the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2020) **103,** 1420–1432 doi: 10.1111/tpj.14809

A draft genome of sweet cherry (*Prunus avium* L.) reveals genome-wide and local effects of domestication

Sara Pinosio[1,2]*, Fabio Marroni[2,3], Andrea Zuccolo[4], Nicola Vitulo[5], Stephanie Mariette[6], Gabriella Sonnante[7], Filippos A. Aravanopoulos[8], Ioannis Ganopoulos[9], Marino Palasciano[10], Michele Vidotto[2], Gabriele Magris[2,3], Amy Iezzoni[11], Giovanni G. Vendramin[1] and Michele Morgante[2,3]

[1]*Institute of Biosciences and Bioresources (IBBR), National Research Council, Via Madonna del Piano 10, Sesto Fiorentino 50019, Italy,*

[2]*Istituto di Genomica Applicata (IGA), Via Jacopo Linussio 51, Udine 33100, Italy,*

[3]*Dipartimento di Scienze Agro-alimentari Ambientali e Animali (DI4A), Università di Udine, via delle Scienze 206, Udine 33100, Italy,*

[4]*Institute of Life Sciences, Scuola Superiore Sant'Anna, Pisa 56124, Italy,*

[5]*Dipartimento di Biotecnologie, Università degli Studi di Verona, Strada Le Grazie 15, Verona 37134, Italy,*

[6]*BIOGECO, INRA, University of Bordeaux, route d'Arcachon 69, Cestas 33612, France,*

[7]*Institute of Biosciences and Bioresources (IBBR), National Research Council, via Amendola 165/A, Bari 70126, Italy,*

[8]*Faculty of Forestry and Natural Environment, Laboratory of Forest Genetics and Tree Breeding, Aristotle University of Thessaloniki, Thessaloníki 54124, Greece,*

[9]*Institute of Plant Breeding and Genetic Resources, Hellenic Agricultural Organization-DEMETER, Thermi 57001, Greece,*

[10]*Dipartimento di Scienze del Suolo, Università degli Studi di Bari Aldo Moro, della Pianta e degli Alimenti, Piazza Umberto I, Bari 70121, Italy, and*

[11]*Department of Horticulture, Michigan State University, 1066 Bogue Street, East Lansing, MI, 48824-1325, USA*

Received 12 February 2020; revised 24 April 2020; accepted 1 May 2020; published online 11 May 2020.
*For correspondence (e-mail sara.pinosio@ibbr.cnr.it).

SUMMARY

Sweet cherry (*Prunus avium* L.) trees are both economically important fruit crops but also important components of natural forest ecosystems in Europe, Asia and Africa. Wild and domesticated trees currently coexist in the same geographic areas with important questions arising on their historical relationships. Little is known about the effects of the domestication process on the evolution of the sweet cherry genome. We assembled and annotated the genome of the cultivated variety "Big Star*" and assessed the genetic diversity among 97 sweet cherry accessions representing three different stages in the domestication and breeding process (wild trees, landraces and modern varieties). The genetic diversity analysis revealed significant genome-wide losses of variation among the three stages and supports a clear distinction between wild and domesticated trees, with only limited gene flow being detected between wild trees and domesticated landraces. We identified 11 domestication sweeps and five breeding sweeps covering, respectively, 11.0 and 2.4 Mb of the *P. avium* genome. A considerable fraction of the domestication sweeps overlaps with those detected in the related species, *Prunus persica* (peach), indicating that artificial selection during domestication may have acted independently on the same regions and genes in the two species. We detected 104 candidate genes in sweep regions involved in different processes, such as the determination of fruit texture, the regulation of flowering and fruit ripening and the resistance to pathogens. The signatures of selection identified will enable future evolutionary studies and provide a valuable resource for genetic improvement and conservation programs in sweet cherry.

Keywords: genome structure and evolution, domestication, genetic diversity, genome assembly, *Prunus avium*.

INTRODUCTION

Domestication is an evolutionary process whereby human selection within animal and plant species determines phenotypic changes that differentiate domesticated taxa from their wild ancestors (Hancock, 2005). Cultivated plants are often the result of multiple rounds of domestication, in which an early domestication event produced landraces that are subsequently exposed to selection to obtain the improved modern varieties. During this process, population bottlenecks and selection of agronomically important traits determine both a global as well as localized progressive reduction of genetic variation in the domesticated forms with respect to the progenitor (Gaut *et al.*, 2018), which result from genetic drift and artificial positive selection respectively.

Sweet cherry (*Prunus avium* L.) is one of the most popular temperate fruit crops. It belongs to the genus *Prunus* together with other economically important fruit trees such as peach, apricot, plum and almond. Sweet cherry is the domesticated form of wild cherry that originated in the area between the Black and Caspian seas of Asia Minor. Birds may have carried it to Europe before human civilization but its cultivation was introduced in to Europe by the Romans (Zohary, 2012). Like most of the edible species, the domestication of sweet cherry started several thousand years ago with the cultivation and improvement of plants directly sampled from wild populations. Previous studies indicated that along the evolutionary history of the species, several domestication events might have happened and gene flow between cultivated and wild forms might have occurred (Mariette *et al.*, 2010). One of the most evident selected traits in sweet cherry is the increase in fleshy fruit size: wild, landraces and modern varieties usually exhibit fruit weights of 2, 6 and up to 14 g respectively (De Franceschi *et al.*, 2013). Other important traits that have been selected during domestication are related to fruit quality, such as fruit texture, color, flavor and ripening. Sweet cherry provides a unique system to understand the genomic consequences of domestication because, unlike in many other field and fruit crops, wild and domesticated forms still largely coexist in the same areas, with the wild forms still distributed over a wide geographic range, and because, unlike other crop species, where the domestication effects have been looked at in detail, it is currently largely vegetatively propagated. In addition, investigating the molecular genetics underlying domestication in cherry is essential to understand the molecular basis of fruit quality traits to inform new DNA-based breeding strategies.

To study the effects of domestication of cherry at a genome-wide scale, we set out to reconstruct the genome sequence of a modern cherry variety ("Big Star*") and to analyze genetic variation and differentiation in accessions exposed to different levels of artificial selection, such as wild cherry trees, landraces and the more recent modern varieties. Several studies have been performed with the aim of mapping quantitative trait loci (QTL) for fruit quality in *Prunus*. However, to our knowledge this is the first attempt to leverage genome-wide evolutionary signatures for the understanding of the genetic consequences of the domestication process and for the identification of candidate genes (CGs) involved in the domestication of sweet cherry.

RESULTS

The genome of *P. avium* variety "Big Star*"

In total, we obtained 350 million paired-end reads and 301 million mate-pair reads from the *P. avium* modern variety "Big Star*" corresponding to an estimated genome coverage of about 60× and 45× respectively (Table S1). Based on k-mer statistics, we estimated a genome size of 322 Mb, which is in agreement with the previously reported size of 338 Mb (Arumuganathan and Earle, 1991). We obtained a genome assembly composed of 21 878 contigs covering 203.4 Mb (Table 1). Mate pairs were employed to assemble contigs into 4240 scaffolds covering in total 272 Mb and corresponding to about 84% of the estimated genome size (Table 1). The longest scaffold was about 1.4 Mb in size and 50% of the assembled genome was represented by 389 scaffolds (N50), of which the shorter was 191 715 bp long (L50). Sweet cherry genetic maps revealed high conservation in genome structure between the cherry and peach (Guajardo *et al.*, 2015). Thus, using the *Prunus persica* v2.0 genome (Verde *et al.*, 2013; Verde *et al.*, 2017) as a guide, we anchored 2142 scaffolds, corresponding to 226.6 Mb (83.3% of the assembly), to the eight pseudomolecules. The structure of the resulting assembly was highly concordant with that recently published (Shirasawa *et al.*, 2017) for the Japanese cherry variety "Satonishiki" (Figure S1). Moreover, analysis of duplicated regions in the assembly revealed that *P. avium* and *P. persica* share a very similar duplication pattern (Figure S2), characterized by seven major triplicated subgenomic regions (Verde *et al.*, 2013), corresponding to seven paleosets of paralogs from the putative paleoancestor originally identified in grapevine (Jaillon *et al.*, 2007).

In total, we predicted 29 487 protein-coding genes and we assigned 93 850 Gene Ontology (GO) terms to 21 527 of them. The number of predicted genes and general gene metrics are very similar when compared with the closely related species *P. persica* (Table S2) and about 96% of the predicted genes have a *P. persica* homolog. Moreover, predicted genes cover 95.6% of the plant ortholog set of BUSCO (Simão *et al.*, 2015), suggesting a high completeness of the assembly and gene prediction (Table S3). In total, 36.7% of the genome consists of repetitive sequences, with known

Table 1 Statistics for the *Prunus avium* "Big Star*" genome assembly

Assembly statistics	
Number of contigs	21 878
Total length of contigs (bp)	203 410 967
Contig N50 length (bp)	21 521
Mean contig length (bp)	9298
Number of scaffolds	4240
Total length of scaffolds (bp)	272 047 324
Scaffold N50 length (bp)	191 715
Mean scaffold length (bp)	64 162
Longest scaffold length (bp)	1 381 054

transposable elements (TEs) accounting for 21.8% of it (Table S4). The total percentage of the genome composed by repeats is very similar to that reported for the related species *P. persica* (37.1%) (Verde *et al.*, 2013). Long terminal repeat retrotransposons represent almost 15% of the whole genome with Ty3-gypsy superfamily being the most represented (8.99%), followed by Ty1-copia (5.66%). DNA transposons cover 5.7% of the assembly, while an additional 15% of the assembly is composed of unclassified repeated sequences related to, in total, 2779 highly divergent and heterogeneous elements.

The *Prunus avium* genome went through a massive satellite DNA expansion

The GC content of "Big Star*" short reads showed a bimodal distribution with one peak at 37% and a second, more pronounced, at 43% (Figure 1a, maroon line). However, by analyzing the GC content on a dataset of short reads simulated from the "Big Star*" assembly, we obtained a normal distribution characterized by a single peak at 37% (Figure 1a, pink line). The same analysis performed on real and simulated reads of the *P. persica* reference sample Lovell resulted in two highly similar distributions with a single peak at about 38% (Figure 1a), which is consistent with the *P. persica* GC content previously reported (Verde *et al.*, 2013). The inconsistency between the GC content distributions in real and simulated *P. avium* reads suggested that the assembly lacks a genomic fraction characterized by a higher GC content than the genome average. Considering that satellite DNA (satDNA) is often characterized by a different GC content with respect to the rest of the genome, we searched the *P. avium* real and simulated reads for the presence of *P. avium* specific satellite sequences. We found a considerably higher fraction of "Big Star*" reads showing similarity with satDNA (13.8%) if compared with simulated ones (0.6%), suggesting that only a small fraction of the copies of this satDNA have been included in the assembly. Thus, the majority of *P. avium* sequence missing in the *de novo* assembly is composed of satDNA, which is very difficult to assemble with short reads due to its tandemly arranged and repetitive structure. This is in accordance with the high degree of completeness of the assembly despite being about 50 Mb shorter than the expected size. We extended the analysis to a larger dataset of *P. avium* and *P. persica* samples and we observed that *P. avium* samples are characterized by a significantly higher satDNA content with respect to *P. persica* ones (Figure 1b).

Genome-wide effects of domestication on *Prunus avium* genetic diversity

To study the effects of cherry domestication on genetic variation, we resequenced at approximately 30× coverage 22 *P. avium* plants belonging to three distinct categories based on their domestication history: eight modern accessions, six landraces and eight individuals sampled from wild populations (Table S5). In total, we identified 2 037 892 of single nucleotide polymorphisms (SNPs) and 282 290 small INDELs. The average SNP frequency was one per 293 bp, with 197 025 (9.6%) located within the coding sequence in the predicted gene models (Table S6). Wild accessions had a large number of private SNPs compared with modern accessions and landraces, as expected if a major bottleneck was experienced during domestication. The identification of 722 581 SNPs (35.5% of the total) that were exclusive to the wild individuals (Figure 2a) suggests that wild populations of *P. avium* are a very rich source of genetic variability that could be used for the genetic improvement of the cultivated varieties. Another distinctive feature observed in wild individuals in comparison with domesticated ones was a higher proportion of the genome in the heterozygous state (Figures S3–S5). Using SNP information, we explored the genetic structure of the resequenced samples with the principal coordinates analysis. Modern varieties and landraces were not separated by the two first coordinates (explaining about 25% of the total variance) but formed a unique heterogeneous group (Figure 2b). On the other hand, wild samples were far apart from both modern varieties and landraces and were separated in four groups reflecting their geographical origin. Wild individuals sampled in France (5-436 and 9-465) are the most closely related to the modern/landrace group and, in fact, admixture analysis revealed gene flow from these samples to landraces (Figure S6). On the other hand, wild samples collected in Georgia (Douc_4-2, Kwar_13-2, Maia_1-3) and Greece (Hayntou1, Katafito1) formed two well-distinct groups separated on both coordinates.

To obtain a more accurate estimation of population genetics parameters, we resequenced at low-coverage 74 additional *P. avium* accessions (25 modern varieties, 24 landraces and 25 wild individuals) (Table S7). The average nucleotide diversity (π) was significantly higher in wild plants (3.1×10^{-3}) than in both modern (2.0×10^{-3}) and landrace ones (2.2×10^{-3}) (Figure 3a), suggesting the

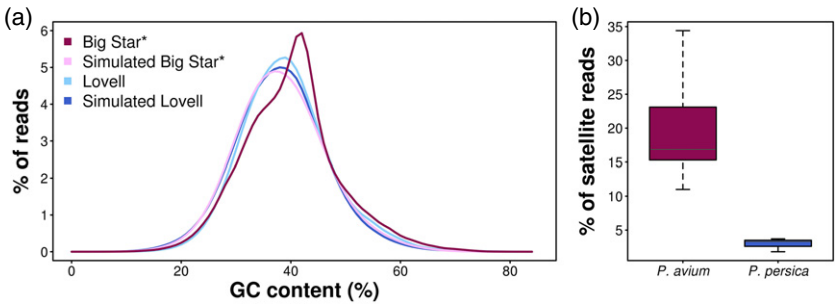

Figure 1. Analysis of satellite DNA content.
(a) GC content distribution in *P. avium* and *P. persica* real and simulated reads.
(b) Distribution of the percentage of satellite reads calculated in *P. avium* and *P. persica* samples. One-sided Wilcoxon rank-sum test P = 6.5e-07.

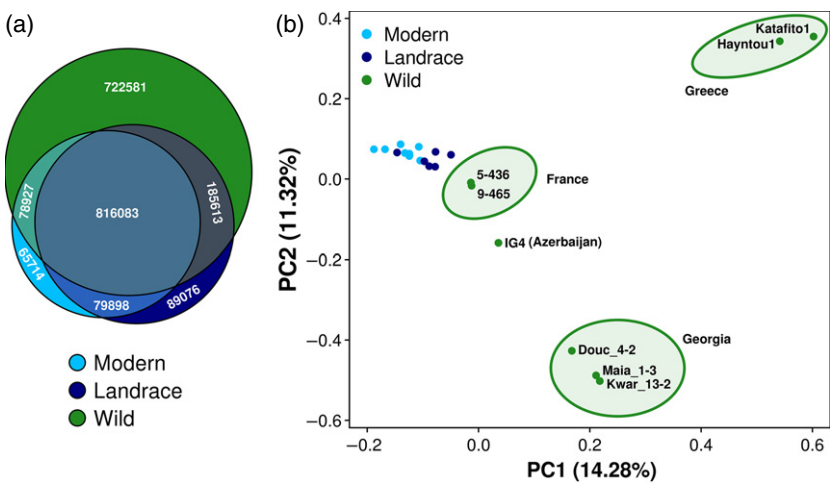

Figure 2. (a) Venn diagram showing the number of single nucleotide polymorphisms (SNPs) shared or not shared between the three domestication categories. Numbers of SNPs shared are indicated at the intersections of the circles in the Venn diagram.
(b) Principal components analysis of the SNP data. Percentage of variance explained by the two first components is reported in brackets.

occurrence of a severe bottleneck during early domestication and a minor one during later stages of breeding. The presence of two successive bottlenecks is more evident when looking at genome-wide data for Tajima's *D* that was higher in modern varieties (0.49), intermediate in landraces (0.16) and lower and close to zero (−0.03) in wild plants (Figure 3b). As expected, based on the severity of the successive bottlenecks, genome-wide linkage disequilibrium (LD) was higher in domesticated samples compared with wild ones (Figure 3c), and high genetic differentiation (F_{ST}) between wild and domesticated cherries was observed (Figure 3d). The same pattern of nucleotide diversity and Tajima's *D* was observed at the individual chromosome level, with some notable exceptions, such as chromosome 3, in which the Tajima's *D* values are very similar in the three categories (Figure S7 and Table S8). The distribution of LD along the single chromosomes (Figure S8) also reflects the genome-wide pattern, with modern accessions and landraces generally having higher LD than wild accessions.

In addition to SNPs, we also analyzed large (>1 kb) structural variants (SVs). We detected 1269 deletions and 8223 insertions with respect to the reference sequence, covering 2.9 and 19.9 Mb of genomic sequence respectively. The number of SVs detected in each sample is in line with what was observed for SNPs: it ranged from 2309 in the modern variety "Summit" to 3331 in the wild sample Hayntou1 (Table S5). Many insertions (77.6%) appear to be due to the movement of class I retroelements (Table S9) and, in agreement to what was observed in other plant species such as poplar (Pinosio *et al.*, 2016) and melon (Sanseverino *et al.*, 2015), long terminal repeat Gypsy were the most active elements.

We searched the set of identified SNPs for variants likely to have functional consequences and we discovered in 3211 genes that 4069 SNPs and 405 small INDELs introduced premature stop codons, extended the open reading frame or affected splice sites of predicted genes. As expected, most of the deleterious SNPs were detected at low frequency and showed a negative Tajima's *D* (Figure 4). In total, 24 276 non-synonymous SNPs affecting 11 217 different genes were predicted *in silico* to have a deleterious effect. Considering that homozygous variants probably have a functional effect, we analyzed the frequency of homozygous non-synonymous and deleterious mutations in the different domestication categories and we

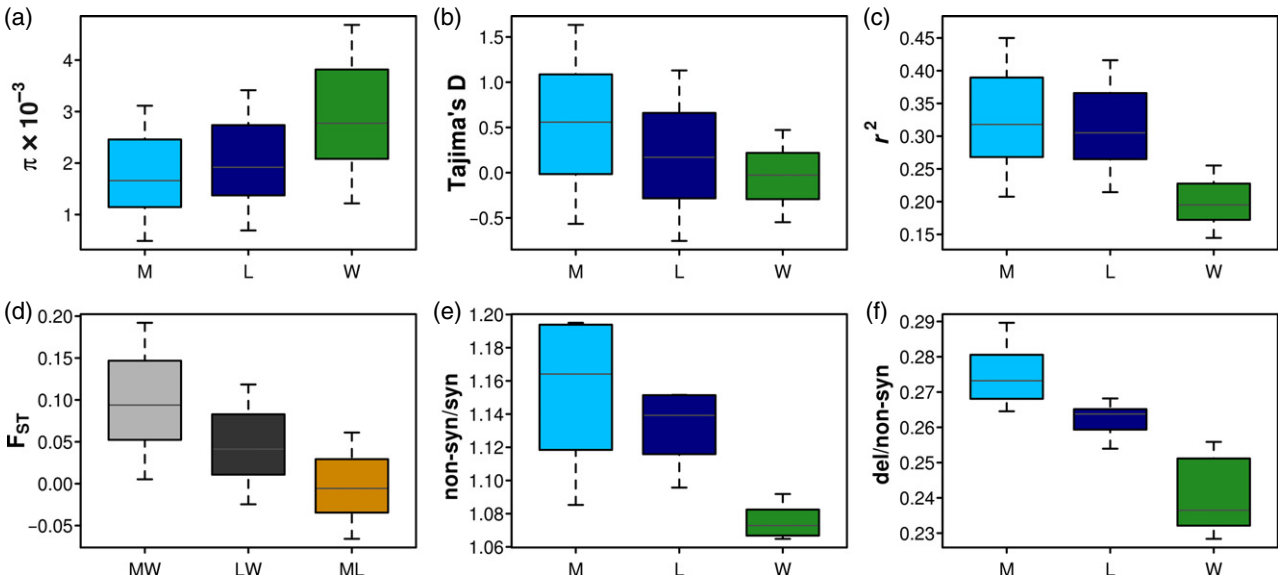

Figure 3. (a) Distribution of nucleotide diversity (π), (b) Tajima's *D* and (c) linkage disequilibrium (r^2) measured at a genome-wide level in the three domestication categories (M, modern; L, landrace and W, wild).
(d) Distribution of the population divergence (F_{ST}) measured by comparing modern varieties with wild plants (MW), landraces with wild plants (LW) and modern varieties with landraces.
(e) Distribution of the number of homozygous non-synonymous mutations divided by the number of homozygous synonymous mutations, and (f) distribution of the number of homozygous non-synonymous deleterious mutations divided by the number of homozygous neutral non-synonymous mutations in the three domestication categories. Six measures were significantly different when comparing both modern varieties and landraces to wild individuals (one-sided Wilcoxon rank-sum test domesticated versus wild: *P*-value < 0.01).

observed an enrichment of both types of variants in modern varieties and in landraces when compared with wild plants (Figure 3e,f).

Local signatures of domestication in *Prunus avium*

We exploited the reduction of diversity (ROD) signature and a cross-population composite likelihood ratio test (XP-CLR) (Chen *et al.*, 2010) to identify selection events that took place in relation to domestication (i.e. "domestication" sweeps) and those related to the modern breeding efforts (i.e. "breeding" sweeps). In total, we identified 11 putative domestication sweeps, encompassing 11.0 Mb of the *P. avium* genome, and five breeding sweeps, covering 2.4 Mb. Putative sweeps are highlighted in Figure 5 and listed in Tables S10 and S11. Notably, we observed that three breeding sweeps overlapped with domestication sweeps, suggesting that a subset of genomic regions subjected to selection during domestication may have undergone a successive round of artificial selection during recent breeding programs. XP-CLR scores were generally lower in domestication sweeps with respect to breeding sweeps (Figure S9). Three domestication sweeps stand out for their large size and strong signal ($\pi_{wild}/\pi_{domesticated} > 4$): one is located at the beginning of chromosome 1, one in the second arm of chromosome 2 and the third in the middle of chromosome 4. These sweeps are all confirmed by high levels of population divergence between domesticated and wild samples (Figure S10). The domestication sweep on chromosome 2 is the one with the strongest signal ($\pi_{wild}/\pi_{domesticated} = 4.9$). In this region, we also observed an increase of LD in landraces and a less pronounced increase of LD in modern varieties while LD levels in wild plants are comparable with the rest of the genome (Figure S8). Further to the domestication sweep located on chromosome 4, we observed a general increase of LD levels and negative values of Tajima's *D* in both modern varieties and landraces (Figures S7 and S8), as expected in case of positive selection for one allele affecting a desired trait. The breeding sweep with the strongest signal ($\pi_{landrace}/\pi_{modern} = 2.4$) is located at the beginning of chromosome 2 and overlaps with a domestication sweep.

We searched for functional CGs in the putative domestication sweeps, by looking for genes whose function may be modified by a mutation specifically detected only in wild or domesticated plants (Table S12). Of the 34 functional CGs identified, the majority (32) were affected by deleterious mutations detected only in wild plants and thus classified as gain-of-function mutations in domesticated ones. In fact, in the present study, genes affected by loss-of-function mutations are difficult to detect because our reference genome was built from a modern variety that underwent selection and thus genes inactivated by loss-of-function mutations may not have been predicted. In addition, using the identity by descent information, we

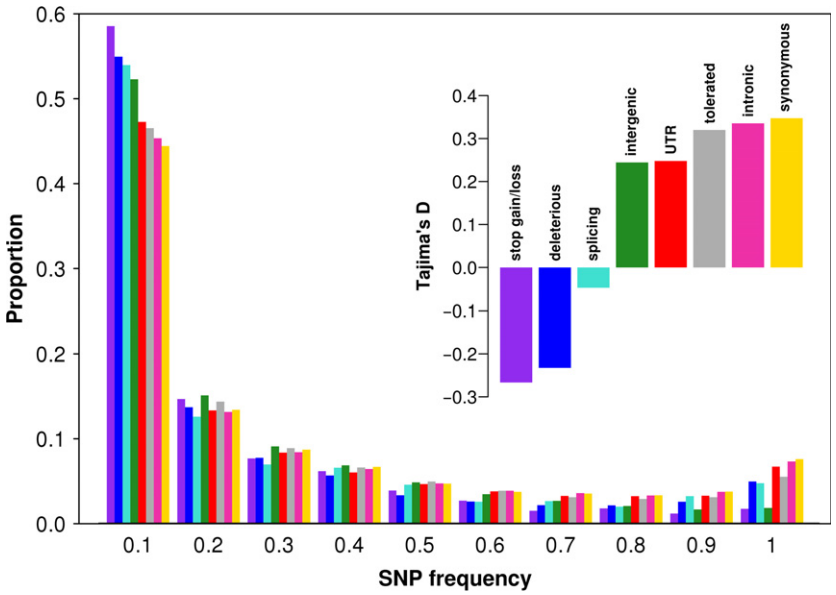

Figure 4. Allele frequency spectrum. Single nucleotide polymorphism (SNP)-derived allele frequency and Tajima's *D* (inner plot) calculated at non-synonymous loci (stop gain/loss, deleterious, splicing), synonymous loci (synonymous), untranslated regions (UTR), intergenic regions (intergenic), introns (intron).

detected 33 tag CGs (Table S13), of which 26 were linked to gain-of-function tag SNPs. Moreover, we identified 41 genes that have been reported as CGs of domestication in previous studies and thus were selected as literature CGs (Table S14).

DISCUSSION

With the aim of identifying genome-wide signatures of domestication in sweet cherry, we report here the *de novo* assembly of the *P. avium* cultivar "Big Star*" and an extensive analysis of genetic diversity across wild plants, landraces and modern accessions. Analysis of the assembly revealed that the difference in genome size between *P. avium* (338 Mbp) and *P. persica* (265 Mb) could be largely due to a massive amplification of satDNA that occurred in *P. avium*. This hypothesis is also supported by the fact that we observed high consistency between both gene metrics and TE-related repeats content between the two species. Genome-size differences between related species due to differences in satDNA content have already been reported for other plant species (Ambrožová *et al.*, 2011; Emadzade *et al.*, 2014). The small fraction of reconstructed satDNA in the "Big Star*" genome was often localized in centromeric regions of the assembly (Figure S11). In plants, centromere expansion due to satellite amplification mediated by segmental duplication has already been reported (Ma and Jackson, 2006). The functional role of satDNA expansion is not completely understood; however, its relevant structural role in vital functions, such as segregation or preservation of the genetic material is recently emerging (Garrido-Ramos, 2017).

Genome-wide analysis of the genetic diversity showed that the domestication of sweet cherry was accompanied by two distinct and successive bottlenecks that led to a decrease in nucleotide diversity across the whole genome, a severe bottleneck during early domestication leading to the so-called landraces and a minor one in later stages of breeding, leading to the currently commercialized modern cultivars. A similar pattern has been recently reported for the related species *P. persica* where a strong loss of genetic diversity was observed during the domestication process from wild progenitors ($\pi = 3.5 \times 10^{-3}$) to landraces ($\pi = 1.2 \times 10^{-3}$) and a weaker loss of genetic variation was observed during peach improvement from landraces to modern cultivars ($\pi = 1.0 \times 10^{-3}$) (Li *et al.*, 2019). However, in cherry the loss of variation in domesticated samples was less marked compared with peach, where about two-thirds of the genetic diversity have been lost (Li *et al.*, 2019) suggesting the bottleneck that occurred in peach was stronger. The presence of two successive bottlenecks is confirmed by the analysis of genome-wide Tajima's *D* values: a highly positive Tajima's D in modern varieties is in accordance with a very recent bottleneck, a lower but still positive D value for landraces is in accordance with a less recent bottleneck (partial recovery with population expansion) while the wild samples appear at equilibrium. LD levels increased in domesticated samples in comparison with wild ones, presumably because of the successive bottlenecks. The observed levels of nucleotide diversity, Tajima's *D* and LD can be explained by a decrease in population size occurring during domestication (Tajima, 1989; Arunyawat *et al.*, 2012).

The analysis of SVs, which in plants are largely due to TE movement (Lisch, 2013), revealed that, as already observed in maize (Vitte *et al.*, 2014) and in grape (Zhou *et al.*, 2019), domestication and breeding bottlenecks do not appear to have coincided with an increase in

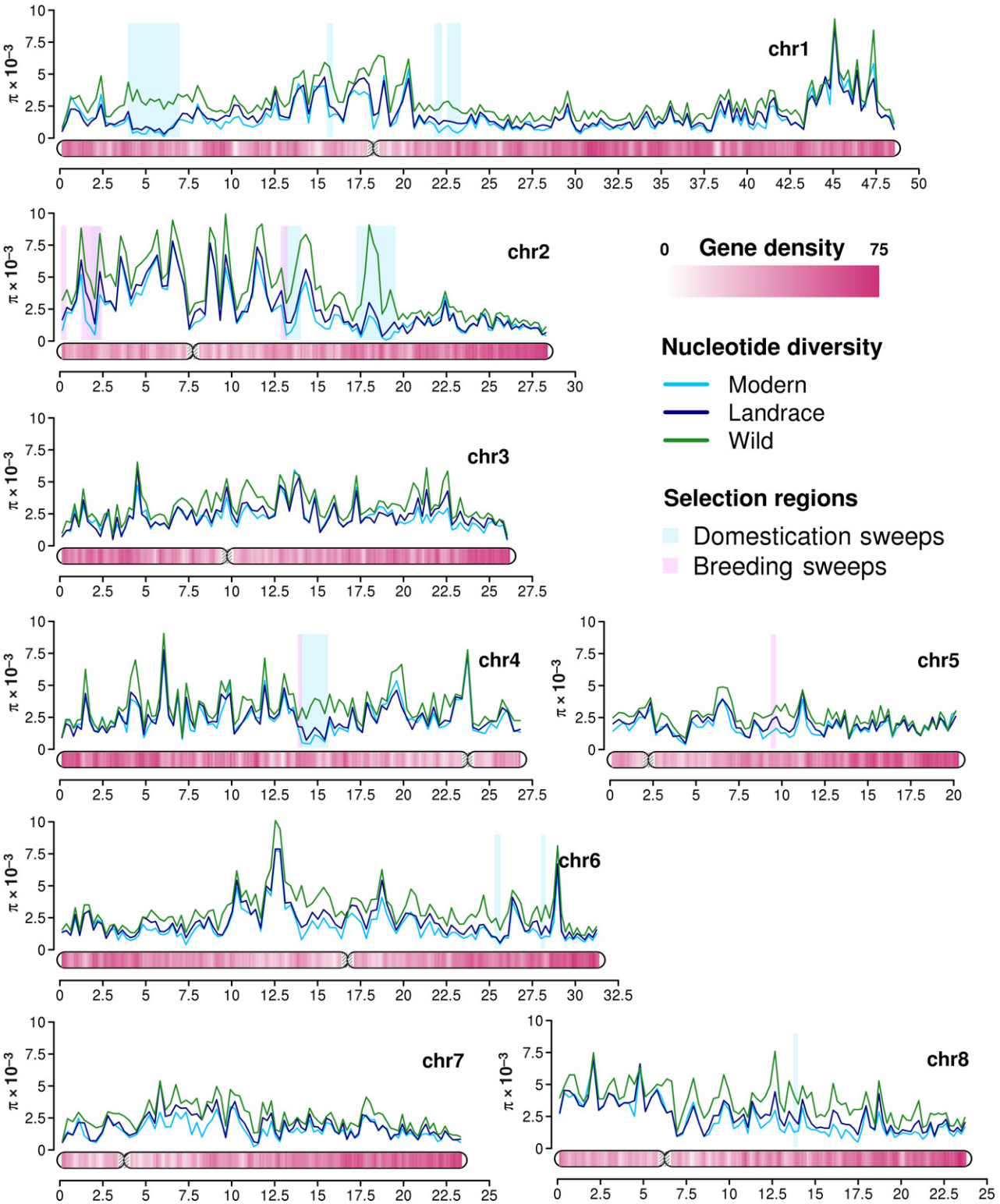

Figure 5. Nucleotide diversity distribution along the eight cherry pseudomolecules in three domestication categories. Domestication sweeps are highlighted in light blue, while breeding sweeps are highlighted in pink.

transposon activity given the progressive decrease in SV number as we go from wild to landrace and to modern accessions. This is in contrast with other examples, where recent amplification of TEs following domestication has been observed (Naito *et al.*, 2006).

An expected additional consequence of the bottlenecks experienced during the domestication process is an enrichment in deleterious mutations in domesticated with respect to wild plants, as already reported in sunflower (Renaut and Rieseberg, 2015) and grape (Zhou *et al.*, 2017). In natural populations, deleterious mutations are purged by negative selection and are expected to be held at low frequencies. During domestication, this mechanism is impaired and deleterious mutations may rise in frequency. Although domesticated samples showed a general reduction of genetic variability due to artificial selection, the observed enrichment of homozygous non-synonymous and more specifically deleterious mutations in modern varieties and in landraces when compared with wild plants could be the result of relaxed selection constraints during domestication.

Artificial selection for desired traits is supposed to lead to an extreme reduction of genetic diversity in specific regions because of selective sweeps around the selected loci (Gaut *et al.*, 2018). The size of the affected region and the extent of diversity reduction may depend on the mode of reproduction of the species, on the number of sexual generations that have occurred after the selection events and on the mode of action of the selected allele (recessive versus dominant). Selective sweeps can alter the allele frequency spectra of SNPs located in the proximity of selected alleles and cause a distorted pattern of genetic variation (Chen *et al.*, 2010).

We detected 11 putative domestication sweeps and five breeding sweeps, encompassing, in total, 13.4 Mbp, which may be due to selection for domestication and breeding-related traits respectively. The levels of nucleotide diversity estimated along the genome in modern varieties were highly concordant with those obtained in seven recently sequenced Japanese varieties (Shirasawa *et al.*, 2017) (Figure S12), suggesting that most of the regions under selection are shared between the two datasets. By comparing the putative sweeps with those reported in a recent work in the related species, *P. persica* (Li *et al.*, 2019), we observed that six domestication sweeps are overlapping with those identified in peach, suggesting that genomics regions under selection are partially shared between the two species (Tables S10 and S11) and indicating that artificial selection during domestication may have acted independently on the same regions and genes in the two species. The most notable peaks of loss of diversity based on size and strength of signal were detected in chromosomes 1, 2 and 4. These regions are enriched in genes affecting traits of agricultural interest such as fruit size, fruit firmness and flowering time. The region of chromosome 2, which has the strongest signal, is a well-known QTL hotspot of *P. avium* (Cai *et al.*, 2017), which contains fruit size, fruit firmness and fruit sweetness QTLs (Zhang *et al.*, 2010; Rosyara *et al.*, 2013; Campoy *et al.*, 2015; Cai *et al.*, 2017), as well as a QTLs for flowering time (Castède *et al.*, 2014) and a candidate fruit size gene associated with domestication (De Franceschi *et al.*, 2013). The region on chromosome 4 has previously been identified as a QTL for fruit firmness and an expansin gene has been proposed as the most promising CG of the region (Cai *et al.*, 2019). In addition, in the corresponding region of the peach genome, a previous work reported the presence of two distinct domestication sweeps and a QTL associated with fruit size (Li *et al.*, 2019). In the region on chromosome 2 where the strongest breeding sweep was detected, a domestication sweep was also observed and we observed a progressive reduction of the nucleotide diversity from wild plants to modern varieties. In addition, modern varieties exhibit a negative Tajima's *D* in this region, confirming the presence of a recent selective sweep due to artificial selection.

Among the CGs we identified based on different criteria, we found the most interesting candidates among those selected as literature CGs (Table S14). Among them, PAV_9578g00030 is of particular interest because it has already been reported as CG involved in fruit ripening, and plant growth and development in three different studies (Pirona *et al.*, 2013; Alkio *et al.*, 2014; Cai *et al.*, 2019), and the orthologous peach gene (ppa007577m) is located within a major QTL controlling maturity date (Pirona *et al.*, 2013). It codes for a NAC domain-containing protein. In plants, NACs form one of the largest families of transcription factors playing vital roles in plant growth regulation and development processes, including abiotic stress responses (Shao *et al.*, 2015). This gene is located in the large domestication sweep detected on chromosome 4, just in correspondence to a negative Tajima's *D* peak. Interestingly, this gene showed very high levels of expression in the fruit tissue of the modern variety "Big Star*" (Table S14). Another literature CG (PAV_7792g00050) is located at the beginning of the large domestication sweep located on chromosome 1. This gene encodes for a pectate lyase involved in cell wall modification processes and in the determination of fruit texture (Ogundiwin *et al.*, 2009; Alkio *et al.*, 2014). In the large domestication sweep on chromosome 2, we detected five different literature CGs putatively involved in flowering and cell wall modification processes. In breeding sweeps, we detected six literature CGs, five of which overlap with those detected in domestication sweeps. One of them (PAV_9514g00020) is located in the largest breeding sweep (chr2:1 255 146–2 454 826) and encodes for carotenoid cleavage dioxygenases involved in the determination of fruit texture (Ogundiwin *et al.*, 2009).

The CGs we identified by leveraging the novel information on gene sequence and function obtained in the present study and information retrieved in public repositories represent likely targets of the domestication process. They deserve further investigation in light of a possible genetic improvement program for sweet cherry.

EXPERIMENTAL PROCEDURES

Study samples

The sweet cherry variety "Big Star*" was selected for *de novo* genome assembly and gene annotation. This variety was obtained by self-crossing of the variety "Lapins" under a broader *P. avium* breeding program (Sansavini *et al.*, 1998) and is therefore expected to exhibit high levels of homozygosity. The analysis of *P. avium* diversity was performed, in total, on 98 *P. avium* individuals belonging to three distinct categories based on their domestication history: modern varieties, landraces and wild plants. In total, 22 samples (eight modern, six landraces and eight wild accessions) were sequenced at high coverage (>20×) and underwent extensive investigation; here they will be referred to as "high-coverage" in subsequent sections (Table S5). Half of the high-coverage wild samples were collected in the Caucasus, i.e. the center of origin of the species. The remaining 74 samples (25 modern 24 landraces and 25 wild) were sequenced at lower coverage (approximately 3–4×) and will be referred to as "low-coverage" (Table S7). Low-coverage samples belonging to each of the three categories were analyzed together as a single pool to estimate nucleotide diversity and Tajima's *D*.

DNA extraction, library preparation and sequencing

For DNA sequencing, leaves were ground in liquid nitrogen and high-molecular-weight genomic DNA was extracted from nuclei as previously described (Zhang *et al.*, 1995). The protocol was improved with the addition of polyvinylpyrrolidone both in the wash (5%) and in the lysis (2%) buffers. For gene prediction, total RNA was extracted from four different tissues of the variety "Big Star*" (flowers, leaves, fruit and wood) using the Spectrum Plant Total RNA kit (Sigma, St. Louis, MO, USA) following the manufacturer's protocol.

For *de novo* genome assembly, six different Illumina libraries were generated: three standard paired-end libraries with different read lengths and insert sizes, one overlapping paired-end library 2× 300 bp with a mean insert size of about 350 bp and two different mate-pair libraries with different read lengths and insert sizes (Table S1). DNA libraries of "Big Star*" and high-coverage samples were prepared using Illumina reagents, according to the manufacturer's specifications (Illumina Inc., San Diego, CA, USA). DNA libraries of low-coverage samples were generated using the ThruPLEX® DNA-seq Kit, according to the manufacturer's specifications (Rubicon Genomics, Ann Arbor, MI, USA). RNA-seq libraries were generated using the TruSeq RNA-Seq Sample Prep kit according to the manufacturer's protocol (Illumina Inc.). Sequencing was performed at the Istituto di Genomica Applicata (IGA, Udine, Italy) facilities using either HiSeq2000 or MiSeq platforms from Illumina Inc. Images from the instruments were processed using the manufacturer's pipeline software to generate FASTQ sequence files. FASTQ files are available at the SRA (Sequence Read Archive) database under the accession no. PRJNA419491.

Genome *de novo* assembly and construction of chromosome pseudomolecules

Adaptor sequences and low-quality 3′ ends were removed from DNA short reads using, respectively, CUTADAPT (Martin, 2011) and ERNE-FILTER (Del Fabbro *et al.*, 2013). After trimming, only pairs with both reads longer than 50 bp were retained. The genome was assembled using ALLPATHS-LG (version 50900), a *de novo* assembler based on a de Bruijn graph algorithm (Gnerre *et al.*, 2011), using the following parameters: HAPLOIDIFY = True, CLOSE_UNIPATH_GAPS = False. The k-mer Analysis Toolkit v1.0 (Mapleson *et al.*, 2016) was employed to assess the completeness of the *P. avium* assembly by comparing the k-mer spectra of the assembly with that obtained from the short reads. k-mer hash files were created with jellyfish (Marçais and Kingsford, 2011), using a *k* value of 19 and used as input for the KAT *comp* utility (see Data S1 for details). In addition, completeness of the assembly was assessed using BUSCO v3.0.2 (Simão *et al.*, 2015). BUSCO was run on a set of 1440 single-copy ortholog genes included in the "Plants set" with default parameters.

Duplicated regions in the cherry genome and orthologous regions among species were analyzed using DAGCHAINER (Haas *et al.*, 2004) with default parameters. Comparative plots were generated using CIRCOS (Krzywinski *et al.*, 2009). Pseudomolecules, as *bona fide* reconstruction of chromosome sequences for *P. avium*, were obtained using the *P. persica* v2.0 reference genome (Verde *et al.*, 2013; Verde *et al.*, 2017) as the synteny proxy. To this scope, an *ad-hoc* pipeline was developed to find placement of *P. avium* scaffolds on *P. persica* genomic coordinates. A BAM file of contig alignments was produced with DENOM software (http://mtweb.cs.ucl.ac.uk/mus/www/19genomes/IMR-DENOM/#DENOM). Contiguous blocks of scaffold alignment were retrieved by parsing contig alignments along with the scaffolding table using a custom python script (see Data S1 for details). The genome sequence has been submitted to the NCBI Genome database (PRJNA419491).

Gene prediction and annotation

Gene prediction was performed using different approaches including RNA-seq data, the alignment at both protein and nucleotide level and *de novo* prediction. In total, four RNA-seq libraries from different tissues (flowers, leaves, fruit and wood) sequenced for the present study, 24 Illumina libraries retrieved from the Short Read Archive (SRP011083), and four Roche (https://www.roche.com/) 454 libraries (SRP020000) were used for gene prediction. Illumina RNA-seq reads were aligned on the masked reference genome using TOPHAT (Trapnell *et al.*, 2009) v2.0.6 setting the option *-b2-sensitive*. Genome-guided transcript reconstruction was performed using CUFFLINKS v2.2.1 (Trapnell *et al.*, 2010) with default options. Transcripts were reconstructed independently for each sample and then merged into a reference transcriptome using Cuffmerge from the CUFFLINKS software suite. The reconstructed transcripts and the 454 sequences were further assembled using PASA (Haas *et al.*, 2008), an eukaryotic genome annotation tool that exploits spliced alignments of expressed transcript sequences to model gene structures automatically. Nucleotide and protein sequences belonging to the genus *Prunus* were downloaded from NCBI and aligned to the reference genome using EXONERATE (https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate). Highly stringent criteria were used to filter the alignments: 30% identity and 70% alignment coverage at protein level, and 50% identity and 70% alignment coverage at nucleotide level. The *Ab-initio* gene prediction was performed using three different programs: AUGUSTUS (Stanke *et al.*, 2006), SNAP

(Korf, 2004) and GENEID (Parra *et al.*, 2000) using Arabidopsis thaliana as the training gene model. Previously collected evidence was combined using EVIDENCEMODELER (Haas *et al.*, 2008) to obtain the final gene model. To remove false positives and low-quality predictions, genes coding for proteins <150 amino acids were aligned against the non-redundant database using an *E*-value threshold of 10^{-6}. Proteins for which no hit was obtained were discarded. Finally, the gene models were processed with PASA to predict alternative spliced isoforms and to add the untranslated regions.

The functional annotation of *P. avium* predicted genes was performed using a similarity search against the non-redundant protein database (NCBI) using Blastp (Altschul *et al.*, 1997) using an *E*-value threshold of 10^{-6}. INTERPROSCAN5 (Jones *et al.*, 2014) was employed to search against different databases (PROSITE patterns, PRINTS, PFAM, PRODOM, SMART, TIGRFAM and PANTHER) to identify conserved protein domains and functional annotation. Gene Ontology and KEGG annotations were performed using BLAST2GO 2.6.0 (Conesa and Götz, 2008).

Repeat content in *Prunus avium* genome

REPEATSCOUT (Price *et al.*, 2005) was used to perform *de-novo* identification of repeats in the *P. avium* assembly. A threshold was set to remove any repeats present in <20 copies. The putative repeats were then characterized using the following analysis steps.

i Sequences were searched using tBlastX (Altschul *et al.*, 1997) against the RepBase database (Bao *et al.*, 2015) with an *E*-value significance threshold of 10^{-6}.
ii Sequences that did not provide significant hits in the previous step were searched using tBlastX against the *nr* division of the GenBank database (Clark *et al.*, 2016). This step allowed to characterize as TEs additional sequences and to identify plastidial contaminants as well as sequences sharing similarity with coding regions not related to TE. These were removed as they were considered members of gene families and/or pseudogenes.
iii The remaining sequences were analyzed using the software TANDEM REPEATS FINDER (Benson, 1999) to identify tandemly arranged satellite sequences.

The obtained *P. avium* library was concatenated with the PRUNUS_PERSICA_TEV2.0 one (Verde *et al.*, 2017) for, in total, 6391 sequences. Sequences <100 bp were considered scarcely informative for repeat searches aimed at evaluating the repeat abundance and were removed. The remaining 6245 putative repeats were clustered based on their similarity using the tool CD-HIT-EST (Fu *et al.*, 2012), collapsing all those sharing at least 80% similarity, and obtaining a library containing 4362 entries. To estimate the abundance of TE-related sequences in *P. avium* genome, the library was used to screen the assembly using the program REPEATMASKER (www.repeatmasker.org) with default parameters.

GC content and satellite DNA analysis

GC content statistics were computed on real and simulated short read datasets of *P. avium* and *P. persica* samples using FASTQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). The analysis on real *P. avium* data was performed on a subset of four million reads randomly selected from the different available libraries to mitigate possible library-specific biases. The analysis on real *P. persica* data was performed on a subset of four million reads randomly selected from a dataset of the peach sample "Lovell" (Clone PLov2-2N), retrieved from the Short Read Archive (SRX150254). The analysis on simulated data for both *P. avium* and *P. persica* was performed using a set of four million 100 bp long single-end reads simulated, respectively, from the "Big Star*" *de novo* assembly and the *P. persica* v2.0 reference genome (Verde *et al.*, 2017) using *wgsim* (https://github.com/lh3/wgsim) with default parameters.

To estimate the abundance of satDNA in *P. avium* and *P. persica* genomes, real and simulated reads of the two species were used as a query for a BLASTN (Altschul *et al.*, 1997) search against species-specific satellite sequences (accession nos AJ011795.1 and AF461123.1 for *P. avium* and accession no. X82222.1 for *P. persica*) using an *E*-value threshold of 10^{-6}. The two *P. avium* satDNAs have the same repeat unit length (166 bp) and a sequence identity of 91% and thus probably originated from the expansion of the same satellite element. To localize the satDNA in the *P. avium* genome, the two *P. avium* satellite sequences AJ011795.1 and AF461123.1 were used as a query for a BLASTN (Altschul *et al.*, 1997) search against the assembly using an *E*-value threshold of 10^{-6}. The dataset used for the estimation of satDNA in short reads was composed by the high-coverage *P. avium* samples, seven additional *P. avium* Japanese varieties (Shirasawa *et al.*, 2017) and 11 *P. persica* accessions (Verde *et al.*, 2013). The difference in satDNA content between *P. avium* and *P. persica* was tested using the non-parametric Wilcoxon rank-sum test.

Short read mapping and variants detection

Adaptor sequences and low-quality 3′ ends were removed from both DNA and RNA short reads using, respectively, CUTADAPT (Martin, 2011) and ERNE-FILTER (Del Fabbro *et al.*, 2013) with default parameters. After trimming, pairs with both reads >50 bp were aligned to the *P. avium* genome using the short read aligner BWA (Li and Durbin, 2009) with default parameters. After alignment, duplicated sequences were removed using the SAMTOOLS *rmdup* utility (Li *et al.*, 2009). The mean individual coverage for each individual was calculated by dividing the total number of uniquely aligned bases by the number of genomic positions covered by at least one read.

High-coverage samples were used to obtain a set of high-quality SNPs. SNP calling was performed using the software package GATK version 3.3-0 (McKenna *et al.*, 2010). The GATK utilities *RealignerTargetCreator* and *IndelRealigner* were used to define the intervals in the proximity of indels and to perform the local realignment of reads spanning small indels respectively. Then, *UnifiedGenotyper* was used to call SNPs in each resequenced sample. Nucleotide differences called by GATK were classified as SNPs if: (i) the position had a coverage ranging between 0.5 and 2.5 times the modal coverage of the sample, (ii) the Phred-scaled quality score was >50, (iii) the variant allele was present in at least 25% of the reads in each sample, and (iv) the SNP position was informative (i.e. had adequate coverage and quality) in at least half of the high-coverage samples.

Low-coverage samples were analyzed to improve frequency estimation of SNPs detected in high-coverage samples. To this aim, the alignment BAM files of the low-coverage samples belonging to each of the three domestication categories were joined using the SAMTOOLS *merge* utility (Li *et al.*, 2009) and polymorphic positions were called in each pool using the software package GATK (McKenna *et al.*, 2010). For each SNP detected in high-coverage samples, its frequency was calculated as a weighted average of the frequency obtained in high- and low-coverage samples for each of the three categories (modern, landrace and wild).

The detection of SVs was performed in high-coverage samples. The detection of deletions was performed by combining the results of the two SV detection tools DELLY (Rausch *et al.*, 2012) and GASV (Sindi *et al.*, 2009) as previously described (Pinosio

et al., 2016). Insertions of TEs were detected exploiting the information carried by read pairs spanning the insertion site as previously described (Pinosio *et al.*, 2016).

Functional annotation of mutations

To predict their effect on protein sequence, SNPs and small INDELs were functionally annotated using ANNOVAR (Wang *et al.*, 2010). SNPs were classified as "non-coding," "coding synonymous," "coding non-synonymous" and "nonsense" (determining a premature STOP codon or a STOP loss), while small INDELs were classified as "non-frameshift," "frameshift" and "nonsense." Non-synonymous SNPs were further categorized as "neutral" or "deleterious" using PROVEAN v.1.1.5 (Choi *et al.*, 2012). To perform this analysis accurately, the ancestral allele corresponding to non-synonymous SNPs was determined using peach as the outgroup (see Data S1 for details). A protein sequence homologs search was performed against the *Viridiplantae* section (Taxid: 33090) of the NCBI *nr* database. Non-synonymous SNPs having a PROVEAN score below −2.5 were classified as deleterious.

Population genetics statistics

Principal components analysis of the high-coverage samples was performed using the R package SNPRELATE (Zheng *et al.*, 2012). To compute genetic parameters on a whole-genome scale, polymorphic sites were analyzed in windows of variable size, containing 200 kb of sequence not annotated as repeats. Nucleotide diversity (π) was calculated in these windows for each of the three categories (wild, landrace and modern) and for all the domesticated plants together (i.e. landraces and modern varieties). For each window, π was computed as the expected heterozygosity per locus ($\hat{H}$) divided by the average number of genotyped positions in the window in high-coverage samples. $\hat{H}$ was calculated as follows

$$\hat{H} = \frac{n}{n-1}\left(1 - \sum_{i=1}^{k} p_i^2\right)$$

where n is the total number of chromosomes in the category, k is the number of alleles and p_i is the frequency of the i-th allele. Tajima's D and population divergence (F_{ST}) were calculated for each category using the *neutrality.stats* method included in the R package *PopGenome* (Pfeifer *et al.*, 2014) on the high-quality set of SNPs detected in high-coverage samples.

LD was computed as the average pairwise r^2 in 100 kb windows using the R function LD of the genetics package (http://cran.uvigo.es/web/packages/genetics/). Differences in π, Tajima's D, F_{ST} and r^2 between domesticated and wild plants were tested using the non-parametric Wilcoxon rank-sum test.

Identification of selection sweeps and domestication candidate genes

Genomic regions selected by domestication and breeding were identified using the combination of two approaches: the ROD signature and an XP-CLR (Chen *et al.*, 2010). To identify "domestication sweeps" in detail, the $\pi_{wild}/\pi_{domesticated}$ ratios were calculated on 200 kb sliding windows (step 25 kb) and the top 5% of the empirical distribution was selected. XP-CLR was run to compare domesticated and wild samples with the parameter "-w1 0.00520020001 -p0 0.95" and windows with the top 5% XP-CLR scores were selected. Finally, regions identified by ROD signature overlapping for at least 50% with those identified by XP-CLR were kept as selective sweep regions. The same approach has been applied to detect "breeding sweeps" by comparing modern varieties with landraces. We analyzed the gene content of the selection regions to identify CGs for domestication. The selection of candidates was performed using three different approaches described in detail in Data S1. Briefly, (i) "functional CGs" were obtained by searching for mutations that are specific to a domestication category and that could affect gene function, (ii) recessive or dominant "tag CGs" were selected in proximity to tag SNPs identified using identity by descent and genotypic information, and (iii) "literature CGs" were obtained by searching for genes linked to *Prunus* domestication in previous studies.

ACCESSION NUMBERS

Genome assembly and the raw sequence data have been deposited at NCBI BioProject under accession no. PRJNA419491. All data that support the findings of this study are also available from the corresponding author upon request.

ACKNOWLEDGEMENTS

Authors thank G. Cipriani (Università di Udine) for his helpful advice on the selection of the analysed varieties and F. Cattonaro (IGATech), I. Jurman (IGA), E. Di Centa (IGATech), N. Felice (Università di Udine) and V. Vendramin (IGATech) for their technical help and support for library preparation and sequencing. Further thanks go to D. Scaglione (IGATech) for technical advice on bioinformatic pipelines used for de novo assembly. This work was supported by the Ministero dell'Economia e delle Finanze under grant agreement B51J10001290001 (Conoscenze Integrate per sostenibilità e innovazione del MADE IN ITALY Agroalimentare) and by the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013) (grant agreement no. 294780, project Novabreed).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

SP, FM, GGV and MM planned and designed the research. SM, GS, FAA, IG and MP contributed to samples selection and collection. SP and MV contributed to sequencing and genome assembly. AZ performed repeat analyses. NV performed gene prediction and annotation. SP, FM and GM contributed to resequencing, diversity and population genetic analyses. SP and AI contribute to selective sweeps and CG identification. SP, FM, GGV and MM wrote the manuscript. All authors revised the manuscript.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Comparative genomics between "Big Star*" and "Satonishiki" genome assemblies.

Figure S2. Duplicated and triplicated regions in the cherry genome.

Figure S3. Distribution of heterozygous and homozygous regions in modern varieties.

Figure S4. Distribution of heterozygous and homozygous regions in landraces.

Figure S5. Distribution of heterozygous and homozygous in wild plants.

Figure S6. Maximum likelihood tree obtained for the high-coverage samples using TREEMIX.

Figure S7. Tajima's *D* distribution.

Figure S8. Linkage disequilibrium (LD) distribution.

Figure S9. Distribution of XP-CLR scores.

Figure S10. Population divergence (F_{ST}) distribution.

Figure S11. Position of satellite sequences in the eight pseudomolecules.

Figure S12. Nucleotide diversity distribution in the Japanese modern.

Figure S13. Assembly copy number plot.

Table S1. Summary of sequencing data generated for "Big Star*" genome assembly.

Table S2. Gene prediction summary statistics.

Table S3. Genome assembly completeness examined with BUSCO.

Table S4. Analysis of repetitive sequences within *P. avium* assembly.

Table S5. DNA libraries and coverage statistics of high-coverage samples.

Table S6. Total number of detected SNPs and small INDELs

Table S7. DNA libraries and coverage statistics of low-depth samples.

Table S8. Chromosome-level nucleotide diversity and Tajima's *D*.

Table S9. Transposable elements classification of insertions.

Table S10. Domestication sweeps.

Table S11. Breeding sweeps.

Table S12. List of functional CGs detected in domestication sweeps.

Table S13. List of tag CGs detected in domestication sweeps.

Table S14. List of literature CGs detected in domestication and/or breeding sweeps.

Table S15. Localization of cherry centromeres.

Table S16. Genotypic states used for IBSRH and genD calculation.

Table S17. Values of IBD required to identify putative tag SNPs.

Data S1. Supplementary methods.

REFERENCES

Alkio, M., Jonas, U., Declercq, M., Nocker, S.V. and Knoche, M. (2014) Transcriptional dynamics of the developing sweet cherry (*Prunus avium* L.) fruit: sequencing, annotation and expression profiling of exocarp-associated genes. *Hortic. Res.* **1**, 11.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.

Ambrožová, K., Mandáková, T., Bureš, P., Neumann, P., Leitch, I.J., Koblížková, A., Macas, J. and Lysak, M.A. (2011) Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of Fritillaria lilies. *Ann. Bot.* **107**, 255–268.

Arumuganathan, K. and Earle, E.D. (1991) Nuclear DNA content of some important plant species. *Plant Mol. Biol. Report.* **9**, 208–218.

Arunyawat, U., Capdeville, G., Decroocq, V. and Mariette, S. (2012) Linkage disequilibrium in French wild cherry germplasm and worldwide sweet cherry germplasm. *Tree Genet. Genomes*, **8**, 737–755.

Bao, W., Kojima, K.K. and Kohany, O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA*, **6**, 11.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580.

Cai, L., Quero-García, J., Barreneche, T., Dirlewanger, E., Saski, C. and Iezzoni, A. (2019) A fruit firmness QTL identified on linkage group 4 in sweet cherry (*Prunus avium* L.) is associated with domesticated and bred germplasm. *Sci. Rep.* **9**, 5008.

Cai, L., Voorrips, R.E., van de Weg, E., Peace, C. and Iezzoni, A. (2017) Genetic structure of a QTL hotspot on chromosome 2 in sweet cherry indicates positive selection for favorable haplotypes. *Mol. Breed.* **37**, 85.

Campoy, J.A., Dantec, L.L., Barreneche, T., Dirlewanger, E. and Quero-García, J. (2015) New insights into fruit firmness and weight control in sweet cherry. *Plant Mol. Biol. Report.* **33**, 783–796.

Castède, S., Campoy, J.A., García, J.Q., Dantec, L.L., Lafargue, M., Barreneche, T., Wenden, B. and Dirlewanger, E. (2014) Genetic determinism of phenological traits highly affected by climate change in *Prunus avium:* flowering date dissected into chilling and heat requirements. *New Phytol.* **202**, 703–715.

Chen, H., Patterson, N. and Reich, D. (2010) Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393–402.

Choi, Y., Sims, G.E., Murphy, S., Miller, J.R. and Chan, A.P. (2012) Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE*, **7**, e46688.

Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W. (2016) GenBank. *Nucleic Acids Res.* **44**, D67–72.

Conesa, A. and Götz, S. (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics*, **2008**, 619832.

Emadzade, K., Jang, T.-S., Macas, J., Kovařík, A., Novák, P., Parker, J. and Weiss-Schneeweiss, H. (2014) Differential amplification of satellite PaB6 in chromosomally hypervariable Prospero autumnale complex (Hyacinthaceae). *Ann. Bot.* **114**, 1597–1608.

Fabbro, C.D., Scalabrin, S., Morgante, M. and Giorgi, F.M. (2013) An extensive evaluation of read trimming effects on illumina NGS data analysis. *PLoS ONE*, **8**, e85024.

Franceschi, P.D., Stegmeir, T., Cabrera, A. *et al.* (2013) Cell number regulator genes in Prunus provide candidate genes for the control of fruit size in sweet and sour cherry. *Mol. Breed.* **32**, 311–326.

Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

Garrido-Ramos, M.A. (2017) Satellite DNA: an evolving topic. *Genes*, **8**(9), 230.

Gaut, B.S., Seymour, D.K., Liu, Q. and Zhou, Y. (2018) Demography and its effects on genomic variation in crop domestication. *Nat. Plants*, **4**, 512–520.

Gnerre, S., Maccallum, I., Przybylski, D. *et al.* (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl Acad. Sci. USA*, **108**, 1513–1518.

Guajardo, V., Solís, S., Sagredo, B., Gainza, F., Muñoz, C., Gasic, K. and Hinrichsen, P. (2015) Construction of High Density Sweet Cherry (*Prunus avium* L.) Linkage maps using microsatellite markers and SNPs detected by genotyping-by-sequencing (GBS) Y. Han, ed. *PLoS ONE*, **10**, e0127750.

Haas, B.J., Delcher, A.L., Wortman, J.R. and Salzberg, S.L. (2004) DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics*, **20**, 3643–3646.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R. and Wortman, J.R. (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7.

Hancock, J.F. (2005) Contributions of domesticated plant studies to our understanding of plant evolution. *Ann. Bot.* **96**, 953–963.

Jaillon, O., Aury, J.-M., Noel, B. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.

Jones, P., Binns, D., Chang, H.-Y. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.

Korf, I. (2004) Gene finding in novel genomes. *BMC Bioinformatics*, **5**, 59.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li, Y., Cao, K., Zhu, G. *et al.* (2019) Genomic analyses of an extensive collection of wild and cultivated accessions provide new insights into peach breeding history. *Genome Biol.* **20**, 36.

Lisch, D. (2013) How important are transposons for plant evolution? *Nat. Rev. Genet.* **14**, 49–61.

Ma, J. and Jackson, S.A. (2006) Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice. *Genome Res.* **16**, 251–259.

Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J. and Clavijo, B.J. (2016) KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics*, btw663.

Marçais, G. and Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.

Mariette, S., Tavaud, M., Arunyawat, U., Capdeville, G., Millan, M. and Salin, F. (2010) Population structure and genetic bottleneck in sweet cherry estimated with SSRs and the gametophytic self-incompatibility locus. *BMC Genet.* **11**, 77.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.J*, **17**, 10–12.

McKenna, A., Hanna, M., Banks, E. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303.

Naito, K., Cho, E., Yang, G., Campbell, M.A., Yano, K., Okumoto, Y., Tanisaka, T. and Wessler, S.R. (2006) Dramatic amplification of a rice transposable element during recent domestication. *Proc. Natl Acad. Sci. USA*, **103**, 17620–17625.

Ogundiwin, E.A., Peace, C.P., Gradziel, T.M., Parfitt, D.E., Bliss, F.A. and Crisosto, C.H. (2009) A fruit quality gene map of Prunus. *BMC Genom.* **10**, 587.

Parra, G., Blanco, E. and Guigó, R. (2000) GeneID in Drosophila. *Genome Res.* **10**, 511–515.

Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S.E. and Lercher, M.J. (2014) PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* **31**, 1929–1936.

Pinosio, S., Giacomello, S., Faivre-Rampant, P. *et al.* (2016) Characterization of the poplar pan-genome by genome-wide identification of structural variation. *Mol. Biol. Evol.* **33**, 2706–2719.

Pirona, R., Eduardo, I., Pacheco, I. *et al.* (2013) Fine mapping and identification of a candidate gene for a major locus controlling maturity date in peach. *BMC Plant Biol.* **13**, 166.

Price, A.L., Jones, N.C. and Pevzner, P.A. (2005) De novo identification of repeat families in large genomes. *Bioinformatics*, **21**(Suppl 1), i351–i358.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V. and Korbel, J.O. (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**, i333–i339.

Renaut, S. and Rieseberg, L.H. (2015) The accumulation of deleterious mutations as a consequence of domestication and improvement in sunflowers and other compositae crops. *Mol. Biol. Evol.* **32**, 2273–2283.

Rosyara, U.R., Bink, M.C.A.M., van de Weg, E. *et al.* (2013) Fruit size QTL identification and the prediction of parental QTL genotypes and breeding values in multiple pedigreed populations of sweet cherry. *Mol. Breed.* **32**, 875–887.

Sansavini, S., Lugli, S., Lugli, A. and Pancaldi, M. (1998) Breeding sweet cherry for self-fertile, compact/spur tree habit and high quality fruits: trait segregation. *Acta Hortic*, 45–52. Available at: https://www.actahort.org/books/468/468_2.htm [Accessed March 13, 2019].

Sanseverino, W., Hénaff, E., Vives, C., Pinosio, S., Burgos-Paz, W., Morgante, M., Ramos-Onsins, S.E., Garcia-Mas, J. and Casacuberta, J.M. (2015) Transposon insertions, structural variations, and SNPs contribute to the evolution of the melon genome. *Mol. Biol. Evol.* **32**, 2760–2774.

Shao, H., Wang, H. and Tang, X. (2015) NAC transcription factors in plant multiple abiotic stress responses: progress and prospects. *Front. Plant Sci.* **6**, 902.

Shirasawa, K., Isuzugawa, K., Ikenaga, M., Saito, Y., Yamamoto, T., Hirakawa, H. and Isobe, S. (2017) The genome sequence of sweet cherry (*Prunus avium*) for use in genomics-assisted breeding. *DNA Res.* **12**, 60–68.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

Sindi, S., Helman, E., Bashir, A. and Raphael, B.J. (2009) A geometric approach for classification and comparison of structural variants. *Bioinformatics*, **25**, i222–i230.

Stanke, M., Schöffmann, O., Morgenstern, B. and Waack, S. (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, **7**, 62.

Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515.

Verde, I., Abbott, A.G., Scalabrin, S. *et al.* (2013) The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat. Genet.* **45**, 487–494.

Verde, I., Jenkins, J., Dondini, L. *et al.* (2017) The Peach v2.0 release: high-resolution linkage mapping and deep resequencing improve chromosome-scale assembly and contiguity. *BMC Genom.* **18**, 1–18.

Vitte, C., Fustier, M.-A., Alix, K. and Tenaillon, M.I. (2014) The bright side of transposons in crop evolution. *Brief. Funct. Genomics*, **13**, 276–295.

Wang, K., Li, M. and Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164.

Zhang, G., Sebolt, A.M., Sooriyapathirana, S.S., Wang, D., Bink, M.C., Olmstead, J.W. and Iezzoni, A.F. (2010) Fruit size QTL analysis of an F1 population derived from a cross between a domesticated sweet cherry cultivar and a wild forest sweet cherry. *Tree Genet. Genomes*, **6**, 25–36.

Zhang, H.B., Zhao, X., Ding, X., Paterson, A.H. and Wing, R.A. (1995) Preparation of megabase-size DNA from plant nuclei. *Plant J.* **7**, 175–184.

Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C. and Weir, B.S. (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, **28**, 3326–3328.

Zhou, Y., Massonnet, M., Sanjak, J.S., Cantu, D. and Gaut, B.S. (2017) Evolutionary genomics of grape (*Vitis vinifera* ssp. *vinifera*) domestication. *Proc. Natl Acad. Sci. USA*, **114**, 11715–11720.

Zhou, Y., Minio, A., Massonnet, M., Solares, E., Lv, Y., Beridze, T., Cantu, D. and Gaut, B.S. (2019) The population genetics of structural variants in grapevine domestication. *Nat. Plants*, **5**, 965–979.

Zohary, D. (2012) *Domestication of plants in the Old World : the origin and spread of domesticated plants in south-west Asia, Europe, and the Mediterranean Basin*, 4th edn. Oxford: Oxford University Press.