

ARTICLE

DOI: 10.1038/s41467-017-00336-7

OPEN

# Genome re-sequencing reveals the history of apple and supports a two-stage model for fruit enlargement

Naibin Duan<sup>1,2</sup>, Yang Bai<sup>3</sup>, Honghe Sun<sup>3</sup>, Nan Wang<sup>1</sup>, Yumin Ma<sup>2</sup>, Mingjun Li<sup>4</sup>, Xin Wang<sup>3</sup>, Chen Jiao<sup>3</sup>, Noah Legall<sup>3</sup>, Linyong Mao<sup>3</sup>, Sibao Wan<sup>5</sup>, Kun Wang<sup>6</sup>, Tianming He<sup>7</sup>, Shouqian Feng<sup>1</sup>, Zongying Zhang<sup>1</sup>, Zhiquan Mao<sup>1</sup>, Xiang Shen<sup>1</sup>, Xiaoliu Chen<sup>1</sup>, Yuanmao Jiang<sup>1</sup>, Shujing Wu<sup>1</sup>, Chengmiao Yin<sup>1</sup>, Shunfeng Ge<sup>1</sup>, Long Yang<sup>1</sup>, Shenghui Jiang<sup>1</sup>, Haifeng Xu<sup>1</sup>, Jingxuan Liu<sup>1</sup>, Deyun Wang<sup>1</sup>, Changzhi Qu<sup>1</sup>, Yicheng Wang<sup>1</sup>, Weifang Zuo<sup>1</sup>, Li Xiang<sup>1</sup>, Chang Liu<sup>8</sup>, Daoyuan Zhang<sup>9</sup>, Yuan Gao<sup>6</sup>, Yimin Xu<sup>3</sup>, Kenong Xu<sup>5</sup>, Thomas Chao<sup>10</sup>, Gennaro Fazio<sup>10</sup>, Huairui Shu<sup>1</sup>, Gan-Yuan Zhong<sup>10</sup>, Lailiang Cheng<sup>5</sup>, Zhangjun Fei<sup>3,11</sup> & Xuesen Chen<sup>1</sup>

Human selection has reshaped crop genomes. Here we report an apple genome variation map generated through genome sequencing of 117 diverse accessions. A comprehensive model of apple speciation and domestication along the Silk Road is proposed based on evidence from diverse genomic analyses. Cultivated apples likely originate from *Malus sieversii* in Kazakhstan, followed by intensive introgressions from *M. sylvestris*. *M. sieversii* in Xinjiang of China turns out to be an “ancient” isolated ecotype not directly contributing to apple domestication. We have identified selective sweeps underlying quantitative trait loci/genes of important fruit quality traits including fruit texture and flavor, and provide evidences supporting a model of apple fruit size evolution comprising two major events with one occurring prior to domestication and the other during domestication. This study outlines the genetic basis of apple domestication and evolution, and provides valuable information for facilitating marker-assisted breeding and apple improvement.

<sup>1</sup>State Key Laboratory of Crop Biology, Shandong Agricultural University, Tai’an, Shandong 271000, People’s Republic of China. <sup>2</sup>Shandong Centre of Crop Germplasm Resources, Shandong Academy of Agricultural Sciences, Jinan, Shandong 250100, People’s Republic of China. <sup>3</sup>Boyce Thompson Institute, Cornell University, Ithaca, NY 14853, USA. <sup>4</sup>State Key Laboratory of Crop Stress Biology in Arid Areas, College of Horticulture, Northwest A&F University, Yangling, Shaanxi 712100, People’s Republic of China. <sup>5</sup>Section of Horticulture, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853, USA. <sup>6</sup>The Institute of Pomology, Chinese Academy of Agricultural Sciences, Xingcheng, Liaoning 125100, People’s Republic of China. <sup>7</sup>College of Forestry and Horticulture, Research Centre of Specialty Fruits, Xinjiang Agricultural University, Urumqi, Xinjiang 830000, People’s Republic of China. <sup>8</sup>Mudanjiang Branch of Heilongjiang Academy of Agricultural Science, Mudanjiang, Heilongjiang 157500, People’s Republic of China. <sup>9</sup>Key Laboratory of Biogeography and Bioresource in Arid Land, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi, Xinjiang 830011, People’s Republic of China. <sup>10</sup>USDA-Agricultural Research Service, Plant Genetic Resources Unit, Geneva, NY 14456, USA. <sup>11</sup>USDA-Agricultural Research Service, Robert W. Holley Center for Plant and Health, Ithaca, NY 14853, USA. Naibin Duan, Yang Bai, Honghe Sun and Nan Wang contributed equally to this work. Correspondence and requests for materials should be addressed to Y.B. (email: [yb63@cornell.edu](mailto:yb63@cornell.edu)) or to Z.F. (email: [zf25@cornell.edu](mailto:zf25@cornell.edu)) or to X.C. (email: [chenxs@sdau.edu.cn](mailto:chenxs@sdau.edu.cn))

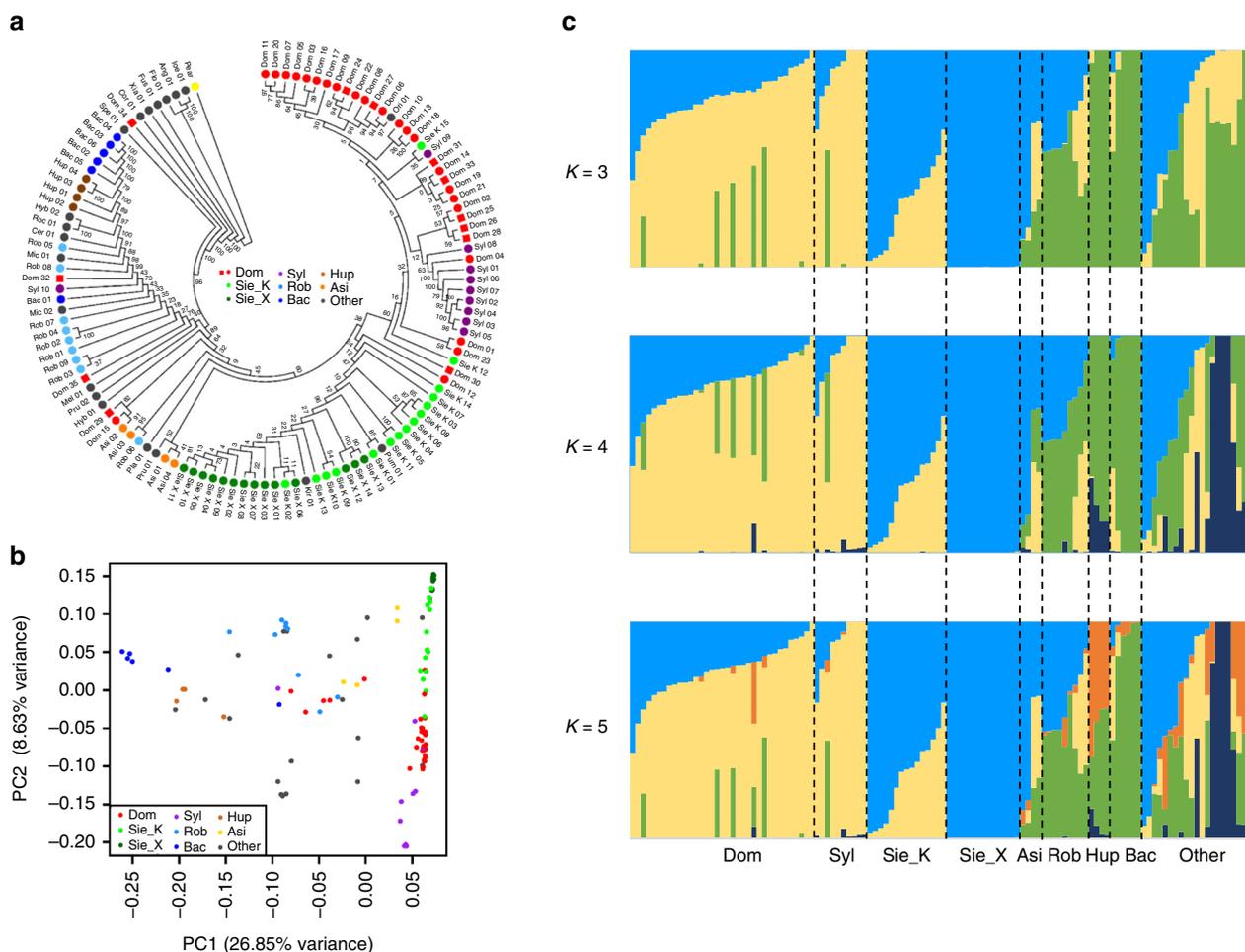
Crop domestication is an artificial selection process, fostering co-dependence between human and crop plants. This process often includes an initial phase of unconscious selection by farmers for several thousand years, followed by a subsequent breeding phase of intentional selection of favorable traits by breeders<sup>1</sup>. As a result, cultivated crops diverge from their wild progenitors morphologically and genetically in response to human selection. Recently, rapid advances in sequencing technologies have allowed researchers to sequence entire crop genomes efficiently and the resultant information can help elucidate how human-involved evolutionary processes have shaped modern crop genomes, as demonstrated in peach<sup>2</sup>, date palm<sup>3</sup>, tomato<sup>4</sup>, maize<sup>5</sup>, soybean<sup>6</sup>, and rice<sup>7</sup>. These studies provide valuable information for facilitating future crop improvement and thus securing global food production.

Cultivated apple (*Malus domestica* Borkh.), one of the most widely produced and economically important fruit crops in temperate regions, has been domesticated from *M. sieversii* in the Tian Shan Mountains for 4000–10,000 years, and dispersed from Central Asia to West Europe along the Silk Road, allowing hybridization and introgression of wild crabapples from Siberia (*M. baccata* (L.) Borkh.), Caucasus (*M. orientalis* Uglitz.), and Europe (*M. sylvestris* Mill.)<sup>8</sup>. It has been reported that *M. sieversii* is the primary progenitor and *M. sylvestris* is a major secondary

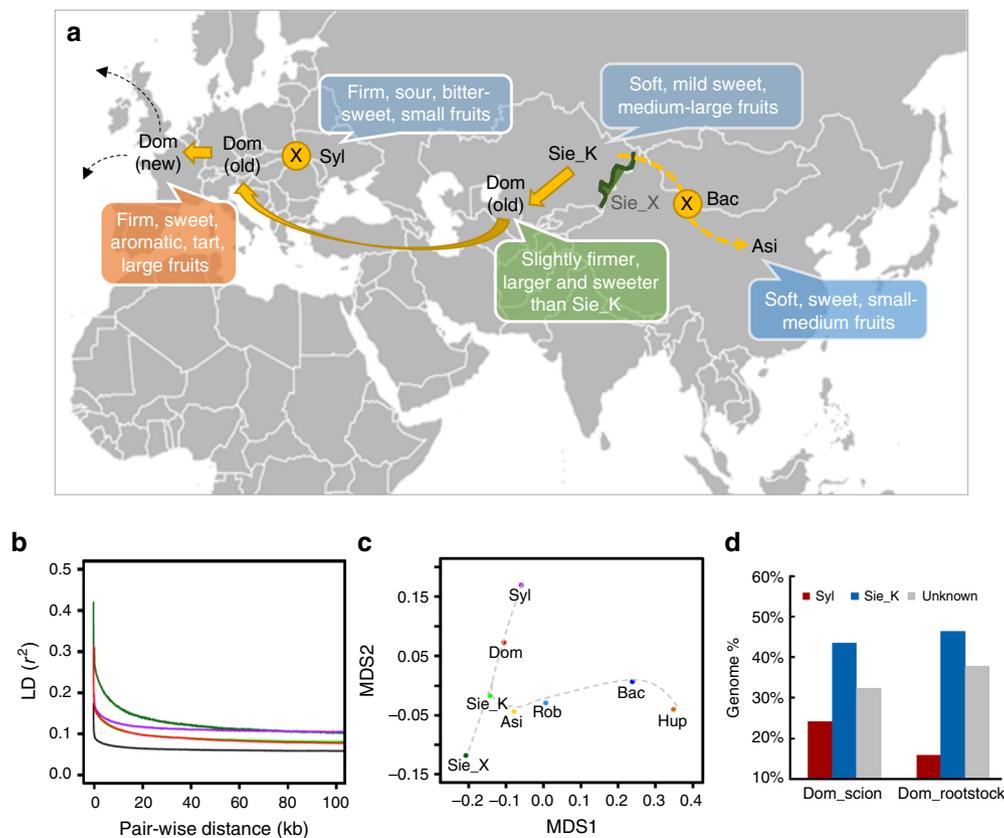
contributor of cultivated apples<sup>9</sup>. Centuries of human exploitation and selection have produced thousands of apple cultivars with diverse fruit sizes and flavors<sup>10</sup>. In addition, apples are characterized with lengthy juvenile phases, self-incompatibility, and vegetative propagation<sup>8, 11</sup>. These characteristics together with its unique domestication history indicate that the impact of domestication on the genomes of apples would be much different from that on the genomes of seed-propagated annual crops as well as self-compatible tree fruits such as peach. However, very limited information is available on how the apple genome has been reshaped by human selections and how small and sour wild ancestors have evolved into modern large and sweet apples. In this study, we performed deep genome sequencing of 117 diverse *Malus* accessions. The large amount of genomic variation resources has not only allowed us to gain new insights into apple evolution and domestication, but also provided invaluable information for accelerating the usage of wild species in apple breeding and improvement.

## Results

**Genome resequencing and variation calling.** A total of 117 *Malus* accessions from 24 species were selected for genome sequencing, including 35 *M. domestica* (24 scion and 11 rootstock



**Fig. 1** Population structure of 117 domesticated and wild apples. **a** Neighbor-joining phylogenetic tree constructed using SNPs at fourfold degenerate sites. Each species group is color coded, with red squares representing rootstocks and red dots scions. **b** Principal component analysis (PCA) of the 117 apple accessions. **c** Bayesian model-based clustering of the 117 apple accessions with the number of ancestry kinship ( $K$ ) from 3 to 5. Each vertical bar represents one apple accession and the x axis shows different apple accessions. Each color represents one putative ancestral background and the y axis quantifies ancestry membership. Asi *M. asiatica*, Bac *M. baccata*, Dom *M. domestica*, Hup *M. hupehensis*, Other other wild species, Rob *M. robusta*, Sie\_K *M. sieversii* in Kazakhstan, Sie\_X *M. sieversii* in Xinjiang, Syl *M. sylvestris*



**Fig. 2** Apple evolutionary map. **a** Apple evolutionary map along the west and east bounds of the Silk Route with center of origin at Kazakhstan in central Asia. **b** Decay of linkage disequilibrium ( $LD$ ) measured as the squared correlation coefficient ( $r^2$ ) by pairwise physical distance in *M. domestica*, *M. sieversii* in Kazakhstan, *M. sieversii* in Xinjiang, *M. sylvestris*, and other wild species. **c** Multidimensional scaling ( $MDS$ ) plot for the pairwise  $F_{ST}$  matrix. The Euclidean distances between each pair of groups significantly represent the corresponding  $F_{ST}$  values (Spearman rank-sum correlation  $\rho = 0.95$ ;  $p < 10^{-14}$ ). **d** Major alleles of scion and rootstock cultivars derived from *M. sieversii* in Kazakhstan and *M. sylvestris*. Asi *M. asiatica*, Bac *M. baccata*, Dom *M. domestica*, Hup *M. hupehensis*, Rob *M. robusta*, Sie\_K *M. sieversii* in Kazakhstan, Sie\_X *M. sieversii* in Xinjiang, Syl *M. sylvestris*

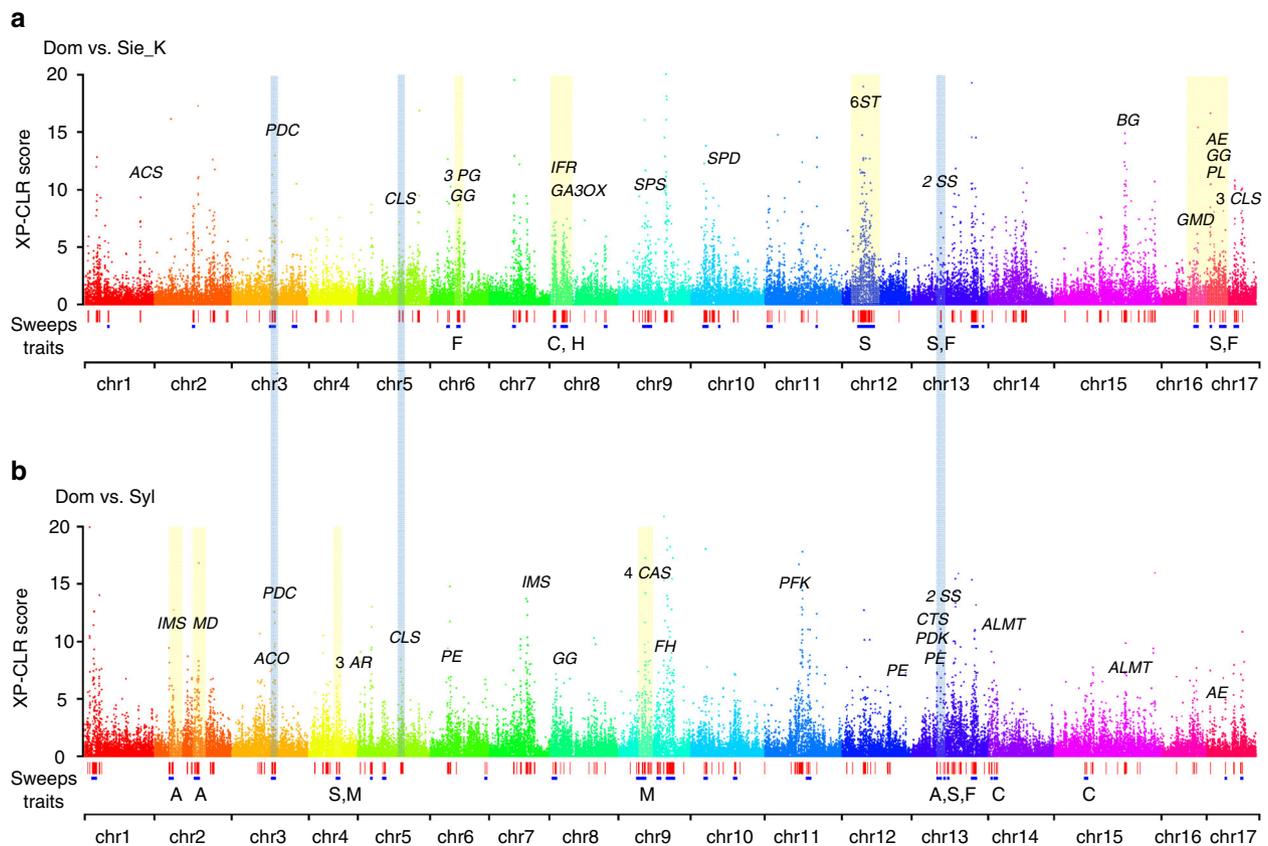
cultivars), 10 *M. sylvestris*, 29 *M. sieversii*, 9 *M. robusta*, 6 *M. baccata*, 4 *M. asiatica*, 4 *M. hupehensis*, and 20 in the remaining 17 wild species with one or two accessions per species (Supplementary Data 1). Among the 29 *M. sieversii* accessions, 15 originated in Kazakhstan, on the west side of the Tian Shan Mountains, and 14 were collected from natural forests in Xinjiang of China, on the east side of the Tian Shan Mountains. Six out of the 24 species are native to China, four native to North America, and two to Europe, while some *Malus* species are considered as intrageneric hybrids, such as *M. robusta* and *M. asiatica* (Supplementary Data 2). Chinese soft apples, such as “Pinpo”, “Xiangguo”, *M. asiatica* and *M. prunifolia* have been cultivated as dessert apples for more than 2000 years in China (Supplementary Note 1). Analysis of the phenotypic data recorded in the USDA-GRIN database (<https://npgsweb.ars-grin.gov/>) indicated that domesticated apples are significantly bigger, firmer, and sweeter than *M. sieversii* in Kazakhstan, while less acidulated and much bigger than *M. sylvestris* (Supplementary Fig. 1).

Resequencing of the 117 apple genomes generated a total of 1060 Gb high-quality cleaned sequences, with an average of 9.06 Gb per accession that represented  $\sim 12.2\times$  of the apple genome (Supplementary Data 1). After aligning the reads to the pseudo-haplotype apple genome<sup>12</sup> (v1.0p), we identified a final set of 7,218,060 single nucleotide polymorphisms (SNPs) (Supplementary Note 2 and Supplementary Tables 1 and 2). Furthermore, we identified 431,597 small insertions and deletions (indels). Polymerase chain reaction (PCR) amplification and Sanger sequencing on genomic regions containing 958 randomly selected

SNP loci in six apple accessions indicated a high accuracy rate (98.1%) for our genotype calling.

**An apple evolutionary map.** We first examined the phylogeny among wild and cultivated apples by constructing a neighbor-joining phylogenetic tree with pear as the outgroup using SNPs at fourfold degenerate sites (4D SNPs). The tree showed that accessions of *M. domestica* and its introgression contributor *M. sylvestris* formed a subclade within a large mixed clade comprising *M. sieversii* accessions that are considered as progenitors of cultivated apples<sup>8,9</sup>, while accessions of other wild species position outside this domestication-related clade (Fig. 1a). Wild species native to North America (*M. ioensis*, *M. angustifolia*, *M. fusca*, and *M. coronaria*) are the closest to pear, followed by Asian wild species *M. baccata* and *M. hupehensis*. Notably, the recent introgression from *M. sylvestris* into *M. domestica* has been so intensive that the cultivated apples now appear to be closer to European crabapple *M. sylvestris* than to their progenitor *M. sieversii*, which is consistent with a previous report<sup>9</sup>. A principal component analysis (PCA) illustrated a similar pattern to the phylogenetic tree in that *M. domestica*, *M. sieversii*, and *M. sylvestris* accessions formed closely related clusters that were clearly separated from the dispersed accessions of other wild species (Fig. 1b).

To further understand the evolutionary history of apple, we used a Bayesian clustering algorithm with admixed models<sup>13</sup> to estimate ancestry proportions for each accession (Fig. 1c).

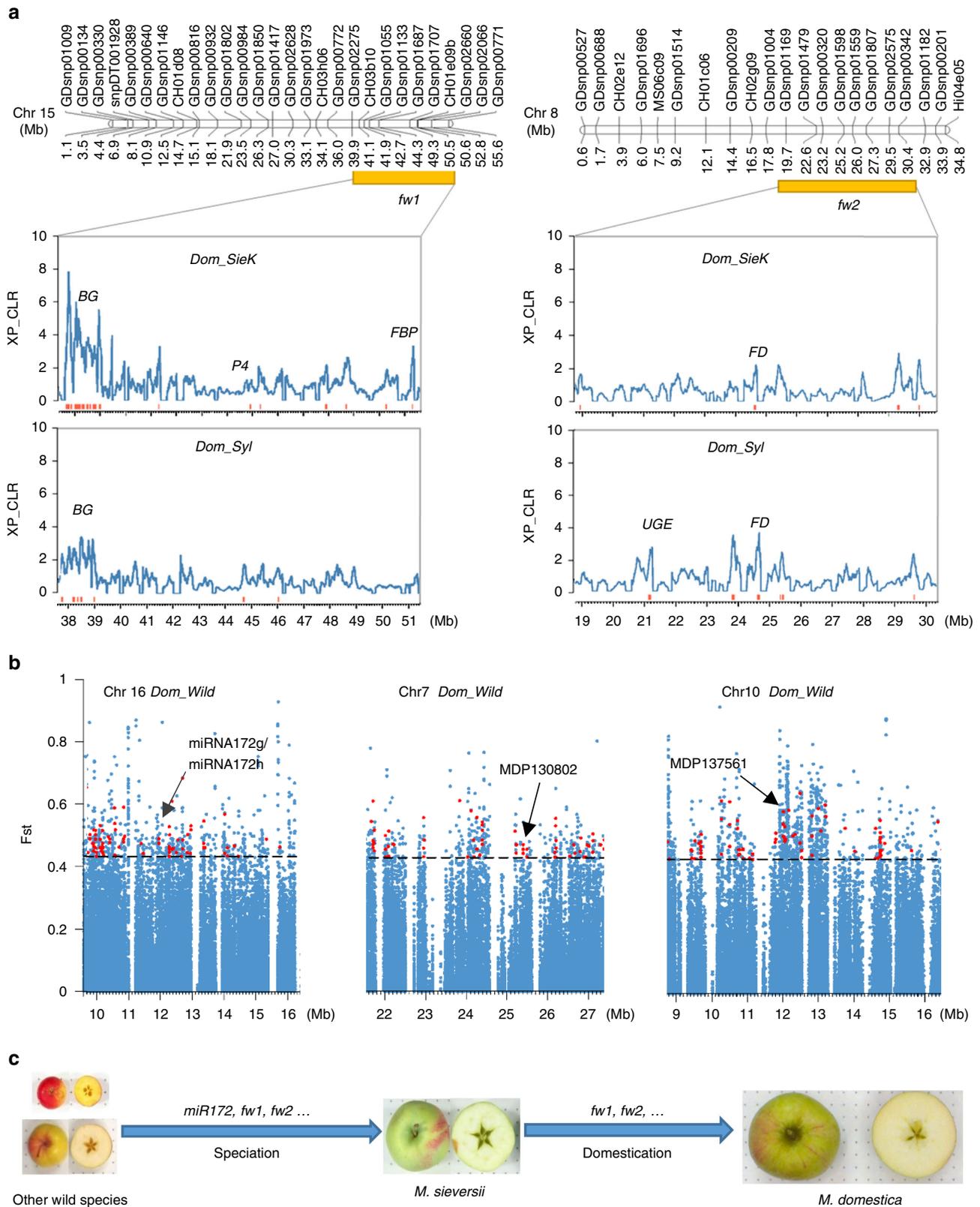


**Fig. 3** Genome-wide distribution of selective sweeps in *M. domestica*. **a** Selective sweeps in *M. domestica* compared with *M. sieversii* in Kazakhstan. **b** Selective sweeps in *M. domestica* compared with *M. sylvestris*. XP-CLR scores are plotted across the 17 chromosomes in the apple genome with key functional enzyme genes labeled above the dot peaks. Red vertical boxes illustrate selective sweeps, and blue boxes represent local GO enriched regions with associated traits labeled below. Selective regions shared by both comparisons are shaded with light blue bars, while interesting regions only identified in one of the two comparisons are shaded with light yellow bars. Traits include fruit acidity (A), color (C), firmness (F), hormone (H), soluble sugar (S), and secondary metabolites (M). Gene abbreviations: ALMT aluminum-activated malate transporter, ACO 1-aminocyclopropane-1-carboxylate oxidase, ACS 1-aminocyclopropane-1-carboxylate synthase, AE aldose 1-epimerase, AR aldose reductase, BG beta-galactosidase, CAS cycloartenol synthase, CTS citrate synthase, CLS cellulose synthase, FH flavanone 3-hydroxylase, GA3OX gibberellin 3-beta-dioxygenase, GG glucan endo-1,3-b-glucosidase, GMD GDP-mannose 4,6-dehydratase, IFR isoflavone reductase, IMS, 2-isopropylmalate synthase, MD malate dehydrogenase, PDC pyruvate decarboxylase, PDK pyruvate dehydrogenase, PE pectin esterase, PFK 6-phospho-fructokinase, PG polygalacturonase, PL pectate lyase, SPD sorbitol 6-phosphate dehydrogenase, SPS sucrose phosphate synthase, SS sucrose synthase, ST sugar transporter

Our  $\Delta K$  analysis revealed that five populations ( $K = 5$ ) represent the best model for these 117 accessions (Supplementary Fig. 2). For  $K = 3$ , *M. domestica* and its wild relatives, *M. sieversii* and *M. sylvestris*, were clearly separated from other wild species, supporting the evolutionary history of domesticated apple<sup>8, 9, 14</sup>. With  $K$  from 4 to 5, two new subpopulations arose from the wild species other than *M. sieversii* and *M. sylvestris*, indicating their high diversity and further distance from domesticated apples. *M. sieversii* accessions from the two sides of the Tian Shan Mountains segregated into two different subpopulations reflecting their geographical distributions. *M. sieversii* accessions in Kazakhstan showed admixed ancestry possibly from hybridizations with wild apples such as *M. orientalis* along the Silk Road, and/or domesticated apples cultivated nearby, while Xinjiang accessions kept their homogeneous genetic background probably due to their geographical isolation that blocks interspecific hybridization. In addition to *M. sieversii* in Xinjiang, six other species in distinct habitats, such as *M. ioensis* and *M. angustifolia* in North America and *M. baccata* in East Asia, were also identified with homogeneous genetic background, which gave rise to other hybrid species and possess tremendous value in apple breeding practices (Supplementary Fig. 3). The structure of several hybrid species was consistent with their known pedigrees

recorded in the USDA-GRIN database, including *M. asiatica*, *M. prunifolia*, *M. robusta*, and several rootstocks, while that of *M. platycarpa* did not agree with its pedigree in the database (Supplementary Figs. 4 and 5 and Supplementary Note 3). Together, these findings prompt us to propose a comprehensive apple evolutionary map across Eurasia continent, illustrating the initial domestication from *M. sieversii* in Kazakhstan, the hybridization between *M. sylvestris* and the ancient domesticated apples spread from Central Asia to Europe via the Silk Road westward, and the rise of orient hybrid species from crosses between *M. baccata* and *M. sieversii* in Kazakhstan distributed and cultivated along the Silk Road eastward (Fig. 2a and Supplementary Note 4). During the domestication process, cultivated apples retained the large fruit size from *M. sieversii*, gained the firm texture and appetizing flavor from the hybridization with *M. sylvestris* and continued to be bred into larger and firmer fruit with better flavor and aroma.

We then evaluated the genetic diversity of different apple subpopulations. The genome-wide nucleotide diversity ( $\pi$ ) of *M. domestica* ( $2.20 \times 10^{-3}$ ) was lower than that of *M. sieversii* in Kazakhstan ( $2.35 \times 10^{-3}$ ), *M. sylvestris* ( $2.55 \times 10^{-3}$ ), and other wild species ( $4.26 \times 10^{-3}$ ), while *M. sieversii* in Xinjiang exhibited the lowest diversity level ( $1.30 \times 10^{-3}$ ). Compared to other



**Fig. 4** Evolution of fruit size during speciation and domestication in apple. **a** Domestication sweeps underlying apple fruit size QTLs. Within the physical intervals (orange boxes) of the two fruit weight QTLs *fw1* and *fw2*, distributions of XP-CLR scores are shown. Selective sweeps are marked with red bars and interesting genes are labeled above peaks. **b** *MiRNA172g/miRNA172h* and the two target genes that contain highly divergent SNPs (pointed by arrows) between *M. domestica* with large fruits and other wild species bearing very small fruits. **c** Schematic diagram of the two-step evolution of apple fruit size. *BG* beta-galactosidase, *FBP* fructose-1,6-bisphosphatase, *FD* ferredoxin, *PPD* pyruvate phosphate dikinase, *P4* patellin-4, *UGE* UDP-glucose 4-epimerase

domesticated perennial crops, the nucleotide diversity in cultivated apple is higher than that of peach<sup>2</sup> ( $1.5 \times 10^{-3}$ ), but lower than that of cassava<sup>15</sup> ( $2.6 \times 10^{-3}$ ) and date palm<sup>3</sup> ( $9.2 \times 10^{-3}$ ). The similar level of nucleotide diversity between *M. domestica* and its progenitor *M. sieversii* in Kazakhstan indicated a very weak bottleneck, if any, during apple domestication, consistent with findings in previous studies<sup>8, 9</sup>. Linkage disequilibrium (LD) analyses for each group further supported a very weak and nearly undetectable domestication bottleneck (Fig. 2b and Supplementary Note 5). The rapid LD decay of domesticated apple suggested that a large set of markers densely covering the genome is preferred for a high-resolution population genetic analysis, as indicated in a recent study<sup>16</sup>. Here we also demonstrated the power of our high-density SNPs in enhancing the resolution of genome-wide association studies (GWAS) (Supplementary Note 6, Supplementary Fig. 6, Supplementary Table 3, and Supplementary Data 3 and 4).

To further investigate population divergence among different species groups, we computed pairwise  $F_{ST}$  values, demonstrating consistent relationships among the subpopulations with our proposed evolutionary scenario (Fig. 2c and Supplementary Fig. 7). Genome-wide inference of major allele origins in *M. domestica* from *M. sieversii* in Kazakhstan or *M. sylvestris* revealed that 46% of the *M. domestica* genome was probably derived from its progenitor *M. sieversii* in Kazakhstan and 21% from its secondary contributor *M. sylvestris*, while the origin of the remaining 33% was uncertain. The genomic introgression from *M. sylvestris* to scion cultivars is about 10% higher than that to rootstock cultivars, raising the possibility that *M. sylvestris* may have contributed important alleles for fruit quality and production traits to dessert apple cultivars (Fig. 2d and Supplementary Fig. 8).

**Differential selection of domesticated apples.** Considering the remarkable role of *M. sylvestris* in shaping modern domesticated apples, genomic regions dramatically affected by selection during apple domestication were identified in two contrasts: *M. domestica* vs. *M. sieversii* in Kazakhstan for initial domestication (*Dom\_SieK*) and *M. domestica* vs. *M. sylvestris* for secondary introgression (*Dom\_Syl*) (Fig. 3 and Supplementary Figs. 9 and 10). The selected regions of *Dom\_SieK* and *Dom\_Syl* had a mean size of 33.1 and 42.8 kb, covered a total length of 13.9 Mb (3.7 % of genome) and 17.2 Mb (4.6%), and harbored 840 and 1089 genes, respectively, among which 246 (29.3%) and 336 (30.9%) showed differential expression during apple fruit development (Supplementary Data 5–8). The identified selective sweeps are enriched with genes associated with fruit sugar content, firmness, color, hormone, and secondary metabolism in both *Dom\_SieK* and *Dom\_Syl* contrasts, while genes related to fruit acidity were only enriched in *Dom\_Syl* (Fig. 3 and Supplementary Data 9). Notably, the enriched genes included those encoding six sugar transporters, several key enzymes in the glycolysis/gluconeogenesis pathway, two sucrose synthases and three cellulose synthases in *Dom\_SieK*, two aluminum-activated malate transporters, a malate dehydrogenase, and a citrate synthase in *Dom\_Syl*, and two sucrose synthases, one pyruvate decarboxylase, and one cellulose synthase in both *Dom\_SieK* and *Dom\_Syl*, highlighting the constant selection of sweet and firm fruits in the history of apple domestication (Fig. 3). Together, these candidate domestication-related genes are indicative of different selective forces for improving different agronomic traits from the two wild contributors during domestication.

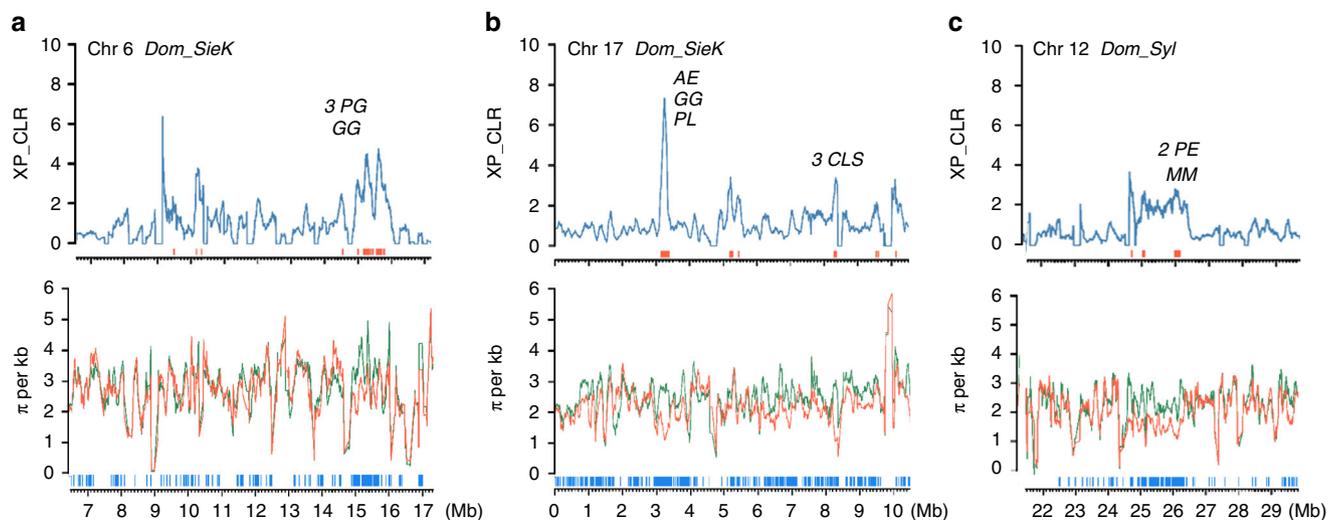
We further scanned SNPs that were highly divergent between *M. domestica* and different wild species groups using the top 1%  $F_{ST}$  values (Supplementary Data 10). A large number of

disease resistance (R) genes and genes involved in various different abiotic stresses were found to contain non-synonymous SNPs highly divergent between *M. domestica* and other wild species groups, suggesting their adaptations to different growth environments. Within *M. domestica* accessions, genes underlying dwarf quantitative trait loci (QTLs) (*Dw1* and *Dw2*)<sup>17, 18</sup>, R genes and receptor kinase genes were identified to have non-synonymous SNPs highly divergent between rootstock and scion cultivars, which could facilitate the study of important apple rootstock traits, such as dwarfing, precocity, and disease resistance (Supplementary Data 11). The highly divergent SNPs and the associated genes discovered in this study provided ample information for broadening our understanding of apple speciation, differentiation, and evolution.

### Increase of fruit size prior to and during apple domestication.

One essential aspect of domestication process for most crop species is to increase fruit and/or seed size, which is also referred to as “domestication syndrome”<sup>19</sup>. Fruit sizes of *M. sylvestris* are significantly smaller than those of both *M. domestica* and *M. sieversii*, while domesticated apples have larger fruits than *M. sieversii* (Supplementary Fig. 1d, e). Two previously reported fruit weight QTLs<sup>20</sup> (designated as *fw1* on chromosome 15 and *fw2* on chromosome 8) were found to be co-located with selective sweeps (Fig. 4a). QTLs *fw1* and *fw2* harbor 11 and 7 genes, respectively, in selective regions of *M. domestica* from *M. sieversii*, and 8 and 21 genes, respectively, in sweeps from *M. sylvestris* (Supplementary Data 12). Genes encoding regulators of cell division, such as *fw2.2* in tomato<sup>21</sup> and *CNR1* in maize<sup>22</sup>, have been reported to control organ size, and recently a  $\beta$ -galactosidase gene was found to be involved in the regulation of fruit weight and size in strawberry, besides fruit softening<sup>23</sup>. We found that one cell division regulatory gene (MDP0000223854) and two  $\beta$ -galactosidase genes (*MDP0000921848* and *MDP0000179821*) in *fw1* were under human selection. Interestingly, the  $\beta$ -galactosidase gene, *MDP0000179821*, was in selected regions of both *Dom\_SieK* and *Dom\_Syl* and showed differential expression during apple fruit development with highest expression at the active cell division stage (Supplementary Data 12). Furthermore, several cell division regulatory genes (e.g., *MDP000055176*, *MDP0000802780*, *MDP0000846861*, and *MDP0000681201*) and a gene (*MDP0000241347*) homologous to rice *GS3* that controls grain size<sup>24</sup> were found in several other selected regions, and they all showed highest expression at the active cell division stage of fruit development (Supplementary Data 6 and 8), suggesting their potential contribution to the increase of fruit size during apple domestication.

*MiRNA172p* was recently reported to regulate apple fruit size by targeting AP2 transcription factors<sup>25</sup>. Using their precursor sequences, the 16 miRNA172 genes identified in apple<sup>26</sup> were clustered into four clades (Supplementary Fig. 11). We detected a highly differentiated SNP only between domesticated and other wild apples in the precursor sequences of *miRNA172g* and *miRNA172h*, two miRNAs sharing the same precursor sequences (Fig. 4b). No differentiated SNPs were detected in the 16 miRNA172 genes between *M. domestica* and *M. sieversii*. Furthermore, we found that five out of the eight target AP2 genes of *miRNA172g* and *miRNA172h* carried highly differentiated SNPs in the comparison between *M. domestica* and other wild species, two of which comprised non-synonymous SNPs (Fig. 4b and Supplementary Fig. 11). Therefore, besides the previously discovered *miRNA172p*<sup>25</sup>, we identified two additional miRNA172 genes (*miRNA172g* and *miRNA172h*) that might have contributed to the increase of fruit size during *Malus* speciation prior to domestication.



**Fig. 5** Domestication sweeps underlying apple fruit firmness. Distributions of XP-CLR scores and nucleotide diversity ( $\pi$ ) in selective regions on chromosomes 6 (**a**) and 17 (**b**) in *M. domestica* from *M. sieversii* in Kazakhstan (*Dom\_SieK*) and chromosome 12 (**c**) from *M. sylvestris* (*Dom\_syl*), which harbor genes known to be associated with fruit firmness. Distributions of XP-CLR scores are shown in *top panels* with selective sweeps marked with *red bars* and interesting genes labeled above peaks. Distributions of  $\pi$  are shown in *bottom panels* with *M. domestica* in *orange lines*, and *M. sieversii* and *M. sylvestris* in *green lines*. AE aldose 1-epimerase, CLS cellulose synthase, GG glucan endo-1,3-b-glucosidase, MM endo-beta-1,4-mannase, PE pectinesterase, PG polygalacturonase, PL pectate lyase

Taken together, we propose a two-step evolution model for fruit size enlargement in apple to characterize its unique evolution process (Fig. 4c). Unlike modern maize and tomato whose domestication process started with ancestral species bearing very small seeds or fruits<sup>4, 5</sup>, apple domestication was initiated from *M. sieversii* whose fruits are larger than all other wild apples<sup>8</sup>. As fruit size is one of the most desirable traits for crop domestication and improvement, apple domestication started with a great advantage and much lower evolutionary pressure than other crops. The finding of fruit weight QTLs, and *miRNA172s*, and their target genes from comparisons between small-fruited wild apples and large-fruited cultivars helped explain why a weak selection in a highly heterozygous perennial crop can still yield favorable large fruits. The mild increase in fruit size during several thousand years of domestication after speciation, partially contributed by QTLs *fw1* and *fw2*, suggests that apple fruit size has great potential to be increased in future breeding practices, considering that the modern tomato fruit is approximately 100 times larger than its direct wild progenitor.

#### Enhancement of fruit firmness during apple domestication.

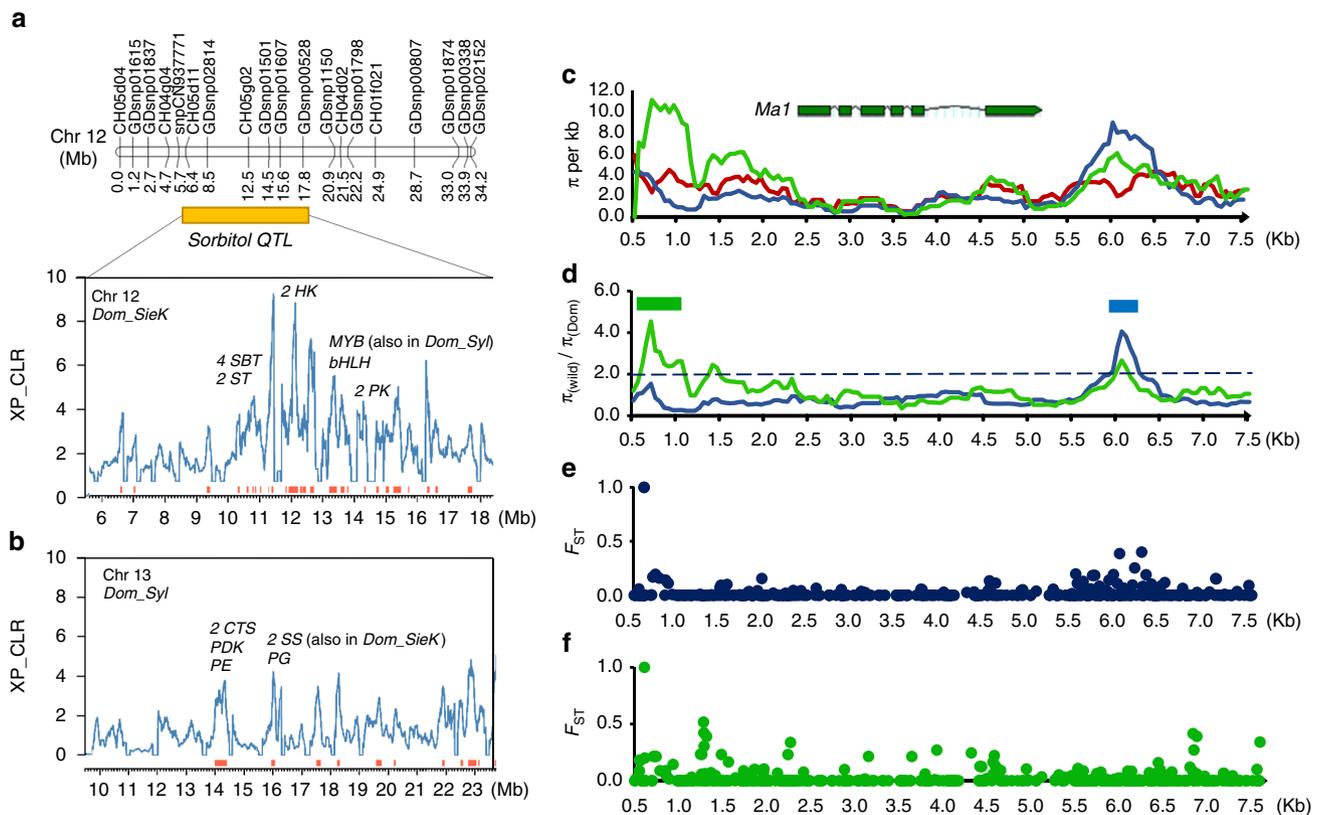
Besides large fruits, humans also have been selecting firm flesh texture, not only for crispy taste, but also for a longer shelf life, better post-harvest disease resistance, and reduced bruising during harvest and transportation. Fruit firmness is directly linked to enzyme-mediated cell wall modification<sup>27</sup>. To decipher the genetic mechanism underlying the selection for firm apples, we mined selective sweeps for genes that were potentially involved in regulating fruit firmness. We found a region on chromosome 16 was under intensive human selection in the *Dom\_SieK* contrast, harboring genes encoding key cell wall modifying enzymes including three polygalacturonases (*MDP0000512850*, *MDP0000939625*, and *MDP0000873268*), and one glucan endo-1,3-beta-glucosidase (*MDP0000295938*) (Fig. 5a). Furthermore, several other selective regions were also found to contain genes related to cell wall modifications. For example, one selective sweep on chromosome 17 from the *Dom\_SieK* comparison comprised three cellulose synthase genes (*MDP0000190520*, *MDP0000289339*, and *MDP0000184309*),

and another sweep from the same chromosome harbored one pectate lyase (*MDP0000301545*), one glucan endo-1,3-beta-glucosidase (*MDP0000206670*), and one aldose 1-epimerase (*MDP0000737131*) (Fig. 5b). Similarly, another sweep on chromosome 12 from the *Dom\_Syl* contrast harbored one endo-beta-1,4-mannase (*MDP0000832632*) and two pectinesterase genes (*MDP0000278119* and *MDP0000278118*) (Fig. 5c). A number of these cell wall-related genes were differentially regulated during fruit development (Supplementary Data 6 and 8). Therefore, the evolution of these genes might have contributed to the firm fruit texture of domesticated apples.

#### Enrichment of fruit flavor during apple domestication.

Fruit flavor, mainly a balance between sugars and acids, is another important trait under human selection. Although it has been reported that fruit acidity rather than sweetness is likely to undergo artificial selection<sup>28</sup>, we found that both fruit sugar content and acidity have been altered during apple domestication using historical phenotype data in the USDA-GRIN database (Supplementary Fig. 1). During the domestication from *M. sieversii* in Kazakhstan, a region on chromosome 12 has undergone intensive selection and co-localizes with a sorbitol QTL<sup>29</sup>. Interestingly, genes encoding four sorbitol transporters and two sugar transporters cluster within this sorbitol QTL under selection, and all four sorbitol transporter genes were differentially expressed during apple fruit development (Fig. 6a and Supplementary Data 13). Similarly, another genomic region on chromosome 13 enriched with genes encoding key enzymes of sugar metabolism is under intensive selection from *M. sylvestris* (Fig. 6b). Notably, the two sucrose synthase genes in this region are also in selective sweeps from *M. sieversii* in Kazakhstan.

To study fruit acidity during apple domestication, we investigated the nucleotide diversity of gene *Ma1*, which underlies the major fruit acidity QTL *Ma* and encodes an aluminum-activated malate transporter<sup>30</sup>. We identified a region around 1.5 kb upstream of *Ma1* coding sequence (CDS) with substantially reduced nucleotide diversity in *M. domestica* compared to *M. sylvestris*, and another region about 750 bp downstream of *Ma1* CDS when compared to *M. sieversii* in Kazakhstan



**Fig. 6** Domestication of fruit sweetness and acidity in apples. **a** Selective sweeps from *M. sieversii* in Kazakhstan that co-localize with a sorbitol QTL. **b** Domestication sweeps from *M. sylvestris* that contain key genes for sugar metabolism. Distributions of XP-CLR scores are shown. Selective sweeps are marked with red bars and interesting genes are labeled above peaks. **c–f** Domestication of the *Ma1* gene that regulates apple fruit acidity. **c** Distributions of nucleotide diversity ( $\pi$ ) of *M. domestica* (red), *M. sieversii* in Kazakhstan (blue), and *M. sylvestris* (green) in the *Ma1* genome region. **d** *Ma1* selective sweeps during domestication from *M. sieversii* in Kazakhstan (blue) and *M. sylvestris* (green). Sweep regions are marked with filled boxes. **e** Distribution of  $F_{ST}$  between *M. domestica* and *M. sieversii* in Kazakhstan in the *Ma1* genome region. **f** Distribution of  $F_{ST}$  between *M. domestica* and *M. sylvestris* in the *Ma1* genome region. *bHLH* transcription factor, *CTS* citrate synthase, *HK* hexokinase, *MYB* transcription factor, *PDK* pyruvate dehydrogenase kinase, *PE* pectin esterase, *PG* polygalacturonase, *PK* pyruvate kinase, *SBT* sorbitol transporter, *SS* sucrose synthase, *ST* sugar transporter

(Fig. 6c–f), indicating that the *Ma1* might be under strong selection during apple domestication. The distributions of  $F_{ST}$  of SNPs surrounding the *Ma1* gene region further underpinned the selective sweep regions. In addition to *Ma1*, two other aluminum-activated malate transporter genes and a gene encoding malate dehydrogenase were identified in selection sweeps from *M. sylvestris*, assisting in explaining the significantly reduced acidity level in domesticated apples when compared to *M. sylvestris* (Fig. 3 and Supplementary Fig. 1).

## Discussion

We report here a comprehensive genome variation map of apple through deep resequencing the genomes of 117 wild and cultivated apples. The population analyses reveal that modern cultivated apples originate from *M. sieversii* in Kazakhstan with intensive recent introgressions from *M. sylvestris*, while *M. sieversii* in Xinjiang is an isolated ecotype with a homogeneous genetic background that holds great potential for future apple improvement. Chinese native species, such as *M. asiatica* and *M. prunifolia*, are very likely to be hybrids between *M. baccata* and *M. sieversii* in Kazakhstan. The similar level of genetic diversity between *M. domestica* and its direct wild progenitor *M. sieversii* in Kazakhstan is consistent with the comparably fast LD decays in these two species and an undetectable bottleneck in apple domestication. The weak selection during apple domestication might be explained by its unique physiological characteristics, such as self-incompatibility, lengthy juvenile

phases, and limited sexual reproduction. However, several thousand years of selection still has accumulated a considerable amount of desirable fruit traits that make apple one of the most important fruit crops worldwide. We identified selective sweeps underlying fruit size, flesh firmness, sweetness, and acidity, as well as highly divergent SNPs between rootstock and scion cultivars that are associated with dwarfing and disease resistance. The identified large set of SNPs, highly confident candidate genes associated with important apple traits, and the insight gained regarding apple evolution and domestication provide valuable information for facilitating future apple marker assisted breeding, QTL mapping, and population-structured association studies.

## Methods

**Sample preparation and sequencing.** A total of 117 apple accessions were included in this study, of which 70 were provided by USDA-ARS Plant Genetics Resources Unit (Geneva, NY, USA), 33 by Institute of Pomology of Chinese Academy of Agricultural Sciences (Xingcheng, Liaoning, China), and 14 were collected from the Tian Shan Mountains (Xinjiang, China). Genomic DNA was extracted from young leaves using the DNeasy Plant Mini Kit (Qiagen, Venlo, the Netherlands). Paired-end Illumina genomic libraries were prepared and sequenced on an Illumina HiSeq 2500 platform following the manufacturer's instructions (Illumina, San Diego, CA).

**Mapping and variant calling.** The Illumina paired-end reads from each apple accession were first processed to collapse duplicated read pairs into unique read pairs. Duplicated read pairs were defined as those having identical bases at positions of 15–90 in both left and right reads. The resulting unique reads were processed to remove adaptor and low-quality sequences using Trimmomatic<sup>31</sup>.

The cleaned unique reads were aligned to the apple reference genome<sup>12</sup> (v1.0p, [https://www.rosaceae.org/species/malus/malus\\_x\\_domestica/genome\\_v1.0p](https://www.rosaceae.org/species/malus/malus_x_domestica/genome_v1.0p)) using BWA<sup>32</sup>, and only uniquely mapped reads were retained. Following mapping, genotypes were assigned to each genomic position for each accession based on the alignment mpileup files generated by SAMtools<sup>33</sup>. SNPs and small indels in the population of 117 apple accessions were then identified if they were supported by at least two mapped reads, and had an allele frequency  $\geq 0.3$ . Furthermore, SNP sites with the nearest SNPs less than 5 bp away, containing only heterozygous SNPs, or with more than 50% missing data were excluded.

To validate the SNP calling, we randomly selected 958 homozygous SNP loci across the genome and PCR-amplified and Sanger-sequenced the corresponding genomic regions in six apple accessions (*M. domestica*, Dom\_01; *M. sieversii*, Sie\_K\_04, Sie\_X\_03 and Sie\_X\_09; *M. sylvestris*, Syl\_01; and *M. hupehensis*, Hup\_01).

**Phylogenetic and population structure analyses.** We selected a total of 24,326 SNPs at fourfold degenerate sites (4D SNPs) with minor allele frequency (MAF)  $\geq 5\%$  and missing rate per site  $\leq 10\%$  for phylogenetic and population structure analyses. The rationale for choosing 4D SNPs is that they do not cause amino-acid changes and thus, should be under less selective pressure and more reliably reflect population structure and demography<sup>18</sup>. PCA was performed using program EIGENSOFT<sup>34</sup>. A neighbor-joining phylogenetic tree was constructed using MEGA6<sup>35</sup> in 1000 bootstrap replicates with pear "*Pyrus × bretschneideri*" as the outgroup. Pear genomic sequence reads<sup>36</sup> were downloaded from the NCBI SRA database (accession no. SRP016889) and aligned to the apple genome v1.0p for SNP calling. Population structure was investigated using STRUCTURE<sup>13</sup>. To determine the most likely number of ancestral kinships (*K*) in the apple population, STRUCTURE was run 20 times on 2000 randomly selected 4D SNPs for each *K* value from 3 to 19. The statistic " $\Delta K$ ", which indicates the change in likelihood of different numbers of clusters, was calculated, and the cluster number with the highest  $\Delta K$  value, which indicated the most likely number of clusters in the population, was obtained. We then used the whole set of 4D SNPs to calculate the cluster membership for each accession with 10,000 iterations. According to the phylogenetic and population structure analyses, nine accessions (*M. domestica*, Dom\_15, Dom\_29, Dom\_32, Dom\_34, and Dom\_35; other wild species, Kir\_01, Ori\_01, and Pum\_01; *M. sylvestris*, Syl\_10) were positioned into unexpected species groups and thus excluded from downstream population genetics analyses.

**Inferring population genetic statistics.** Nucleotide diversity ( $\pi$ ) is often used as a measurement of the degree of the genotype variability within a population or species<sup>37</sup>. Values of  $\pi$  were calculated for 10-kb non-overlapping windows across the apple genome with the final set of 7,218,060 SNPs using the BioPerl module PopGen<sup>38</sup>.

Population differentiation is evaluated by the fixation index  $F_{ST}$ <sup>39</sup>. Pairwise  $F_{ST}$  values were calculated among the eight species groups listed in the Supplementary Data 1 for the final SNP set using a variance component approach implemented in the R package HIERFSTAT<sup>40</sup>. To visualize the  $8 \times 8$  pairwise  $F_{ST}$  matrix, multidimensional scaling (MDS) was performed using an R function *cmdscale* to transfer  $F_{ST}$  values into two dimensional values for plotting. To test the accuracy of this MDS visualization, the Spearman rank-sum correlation  $\rho$  was calculated between Euclidean distances between all pairs of values in the MDS plot and the corresponding  $F_{ST}$  values. SNPs with top 1%  $F_{ST}$  values were identified as highly divergent variations in comparisons between *M. domestica* and each of the four wild groups, including *M. sieversii* in Kazakhstan, *M. sieversii* in Xinjiang, *M. sylvestris*, and other wild species, as well as between rootstock and scion cultivars. To better reflect selective pressure on nucleotide variants, non-synonymous SNPs were further identified from the highly divergent SNPs with top 1%  $F_{ST}$  values, and the corresponding genes containing these non-synonymous SNPs were determined.

**LD.** Correlation coefficient ( $r^2$ ) of genotypes was calculated to measure LD level in each of the five groups using Haploview<sup>41</sup> with the following parameters: -nogui -minMAF 0.10 -hwcutoff 0.001 -dprime. SNPs within each group were extracted for the analysis. SNPs within 500-kb sliding windows were used to estimate average  $r^2$  at various physical distance classes, and the LD decay was plotted as a function of the derived average  $r^2$  and physical distances. Tag SNPs were also identified using Haploview with parameters set as follows: -nogui -minMAF 0.05 -hwcutoff 0.001 -dprime -blockoutput GAB -minGeno 0.8 -pairwiseTagging.

**Inferring-derived alleles.** To evaluate whether major alleles of *M. domestica* (*Dom*) are derived from those of *M. sieversii* in Kazakhstan (*Sie\_K*) or *M. sylvestris* (*Syl*), the entire genome was scanned using SNPs with MAF  $\geq 5\%$  and missing rate per site  $\leq 10\%$ . First, allele frequency per site was computed in each species group and the allele with the frequency  $> 0.6$  was assigned as the major allele. For each site, we considered the major allele in *Sie\_K* as the ancestral allele in *Dom* if these two groups share consistent major alleles; similarly, *Syl* was considered as the allelic ancestor for *Dom* when they had the same major alleles. For each non-overlapping 100-kb window, if the count of ancestral alleles as *Sie\_K* was greater than two times of that as *Syl* and the total SNP count was larger than five, we regarded this region

as being derived from *Sie\_K*; similarly, if the ancestral-allele count ratio between *Syl* and *Sie\_K* was larger than two and the window contained more than five SNPs, we considered this window as originating from *Syl*.

**Identification of selective sweeps.** Selective sweeps across the genome were identified using a XP-CLR method that models the allele frequency spectrum to search for signals of differentiation between two populations<sup>42</sup>, following previous studies<sup>43, 44</sup> with modifications. Briefly, selection screening was performed in two contrasts: *M. domestica* vs. *M. sieversii* in Kazakhstan for initial domestication (*Dom\_SieK*) and *M. domestica* vs. *M. sylvestris* for introgressive domestication (*Dom\_Syl*). Genetic distances between adjacent SNPs were calculated according to their physical distances in a widely used genetic map reported with the apple reference genome<sup>12</sup>. Parameters "-w1 0.0005 100 100 -p0 0.7" were adopted to run the program for each chromosome, allowing XP-CLR scores to be calculated for each 100-bp region using unphased SNPs with correlation level at 0.7 and no more than 100 SNPs per window (in size of 0.05 cM). The XP-CLR scores per 100 bp were averaged across non-overlapping 10-kb windows on each chromosome. Adjacent 10-kb windows with an average XP-CLR score no less than 80% of the genome-wide average XP-CLR score were joined as putative selective regions, which were further merged together if two regions were separated by only one low-score 10-kb window. The maximum of window-wise average XP-CLR scores in a merged region was assigned as the region-wise XP-CLR score. Merged regions with top 10% of region-wise XP-CLR scores were considered as candidate selective sweeps. To improve the prediction accuracy, only candidate selective sweeps having top 50% of  $\pi$  ratios between the ancestral and domesticated species were kept.

Genes in the candidate selective sweeps were grouped by their physical locations into each 1-Mb non-overlapping segments throughout the genome. Adjacent segments and segments separated by no more than 1-Mb were merged together, while segments containing no more than five genes were discarded. GO term enrichment analyses were conducted for genes in each merged segment along the genome using GO::TermFinder<sup>45</sup>.

**Genome-wide association study.** The USDA-GRIN database archives phenotypic observations for 105 different traits in 4123 apple accessions from 1995 to 2014 (<https://npgsweb.ars-grin.gov/gringlobal/cropdetail.aspx?type=descriptor&id=115>).

The phenotypic scores of fruit flavor, weight, width, length, flesh firmness, skin color, and soluble solid content were downloaded for 66 out of the 70 accessions from USDA-ARS that are included in the USDA-GRIN database. Phenotype scores recorded in multiple years for a particular accession were screened to exclude outliers (more than 1.5 times the interquartile range below the first quartile or above the third quartile) and then averaged across the years. Tukey's HSD mean separation tests were conducted to analyze these phenotypes in pairwise comparisons between *M. domestica*, *M. sieversii* in Kazakhstan, and *M. sylvestris*.

Phenotypic data of skin color was used for the association study. To improve phenotyping of fruit skin color, we manually examined fruit pictures in the USDA-GRIN database and assigned a score of fruit skin color for each accession using the following scoring system: 0, *green*; 1, *yellow*; 2, *light red* (*green base* with *pink-red* area and/or stripes less than half of the fruit); 3, *medium red* (*green-yellow base* with blush more than half of the fruit); 4, *red* (without *green* or *yellow base*); 5, *dark red*. Same was done on 13 accessions collected from China (Supplementary Table 3). The final set SNPs were filtered by keeping those with missing rate  $\leq 10\%$ , MAF  $\geq 5\%$ , SNP calling rate per accession  $\geq 80\%$ , and *P* value of Hardy-Weinberg equilibrium test  $\leq 0.001$ . The resulting SNPs were used to perform GWAS for fruit skin color using a linear mixed model (LMM) implemented in the FaST-LMM program<sup>46</sup>. Raw *P* values were adjusted for multiple testing using the Benjamini-Hochberg procedure<sup>47</sup> and a significant association was based on a threshold of false discovery rate of 0.1 (corresponding to raw *P* value of  $3.03 \times 10^{-6}$ ).

**RNA-seq data generation and gene expression analysis.** Fruit samples were taken at five stages of fruit development from the apple cultivar "Greensleeves": 18 days after bloom (DAB) (active cell division), 37 DAB (the end of cell division), 67 DAB (rapid cell expansion), 90 DAB (rapid cell expansion), and 132 DAB (ripening). Three biological replicates were collected for each stage. Fresh tissues were immediately frozen in liquid nitrogen and ground to fine powder. Total RNA was extracted using the QIAGEN RNeasy Plant Mini Kit (Qiagen, USA) followed by treatment with RNase-free DNase I (Promega, USA) according to the manufacturers' protocols. Strand-specific Illumina RNA-Seq libraries were constructed following the protocol described in Zhong et al.<sup>48</sup> and sequenced on the Illumina HiSeq 2000 system using single-end 51-bp mode. Around 20–50 million reads were generated for each of the 15 libraries. Raw reads were first processed to remove adapter and low-quality sequences using Trimmomatic<sup>31</sup>. Reads shorter than 40 bp after trimming were discarded. The resulting reads were then aligned to the ribosomal RNA database<sup>49</sup> using Bowtie<sup>50</sup>, allowing up to three mismatches. The aligned reads were discarded and the remaining reads were aligned to the apple genome (v1.0) using HISAT<sup>51</sup>. Raw counts for each apple gene were derived from the read alignments and normalized to reads per kilobase of exon model per million mapped reads (RPKM). Differential expression analyses

were performed using edgeR<sup>52</sup>. Genes with adjusted *P* values less than 0.01, expression changes of at least twofold, and expression levels of at least 1 RPKM in at least one of the five stages were identified as differentially expressed genes during fruit development.

**Data availability.** Sequence reads of genome and transcriptome resequencing have been deposited into the NCBI sequence read archive under accessions SRP075497 and SRP102350, respectively. SNP data sets can be accessed at [ftp://bioinfo.bti.cornell.edu/pub/Apple\\_SNP/](ftp://bioinfo.bti.cornell.edu/pub/Apple_SNP/).

Received: 24 March 2017 Accepted: 20 June 2017

Published online: 15 August 2017

## References

- Gross, B. L. & Olsen, K. M. Genetic perspectives on crop domestication. *Trends Plant. Sci.* **15**, 529–537 (2010).
- Cao, K. et al. Comparative population genomics reveals the domestication history of the peach, *Prunus persica*, and human influences on perennial fruit crops. *Genome Biol.* **15**, 415 (2014).
- Hazzouri, K. M. et al. Whole genome re-sequencing of date palms yields insights into diversification of a fruit tree crop. *Nat. Commun.* **6**, 8824 (2015).
- Lin, T. et al. Genomic analyses provide insights into the history of tomato breeding. *Nat. Genet.* **46**, 1220–1226 (2014).
- Jiao, Y. et al. Genome-wide genetic changes during modern breeding of maize. *Nat. Genet.* **44**, 812–815 (2012).
- 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).
- Xu, X. et al. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotechnol.* **30**, 105–111 (2012).
- Cornille, A., Giraud, T., Smulders, M. J., Roldan-Ruiz, I. & Gladieux, P. The domestication and evolutionary ecology of apples. *Trends Genet.* **30**, 57–65 (2014).
- Cornille, A. et al. New insight into the history of domesticated apple: secondary contribution of the European wild apple to the genome of cultivated varieties. *PLoS Genet.* **8**, e1002703 (2012).
- Morgan, J. & Richards, A. *the New Book of Apples: The Definitive Guide to Over 2,000 Varieties* (Ebury Press, 2002)
- Pickersgill, B. Domestication of plants in the Americas: insights from mendelian and molecular genetics. *Ann. Bot.* **100**, 925–940 (2007).
- Velasco, R. et al. The genome of the domesticated apple (*Malus x domestica* Borkh.). *Nat. Genet.* **42**, 833–839 (2010).
- 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).
- Harris, S. A., Robinson, J. P. & Juniper, B. E. Genetic clues to the origin of the apple. *Trends Genet.* **18**, 426–430 (2002).
- Kawuki, R. S., Ferguson, M., Labuschagne, M., Herselman, L. & Kim, D.-J. Identification, characterisation and application of single nucleotide polymorphisms for diversity assessment in cassava (*Manihot esculenta* Crantz). *Mol. Breeding* **23**, 669–684 (2009).
- Migicovsky, Z. et al. Genome to phenome mapping in apple using historical data. *Plant Genome* **9**, doi:10.3835/plantgenome2015.11.0113 (2016).
- Foster, T. M., Celton, J.-M., Chagné, D., Tustin, D. S. & Gardiner, S. E. Two quantitative trait loci, *Dw1* and *Dw2*, are primarily responsible for rootstock-induced dwarfing in apple. *Hortic. Res.* **2**, 15001 (2015).
- Fazio, G. et al. *Dw2*, a new dwarfing locus in apple rootstocks and its relationship to induction of early bearing in apple scions. *J. Am. Soc. Hortic. Sci.* **139**, 87–98 (2014).
- Doebley, J. F., Gaut, B. S. & Smith, B. D. The molecular genetics of crop domestication. *Cell* **127**, 1309–1321 (2006).
- Devoghalare, F. et al. A genomics approach to understanding the role of auxin in apple (*Malus x domestica*) fruit size control. *BMC Plant Biol.* **12**, 1–15 (2012).
- Frary, A. et al. *fw2.2*: a quantitative trait locus key to the evolution of tomato fruit size. *Science* **289**, 85–88 (2000).
- Guo, M. et al. Cell Number Regulator1 affects plant and organ size in maize: implications for crop yield enhancement and heterosis. *Plant Cell* **22**, 1057–1073 (2010).
- Paniagua, C. et al. Antisense down-regulation of the strawberry  $\beta$ -galactosidase gene *Fa $\beta$ Gal4* increases cell wall galactose levels and reduces fruit softening. *J. Exp. Bot.* **67**, 619–631 (2016).
- Takano-Kai, N. et al. Evolutionary history of GS3, a gene conferring grain length in rice. *Genetics* **182**, 1323–1334 (2009).
- Yao, J.-L. et al. A *microRNA* allele that emerged prior to apple domestication may underlie fruit size evolution. *Plant J.* **84**, 417–427 (2015).
- Xia, R., Zhu, H., An, Y. Q., Beers, E. P. & Liu, Z. Apple miRNAs and tasiRNAs with novel regulatory networks. *Genome Biol.* **13**, R47 (2012).
- Brummell, D. A. Cell wall disassembly in ripening fruit. *Funct. Plant Biol.* **33**, 103–119 (2006).
- Ma, B. et al. Comparative assessment of sugar and malic acid composition in cultivated and wild apples. *Food Chem.* **172**, 86–91 (2015).
- Kunihisa, M. et al. Identification of QTLs for fruit quality traits in Japanese apples: QTLs for early ripening are tightly related to preharvest fruit drop. *Breed. Sci.* **64**, 240–251 (2014).
- Bai, Y. et al. A natural mutation-led truncation in one of the two aluminum-activated malate transporter-like genes at the *Ma* locus is associated with low fruit acidity in apple. *Mol. Genet. Genomics* **287**, 663–678 (2012).
- Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
- Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
- Li, H. et al. The sequence alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
- Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
- Tamura, K., Stecher, G., Peterson, D., Filipiński, A. & Kumar, S. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729 (2013).
- Wu, J. et al. The genome of the pear (*Pyrus bretschneideri* Rehd.). *Genome Res.* **23**, 396–408 (2013).
- Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
- Stajich, J. E. & Hahn, M. W. Disentangling the effects of demography and selection in human history. *Mol. Biol. Evol.* **22**, 63–73 (2005).
- Nei, M. Analysis of gene diversity in subdivided populations. *Proc. Natl Acad. Sci. USA* **70**, 3321–3323 (1973).
- Goudet, J. HIERFSTAT, a package for R to compute and test hierarchical *F*-statistics. *Mol. Ecol. Notes* **5**, 184–186 (2005).
- Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
- Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393–402 (2010).
- Hufford, M. B. et al. Comparative population genomics of maize domestication and improvement. *Nat. Genet.* **44**, 808–811 (2012).
- Xie, W. et al. Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. *Proc. Natl Acad. Sci. USA* **112**, E5411–E5419 (2015).
- Boyle, E. I. et al. GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20**, 3710–3715 (2004).
- Lippert, C. et al. FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8**, 833–835 (2011).
- Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
- Zhong, S. et al. High-throughput Illumina strand-specific RNA sequencing library preparation. *Cold Spring Harb. Protoc.* **2011**, 940–949 (2011).
- Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
- Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
- Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
- Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

## Acknowledgements

We thank Dr Eric Richards for critical reading of this manuscript. This research was supported by grants from the National Natural Science Foundation of China (31572091), the Special Fund for Agro-Scientific Research in the Public Interest of China (201303093), and the US National Science Foundation (IOS-0923312, IOS-1339287, and IOS-1539831).

## Author contributions

X.C., Z.F., Y.B., N.D., L.C., G.Z., and H.S. designed and managed the project. N.D., N.W., Y.M., C.J., S.W., Y.X., C.T.C., K.W., T.H., Z.Z., S.J., H.X., J.L., D.W., C.Q., Y.W., L.X., W.Z., C.L., D.Z., and Y.G. collected samples and performed biological experiments for DNA sequencing. N.D., Y.B., H.S., X.W., N.L., L.M., Z.M., X.S., X.C., S.W., S.F., C.Y.,

S.G., L.Y., and K.X. performed data analyses. Y.B. and N.D. wrote the manuscript. Z.F., L.C., G.-Y.Z., and K.X. revised the manuscript. All authors read and approved the manuscript.

### Additional information

**Supplementary Information** accompanies this paper at doi:[10.1038/s41467-017-00336-7](https://doi.org/10.1038/s41467-017-00336-7).

**Competing interests:** The authors declare no competing financial interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017