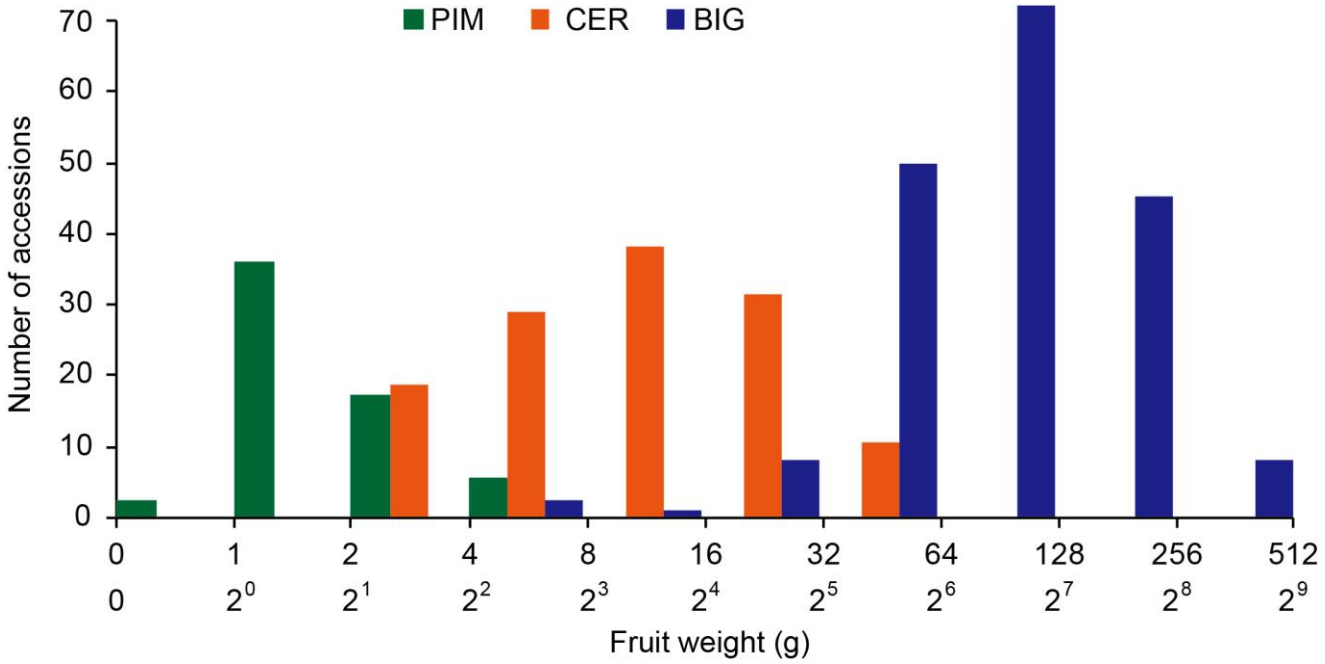
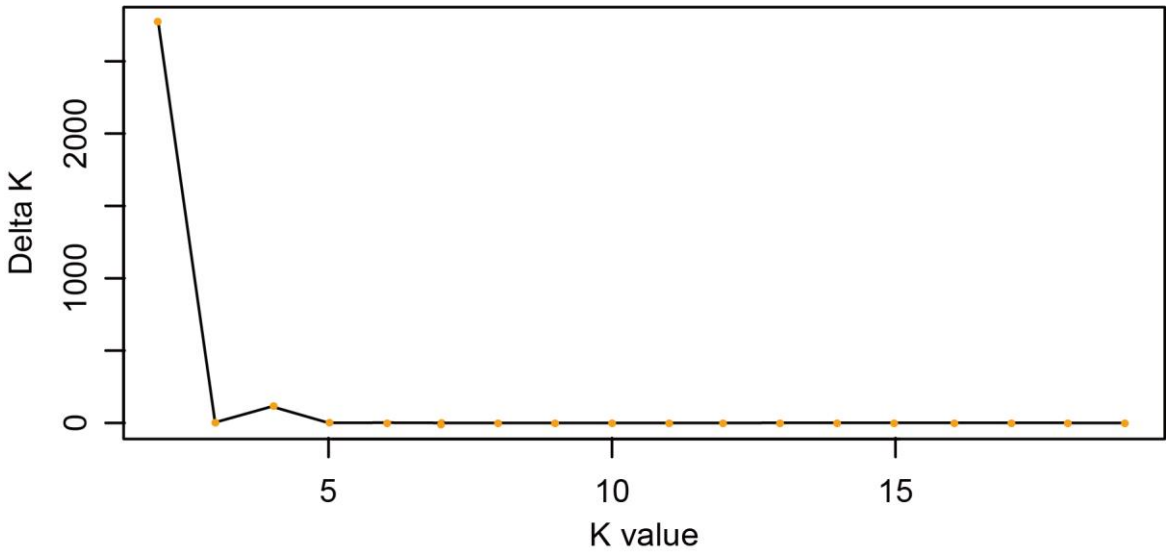



**Supplementary Figure 1**

**Spectra of fruit weight for three tomato groups.**



Nature Genetics: doi:10.1038/ng.3117



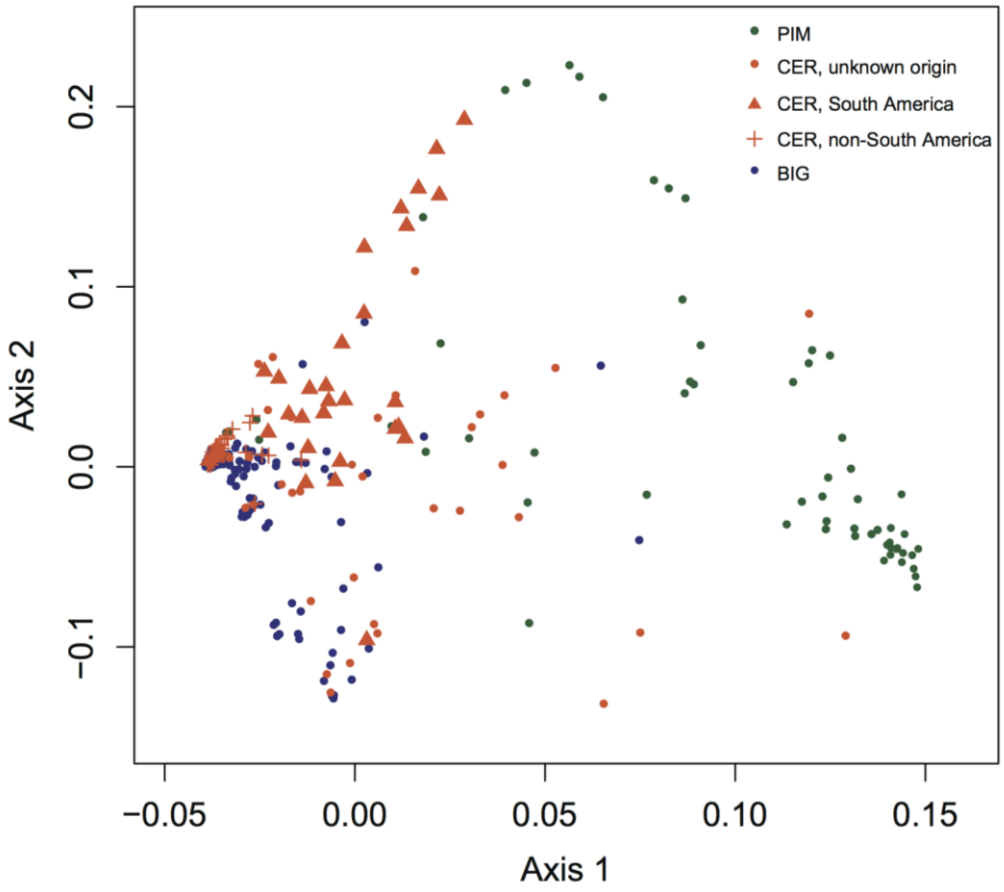

**Supplementary Figure 2**

**Determination of Δ*K* using STRUCTURE.**

Δ*K* analysis for a different number of clusters (*K*) for a tomato population consisting of 331 accessions (excluding 10 wild accessions). Δ*K* showed a peak at 2, suggesting two clusters as the optimal option.



Nature Genetics: doi:10.1038/ng.3117



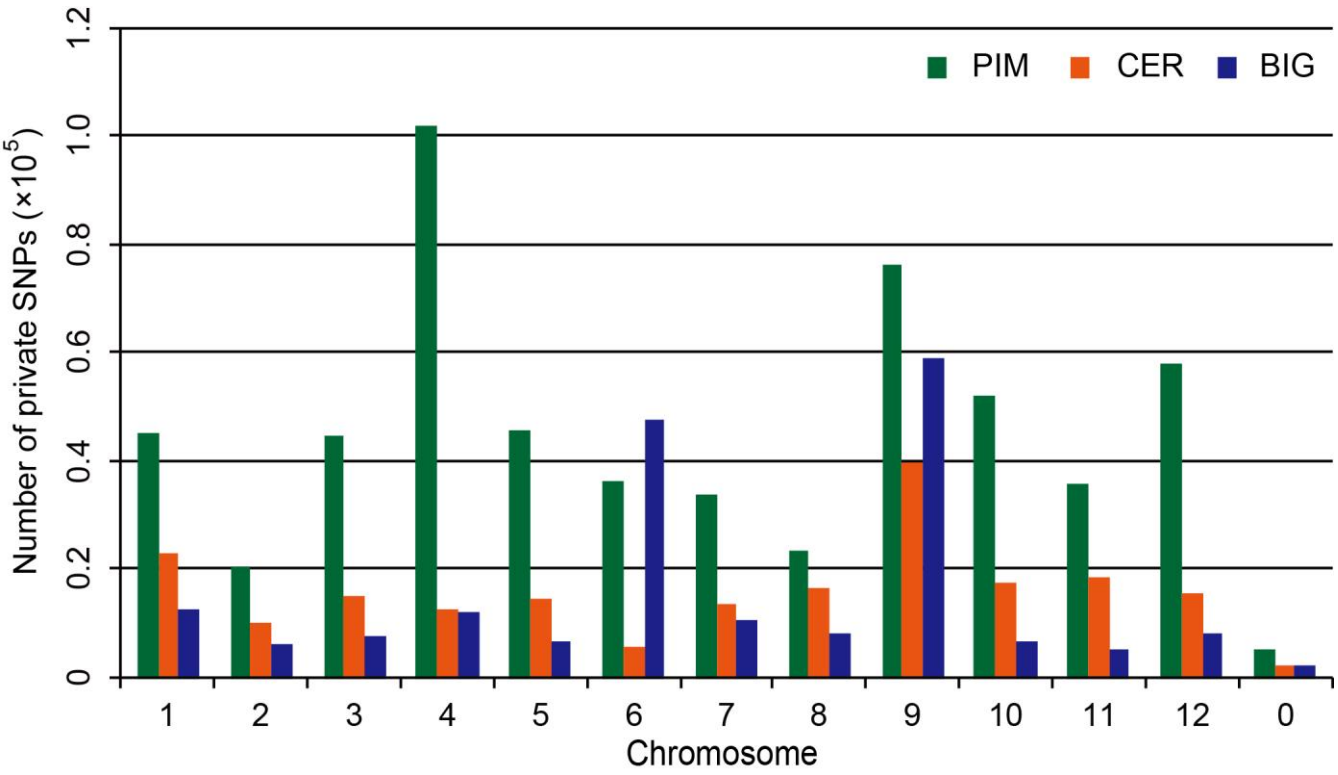

**Supplementary Figure 3**

**Principal-component analysis (PCA) of 331 tomato accessions.**

A total of 2,340,973 whole-genome SNPs (MAF > 10%, missing ≤ 5%) were used for PCA. Two-dimension coordinates were plotted for the 331 tomato accessions. CER (orange) and BIG (blue) accessions have a relatively concentrated distribution, whereas PIM accessions (green) are dispersed widely.



Nature Genetics: doi:10.1038/ng.3117



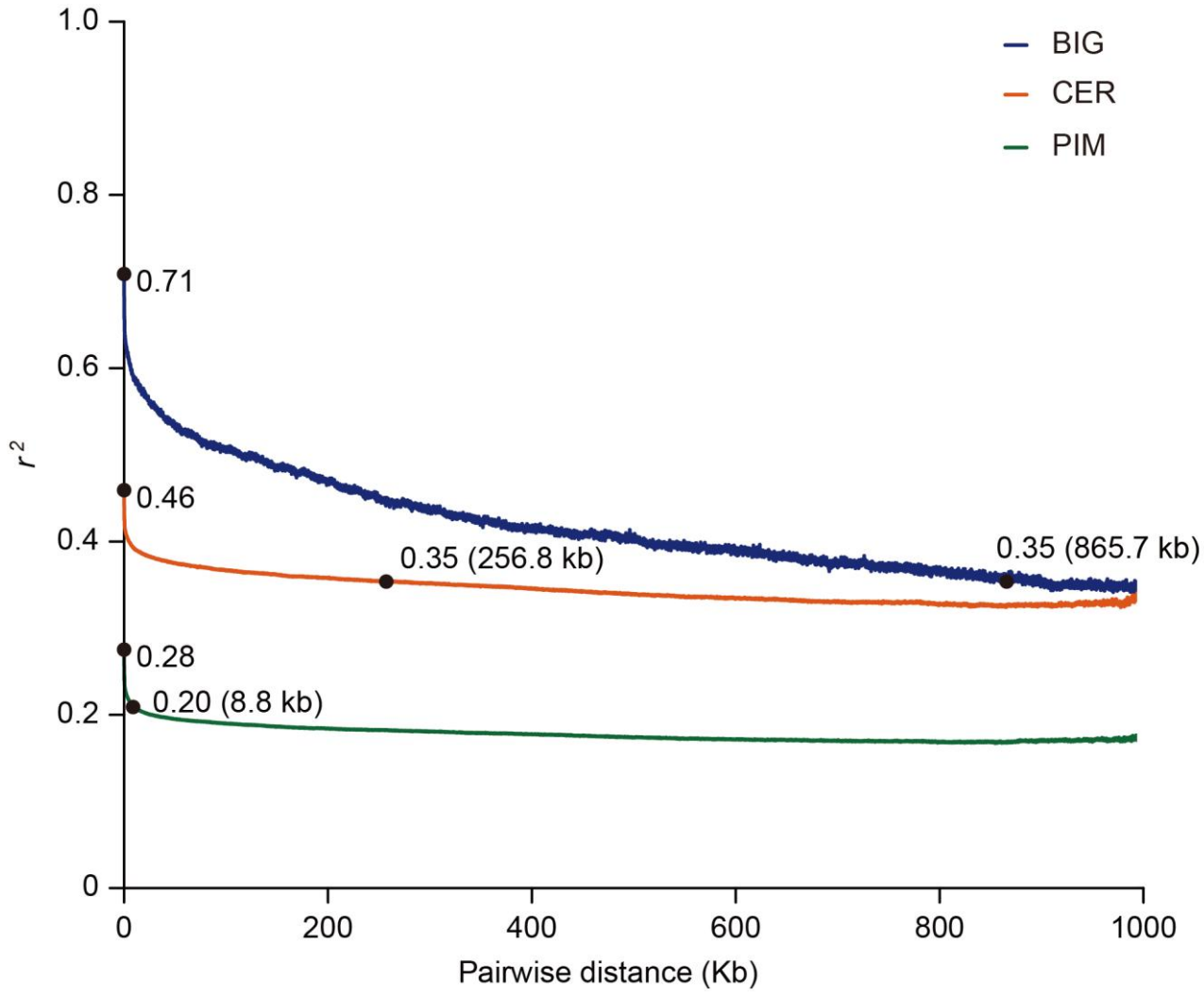

**Supplementary Figure 4**

**Distribution of private SNPs in three tomato groups.**

Private SNPs are presented for each chromosome in PIM (green), CER (orange) and BIG (blue).



Nature Genetics: doi:10.1038/ng.3117



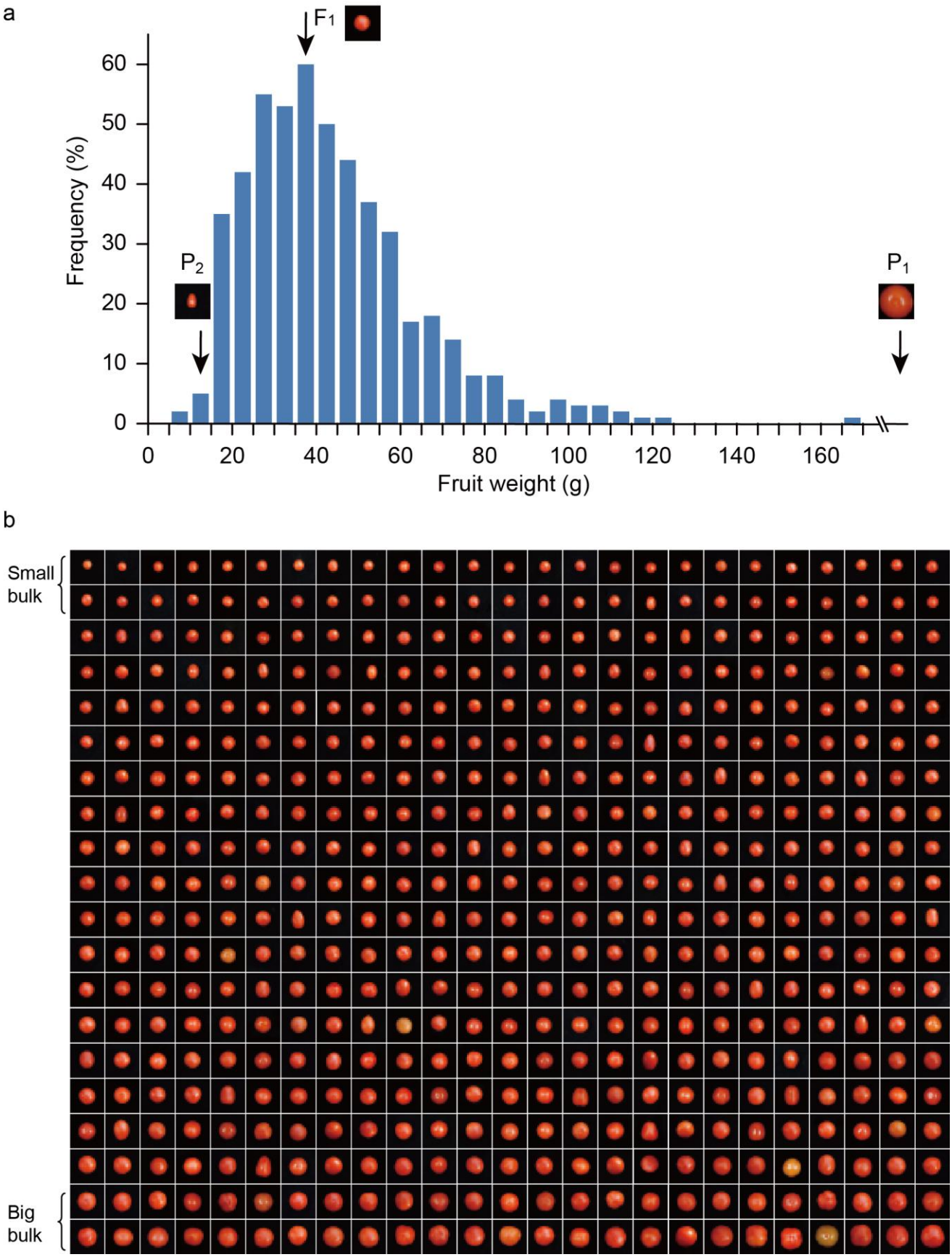

**Supplementary Figure 5**

**Genome-wide average LD decay in three tomato groups.**

LD decay is estimated by the squared correlations of allele frequency ($r^2$) against distance between polymorphic sites in PIM (green), CER (orange) and BIG (blue).



Nature Genetics: doi:10.1038/ng.3117






Nature Genetics: doi:10.1038/ng.3117

**Supplementary Figure 6**

**Distribution of fruit weight and size of the 500 $F_2$ individuals.**

(**a**) Frequency distribution of fruit weight in the $F_2$ population. The fruit weight of both parents and the $F_1$ are shown. (**b**) Fruit appearances of individuals from the $F_2$ population. The two bulks (big bulk and small bulk) are constructed by selecting the fruits shown in the first and last two rows, respectively.



Nature Genetics: doi:10.1038/ng.3117

**Supplementary Material index**



Nature Genetics: doi:10.1038/ng.3117

**Supplementary Tables**

**Supplementary Table 1.** Summary of the sampled collection of tomato (available in separate file.xls).

**Supplementary Table 2.** Distribution of SNPs within various genomic regions in tomato.

| | SNP | Intergenic | 5' UTR | CDS | Intron | 3' UTR | Synonymous[#] | Non-synonymous[#] |
|---|---|---|---|---|---|---|---|---|
| ch01 | 1,220,209 | 1,042,008 | 7,638 | 42,231 | 119,758 | 8,574 | 17,834 | 25,223 |
| ch02 | 661,454 | 538,279 | 2,303 | 33,348 | 82,576 | 4,948 | 13,903 | 19,567 |
| ch03 | 885,683 | 758,364 | 2,199 | 33,523 | 86,679 | 4,918 | 13,843 | 19,752 |
| ch04 | 1,063,878 | 948,625 | 1,853 | 29,732 | 79,513 | 4,155 | 12,112 | 17,681 |
| ch05 | 1,014,026 | 925,488 | 1,386 | 24,272 | 59,663 | 3,217 | 9,594 | 14,702 |
| ch06 | 813,858 | 690,607 | 2,157 | 31,045 | 85,168 | 4,881 | 12,860 | 18,265 |
| ch07 | 1,131,425 | 1,028,751 | 1,686 | 27,386 | 69,966 | 3,636 | 11,368 | 16,074 |
| ch08 | 1,032,516 | 932,747 | 1,616 | 25,462 | 69,281 | 3,410 | 10,257 | 15,244 |
| ch09 | 1,160,963 | 1,050,579 | 1,796 | 28,495 | 76,229 | 3,864 | 11,773 | 16,798 |
| ch10 | 831,011 | 752,219 | 409 | 23,400 | 53,970 | 1,013 | 9,135 | 14,285 |
| ch11 | 785,079 | 695,596 | 9 | 22,660 | 66,802 | 12 | 9,405 | 13,255 |
| ch12 | 824,248 | 738,611 | 15 | 24,215 | 61,385 | 22 | 9,542 | 14,673 |
| ch00 | 196,167 | 189,245 | 62 | 2,695 | 3,973 | 192 | 909 | 1,787 |
| Total | 11,620,517 | 10,291,119 | 23,129 | 348,464 | 914,963 | 42,842 | 142,535 | 207,306 |

[#] Because SNPs are annotated independently, some SNPs are located in overlapped regions of different transcripts, resulting in synonymous SNPs in one transcript, and non-synonymous SNPs in another transcript. The sum of synonymous and non-synonymous SNPs may be higher than the number of SNPs in the corresponding CDS regions.

Nature Genetics: doi:10.1038/ng.3117

**Supplementary Table 3.** Distribution of indels within various genomic regions in tomato.

| | indel | Intergenic | 5' UTR | CDS | Intron | 3' UTR | Non-frame shift | Frame shift |
|---|---|---|---|---|---|---|---|---|
| ch01 | 152,744 | 117,870 | 1,723 | 1,966 | 28,726 | 2,459 | 1,453 | 513 |
| ch02 | 94,943 | 71,204 | 832 | 1,633 | 19,854 | 1,420 | 1,216 | 417 |
| ch03 | 110,385 | 85,325 | 748 | 1,619 | 21,190 | 1,503 | 1,177 | 442 |
| ch04 | 113,612 | 91,707 | 607 | 1,439 | 18,602 | 1,257 | 1,060 | 379 |
| ch05 | 105,244 | 88,694 | 498 | 1,199 | 13,864 | 989 | 873 | 326 |
| ch06 | 92,855 | 70,513 | 651 | 1,419 | 18,924 | 1,348 | 1,048 | 371 |
| ch07 | 112,547 | 93,379 | 559 | 1,304 | 16,224 | 1,081 | 944 | 360 |
| ch08 | 105,363 | 85,960 | 514 | 1,184 | 16,698 | 1,007 | 852 | 332 |
| ch09 | 109,130 | 89,989 | 559 | 1,212 | 16,228 | 1,142 | 894 | 318 |
| ch10 | 98,091 | 83,801 | 163 | 1,180 | 12,622 | 325 | 899 | 281 |
| ch11 | 90,249 | 73,622 | 1 | 1,025 | 15,595 | 6 | 718 | 307 |
| ch12 | 100,584 | 85,112 | 3 | 1,253 | 14,210 | 6 | 865 | 388 |
| ch00 | 17,466 | 16,392 | 19 | 127 | 881 | 47 | 109 | 18 |
| Total | 1,303,213 | 1,053,568 | 6,877 | 16,560 | 213,618 | 12,590 | 12,108 | 4,452 |

**Supplementary Table 4.** SNP loci selected for validation by PCR and Sanger sequencing (available in separate file.xls).

Nature Genetics: doi:10.1038/ng.3117

**Supplementary Table 5.** SNP loci selected for validation by the SNP array dataset*.

| Accession | Group | Consistent | Inconsistent | Missing Data | Accuracy |
|---|---|---|---|---|---|
| TS-3 | BIG | 6,386 | 13 | 112 | 99.80% |
| TS-6 | BIG | 6,164 | 4 | 343 | 99.94% |
| TS-8 | BIG | 6,337 | 4 | 170 | 99.94% |
| TS-9 | BIG | 6,007 | 248 | 256 | 96.04% |
| TS-10 | BIG | 6,383 | 82 | 46 | 98.73% |
| TS-14 | PIM | 6,086 | 146 | 279 | 97.66% |
| TS-15 | PIM | 6,253 | 6 | 252 | 99.90% |
| TS-17 | PIM | 6,143 | 70 | 298 | 98.87% |
| TS-22 | PIM | 6,205 | 28 | 278 | 99.55% |
| TS-31 | CER | 6,436 | 4 | 71 | 99.94% |
| TS-33 | CER | 5,978 | 298 | 235 | 95.25% |
| TS-34 | CER | 6,358 | 8 | 145 | 99.87% |
| TS-43 | BIG | 6,345 | 24 | 142 | 99.62% |
| TS-49 | BIG | 6,248 | 2 | 261 | 99.97% |
| TS-72 | CER | 5,933 | 378 | 200 | 94.01% |
| TS-87 | CER | 5,199 | 397 | 915 | 92.91% |
| TS-92 | PIM | 4,453 | 241 | 1,817 | 94.87% |
| TS-98 | CER | 6,287 | 5 | 219 | 99.92% |
| TS-107 | CER | 3,971 | 255 | 2,285 | 93.97% |
| TS-122 | BIG | 3,586 | 22 | 2,903 | 99.39% |
| TS-123 | PIM | 5,871 | 287 | 353 | 95.34% |
| TS-128 | BIG | 6,353 | 99 | 59 | 98.47% |
| TS-133 | BIG | 5,351 | 13 | 1,147 | 99.76% |
| TS-145 | PIM | 5,690 | 15 | 806 | 99.74% |
| TS-151 | BIG | 6,358 | 2 | 151 | 99.97% |
| TS-154 | CER | 6,433 | 2 | 76 | 99.97% |
| TS-155 | BIG | 5,894 | 3 | 614 | 99.95% |
| TS-156 | PIM | 6,316 | 34 | 161 | 99.46% |
| TS-158 | CER | 5,844 | 585 | 82 | 90.90% |
| TS-163 | BIG | 6,382 | 2 | 127 | 99.97% |
| TS-171 | BIG | 5,246 | 10 | 1,255 | 99.81% |
| TS-181 | CER | 6,170 | 6 | 335 | 99.90% |
| TS-192 | BIG | 6,340 | 3 | 168 | 99.95% |
| TS-221 | CER | 6,300 | 2 | 209 | 99.97% |
| TS-224 | BIG | 4,677 | 118 | 1,716 | 97.54% |
| TS-228 | BIG | 6,000 | 2 | 509 | 99.97% |
| TS-229 | CER | 6,457 | 2 | 52 | 99.97% |
| TS-237 | BIG | 4,427 | 8 | 2,076 | 99.82% |
| TS-239 | BIG | 4,905 | 861 | 745 | 85.07% |
| TS-240 | CER | 5,999 | 232 | 280 | 96.28% |
| TS-244 | PIM | 5,060 | 20 | 1,431 | 99.61% |
| TS-264 | BIG | 6,242 | 8 | 261 | 99.87% |
| TS-266 | PIM | 5,001 | 1,050 | 460 | 82.65% |
| TS-267 | CER | 5,915 | 182 | 414 | 97.01% |
| TS-291 | PIM | 4,773 | 6 | 1,732 | 99.87% |
| TS-298 | CER | 6,284 | 2 | 225 | 99.97% |
| TS-299 | CER | 5,731 | 617 | 163 | 90.28% |
| TS-420 | PIM | 5,744 | 581 | 186 | 90.81% |

* The array dataset is from S. C. Sim *et al*. 2012

**Supplementary Table 6.** Best fitting parameters for the three-population model in three tomato groups by demographic analysis.

| | CER | | | | BIG | | | |
|---|---|---|---|---|---|---|---|---|
| | **Lower** | **Higher** | **Mean** | **Mean[a]** | **Lower** | **Higher** | **Mean** | **Mean[a]** |
| Population size at bottleneck | 0.0017 | 0.0019 | 0.0018 | 295 | 0.050 | 0.102 | 0.076 | 12,217 |
| Final size of population | 0.081 | 0.092 | 0.086 | 13,941 | 0.057 | 0.062 | 0.059 | 9,558 |
| Duration of bottleneck (years) | 0.00066 | 0.00071 | 0.00068 | 109.8 | 0.00019 | 0.00024 | 0.00021 | 34.4 |
| Time after bottleneck to final population (years) | 0.039 | 0.052 | 0.045 | 7,295.8 | 0.0025 | 0.0032 | 0.0028 | 458.6 |

We estimate the ancient population size by the formula $4N_e \times \mu \times L = \theta$, where $\mu$ is the mutation rate, L is the generation time and $\theta$ is the genetic diversity.

[a] Values are calculated by multiplying $2N_e$ ($1.615 \times 10^5$) and the best fitting parameters.

**Supplementary Table 7.** Putative domestication sweeps (available in separate file.xls).

**Supplementary Table 8.** Putative improvement sweeps (available in separate file.xls).

**Supplementary Table 9.** Genes within the putative domestication sweeps (available in separate file.xls).

**Supplementary Table 10.** Genes within the putative improvement sweeps (available in separate file.xls).

**Supplementary Table11.** Summary of known fruit weight loci during domestication and improvement.

| Genes/QTLs | Linked Marker | Chr. | Type | Sweep number | Reference |
|---|---|---|---|---|---|
| *fw1.1* | TG301 | ch01 | domestication | IS003 | 1 |
| *fw2.1* | cTOD_16_E7, TG151 | ch02 | improvement | IS029-IS034 | 2 |
| *fw2.2** | | ch02 | improvement | IS033 | 3 |
| *fw2.3* | CG59, TG154 | ch02 | improvement | IS035 | 1 |
| *lc*(*lcn2.1*)* | | ch02 | improvement | IS29-IS30 | 4 |
| *lcn2.2* | T347, TG167 | ch02 | improvement | IS033 | 5 |
| *fw3.1* | TG246, SSR320 | ch03 | improvement | IS054-IS056 | 2,6 |
| *fw3.2** | | ch03 | improvement | IS054 | 7 |
| *fw5.2* | CT118, cLEX-13-G5 | ch05 | domestication | DS065 | 8 |
| *fw7.2* | TG183 | ch07 | domestication | DS126-DS133 | 8 |
| *fw9.1* | TG254, TG9 | ch09 | improvement | IS098-IS102 | 1 |
| *fw9.3* | CT74, TG328 | ch09 | improvement | IS110-IS113 | 1 |
| *lcn10.1* | CT234 | ch10 | improvement | IS114-IS118 | 8 |
| *fw11.1* | TG400, TG26 | ch11 | improvement | IS128-IS130 | 2 |
| *fw11.2* | TG46 | ch11 | improvement | IS128 | 1 |
| *fw11.3* | EP1057, EP1573 | ch11 | improvement | IS131 | 9 |
| *fw12.1* | TG180, cLEC-67-B16 | ch12 | domestication | DS174-DS177 | 2 |
| *lcn12.1* | TG167, TG151 | ch12 | domestication | DS180 | 8 |

* It shows these genes were cloned.

**Supplementary Note**

**Validation of SNP calling**

To evaluate the accuracy of SNPs, we first randomly selected 349 SNPs in three accessions for PCR and Sanger sequencing (**Supplementary Table 4**). As a result, 347 SNPs could be amplified and sequenced successfully in all three accessions and two SNPs in one accession, resulting in 1,043 genotypes. Some genotypes could not be amplified and sequenced successfully, and presumably other additional variants may be present in the primer sequences. As a result, we found only 1.6% of genotypes were different from those for PCR and Sanger sequencing. Secondly, we compared 285,508 SNPs from 48 tomato accessions to a previously published tomato SNP array dataset. As a result, only 6,987 SNPs (2.4%) are different from those determined using re-sequencing data (**Supplementary Table 5**). Therefore, the accuracy of our SNP calling is estimated to be 98.4% and 97.6% based on the two different validation approaches, respectively.

**Private SNP distribution in different groups**

The private SNPs of different groups were identified, and we found PIM has substantially more private SNPs (582,954) than those in CER (207,892) and BIG (194,919). Meanwhile, we found that ~55% of private SNPs were located on chromosomes 6 and 9 in BIG (**Supplementary Figure 4**).

**Ancestral SNP distribution in three groups**

We identified a total of 23,509,525 SNPs using the MUMmer package[10] (version 3.0, nucmer and show-snps programs with default parameters) between *S. pennellii* and the reference Heinz 1706 genomes. Among these SNPs, 3,480,199 (14.8%) can be reliably recovered in the ~11.6 million SNP sites. Among the ~3.5 million SNPs, those identical to SNPs in *S. pennellii*, representing presumably ancestral alleles, were identified in each accession. As a result, on average PIM accessions have a higher percentage (30.4%) of ancestral alleles than CER (6.6%) and BIG (2.8%).

**Determination of the cutoffs for selective sweep identification**

The genetic diversity ($\pi$) at the whole-genome level was calculated for each of the three groups (PIM, CER and BIG), using 100 kb window with a step size of 10 kb. High positive correlation was found between PIM and CER ($R^2$ = 0.6), as well as CER and BIG ($R^2$ = 0.4). By scanning the ratios of the genetic diversity ($\pi_{PIM}/\pi_{CER}$ and $\pi_{CER}/\pi_{BIG}$), we selected windows with the top 5% ratio cutoffs (3.0 and 6.9) as candidate swept regions. Both cutoffs were confirmed based on the permutation tests.

1. To confirm the appropriate cutoffs, all the accession genotypes in the combined groups of PIM and CER, as well as CER and BIG were reshuffled and then we performed selective sweep detection with the same parameters ten times. No positive signals were found to pass the empirical cutoffs.
2. As a negative control, we used the value of $\pi_{CER}/\pi_{PIM}$ to detect the low diversity regions in PIM. With the same method, the low diversity regions of CER were also scanned. We found that the highest value is 1.6 and 1.5, both far lower than positive signals.

Both the positive and negative tests indicated that the empirical cutoffs selected could generate selection signals with very low false discovery rates.

**Bulked-segregant analysis and whole-genome resequencing of the $F_2$ population**

Approximately 51 Gb and 54 Gb data for big and small bulks were generated through Illumina high-throughput sequencing, respectively (**Fig. 2e and Supplementary Fig. 6**). The reads of Big-bulk (48× depth) and Small-bulk (51× depth) covered 93.3% and 93.8% of the assembled genome, respectively. Approximate 2.3 million SNPs between the parents were identified with the base quality value >= 20 and the SNP quality value >= 20. The SNP-index for both bulks was calculated and ΔSNP-index was used for fruit QTL detection. We found the peak regions (out of 95% confidence values) resided on three chromosomes, including the distal end of the long arm of chromosome 2, both distal ends of chromosome 9 and the distal end of the long arm of chromosome 11. As a result, 10 QTLs (*fw2.1*, *fw2.2*, *fw2.3*, *lcn2.1*, *lcn2.2*, *fw9.1*, *fw9.3*, *fw11.1*, *fw11.2* and *fw11.3*) for improvement were found in the $F_2$ population (**Fig. 2f,g**). These results indicate that improvement sweeps play important roles in controlling fruit size during tomato evolution. However, minor QTLs are less easily detected using this method, as shown in the

simulation test in a previous study[11].

**Supplementary References:**


1. 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).
2. Ashrafi, H., Kinkade, M.P., Merk, H.L. & Foolad, M.R. Identification of novel quantitative trait loci for increased lycopene content and other fruit quality traits in a tomato recombinant inbred line population. *Mol. Breeding* **30**, 549-567 (2012).
3. Frary, A. *et al. fw2.2*: a quantitative trait locus key to the evolution of tomato fruit size. *Science* **289**, 85-88 (2000).
4. Munos, S. *et al.* Increase in tomato locule number is controlled by two single-nucleotide polymorphisms located near *WUSCHEL*. *Plant Physiol.* **156**, 2244-2254 (2011).
5. Barrero, L. & Tanksley, S. Evaluating the genetic basis of multiple-locule fruit in a broad cross section of tomato cultivars. *Theor. Appl. Genet.* **109**, 669-679 (2004).
6. Doganlar, S., Frary, A., Ku, H.-M. & Tanksley, S.D. Mapping quantitative trait loci in inbred backcross lines of Lycopersicon pimpinellifolium (LA1589). *Genome* **45**, 1189-1202 (2002).
7. Chakrabarti, M. *et al.* A cytochrome P450 regulates a domestication trait in cultivated tomato. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 17125-17130 (2013).
8. Van der Knaap, E. & Tanksley, S. The making of a bell pepper-shaped tomato fruit: identification of loci controlling fruit morphology in Yellow Stuffer tomato. *Theor. Appl. Genet.* **107**, 139-147 (2003).
9. Huang, Z. & van der Knaap, E. Tomato fruit weight 11.3 maps close to fasciated on the bottom of chromosome 11. *Theor. Appl. Genet.* **123**, 465-474 (2011).
10. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
11. 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).


Nature Genetics: doi:10.1038/ng.3117