP index was obtained by subtracting the SNP index for the small-fruit bulk sample from that for the big-fruit bulk sample. An average SNP index for the big-fruit and small-fruit bulk samples was calculated using a 1,000-kb sliding window with a step size of 10 kb (**Fig. 2f,g**). We also calculated the statistical confidence intervals of the ΔSNP index under the null hypothesis of no QTLs. For each position, the 95% confidence intervals of the ΔSNP index were obtained following the method described in Takagi *et al.*[49].

Genome-wide association studies for fruit color. We used 10,990,318 high-quality SNPs (MAF > 0.05) to perform GWAS for fruit color in 231 accessions (205 red- and 26 pink-fruit accessions). The association analyses were performed using the compressed MLM[50,51] (**Fig. 3b**) with TASSEL 4.0 (ref. 52). To further detect the causative variant in the significantly associated region (chromosome 1: 71,229,871–71,258,882), we analyzed the discordant paired-end reads between pink and red tomatoes by aligning the resequenced reads of 20 pink- and 20 red-fruit accessions against the reference genome (release SL2.40) using BWA and SAMtools. We further checked the variants in 205 red- and 122 pink-fruit accessions from our tomato germplasm and IVF-CAAS accessions by direct PCR and Sanger sequencing.

Detection of introgression status from wild germplasm in cultivars. Some wild accessions harbor important loci, including disease resistance genes (*R* genes), and have been widely applied in modern tomato breeding programs. To detect introgressions in modern cultivars, we analyzed the introgression pattern by calculating the ratio of identical SNPs between cultivated accessions (including 112 CER, 166 BIG and 17 F_1 accessions) and wild donor accessions (**Fig. 4**). The ratio from each chromosome was plotted in a 100-kb sliding window with a step size of 10 kb.

39. Gawel, N. & Jarret, R. A modified CTAB DNA extraction procedure for *Musa* and *Ipomoea*. *Plant Mol. Biol. Rep.* **9**, 262–266 (1991).
40. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).
41. Li, R. *et al.* SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **19**, 1124–1132 (2009).
42. Felsenstein, J. PHYLIP—phylogeny inference package (version 3.2). *Cladistics* **5**, 164–166 (1989).
43. Falush, D., Stephens, M. & Pritchard, J.K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
44. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
45. Gaut, B.S. Molecular clocks and nucleotide substitution rates in higher plants. *Evol. Biol.* **30**, 93–120 (1998).
46. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
47. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
48. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
49. Takagi, H. *et al.* 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 (2013).
50. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208 (2006).
51. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).
52. Bradbury, P.J. *et al.* TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).