Horticulture Research

Horticulture Research, 2023, **10**: uhad097

https://doi.org/10.1093/hr/uhad097

Article

Genome of tetraploid sour cherry (Prunus cerasus L.) 'Montmorency' identifies three distinct ancestral Prunus genomes

Charity Z. Goeckeritz[1], Kathleen E. Rhoades[1], Kevin L. Childs[2], Amy F. Iezzoni[1], Robert VanBuren[1,*] and Courtney A. Hollender[1,*]

[1]Department of Horticulture, Michigan State University, 1066 Bogue St, East Lansing, MI 48824, USA
[2]Department of Plant Biology, Michigan State University, 612 Wilson Road, East Lansing, MI 48824, USA
*Corresponding authors. E-mail: vanbur31@msu.edu; chollend@msu.edu

Abstract

Sour cherry (*Prunus cerasus* L.) is a valuable fruit crop in the Rosaceae family and a hybrid between progenitors closely related to extant *Prunus fruticosa* (ground cherry) and *Prunus avium* (sweet cherry). Here we report a chromosome-scale genome assembly for sour cherry cultivar Montmorency, the predominant cultivar grown in the USA. We also generated a draft assembly of *P. fruticosa* to use alongside a published *P. avium* sequence for syntelog-based subgenome assignments for 'Montmorency' and provide compelling evidence *P. fruticosa* is also an allotetraploid. Using hierarchal k-mer clustering and phylogenomics, we show 'Montmorency' is trigenomic, containing two distinct subgenomes inherited from a *P. fruticosa*-like ancestor (A and A') and two copies of the same subgenome inherited from a *P. avium*-like ancestor (BB). The genome composition of 'Montmorency' is AA'BB and little-to-no recombination has occurred between progenitor subgenomes (A/A' and B). In *Prunus*, two known classes of genes are important to breeding strategies: the self-incompatibility loci (S-alleles), which determine compatible crosses, successful fertilization, and fruit set, and the *Dormancy Associated MADS-box* genes (*DAMs*), which strongly affect dormancy transitions and flowering time. The S-alleles and *DAMs* in 'Montmorency' and *P. fruticosa* were manually annotated and support subgenome assignments. Lastly, the hybridization event 'Montmorency' is descended from was estimated to have occurred less than 1.61 million years ago, making sour cherry a relatively recent allotetraploid. The 'Montmorency' genome highlights the evolutionary complexity of the genus *Prunus* and will inform future breeding strategies for sour cherry, comparative genomics in the Rosaceae, and questions regarding neopolyploidy.

Introduction

Sour cherry (*Prunus cerasus* L.) is an important temperate tree crop whose fruit is valued for its uniquely sweet and acidic flavor and superior processing characteristics for products, such as jam, juice, compote, and pie. Sour cherry is a member of the economically important Rosaceae family, which includes other cultivated *Prunus* species, such as peach, sweet cherry, apricot, almond, and plum, as well as apples, pears, roses, strawberries, and various cane fruits [1, 2]. The evolutionary history of *Prunus* has been historically difficult to resolve as hybridization, polyploidy, and incomplete lineage sorting is rampant throughout the genus [1, 3]. Several studies have demonstrated that sour cherry is an allotetraploid ($2n = 4x = 32$) and shown to be a hybrid between a tetraploid resembling *Prunus fruticosa* Pall. (ground cherry) and a diploid resembling *Prunus avium* L. (sweet cherry) [4–8]. In 1968, Olden and Nybom crossed synthetic tetraploid *P. avium* and tetraploid *P. fruticosa* in an attempt to resynthesize the species. As hypothesized, the resulting offspring were extremely similar to known *P. cerasus* accessions [4]. Chloroplast evidence suggests a *P. fruticosa*-like maternal ancestor for most sour cherry varieties, whereas S-alleles imply at least four *P. avium* individuals contributed to the *P. cerasus* gene pool [1, 4, 6, 9–14]. More recently, Bird et al. (2022) used transcriptomics to compare the nuclear and plastid genomes of eight *Prunus* species and three sour cherry cultivars. The results placed *P. avium* and *P. fruticosa* as most related to *P. cerasus* [8]. Cytological and genetic data suggest that sour cherry could be considered a segmental allotetraploid [5, 9, 15]. Trivalents and quadrivalents are common at meiosis, and although disomic inheritance is more common, tetrasomic inheritance has also been observed [5, 9, 15, 16]. The native distributions for both *P. fruticosa* and *P. avium* overlap in central and eastern Europe and sour cherry exhibits intermediate phenotypes between these progenitor species [4, 7, 10, 16]. This hybridization event likely happened multiple times as *P. fruticosa*, *P. avium*, and *P. cerasus* all hybridize in the wild, and sour cherries with substantially different phenotypes occupy dissimilar hardiness zones [7, 11]. The *P. fruticosa*-like progenitor has been shown to be the more common maternal parent of sour cherry; however, some accessions resulted from the *P. avium*-like progenitor as the maternal parent [6, 7, 12]. To add to the polyploid complexity, it is unclear if *P. fruticosa* is an allotetraploid or autotetraploid as to our knowledge, no rigorous study has been reported and the recently published genome of the species was collapsed into a simple monoploid representation [11, 13, 17, 18].

Received: 22 March 2023; **Accepted:** 4 May 2023; **Published:** 10 May 2023; **Corrected and Typeset:** 1 July 2023

Early spring freezes are major contributors to crop loss in the temperate fruit tree industry. For example, in 2012, an unseasonably warm March followed by an April freeze decimated tree fruit production throughout the midwestern USA [19]. Michigan, the number one producer of sour cherry in the USA, lost more than 90% of its crop [19]. Depending on the region, climate change is increasing the frequency of these events worldwide [19–21]. A better understanding of the genetic control of bloom time in fruit trees and breeding cultivars with later bloom times would reduce floral death and subsequent crop loss; as such, this has been a major goal for the sour cherry breeding program at Michigan State University (MSU). Sour cherries exhibit bloom times that span those of its two progenitor species, and it is hypothesized the alleles conferring later bloom time are derived from the *P. fruticosa*-like progenitor since *P. fruticosa* inhabits more northern latitudes compared to *P. avium* [22]. Development of a sour cherry genome resource would support breeding efforts by enabling gene discovery and an understanding of the genetic basis of agronomic traits in this complex tetraploid. Until now, genetic studies have depended on traditional methods involving linkage maps and common markers and synteny between *Prunus* species, since no public sour cherry genome sequence is available [9, 16, 22].

In this work, we constructed and annotated the first *P. cerasus* reference genome for the cultivar Montmorency, an ~400-year-old French amarelle sour cherry selection of unknown origin but the most widely grown cultivar in the USA. We also sequenced, assembled, and annotated sequences from a *P. fruticosa* accession present in the MSU germplasm collection, which was used, along with a published *P. avium* genome, to assign subgenomes to the 'Montmorency' superscaffolds [23]. To demonstrate the utility of the genome, we identified, manually annotated, and assigned progenitor subgenomes to two sets of genes present in the *Prunus* lineage. The first set includes the *Dormancy-Associated MADS box* genes (*DAMs*): highly conserved genes initially discovered in peach (*Prunus persica*) with major effects on dormancy transitions and flowering time in *Prunus* species [24–31]. The second set includes the self-incompatibility S-allele genes that make up the S-haplotype, consisting of S-ribonuclease (S-RNase) and S-locus F-box (SFB) [32–42]. The four S-haplotypes in 'Montmorency' have been previously characterized, but no S-haplotypes in *P. fruticosa* have been thoroughly described [37]. Our findings for these two sets of genes are discussed in the context of their putative subgenome origin and possible influence on flowering time and floral self-compatibility.

Results

Assembly of the sour cherry 'Montmorency' genome reveals 3 distinct Prunus subgenomes

We generated a chromosome-scale reference genome for sour cherry 'Montmorency' using a combination of PacBio long-read and Illumina short-read sequencing for genome assembly and chromosome conformation capture (Hi-C) for scaffolding (Figure S1). PacBio reads were assembled using Canu and polished with Pilon. The polished 'Montmorency' contigs have a total assembly size of 1066 Mb, or 172% of the estimated genome size of 621 Mb according to a k-mer analysis (k = 25). The size of the assembly in conjunction with the abundant estimated heterozygosity (4.90%) implied multiple haplotypes were assembled. The Merqury plot (Figure S2) indicated high genome completeness as most k-mers in the Illumina dataset are found in the assembly [43]. Additionally, k-mers from the Illumina reads are found in the assembly at the expected relative frequencies and there are four distinct peaks, indicating haplotypes are well phased. A BUSCO assessment demonstrated the assembly contains a suitable representation of the gene space (>98% complete BUSCOs) and most of them are duplicated (>93%), as expected of a polyploid [44].

As sour cherry is an allotetraploid derived from two progenitor species, we expected to assemble two full haplotypes for 'Montmorency' (*Prunus* subgenomes; n = 2x = 16) with additional alleles being unanchored. To our surprise, initial scaffolding resulted in 24 linkage groups (chromosomes), 8 of which experienced sudden drops in signal along the Hi-C diagonal and were made up of significantly smaller contigs than the other 16. Preliminary phylogenomic analyses (see Methods) showed genes on these eight poorly scaffolded linkage groups were most likely derived from the *P. avium*-like ancestor. Therefore, we posited there was a disproportionate collapse of *P. avium*-like haplotypes compared to the others, with the poorer scaffolding being a result of collapsing and haplotype switching. From here, efforts were made to reassemble the genome with the goal of either 1) better phasing the *P. avium*-like sequences or 2) forcing them to collapse into a monoploid representation. Unfortunately, due to the heterogeneous nature of all four haplotypes' sequences, the first scenario resulted in poor sequence correction and assembly inflation while the second negatively affected the assembly of the haplotypes that had previously been intact. Instead, Purge Haplotigs was used to set aside one of the two possible alleles of the *P. avium*-like sequences [45]. Since we knew these sequences consisted of smaller contigs, those greater than 400 kb that had been removed by Purge Haplotigs were added back into the assembly prior to scaffolding. Figure S3 showed promising results: k-mers from the Illumina dataset were found only once in the purged (removed) assembly portion. In other words, few to no sequences common to multiple alleles (2×, 3×, 4×) were in this portion of the assembly.

The scaffolding pipeline was subsequently rerun, and 771.8 Mb (124% of the estimated haploid genome size) was scaffolded into 24 linkage groups (Figure S4). These chromosomes were numbered according to their synteny with *Prunus persica* chromosomes 1–8 (Figure 1a) [46, 47]. These results and the Hi-C matrix suggested three homoeologs, instead of the predicted two, were assembled for each *Prunus* ancestral linkage group. We subsequently named and clustered these 24 chromosomes into three groups of 8 (A, A', and B) based on their 25-mer signatures and progenitor assignments (details to follow). A comparison between the 24 linkage groups and the recently published *P. avium* 'Tieton' genome indicated substantial synteny amongst the 'Montmorency' chromosomes themselves as well as *P. avium* (Figure 1b), in concordance with numerous reports that *Prunus* genomes are highly syntenic [23, 48–50]. When the scaffolded linkage groups were compared with a sour cherry genetic map of 545 unique markers, 78% mapped exactly once to each homoeolog [16]. Furthermore, all markers showed nearly perfect linearity with the assembly (Figure 2, Figure S5).

To identify subgenome groups for the 24 chromosomes, we used a strategy based on the rationale that chromosomes enriched for the same set of repetitive elements, represented by k-mer type and abundance, will have a more recent common ancestor. Thus, distinct groups of chromosomes with more k-mers in common may represent subgenomes with the same origin. Unsupervised k-mer clustering to identify allopolyploid subgenomes has been successfully applied to *Miscanthus sinensis*, *Nicotiana tobacum*, *Triticum aestivum*, *Eragrostis tef*, and *Panicum virgatum* [51–54]. Using this strategy, we identified 840 25-base pair sequences (25-mers) with more than 10 copies on each

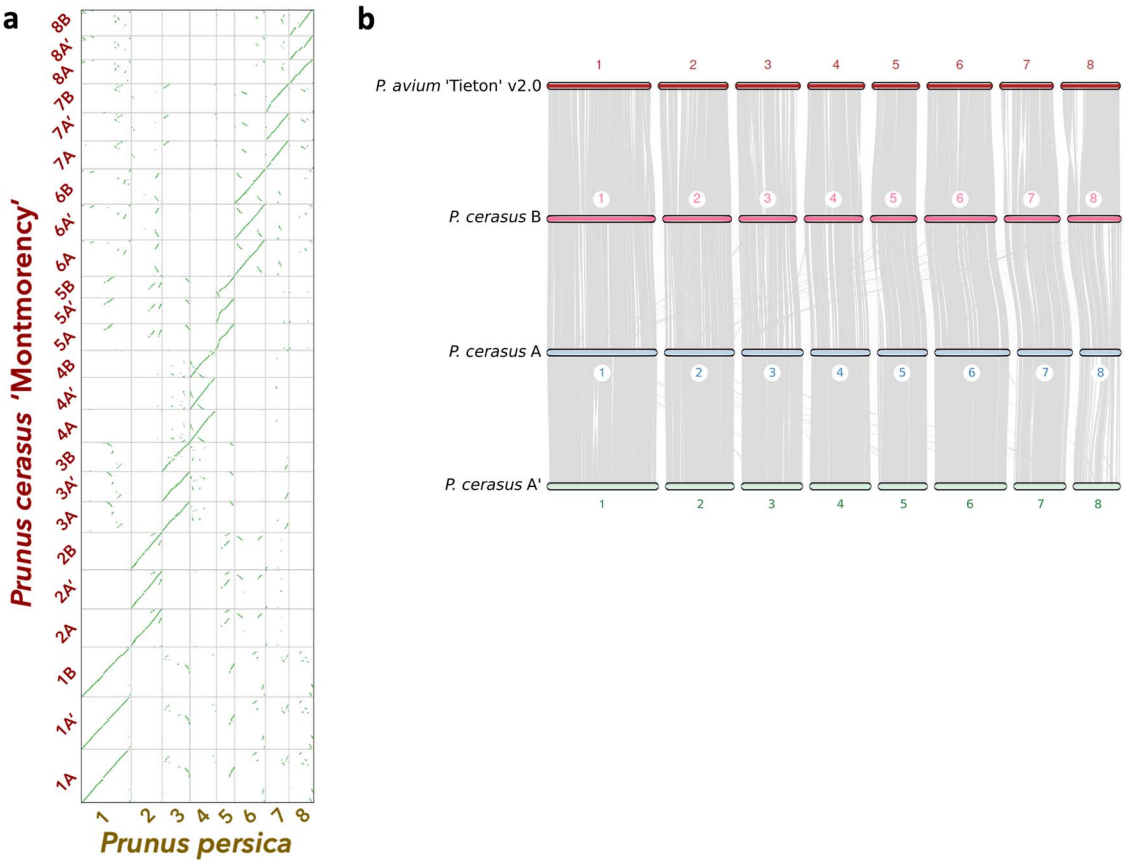

Figure 1. Syntenic relationships between 'Montmorency' subgenomes A, A', and B and other *Prunus* species. (a) A syntenic dotplot showing the results of a synteny comparison of the 24 superscaffolds (chromosomes) of *P. cerasus* 'Montmorency' and the eight chromosomes of *P. persica* "Lovell" v2.0 [47]. The figure was generated using unmasked coding sequences for each species with the CoGe platform. The prominent linearity for chr #[A, A', B] for *P. cerasus* vs the respective chr # in *P. persica* highlights the collinearity of this genus and supports the integrity of the assembly. (b) Macrosynteny determined with coding sequences shows the three subgenomes of sour cherry are highly syntenic with each other and the published *P. avium* 'Tieton' v2.0 genome [23]. Each gray line represents a syntenic block between the genomes. Small rearrangements between *P. avium* and each of the 'Montmorency' subgenomes are evident.

chromosome and twice the abundance on any one homoeolog compared to one or both of its sisters. These 25-mers were used to conduct a clustering analysis with all 24 chromosomes of the 'Montmorency' assembly, which resulted in two distinct clades: one consisting of 16 chromosomes and the other consisting of the remaining 8 (Figure 3a, Figure S6). These groupings were mainly attributed to the differential abundance of two 25-mer clusters, designated Group 2 (consisting of 48 25-mers) and Group 3 (consisting of 278 25-mers) (Figure S6).

We further posited the set of 16 chromosomes could be subdivided into additional groups given the ease at which most of these chromosomes had been phased. Upon closer examination of the Hi-C matrix of homoeologous group 8, the two included in the group of 16 chromosomes contained many sequences interchangeable with either homoeolog (Figure 3b). This may have resulted from the collapsing of similar sequences during assembly, homoeologous exchange, or both. Nonetheless, it was indicative of poorer assembly quality compared with the seven other homoeologous groups and could be clouding the complete separation of these 24 linkage groups into three sets of eight chromosomes. Indeed, when these two linkage groups from homoeologous set 8 were excluded from the clustering, the 22 chromosomes neatly grouped into three clades (Figure 3c).

Therefore, a separate clustering analysis including only 14 chromosomes (without homoeologous chromosome 8s) was conducted to better visualize the 25-mers separating the two sets of seven chromosomes. A total of 481 25-mers in three clusters differentiated the two chromosome sets (Figure S7; Groups 5 [n = 44], 6 [n = 148], and 7 [n = 289]). Based on their distinct 25-mer signatures and subgenome assignments (see below), the three chromosome sets were named A, A' (denoted as A__ in all data files), and B. 8A and 8A' were grouped arbitrarily.

To further investigate the locations of the 25-mer groups contributing to the chromosome subgenome separation, we plotted the densities of several of these 25-mer groups across each chromosome and marked putative centromere locations based on gene and transposable element (TE) densities (Figure 4a, Figure S8a). Peak 25-mer densities colocalized with the lowest gene content and highest TE density, demonstrating these 25-mer groups are most abundant at or near putative centromeres. These estimated centromeric regions agree with a previously published *P. avium* genome [55]. Group 2 and Group 3 densities, which distinguished A/A' subgenomes from subgenome B in the 24-chromosome clustering analysis (Figure 3a; Figure S6), clearly exhibited contrasting peak 25-mer densities at approximate centromeres (Figure 4b, Figure S8b). Likewise, Group 5 and Group 6 25-mers, which differentiated the A and A' subgenomes (Figure S7), showed a similar pattern when their densities were aligned to each of the homoeologous chromosome sets (Figure 4c, Figure S8c). Since highly variable, repetitive satellite

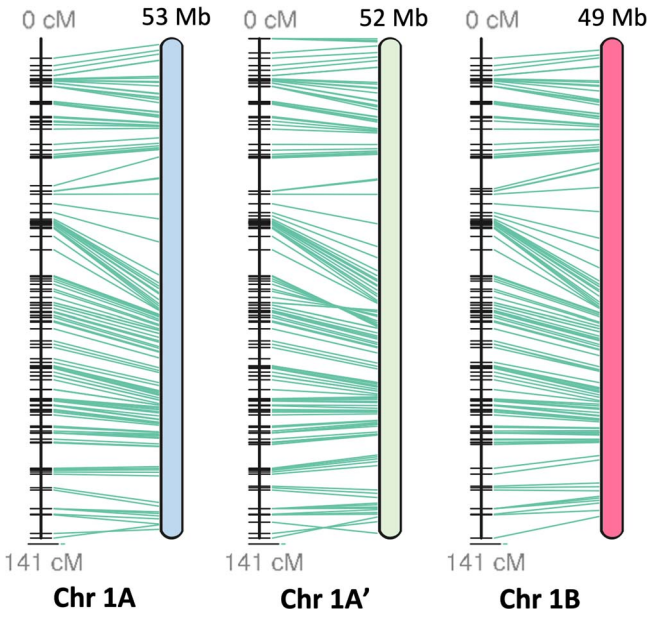

Figure 2. Linearity comparison of linkage group 1 and a published sour cherry genetic map [16]. A total of 545 markers from an F1 sour cherry cross in which 'Montmorency' was the female parent were mapped to the assembly, and the results demonstrate the high collinearity between the linkage map and assembly. Green lines connect the markers on the genetic map (left) to the physical location in the assembly (right). Each horizontal black line on the genetic map represents one marker. Postfiltering, 426 of the 545 markers mapped exactly once to each subgenome. Subgenome B is a representative of two possible haplotypes. This figure was generated with ALLMAPS [103]. Other chromosome sets (2–8) are shown in Figure S5.

DNA is species-specific and often associated with centromeric and pericentromeric regions in eukaryotes, these observations support the claim that A, A', and B represent distinct subgenomes derived from separate *Prunus* species [56–58].

As it is well-established sour cherry is a tetraploid, we explored the relative dosage of each 'Montmorency' subgenome by conducting a depth analysis of the Illumina reads against the assembly [5, 11, 15]. As suspected earlier, this analysis revealed subgenome B had twice the genome dosage of either subgenomes A or A' as regions along the length of subgenome B frequently showed about twice the read depth of subgenomes A and A' **(**Figure S9). Therefore, the genome structure of *P. cerasus* 'Montmorency' is AA'BB. A summary of assembly and scaffolding statistics is given in Table 1.

Annotation of the *P. cerasus* 'Montmorency' assembly

For structural annotation of the gene space, RNA sequencing of a variety of tissues and long-read cDNA datasets were generated for 'Montmorency' and used as transcript evidence for de novo gene annotation via MAKER [59, 60]. MAKER predicted a total of 92 783 protein-coding genes in the full assembly of 'Montmorency' after filtering for gene predictions with known Pfam domains (known as the Standard MAKER gene set). The standard MAKER gene set was the input for the first step of defusion, a MAKER-compatible software designed to disentangle one or more adjacent gene models that are erroneously fused [61]. Defusion detected 906 potentially fused genes in the full assembly; 707 of these were on the scaffolded assembly [chr1A/A'/B–chr8 A/A'/B]. In addition to identifying these candidate fusions automatically with defusion, MAKER was run using only protein evidence to find gene models with more than one protein hit aligning to them—suggesting a possible fusion. Using a custom script, 9867 gene models fitting this criterion were found in the full assembly and 7537 gene models in the scaffolded assembly. Gene models tagged as fusions using these two approaches on the scaffolded assembly were manually checked against protein, RNA, and nanopore cDNA alignments in IGV and added to the breakpoint file if identified as a true fusion (Figure S10). In total, 4481 genomic regions (gene models) on the scaffolded assembly were locally reannotated using MAKER within defusion. After another Pfam domain search, 9777 of the defused gene models contained known protein domains. These gene models were added back into the final annotation.

Summary statistics for the annotated gene set are given in Table S1, and Figure S11 shows the cumulative distribution of all gene models' AED values on the 24 chromosomes of the assembly. Each of the three 'Montmorency' subgenomes have similar gene prediction counts and high BUSCO completion scores (85–90% for both transcripts and proteins). The shape of AED distribution, with well over half of the gene models having AED values <0.2, suggests a high-quality annotation that is very agreeable with protein and expressed sequence data.

Assembly of a draft genome for *Prunus fruticosa*, a probable allotetraploid

PacBio long-reads and Illumina short-reads were also generated for a *P. fruticosa* accession in the Michigan State University (MSU) germplasm. Similarly to *P. cerasus* 'Montmorency', Canu and Pilon were used to create a polished draft assembly. The primary reason for generating this *P. fruticosa* resource was for subgenome assignments and divergence estimates between progenitors and A, A', and B in 'Montmorency'. Given 'Montmorency' was verified to contain three subgenomes and this sour cherry and *P. fruticosa* share a recent maternal ancestor, we suspected *P. fruticosa* may also be an allotetraploid [6–8, 12]. This was important to ascertain before downstream analyses since a collapsed representation of the genome would affect syntelog-based subgenome assignments and the divergence estimates of each subgenome. A k-mer assessment using the Illumina short-reads showed the predominant class of heterozygosity of the *P. fruticosa* accession is *aabb*, typical of allopolyploids with strict homologous pairing of subgenomes at meiosis (Table S2) [62]. Further, a K_s analysis of 102 293 gene pairs in the draft assembly showed two distinct peaks at frequencies 0.003 and 0.022, suggesting homologous and homoeologous gene comparisons, respectively (Figure S12) [63]. Thus, we had reason to believe *P. fruticosa* is an allotetraploid and it was best to have an assembly representing the most allele diversity possible for accurate syntelog comparisons. A Merqury plot of the *P. fruticosa* assembly revealed common issues associated with polyploid assemblies, namely the collapsing of some haplotypes (red and blue overlap, green and purple overlap) and possible artificial duplication of others (small green overlap with blue and red; Figure S13). The most notable overlap (red and blue; *aabb* k-mers) likely represents some collapse of homologs within subgenomes.

The *P. fruticosa* draft assembly contains 986 Mb in 3932 contigs, while the genome size was predicted to be 532 Mb. The assembly contains >99% complete BUSCOs, of which 92.7% are duplicated. A syntenic analysis of the *P. fruticosa* draft assembly's gene predictions with *P. avium* 'Tieton' showed a 4:1 pattern (Figure S14) [23, 64] . These three metrics verified multiple haplotypes are assembled in the *P. fruticosa* draft assembly and the gene space is well represented. Although the draft assembly is not chromosome scale, the contigs are sufficiently large enough for syntenic and

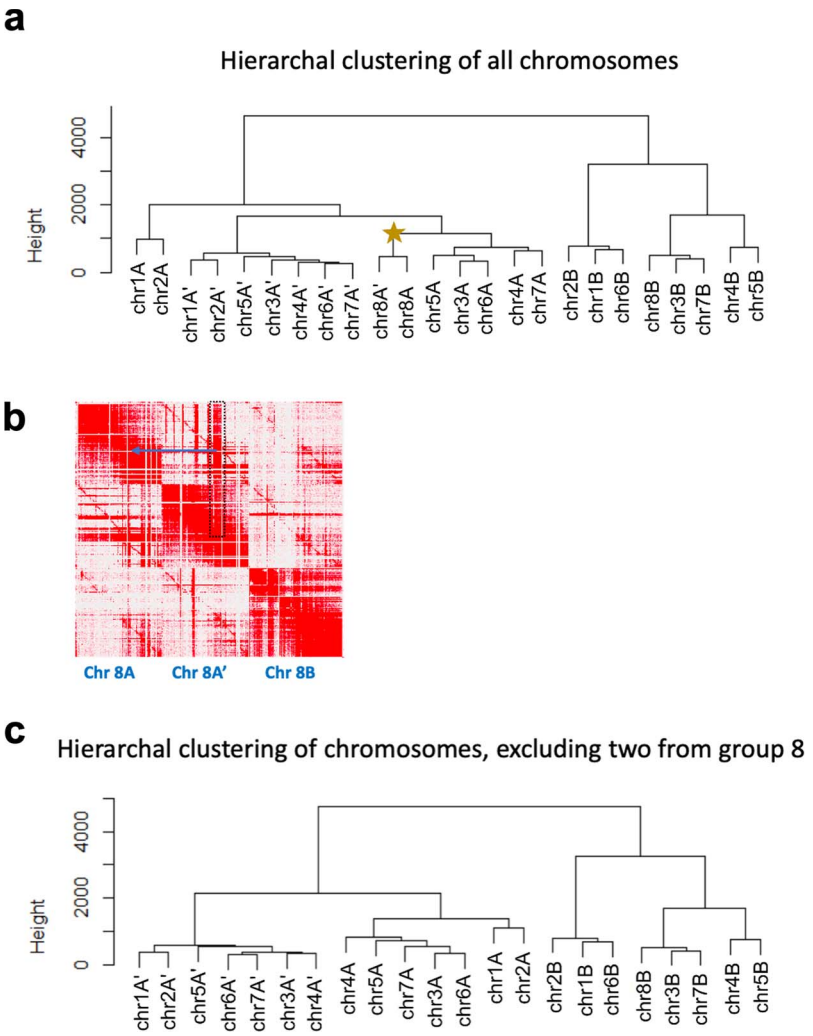

Figure 3. Chr8A and chr8A' affect the k-mer groupings that differentiate subgenomes. (a) Hierarchal clustering of all 24 chromosomes based on 25-mer abundance (present 10 times or more on every chromosome and at least twice as abundant on one homoeolog compared to its sisters). The star indicates chr8A and chr8A'. A corresponding heat map is shown in Supplementary Figure S6. (b) An enlarged section of the Hi-C matrix for homoeologous chromosome set 8. The dark red signal in the dotted black box indicates an example of sequence on chromosome 8A' that could have been placed on chromosome 8A in the region indicated with a blue arrow. This could be due to an assembly artifact (collapse of highly similar sequences in these regions), homoeologous exchange between these two chromosomes, or a combination of both. (c) Same hierarchal clustering analysis as in a) but excluding the two 8-chromosome groups.

gene orthology comparisons as several megabases in small *Prunus* genomes may contain hundreds of genes (Table S2**,** see NG50).

Annotation of the *P. fruticosa* draft assembly

RNA sequences from five separate tissues from the same *P. fruticosa* accession used for genome assembly, gene models from 'Montmorency', and manually curated protein datasets were used as evidence for annotation of the *P. fruticosa* assembly with MAKER [65]. MAKER predicted a total of 102 361 protein-coding genes for the *P. fruticosa* contigs after filtering to select genes with known Pfam domains and excluding those with known TE domains. The BUSCO completion score for the annotated transcripts of the draft assembly is 97.10% with 92.20% of those duplicated. A summary of statistics for the annotation of the *P. fruticosa* contigs is shown in Table S3**.** Like the *P. cerasus* 'Montmorency' annotation, well over half of the cumulative fraction of AED values assigned for all gene models (excluding those edited with Apollo), have AED values <0.2, suggesting most genes are well supported by transcript and protein homology evidence (Figure S15**).**

Progenitor assignments of the subgenomes in *P. cerasus* 'Montmorency'

There were 267 936 orthologs, or 93.1% of genes in all seven "species" (*Malus domestica*, *P. persica*, *P. avium*, *P. fruticosa*, 'Montmorency' subgenome A, subgenome A', and subgenome B) were assigned to orthogroups using OrthoFinder v 2.5.4 [66]. Of all orthogroups, 12 051 included at least one ortholog for every species. There were 336 identified as single-copy orthologous groups. After multiple sequence alignment, trimming, phylogenetic tree construction for every orthogroup, and extraction of progenitor–subgenome gene relationships based on a bootstrap support value >80%, we identified 6797, 7036, and 10 398 relationships in subgenome A, A', and B, respectively. The vast majority of these were syntelogs, and marking their locations colored by representative progenitors on the 'Montmorency' chromosomes showed each chromosome was predominantly derived from one progenitor (Figure 5). In total, 98.9% (6664/6739) and 99.6% (6957/6984) of syntelog relationships indicated subgenomes A and A', respectively, were derived from a *P. fruticosa*-like progenitor. Conversely, 98.8% (10 245/10370) of syntelog

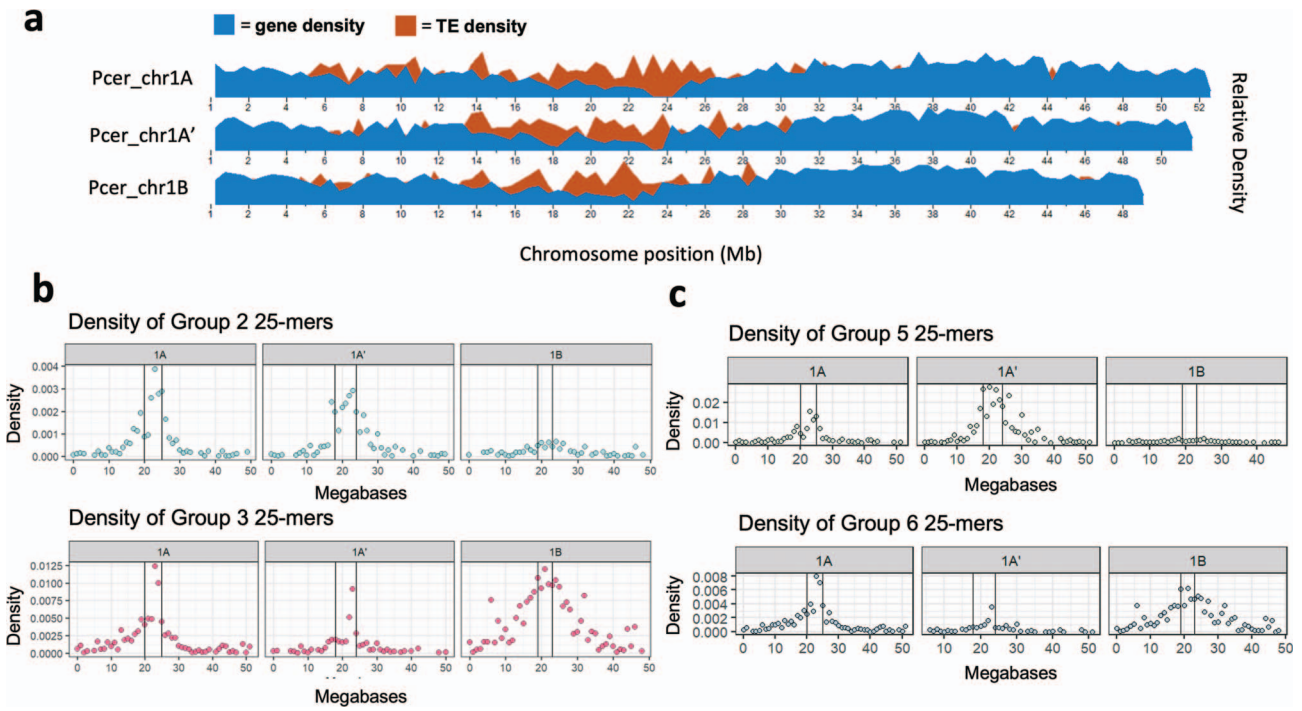

Figure 4. 25-mer group densities differentiating the 'Montmorency' subgenomes peak at approximate centromeres. Only chr1 is shown for clarity. (a) Gene and transposable element (TE) densities plotted along the three chromosome 1 homoeologs. The centromeres are estimated to be regions that coincide with relatively low gene and high TE densities. (b) Group 2 and Group 3 25-mer densities (from Figure S6) plotted along the length of chromosome 1. These distinguish the A/A' subgenomes from subgenome B, when all 24 chromosomes are included in this clustering (Figure 3a). (c) Group 5 and Group 6 25-mer densities (from Figure S7) along the length of chromosome 1. These distinguish A and A' from one another. In both b) and c), the region between the vertical lines along the density plots designates the approximate location of the centromeres. Corresponding figures for the seven other chromosome sets are in Figure S8.

Table 1. Summary of the Montmorency assembly metrics

Estimated haploid genome size (k-mer analysis; k = 25)					**621 Mb**
Estimated Heterozygosity (total)					4.90%
class	aaab 2.430%	aabb 2.060%	aabc 0.001%	abcd 0.451%	
Full assembly size					1066 Mb
Scaffolded assembly size					771.8 Mb
NG50					11.56 Mb
Number of contigs					3592
Linkage Groups					24
BUSCO (viridiplantae db10)	complete	singletons	duplicates	missing	fragmented
Scaffolded assembly (24 LGs)	98.6%	5.4%	93.2%	0.9%	0.5%
subgenome A	91.5%	89.6%	1.9%	6.4%	2.1%
subgenome A'	94.8%	92.9%	1.9%	4.0%	1.2%
subgenome B	90.8%	88.2%	2.6%	7.3%	1.9%
Estimated % repeats. (Full assembly)	LTR 35.6%	TIR 11.6%	Helitron 1.3%		Total 48.5%
LAI	Full assembly (incl. unanchored) Scaffolded assembly				14.74 17.09

Mb, megabases; LG, Linkage Group; NG50, 50% of the estimated genome size is contained in contigs of equal or greater value; BUSCO, Benchmarking Universal Single-Copy Orthologs [44]; LTR, Long Terminal Repeat; TIR, Terminal Inverted Repeat; LAI, LTR Assembly Index [96].

relationships identified for subgenome B supported it as *P. avium*-like. These results also suggested little-to-no homoeologous recombination had occurred between the A/A' and B subgenomes.

To identify progenitor relationships of the unanchored genes, we conducted a parallel analysis including the unanchored sequences. However, due to the fragmented nature of these scaffolds, we did not attempt to identify which orthologs were syntelogs. Despite the low number of high-confidence relationships identified in the unanchored sequences (n = 858), 87.6% of these genes are *P. avium*-like, further supporting the claim that subgenome B is at twice the dosage of subgenome A and A'.

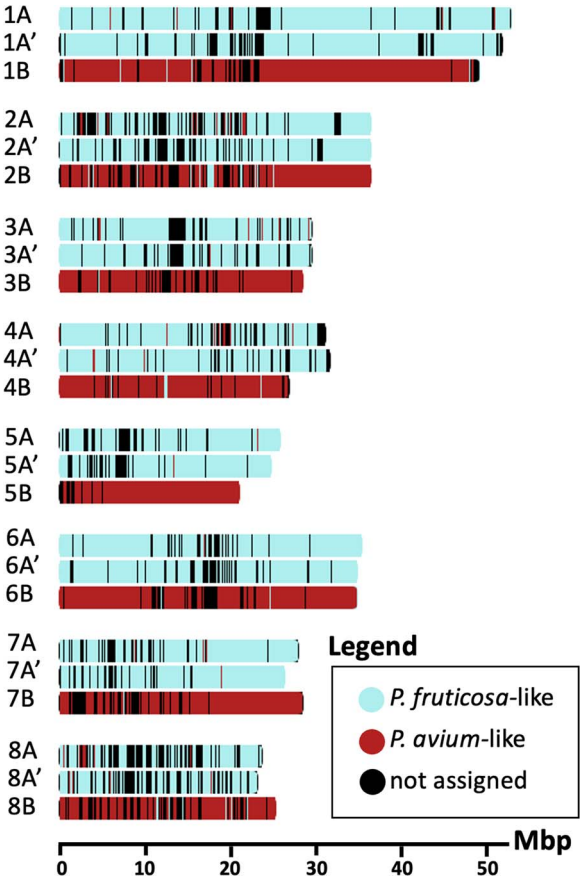

Figure 5. Subgenome assignment using syntelogs reveals little-to-no recombination has occurred between progenitor genomes in 'Montmorency'. 24 093 syntelogs that were identified with phylogenomic ortholog comparisons and synteny analyses are plotted along the lengths of all eight chromosome sets and colored by the progenitor they are most likely derived from. Window size for each tick mark ranges between 130 and 133 Kb and is automatically optimized in chromoMap [142] based on the largest chromosome's size. Mbp, Megabase pairs.

DAM gene haplotypes identified in 'Montmorency' and *P. fruticosa*

The *Dormancy-Associated MADS-box* genes (*DAMs*) are six tandemly arrayed, type II MIKC^c MADS-box genes required for proper dormancy transitions and bloom time in *Prunus* [25, 29]. Given the agronomic significance of bloom time, we sought to identify and polish the annotation of these genes in the 'Montmorency' and *P. fruticosa* assemblies. BLAST+ analyses revealed *DAM* gene candidates on 'Montmorency' chromosomes 1A, 1A', and 1B [67]. Upon manual inspection and correction of these gene models using Apollo v 2.6.5, three full haplotypes of *DAM1–DAM6* were found on chr1A, chr1A', and chr1B [68]. All 18 *DAM* genes have open reading frames, and all have the characteristic intron–exon structure of these MADS-box genes except for *DAM6* on chr1A'. This gene had extremely low expression in the tissues sampled for annotation and only a single, full-length Nanopore cDNA read supported the gene model. It is possible this transcript represents a splice variant, as it is missing three of the nine exons typical of the other *DAM* genes [29]. For the *P. fruticosa* genome, 23 gene models on seven different contigs representing partial and full *DAM* haplotypes were found; however, only contig 8 and contig 33 contained full *DAM* haplotypes (i.e. all six *DAM* genes in tandem), and these regions were confirmed to be syntenic with *DAM* haplotypes found on 'Montmorency' chromosomes 1A, 1A', and 1B (Figure 6a).

We next carried out a phylogenetic analysis of all 18 and 12 *DAM* genes in 'Montmorency' and *P. fruticosa*, respectively, with manually annotated *DAM* genes from *P. avium* (Figure 6b) [31]. Gene numbers and NCBI identifiers associated with this analysis can be found in Table S4. All *DAM1–DAM6* genes formed well-supported monophyletic clades (BSV >80%) consistent with their order in the tandem array (Figure S16), garnering further support these genes were correctly identified. Furthermore, each 'Montmorency' *DAM* gene was sister to the syntelog of its respective progenitor. In other words, all chr1A *DAM* genes were sister to *P. fruticosa* contig 8 *DAM* genes, chr1A' *DAM* genes were sister to *P. fruticosa* contig 33 *DAM* genes, and chr1B *DAM* genes were sister to *P. avium DAM* genes (Figure 6b). Interestingly, the gene downstream from the *DAM* haplotype for 'Montmorency' chr1A' (Pcer_010093) and *P. fruticosa* contig 33 (Pfrut_003731; note that two gene models were created for this region) contain similar insertions of nearly 9 kb in the sixth intron (99.87% identical), providing higher confidence these haplotypes have shared ancestry. The size of this intron on chr1A, chr1B, and *P. fruticosa* contig 8 is approximately 767 bp.

S-alleles identified in 'Montmorency' and *P. fruticosa* assemblies

S-alleles, consisting of an RNase (S-RNase) and linked F-box protein (SFB), are responsible for gametophytic self-incompatibility in *Prunus* [41]. Four S-alleles identified in the 'Montmorency' assembly on chromosomes 6A, 6A', and 6B, matched the 'Montmorency' S-alleles previously reported, i.e. S_6, S_{13m}, S_{35}, and S_{36a} [37, 39]. The linked SFBs and S-RNases comprising these four S-allele haplotypes all have >99% identities to their published sequences [37, 39, 42]. BLAST+ results indicated S_{36a} is on chr6A, S_{35} is on chr6A', and both S_{13m} and S_6 are on chr6B. The two S-alleles on the same chromosome (6B) is likely an assembly artifact, since these alleles have been demonstrated to segregate independently in sour cherry crosses with 'Montmorency' as a parent [34, 39]. The placement of the S_6 and S_{13m} alleles on the *P. avium*–derived subgenome is consistent with the known prevalence of S_6 and S_{13} in *Prunus avium* [42]. However, the S_{13m} in 'Montmorency' is a stylar-part mutation of the *P. avium* S_{13} allele that has only been identified in sour cherry [36]. The other S-alleles in 'Montmorency', S_{35} and S_{36a}, have not been identified in *P. avium* and are thought to be derived from the *P. fruticosa*-like progenitor [39]. In sour cherry, there are four variants of the S_{36} haplotype based on minor sequence differences within and/or flanking the S-RNase and SFB coding regions. These four S_{36} variants (S_{36a}, S_{36b}, S_{36b2}, and S_{36b3}) collectively are the most widespread S-alleles in sour cherry as all genotypes examined to date have at least one or two S_{36} variants [14, 39]. In the *P. fruticosa* draft assembly, S-allele candidates were found on contigs 53 and 1100. Contig 53 was verified to be syntenic to the regions containing the S-alleles in 'Montmorency' (Figure S17); however, contig 1100 was too small (130 kb) to do a macrosyntenic analysis using MCScan. Of the sequences used as queries (Table S5), the closest S-RNase and SFB matches for the *P. fruticosa* haplotype on contig 53 were *P. cerasus* S_{36b} variants, with >99% shared sequence identity. All domains characteristic of S-RNases were found in the *P. fruticosa* S-RNase identified on contig 53, but a premature stop codon is predicted at amino acid position 67, where a W resides for all other S-RNase 36a and b variants (Figure S18a; note the amino acid fasta sequence reads through the stop codon to indicate conserved domains). The SFB protein on contig 53 has an open

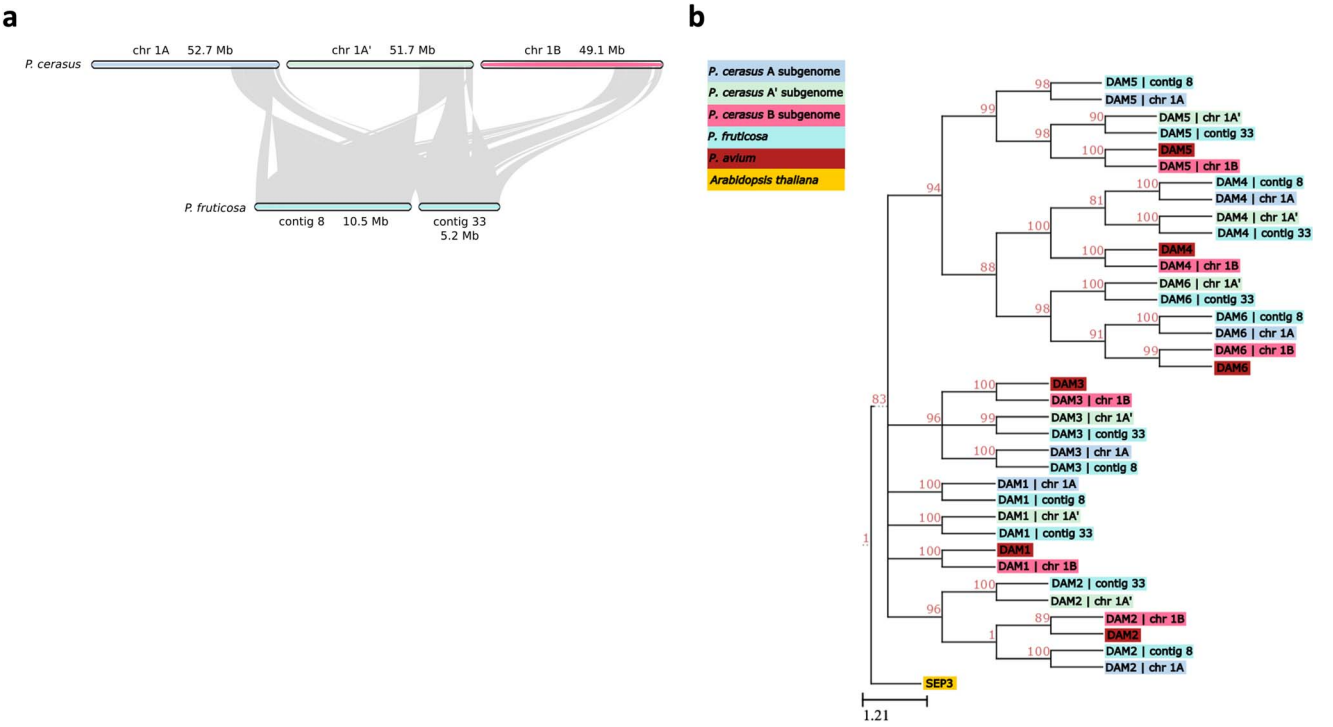

Figure 6. Synteny and phylogeny of the *Dormancy Associated MADS-box* (*DAM*) genes in 'Montmorency' and *Prunus fruticosa*. (a) The genomic regions where full *DAM* haplotypes were identified in both species show high macrosynteny. (b) Phylogeny of the *DAM* gene coding sequences of 'Montmorency', *P. fruticosa*, and *P. avium* [31]. *Arabidopsis thaliana SEP3* was used as an outgroup. The clustering of *DAMs* together by number suggests correct identification of these genes. The clustering shows the *DAM* genes from subgenomes A and A' are most closely related to *P. fruticosa DAM* genes while *DAM* genes from subgenome B are most closely related to *P. avium DAMs*. This agrees with each subgenome's prior assignment to a *P. avium*-like or *P. fruticosa*-like progenitor. Nodes below an 80% bootstrap value were collapsed.

reading frame, expected protein domains, and an amino acid sequence identical to the *P. cerasus* SFB_{36b} sequence (Figure S18b) [39]. In summary, S_{36} variants have not been previously confirmed in *P. fruticosa*, so the S_{36} variant found in this *P. fruticosa* draft assembly supports the hypothesis that sour cherries, including 'Montmorency', inherited their S_{36} haplotype(s) from a *P. fruticosa*-like progenitor.

P. cerasus 'Montmorency' is descended from a recent hybridization event

Lastly, we sought to estimate divergence time of each 'Montmorency' subgenome from its most closely related representative progenitor. First, we examined the individual topologies of the 336 phylogenetic trees of single-copy orthologs. Only single-copy orthologs were used in these analyses to diminish the effect recent gene duplication events would have on divergence estimates. Based on previous phylogenetic assessments and the 'Montmorency' subgenome assignments, a topology with A or A' sister to the *P. fruticosa* ortholog, B sister to *P. avium*, *P. persica* sister to the cherries, and *Malus* × *domestica* sister to *Prunus* would most accurately estimate when each subgenome last shared a common ancestor with its representative progenitor (i.e. when these lineages began diverging) [1–3, 69]. The hybridization event 'Montmorency' is descended from would have occurred sometime after this estimate.

P. fruticosa is tetraploid; therefore, all single-copy orthologs used in these analyses are an assembly collapse of four possible alleles. If *P. fruticosa* is indeed an allotetraploid with two intact and divergent subgenomes, approximately half of the collapsed single-copy orthologs would be expected to come from the first subgenome while the other half would come from the second subgenome. If A and A' share a more recent common ancestor with one *P. fruticosa* subgenome compared to the other, we would expect the orthologs from A and A' to be sister to the collapsed *P. fruticosa* ortholog in an approximately equal number of trees. Moreover, if A and A' are descended from divergent *Prunus* ancestors, we would rarely expect their single-copy orthologs to be sister to one another.

The most frequent topologies observed among the single-copy orthogroups include one where A is sister to *P. fruticosa* (n = 43) and the other where A' is sister to *P. fruticosa* (n = 70; Figure S19). In all single-copy ortholog topologies, A was sister to the *P. fruticosa* ortholog 42% of the time (n = 142), whereas A' was sister to the *P. fruticosa* ortholog 49% of the time (n = 165). A was sister to A' in only 1.8% of topologies (n = 6). These results are consistent with *P. fruticosa* being an allotetraploid, and that A and A' are diverged enough to be considered derived from separate *Prunus* species. Subgenome B was sister to *P. avium* in nearly all single-copy ortholog trees (n = 310; 92%).

The r8s analysis of the most frequent topologies suggested 'Montmorency' subgenome A and *P. fruticosa* began diverging 1.61–1.63 mya (topology B, node 6), subgenome A' and *P. fruticosa* began diverging 4.48 to 4.51 mya (topology A, node 4), and subgenome B and *P. avium* began diverging <1.72 mya (topology A, node 6; topology B, node 5; Figure S19). Taken together, these results suggest the hybridization event *P. cerasus* 'Montmorency' is descended from occurred less than 1.61 mya.

Discussion

Here we report the genome assembly for sour cherry (*P. cerasus* L.) cultivar Montmorency, which to our knowledge is the first published genome sequence for sour cherry and the onlymes

to be published to date. Recently, a genome sequence for the cultivar "Schattenmorelle" became available as a preprint; these authors distinguished scaffolds as *P. avium*-derived and *P.fruticosa*-derived, but did not identify subgenomes within the latter. Additional references like these will be useful in understanding the evolutionary history of sour cherry [70]. We chose to sequence 'Montmorency' as it is the most grown cultivar in the USA. This chromosome-scale assembly is highly collinear with a published sour cherry genetic map, is syntenic with other *Prunus* species, and has a quality annotation with thorough manual curation [16, 23, 47]. Additionally, we assembled and annotated a draft genome of allotetraploid *P. fruticosa*, the closest extant relative of one of sour cherry's proposed progenitor species. We expect both resources to be informative for future sour cherry breeding strategies and comparative genomic studies in *Prunus*.

The allotetraploid origin of sour cherry is well established and further supported by this work; however, the three subgenome composition of 'Montmorency', AA'BB, with two divergent genomes likely contributed by the *P. fruticosa*-like ancestor, was an unexpected result. The separation of subgenomes A and A' using k-mers required the exclusion of chromosomes 8A and 8A', which showed lower assembly quality compared with other chromosomes. In addition to k-mer evidence, the low frequency in which A and A' genes were sister to one another in single-copy ortholog trees lends strong support for treating them as separate subgenomes derived from different *Prunus* species. Until now, concrete evidence of the origin of *P. fruticosa* has been lacking despite the recent publication of its genome sequence [17]. This *P. fruticosa* sequence was published while the present study was underway, and these authors did not attempt to distinguish whether the species was an allotetraploid or autotetraploid. Our results strongly imply *P. fruticosa* is an allopolyploid hybrid of two distinct *Prunus* species. However, since large-scale evolutionary studies of *Prunus* rarely include this species, the extant relatives of its progenitor species are unknown and is a question for future research [1–3]. Identification of the *P. fruticosa* progenitor species (or, more probably, their extant relatives) would allow for greater resolution of the dynamics of subgenome A and A' within sour cherry.

The subgenome assignments for 'Montmorency' were further supported by our results for two sets of biologically significant genes for *Prunus* in the 'Montmorency' reference and *P. fruticosa* draft genomes: the *DAM* genes and *S*-alleles. Alleles of both gene sets were consistent with the conclusion that 'Montmorency' has an AA'BB subgenome structure where A/A' and B are derived from *P. fruticosa*-like and *P. avium*-like ancestors, respectively. In this work, we document the first discovery of an S_{36} variant in *P. fruticosa*, supporting the previous hypothesis that S_{36} variants identified in sour cherry are derived from a *P. fruticosa*-like progenitor [39]. An interesting question moving forward is how the progenitor origins of the *DAM* genes may relate to bloom time in diverse sour cherry accessions. Sour cherry exhibits a transgressive range in bloom time: some genotypes bloom earlier or later than either progenitor [16]. *P. fruticosa* is native to colder northern latitudes in eastern Europe and flowers later than *P. avium*, which originates from the Mediterranean region south of the Black Sea [71]. One might expect to see selective pressure on the *DAM* genes depending on whether a *P. cerasus* genotype distributes to more northern or southern latitudes, tailoring bloom time to maximize reproductive potential. However, this question remains largely unexplored.

Our subgenome assignments for 'Montmorency' also provide insights into the possible gametes that formed 'Montmorency'. First, we did not find obvious evidence of recombination between the 'Montmorency' subgenomes derived from the *P. fruticosa*-like progenitor (A/A') and the subgenome derived from the *P. avium*-like progenitor (B). Since younger polyploids with more recent homoeologous exchanges tend to show them in large chromosomal arm segments, the small sections of subgenomes A/A' within B and vice versa may represent more ancient exchanges in the lineage and/or incomplete lineage sorting (Figure 5) [72, 73]. This lack of recombination supports the theory that 'Montmorency' may have formed by one gamete from a *P. fruticosa*-like ancestor and one gamete from a *P. avium*-like ancestor. Second, since A and A' were readily distinguishable according to gene and repetitive sequence differences (with the exclusion of chr8A and chr8A'), this implies very few homoeologous exchanges (mixing of genomes) have occurred between these more similar subgenomes. This finding is once again consistent with the *P. fruticosa*-like progenitor being an allopolyploid, and the gamete that gave rise to 'Montmorency' resulted from preferential chromosome pairing of the two subgenomes. Chr8A and 8A' may be an exception; however, it is unclear whether this is due to subgenome mixing, disproportionately high sequence similarity among these chromosomes, assembly artifacts, or a combination of one or more scenarios. Nevertheless, the observations herein are consistent with the possibility that 'Montmorency' was formed from the fusion of a reduced gamete from a *P. fruticosa*-like ancestor and an unreduced gamete from a *P. avium*-like ancestor [8]. In the case of 'Montmorency,' chloroplast data identified the *P. fruticosa*-like progenitor as the maternal parent [6].

It follows the integrity of the three 'Montmorency' subgenomes is not likely to be transmitted to the next generation. If the A and A' chromosomes preferentially pair, crossing over will result in a patchwork of regions exchanged between these two homoeologous genomes. Additionally, cytological analysis of meiotic pairing for 'Montmorency' shows lack of complete bivalent pairing: quadrivalents, trivalents, and univalents occur [15, 74]. Segregation data support primarily "homologous" pairing (A with A' and B with B); however genetic results are consistent with occasional homoeologous pairing [5, 16]. This suggests that homoeologous exchanges could occur between the A and A' subgenome chromosomes and the B subgenome chromosomes, further eroding the subgenome integrity in 'Montmorency' offspring.

Poor fertility supported by evidence of irregular meiosis and occasional tetrasomic inheritance are prevalent in sour cherry germplasm [75]. In general, the quantity of fruit set is vastly below what the plant could support. Commercial cultivars, such as "Montmorency," are the exception as this cultivar can achieve a "full crop" by setting approximately 30% of its fruit. Indeed, it is the high fertility in 'Montmorency' that led it to be the predominant cultivar in the USA. However, even in crosses between two productive cultivars, the low fertility in the offspring is clear. For example, in one study, the mean fruit set and pollen germination for the German sour cherry cultivar Schattenmorelle and the Hungarian cultivar Érdi Bötermö was 16.0% and 13.4%, respectively, for fruit set, and 18.5% and 8.0% respectively for pollen germination [9]. When these two cultivars were crossed, the mean values for fruit set and pollen germination of the 86 offspring were just 6.8% and 6.6%, respectively. Both these values are far below what is needed for a commercial crop. As such, breeding for higher fruit set is a major but challenging goal for the MSU sour cherry breeding program.

In the present study, we provide evidence the allotetraploid event from which the 'Montmorency' lineage is descended occurred less than 1.61 million years ago (mya)—the lowest

divergence estimate between a 'Montmorency' subgenome and a representative progenitor (Figure S19). Peculiarly, the estimates range from 1.61 to 2.07 mya between 'Montmorency' subgenomes A and B and *P. fruticosa* and *P. avium*, respectively, but the estimate between *P. fruticosa* and A' is 4.48 to 4.51 mya. Compared to the first *P. fruticosa* subgenome and 'Montmorency' subgenome A, this would indicate the second subgenome within this *P. fruticosa* accession and A' are far more divergent from each other. It is readily understood individual accessions used in such estimates can significantly impact results. Thus, to understand this large difference in 'Montmorency' subgenome A' and *P. fruticosa*, further study of population structure and genotypic diversity in *P. fruticosa* and sour cherry is needed.

The estimates reported herein are largely in line with previous molecular dating analyses for *Prunus* species. The present study used calibration points informed by Xiang et al. These authors included eight *Prunus* species and several other taxa in the Amygdaleae in a large-scale transcriptomic analysis of divergence times in the Rosaceae. The youngest *Prunus* node of this study, which was the most recent common ancestor (MRCA) between *P. persica* (peach) and *Prunus dulcis* (almond), was estimated at about 8 million years [2]. Yet, these species are close enough in relation to form fertile, interspecific hybrids [48]. The MRCA for the entire Amygdaleae was determined to be approximately 30 mya [2]. In separate analyses, Chin et al. and Baek et al. estimated the maximum divergence of *Prunus* to be much higher, at roughly 56 and 66 mya, respectively; however, these studies used either limited sequence data or species in comparison to Xiang et al. [1, 76]. Despite the sizeable ranges in estimates between studies, these comparisons underline that sour cherry is a relatively young hybrid.

Though not considered recent by some standards, sour cherry is a long-lived perennial species; therefore, while 1.61 million years represents many generations and opportunities for recombination for an annual species, the same time span equates to far fewer generations for sour cherry. For example, the first written record of 'Montmorency' was in the 17th century, making it ~400 years old [77]. Thus, this sour cherry lineage may be considered younger than short-lived polyploids in terms of absolute number of generations. It must be reemphasized, however, that progenitor representatives used in dating analyses may greatly affect estimates. Additionally, since sour cherry has formed multiple times, the timing of hybridization for certain lineages would be expected to vary [6]. Still, given these results and speculations, it would be reasonable to suggest sour cherry exhibits the behavior of a neopolyploid actively undergoing the process of diploidization. Cytological and genetic data are consistent with a neopolyploid as irregular meiosis visualized as trivalents and quadrivalents is documented in several sour cherry genotypes, and genetic data indicate that although disomic inheritance is more common, tetrasomic inheritance also occurs [9, 15, 16, 74, 78]. Such events can result in aneuploid gametes, endosperm imbalance, embryo abortion, and unsuccessful seed and fruit development. Neoallopolyploids are especially prone to such issues depending on the likeness of their progenitors' genomes [79]. If progenitors are relatively divergent, preferential pairing of homologous chromosomes and *not* homoeologous chromosomes will more likely occur, and the polyploid will exhibit frequent bivalent formation during meiosis and diploid-like segregation. However, if progenitors are more similar in terms of their sequence divergence and collinearity, homoeologous chromosomes may exchange genetic information and create imbalanced genomic combinations. This could result in pairing irregularities during meiosis and reduced fertility in subsequent generations.

Theoretically, over time, selection would act promptly against these infertile genetic combinations and the polyploid would eventually "diploidize" [79–81]. However, the diploidization process in sour cherry has likely been repressed by the prevalent intercrossing with both its progenitor species, as natural hybrids among the three cherry species (*P. cerasus*, *P. avium*, and *P. fruticosa*) are common [11]. In certain scenarios, progeny from a cross of sour cherry (AA'BB) and *P. avium* (BB) or *P. fruticosa* (AAA'A') would likely inherit imbalanced subgenome combinations. Furthermore, human selection may also have constrained the diploidization process. For example, in Hungary and Romania, human selection of vegetatively propagated century-old landrace cultivars Pándy and Crişana, respectively, favored fruit quality over fruit quantity, as these landrace cultivars have extremely low fruit set but highly desirable fruit quality [82]. Taken together, the evolutionary history of sour cherry illustrates the intersection of fundamental principles of natural selection and human influence.

Finally, an intriguing question moving forward is how many separate allotetraploid lineages led to what we consider to be sour cherry. Chloroplast data indicates sour cherry was independently formed at least twice: although *P. fruticosa* is more commonly the maternal parent, cultivars with *P. avium* as the maternal parent have also been identified [6]. As the native distributions of *P. fruticosa* and *P. avium* overlap, sour cherry could have been formed by a reduced gamete from *P. fruticosa* and an unreduced gamete from *P. avium*, as speculated above for "Montmorency." However, it is also possible sour cherry lineages could be the product of a triploid bridge. In this scenario, a hybrid of a *P. fruticosa*-like and *P. avium*-like species (*Prunus* × *mohacsyana*) would produce an unreduced gamete (3×) that either fertilized or was fertilized by a *P. avium* gamete (1×) to form allotetraploid *P. cerasus*. In accordance with the first scenario, unreduced pollen (based on grain size) has been documented in *P. avium*, although the frequency of this phenomenon has not been estimated [12]. In favor of the second scenario, all seedlings produced in a controlled cross of tetraploid *P. fruticosa* and diploid *P. avium* were triploid, and triploid hybrids have been verified in nature [4, 11, 83]. A thorough survey of the frequency of unreduced gamete formation and further assessment of triploid fertility in natural populations may clarify if one scenario is favored over the other. However, irrespective of these different evolutionary trajectories, it is important to consider the 'Montmorency' subgenome structure reported herein may be unique to this genotype and therefore may not represent the subgenome structure of a range of sour cherry accessions.

In conclusion, here we present the 'Montmorency' reference genome and a *P. fruticosa* draft genome to be used for future comparative studies in the genus *Prunus* and beyond. Our characterization of the 'Montmorency' subgenomes provides a valuable resource for exploring the evolutionary history of sour cherry with wider implications for questions surrounding allopolyploidization and neopolyploidy. These resources will aid in developing targeted breeding strategies for sour cherry and allow investigation into whether an imbalanced subgenome composition leads to the low fruit set prevalent in the species.

Materials & methods

Collection of materials

Young leaves for gDNA libraries were collected fresh (for Hi-C, *P. cerasus* 'Montmorency' only) or flash-frozen in liquid nitrogen and

stored at −80°C until extraction (for both PacBio SMRT sequencing and Illumina HiSeq) from a clone of 'Montmorency' and an accession of *P. fruticosa* growing at Michigan State University's Clarksville Research Center in Clarksville, Mich., in the spring of 2019. Tissues for RNAseq and Nanopore cDNA libraries were collected the same year from 'Montmorency' and flash frozen in liquid nitrogen and stored at −80°C until extraction. For *P. cerasus* "Montmorency," biological replicates were collected for RNA extraction and RNAseq/Nanopore cDNA-sequencing from the tissues indicated in Figure S20. Fruit tissue was collected and staged based on the double-sigmoidal growth curve characteristic of *Prunus* sp. [84, 85]. Vegetative and floral meristems in typical floral positions were broadly characterized as prefloral initiation, transitioning to floral, and organ differentiation according to histological sectioning (data not shown). For *P. fruticosa*, biological replicates were collected for extraction and RNAseq from young leaves, whole flowers at balloon stage, whole fruits in stage I, whole fruits at the end of stage II, and whole fruits in stage III.

DNA extraction, library preparation, and sequencing

Extraction of high molecular weight (HMW) DNA from young leaves was done at the University of Georgia's Genomics and Bioinformatics Core using a nuclei extraction method for both *P. cerasus* 'Montmorency' and *P. fruticosa*. From this HMW DNA, a large SMRTbell library (>30 kb) was prepared and sequenced on six flow cells of a PacBio Sequel II machine for each species, producing 61.98 Gb of data (100× coverage of 621 Mb estimated genome size) for *P. cerasus* 'Montmorency' and 48.13 Gb of data (90.5× of 532 Mb estimated genome size) for *P. fruticosa*. A second batch of young leaves was used to extract DNA for short-read sequencing using the DNEasy Plant kit (Qiagen, Valencia, CA). An Illumina TruSeq gDNA library was prepared and sequenced for both species on a HiSeq4000 at the Research and Technology Facility (RTSF) of Michigan State University (MSU), and approximately 34.7 Gbp of data was produced (56× coverage) for *P. cerasus* 'Montmorency' and 40 Gbp of data was produced for *P. fruticosa* (75.2× coverage). A third collection of fresh young leaves from *P. cerasus* 'Montmorency' was shipped overnight on ice to Phase Genomics (Seattle, WA), where a Hi-C Proximo Library was created with a DpnII restriction enzyme. The Hi-C library was 150-bp paired-end sequenced on a HiSeq4000 instrument at MSU's RTSF, producing 93.3 Gbp of data or 150.5× physical coverage.

RNA extraction, library preparation, and sequencing

All RNA was extracted using a CTAB-based protocol [86]. One Illumina TruSeq Stranded mRNA library was prepared for each of the two to three biological replicates per tissue used for RNAseq (Figure S20, excl. c, d, g, j–l for 'Montmorency'). Young leaves, whole flowers at balloon stage, whole fruits in stage I, whole fruits at the end of stage II, and whole fruits in stage III, or RNA from five tissue-types total, were sequenced for *P. fruticosa* at MSU's RTSF. 24 RNAseq libraries for 'Montmorency' and 14 RNAseq libraries for *P. fruticosa* were 150 bp paired-end sequenced on a HiSeq4000, producing between 25 and 36.6 million reads per library. One library from 'Montmorency' (replicate of whole fruits at the end of stage II) and one library from *P. fruticosa* (replicate of whole fruits at stage I) were deemed contaminants/low-quality based on unusually poor alignment to the respective assemblies and were excluded from downstream analyses. In addition to most of the tissue types used for RNAseq (excluding mature fruit mesocarp and mature fruit exocarp), whole fruits at the beginning of phase II, vegetative apices in early summer, transitioning apices in midsummer, mature leaves in late spring, mature leaves in midsummer, and floral apices during organ development (confirmed via histological observations—unpublished results) were included in the Nanopore cDNA sequencing (12 tissue-types total) for 'Montmorency' (Figure S20 c, d, g, j–l).

Assembling the genome of *P. cerasus* 'Montmorency'

Illumina gDNA reads' 25-mers were counted using Jellyfish v 2.2.10 and the resulting histogram was visualized using GenomeScope v2.0 [62, 87]. Canu v 1.9 was used to assemble the PacBio reads [88]. Reads below 5 kb in length were excluded from the assembly process and batoptions were set to "-dg 3 -db 3 -dr 1 -ca 500 -cp 50" to utilize the heterozygosity to assemble all possible haplotypes. The assembly was polished with the Illumina gDNA reads iteratively four times with Pilon v1.23 [89]. Reads were aligned to the assembly with Bowtie2 v 2.3.4.2 until there was no further improvement in read alignment [90]. The polished assembly was visualized with Bandage, a BUSCO analysis was done to assess gene space, and Merqury was used to determine phasing quality and genome completeness [43, 44, 91].

Scaffolding the *P. cerasus* 'Montmorency' assembly

Preliminary scaffolding results indicated two full haplotypes (16 pseudomolecules) had been well assembled while a third (8 pseudomolecules) experienced sudden and frequent drops in Hi-C signal along the diagonal, likely due to haplotype switching and the 3D-DNA software attempting to position multiple, noncollapsed alleles next to collapsed sequence (i.e. heterozygous bubbles). The rest of the assembly was considered unanchored. We posited that removal of similar haplotypes would lead to a tidier representation of the third group. Purge_haplotigs v1.1.2 was used with a cutoff value of 99% alignment to set aside very similar alleles in the assembly prior to scaffolding [45]. Since the contig lengths of the 16 well-assembled pseudomolecules were much larger than the other 8, the lengths of the purged contigs were manually inspected to avoid removal of haplotypes that were formerly intact. As a result, purged contigs greater than 400 kb were added back into the assembly prior to scaffolding. After this size selection, we verified we had removed mostly alternative alleles of the three assembled groups in two ways. First, BUSCO analyses showed completeness to be very high (>90%) and duplication to be very low (<3.0%) for each of the three pseudomolecule groups (Table 1). Second, a k-mer assessment using Merqury indicated 25-mers in the remaining purged contigs were found only once and at all relative multiplicities (1×–4×) in the Illumina read dataset. If 25-mers at any multiplicity had been present in the purged contigs more than once, this would suggest more than one haplotype had been removed from the assembly (Figure S3). After reducing the complexity of the assembly as described above, Hi-C reads were aligned to the assembly using BWA v 0.0.7.17 within the Juicer pipeline [92, 93]. The -S flag was set to exit the pipeline early after production of the merge_nodups.txt file. This file was then used as input for 3D-DNA, which was run with "- - editor-repeat-coverage 5" to prevent areas of the assembly with higher levels of coverage (due to ploidy) being flagged as "junk" [94]. The output was a Hi-C matrix (.hic) that was manually edited in JuiceBox Assembly Tools to correct misassemblies [95]. Following manual editing, the new .hic and

.asm files were used to create the chromosome-level.fasta file with the script run-asm-pipeline-post-review.sh from the 3D-DNA suite of tools. The resulting superscaffolds (chromosomes) were named according to a syntenic comparison with a peach genome and subgenome assignments based on 25-mer groups [46, 47].

Assessing repeat content and quality

The LTR assembly index (LAI) was determined by first identifying TEs with LTR_FINDER_parallel and LTRharvest, then combining the output files and using it as input for LTR_retriever [96–99]. The EDTA pipeline was used to estimate repeat content and produce a custom repeat library [100].

Marker mapping and visualization

582 marker sequences from a genetic map of an F1 cross of two sour cherries ('Montmorency' × 25–02–29, n = 53) were downloaded from the Genomic Database for Rosaceae (GDR; https://www.rosaceae.org/) and mapped to the 24 superscaffolds of the assembly using BLAST+ v 2.2.31 [16, 67, 101, 102]. Markers mapping more than four times or below 80% of their length were filtered from the dataset, resulting in 545 unique markers' mappings visualized in ALLMAPS [103].

Annotation of the genome of *P. cerasus* 'Montmorency'

We used multiple sources of high-quality data to annotate the *P. cerasus* 'Montmorency' genome, including RNAseq and long-read cDNA-PCR sequencing using a GridION machine (Oxford Nanopore Technologies), and manually curated protein databases [65, 104]. All data were processed to produce .gff3 files which were used as input for MAKER [105].

Preparation of RNAseq data for MAKER

Adapters and low-quality bases were removed from RNAseq reads (2–3 reps per tissue, 23 libraries total) with Trimmomatic v 0.39 [106]. Individual libraries, totaling 3.1 billion reads, were aligned to the 'Montmorency' genome assembly using default parameters in STAR v 2.7.3a [107]. Alignment rates were 94%+ per library. Approximately 33–39% of reads mapped to multiple locations, and random checks of several alignments confirmed these reads were aligning to homoeologous chromosomes and/or alleles (e.g. 1A and/or 1A' and/or 1B). SAMtools v1.9 was used to sort and index all .sam/.bam files [108]. All alignments were merged, and a transcriptome assembly was created using StringTie v 2.1.2 [109]. The transcriptome assembly was checked against raw RNAseq and protein alignments for improper fusions and breaks in potential genes with the Integrative Genomics Viewer (IGV) v 2.8.0, and parameters were adjusted accordingly in StringTie v 2.1.2 ("-m 200 -t -c 3 -f 0.05 -g 50") [110]. The final .gtf file was converted to .gff3 using the gffread function in Cufflinks v 2.2.1 [111] and the features "StringTie," "transcript," and "exon," were replaced with "est2genome," "expressed_sequence_match," and "match_part," respectively, for compatibility with MAKER v 2.31.10 [105].

Preparation of long-read RNA sequencing for MAKER

Nanopore reads were demultiplexed, trimmed, and filtered (reads <150 bp were dropped) with Porechop v 0.2.4 and NanoPack [112, 113]. 4.8 million reads were aligned at a rate of 89% to the 'Montmorency' genome assembly using minimap2 v 2.15 with the following parameters: "-N 5 -ax splice -g2000 -G10k" [114]. Sorting and indexing of .sam and .bam files was done with SAMtools v. 1.9. The transcriptome assembly was built using StringTie2 ("-m 150 -t -c 1 -f 0.05 -g 50"), and the .gtf was converted to .gff3 and features changed similarly to the RNAseq data prior to giving the data to MAKER.

Preparation of protein data for MAKER

Manually curated Uniprot viridiplantae protein sequences and Arabidopsis protein sequences from TAIR10 were downloaded in fasta format on 17 April 2021 and 26 February 2021, respectively [65, 104]. Sequences were aligned using Exonerate v 2.2.0 and the five best matches for each alignment were kept in the following format: —ryo ">%qi length=%ql alnlen=%qal\n>%ti length=%tl alnlen=%tal\n" [115]. The resulting .gff2 was converted to a .gff3 using the script process_exonerate_gff3.pl, and the features "exonerate:protein2genome:local," "mRNA," and "CDS" were changed to "protein2genome," "protein_match," and "match_part," respectively, for compatibility with MAKER [116].

Running MAKER iteratively

MAKER was run similarly to Bowman et al. with the evidence detailed above and the custom repeat library created from the EDTA pipeline for masking [60, 100, 105]. The output transcript and protein fasta files were extracted from MAKER's first run and gene predictions with AED (Annotation Edit Distance) values <= 0.2 were used to train AUGUSTUS v 3.3.2 [117]. Subsequently, MAKER was run a second time, with features from the first run's .gff3 file being passed as hints to AUGUSTUS. After the run was complete, the resulting .gff3 and transcript and protein fasta files were again extracted as previously detailed [60].

Polishing and filtering the annotation

Gene predictions output by MAKER's second run were additionally processed to improve the annotation. First, the protein sequences were searched against the Pfam-A database using hmmscan v 3.1b2, and the predictions containing no known protein domains were removed [118–120]. Second, defusion was run to identify putatively fused genes on the 24 chromosomes of the assembly (chr1[A, A', B] – chr8[A, A', B]) [61]. Defusion specializes in identifying potential tandem duplicates but does not typically identify fusions of genes with divergent intron–exon structures (chimeric fusions). However, it can extract and locally reannotate any sequences when given coordinates and breakpoint(s). Therefore, in addition to automatically identifying candidate gene fusions with defusion, we used an alternative method to identify candidates of the second class of gene fusions. Putatively fused loci from the initial MAKER gene set were identified when two or more distinct proteins-only gene predictions overlapped with gene predictions from the transcript plus protein MAKER run (see https://github.com/goeckeritz/Montmorency_genome, identify_fusion_candidates_w_PROTEIN_ONLY_datasets.bash). These candidate fusions were checked alongside those identified by defusion in IGV and break points were manually added to the .brk file as necessary (Figure S10). The defused annotation was then filtered again to remove predictions lacking a Pfam domain. Lastly, putative *DAM* genes, *S*-RNases, and SFBs were manually annotated with Apollo v. 2.6.5 [68]. The old gene models were then removed with agat and replaced with the corrected models to produce the final annotation file [121].

Assigning functions to genes

Functional information was assigned to the .gff and protein and transcript .fasta files via a BLAST+ v 2.9.0 comparison

of amino acid sequences to a Uniprot database and several accessory scripts within MAKER (maker_functional_gff, maker_functional_fasta) [59, 65, 67]. Moreover, Pfam, PANTHER, TIGRFAM, InterProScan, and Gene Ontology (GO) database reference numbers or IDs were added to the .gff file by scanning the amino acid sequences with InterProScan and using the MAKER accessory script ipr_update_gff [59, 118, 122–126]. Only hits with p values $<1.0 \times 10^{-10}$ were kept. Pfam, PANTHER, and TIGRFAM hits are also provided as separate .csv files that include the gene ID and functional descriptions.

Draft genome assembly and polyploid type determination for *P. fruticosa*

A similar approach to the *P. cerasus* 'Montmorency' genome was taken to assemble a draft genome of *P. fruticosa* but with some differences. Illumina gDNA reads' 25-mers were counted using Jellyfish v 2.2.10 and the resulting histogram was visualized using GenomeScope v2.0 [62, 87]. The draft genome was created for the sole purposes of identifying *P. fruticosa*-like regions of the *P. cerasus* 'Montmorency' genome and estimating a divergence date between the two species via ortholog analyses. Parameters in Canu v 1.9 were set similarly to those used for *P. cerasus* 'Montmorency' so that multiple haplotypes could be assembled from the PacBio reads [88]. This provided assurance that if *P. fruticosa* comprises two different ancestral genomes, alleles from both subgenomes would most likely be assembled and included in ortholog analyses. The draft assembly was polished with the Illumina gDNA reads iteratively three times with Pilon v1.23 [89]. A BUSCO analysis was done to assess gene space, and Merqury was used to determine phasing quality and genome completeness [43, 44].

The K_s analysis was done using a previously developed pipeline [63]. First, syntenic blocks and homo/homoeologs were identified using JCVI-MCScan v1.2.4 [64]. K_a (dN) and K_s (dS) values for gene pairs were obtained with MUSCLE v.3.8.31, PAL2NAL v14, and PAML v4.9h [127–129]. The dataset was then imported into R version 4.2.2, and comparisons of a gene with itself ($K_s = 0$) as well as gene pairs with a K_s or $K_a > 3.0$ were removed. Ggplot2 was used to plot a histogram of K_s frequencies, and function ggplot_build was used to access the frequency data and precisely determine at which K_s values the frequencies peaked [130]. The graph axes were limited for readability.

Annotation of a *P. fruticosa* draft genome

The polished *P. fruticosa* contigs were annotated using a similar pipeline to the one described for *P. cerasus* 'Montmorency'. The RNAseq reads from *P. fruticosa* leaves, flowers, and developing fruits at 3 stages were processed in an identical manner to the *P. cerasus* RNAseq reads to produce a .gff3 file as input for MAKER. Uniprot and TAIR10 protein databases were aligned to the *P. fruticosa* contigs and processed similarly as well. Additionally, predicted proteins from *P. cerasus* 'Montmorency' with AED values <0.3 were also used as evidence for MAKER. Gene finders SNAP and AUGUSTUS were trained on the .gff3 from the first MAKER run before using them for the second run [117, 131]. The final predicted gene set was filtered to keep predictions with known Pfam domains but lacking known TE domains. Due to limited resources, no manual annotation was performed on the *P. fruticosa* contigs. Assigning gene function to *P. fruticosa* gene predictions was done similarly as the *P. cerasus* 'Montmorency' annotation.

Syntenic comparison of the *P. cerasus* 'Montmorency' assembly with *P. persica*

A synteny analysis was conducted between the 24 superscaffolds (chromosomes) of the *P. cerasus* 'Montmorency' assembly and a *P. persica* genome using the SynMap tool within the Comparative Genomics Platform (CoGe) [46, 47]. Coding sequences (unmasked) of 'Montmorency' and *P. persica* were compared with default settings. Based on these synteny results (Figure 1a**),** k-mer clustering, and phylogenomic comparisons of syntelogs as described below, the 24 superscaffolds of the *P. cerasus* 'Montmorency' were named chr1[A, A', B]–chr8[A, A', B].

Syntenic comparisons of the *P. cerasus* 'Montmorency' assembly and representative progenitor genomes

Macrosyntenic comparisons of 'Montmorency' with the *P. avium* 'Tieton' v2.0 genome, the *P. fruticosa* draft genome, and with itself were done using the MCScan package from JCVI [64, 132]. We first built .cds and .bed files for each genome from the coding sequence and .gff files, respectively, then used command "jcvi.compara.catalog ortholog" to generate a list of syntenic blocks between either *P. avium* and 'Montmorency', *P. fruticosa* and "Montmorency," or between 'Montmorency' and itself. All karyotype figures were constructed with the "jcvi.graphics.karyotype" command.

The macrosyntenic comparison of the *P. fruticosa* draft assembly with the *P. avium* 'Tieton' genome was done similarly. The command "jcvi.compara.synteny depth–histogram" was used to create the 4:1 histogram pattern figure (Figure S14).

K-mer clustering

We used Jellyfish 2.2.10 to count 25-mers on each chromosome of "Montmorency." 25-mers with fewer than 10 occurrences per chromosome were removed from the dataset and files were imported into R 4.2.0 [87, 133]. Further filtering was done if the 25-mers were not 1) present at two times or more abundance in a homoeolog than in one of its sisters, and 2) present on all 24 chromosomes. We used the R function hclust() method "complete" to hierarchically cluster the 25-mers in the 24 and 22 chromosomes (excluding chromosome 8A and 8A') and to construct dendrograms. The package "pheatmap" was used to create heat maps [134]. For the 25-mer clustering analysis to differentiate the A and A' subgenomes only, the analysis was completed as described above with only 14 chromosomes (1[A, A']–7[A, A']).

25-mer density per 1 Megabase (Mb) window of each chromosome was calculated as follows: (Number of group "X" 25-mers in a 1 Mb window × 25 bp) / (1 Mb). This is equivalent to the proportion of bases occupied by group "X" 25-mers in each 1 Mb window. Chromosome plots of 25-mer group densities were made in ggplot2 [130].

Assessing read depth of the 'Montmorency' subgenomes

Read depth per position was assessed by mapping the 'Montmorency' Illumina reads to the assembly with Bowtie2 v. 2.3.4.2 default settings [90]. SAMtools v 1.9 was then used to sort and calculate depth at every position [108]. From there, subgenomes were separated and concatenated end-to-end from chr1[A, A', B] to chr8[A, A', B]. For data reduction purposes, the average read depth per 1000 sites was plotted along the length of each subgenome and the median read coverage of the full genome was overlaid on

the data (Figure S9). Plots were created with R v. 4.2.1 and ggplot2 [130, 133].

Phylogenomic comparisons of syntelogs to identify progenitor relationships

We used syntelogs (syntenic orthologs) between 'Montmorency' and each progenitor to assign regions of the assembly as either *P. fruticosa*-like or *P. avium*-like. Peptide and coding sequences of *Malus* × *domestica* "Gala," *P. persica* "Lovell," *P. avium* "Tieton," *P. fruticosa* from the present study, *P. cerasus* 'Montmorency' subA, subA', and subB were either downloaded from GDR or generated as described above [23, 47, 135]. Orthogroups (groups of orthologous genes) were identified between these seven "species" using OrthoFinder v. 2.5.4 [66]. Multiple sequence alignments (MSAs) of orthogroups were done within OrthoFinder using MAFFT v. 7.480 with no alignment trimming (–z) [136]. Only orthogroups including all seven species were used for downstream analyses. Protein sequence alignments were converted to nucleotide alignments using PAL2NAL v. 14.1 [128], and raw cds alignments were trimmed with trimAl v. 1.4.1 using flag-automated1 [128, 137]. Alignments before and after trimming were visualized with MView to ensure high quality of the resulting alignments [138]. A phylogenetic tree for each orthogroup was created with RAxML-NG v 1.0.0 using the gamma + GTR model, 500 bootstrap replications, and an apple ortholog outgroup [139]. A two-column list of 'Montmorency' orthologs was used as input for PhyDS to identify gene–gene sister relationships to either a *P. fruticosa* or *P. avium* ortholog with bootstrap values (BSV) of at least 80% [140]. PhyDS requires a two-column paralog list to extract relationships from phylogenetic trees, but no combination of two genes should be listed more than once. Thus, we extracted the IDs of all 'Montmorency' genes in all orthogroups and simply duplicated this list for the second column. We manually examined at least ten trees with a phylogenetic tree viewer to ensure the paralog list and phyDS scripts were behaving as expected and extracted relationships where a 'Montmorency' ortholog was sister to a single representative progenitor gene (BSV > = 80%) using basic Unix commands (see https://github.com/goeckeritz/Montmorency_genome) [141]. At the same time, syntenic gene pairs between A, A', and B versus each representative progenitor (a total of six comparisons) were identified using the default settings of the python version of MCScan [64]. The orthologous relationships identified with OrthoFinder fitting the above criteria were interjoined with the syntenic orthologous relationships identified by MCScan using R version 4.2.1 [133]. These high-confidence syntenic orthologs were mapped back to the 'Montmorency' assembly and labeled as either "*P. avium*-like" or "*P. fruticosa*-like." The R package chromoMap was used to visualize these results [142].

Estimating divergence time of *P. cerasus* 'Montmorency' subgenomes from representative progenitor species

RAxML-NG was used to create phylogenetic trees for single-copy orthogroups as described above. Based on current knowledge of Rosaceae phylogenetics, the topolog(ies) for most accurately estimating the divergence of 'Montmorency' subgenomes and its representative progenitors from their most recent common ancestor (MRCA) should place apple as the outgroup and peach as sister to the cherry lineage [1–3, 69]. Additionally, to calculate divergence time without error from possible homoeologous recombination, only orthogroups where 'Montmorency' homoeologs from each subgenome were predominantly sister to one progenitor over the other (*P. avium* or *P. fruticosa*) were included in the r8s analysis. This assumed that any previous homoeologous recombination taking place between the subgenomes did not replace >50% of the original sequence contributed by the progenitor. As a result, in any given tree, one homoeolog would not be able to pair with *P. avium* or *P. fruticosa* (i.e. there are two representative progenitors but three 'Montmorency' subgenomes). Thus, we were prepared to calculate node ages of multiple topologies to obtain MRCA divergence estimates for each subgenome and its most closely related progenitor. This required each single-copy orthologous gene tree (n = 336) to be manually inspected and the frequencies of each topology noted. The two most frequent topologies (Figure S19) made it clear which progenitor subgenome A, A', and B was most related to: subgenome B orthologs were almost always sister to *P. avium* while A and A' orthologs were nearly equally likely to be sister to *P. fruticosa*. Orthogroup sequence alignments showing the two most frequent topologies were separately concatenated for all "species" (n = 7) regardless of bootstrap support. Phylogenetic trees for the two concatenated alignments were created with RAxML-NG as described above [139]. Bootstrap replications (500 per topology) were used to calculate node age estimates and confidence intervals with r8s and R v. 4.2.1 [133, 143]. Based on Xiang et al., the *Malus/Prunus* node was fixed at 95 million years ago (mya) and the peach/cherry node was constrained to a minimum age of 10 mya for both analyses. The smoothing parameter was set to 1, rate was set to gamma, and divergence time was set to penalized likelihood with the TN algorithm. 128 633 sites were used to determine divergence times of topology A in Figure S19**,** and 78 183 sites were used for topology B.

Identification of the DAM genes and S-alleles

Dormancy-Associated MADS-box (*DAM*) genes were identified in 'Montmorency' and the *P. fruticosa* contigs with BLAST+ v. 2.9.0, using *P. persica DAM1–DAM6* coding sequences from NCBI's GenBank as query and genomic sequence or transcripts from the MAKER pipeline as the target [144]. The sequence IDs of the *P. persica DAMs* used as query were: DQ863253.2, DQ863254.1, DQ863256.1, DQ863250.1, AB932551.1, AB932552.1. Only matches with p value <1.0e-10 were kept with a max of 24 matches per query. This BLAST+ analysis identified *DAM* candidates in 'Montmorency' on chr1A, chr1A', chr1B, and unanchored scaffold 2998, and in *P. fruticosa* on seven different contigs, some containing only partial haplotypes of the expected six tandem genes with occasional erroneous fusions. Only genomic regions containing full haplotypes (six tandem *DAM* genes) were manually annotated using Apollo. Chr1A, chr1A', and chr1B in 'Montmorency' contained a full haplotype each, and two large contigs (8 and 33) contained a full haplotype each in *P. fruticosa*. Only these genes were used for phylogenetic comparisons with *DAMs* in other *Prunus* sp. These regions were confirmed to be syntenic between 'Montmorency' and *P. fruticosa,* and 'Montmorency' and *P. persica* (Figure 5a**,** Figure 1a).

S-alleles (S-RNase linked with an F-box protein/SFB) were identified in 'Montmorency' and *P. fruticosa* similarly to the *DAMs* with BLAST+ v. 2.9.0 and the following complete cds sequences from NCBI's GenBank were used as query: P.cerasusS_{36b}-RNase, P.cerasusS_{36b3}-RNase, P.cerasusS_{36b2}-RNase, P.aviumS_{6}-RNase, P.cerasusS_{35}-RNase, P.cerasusS_{36}a-RNase, P.cerasusS_{13m}-RNase, *P. cerasus*SFB$_{36b}$, P.cerasusSFB$_{36b3}$, P.cerasusSFB$_{36b2}$, P.aviumSFB$_{13}$, P.cerasusSFB$_{35}$, and P.cerasusSFB$_{36a}$ (Table S5) [144]. To be considered a full allele, an S-RNase and SFB with >90% identity to a query sequence had to be tightly linked (<100 kb apart).

DAM phylogenetic comparisons

Coding sequences for *Arabidopsis thaliana* SEP3 and the six *P. avium DAM* genes were downloaded from NCBI (Table S4). Sequence alignments were done with MUSCLE/3.8.31 using default settings and the phylogenetic tree was constructed using RAxML-NG/1.0.0 [127, 139]. We used the PROTGTR+G model to infer the best maximum likelihood (ML) tree and mapped 500 bootstrap replicates onto the best ML tree to create the final phylogeny.

S-RNase and SFB alignments

We downloaded the protein and coding sequences for the 'Montmorency' S-haplotypes from NCBI (Table S5). *Prunus fruticosa* S-alleles were aligned to their 'Montmorency' counterparts with MUSCLE/3.8.31 and alignment figures were made using the R package gggenes (Figure S16) [130].

Acknowledgements

We thank Dr. Shujun Ou for his assistance in running EDTA, and all members of the Aiden Lab and Dr. Ching Man Wai for their help in operating Juicer and 3D DNA. We are also grateful to Dr. Jose Ramon Planta and Dr. Jie Wang for their help with defusion, and to Dr. Nathan Dunn and Dr. Garrett Stevens for guidance with installing and navigating Apollo. Finally, we appreciate Chloe Grabb for her assistance with manual gene annotation and Dr. Pat Edger for his advice on all phylogenomic analyses. This research was funded by AgBioResearch Project GREEEN grant GR19-046, the United States Department of Agriculture National Institute of Food and Agriculture (USDA-NIFA) project 2014-51181-22378 and USDA-NIFA HATCH project 1013242.

Authors' contributions

C.A.H., A.F.I., and R.V. conceptualized the experiments. C.Z.G. performed genome assembly, annotation, subgenome assignment using orthologs, and divergence time estimate analyses. K.B.R. performed the k-mer hierarchal clustering, K_s analysis, synteny, *DAM* gene phylogenetic analyses, and S-allele alignments. K.L.C. provided expertise and assistance with annotation and contributed code. C.Z.G., K.B.R., and A.F.I. wrote the manuscript. All authors assisted with editing the manuscript.

Data availability

The datasets supporting the conclusions of this article are available in the Genomic Database for Rosaceae (GDR) https://www.rosaceae.org/ and NCBI's Sequence Retrieval Archive (SRA) under BioProject number PRJNA922242 (raw sequence data). Scripts and example files associated with genome assembly, annotation, gene function predictions, subgenome assignment analyses, and r8s estimates can be found at https://github.com/goeckeritz/Montmorency_genome. Scripts for k-mer hierarchal clustering, K_s analysis, synteny, *DAM* gene phylogenetic analyses, and S-allele alignments can be found at https://github.com/KEBRhoades/Montmorency_genome.

Conflict of interests statement

None declared.

Supplementary Data

Supplementary data is available at *Horticulture Research* online.

References

1. Chin S-W, Shaw J, Haberle R *et al*. Diversification of almonds, peaches, plums and cherries—molecular systematics and biogeographic history of *Prunus* (Rosaceae). *Mol Phylogenet Evol*. 2014;**76**:34–48.
2. Xiang Y, Huang CH, Hu Y *et al*. Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Mol Biol Evol*. 2017;**34**:262–81.
3. Hodel RGJ, Zimmer E, Wen J. A phylogenomic approach resolves the backbone of *Prunus* (Rosaceae) and identifies signals of hybridization and allopolyploidy. *Mol Phylogenet Evol*. 2021;**160**:107118.
4. Olden EJ, Nybom N. On the origin of *Prunus cerasus* L. *Hereditas*. 1968;**59**:107118–59.
5. Beaver JA, Iezzoni AF. Allozyme inheritance in tetraploid sour cherry (*Prunus cerasus* L.). *J Am Soc Hortic*. 1993;**118**:873–7.
6. Brettin TS, Karle R, Crowe EL *et al*. Chloroplast inheritance and DNA variation in sweet, sour, and ground cherry. *J Hered*. 2000;**91**:75–9.
7. Iezzoni AF. Acquiring cherry germplasm from central and Eastern Europe. *HortScience*. 2005;**40**:304–8.
8. Bird KA, Jacobs M, Sebolt A *et al*. Parental origins of the cultivated tetraploid sour cherry (*Prunus cerasus* L.). *Plants People Planet*. 2022;**4**:444–50.
9. Wang D, Karle R, Iezzoni AF. QTL analysis of flower and fruit traits in sour cherry. *Theor Appl Genet*. 2000;**100**:535–44.
10. Barać G, Ognjanov V, Vidaković DO *et al*. Genetic diversity and population structure of European ground cherry (*Prunus fruticosa* pall.) using SSR markers. *Sci Hortic*. 2017;**224**:374–83.
11. Macková L, Vít P, Urfus T. Crop-to-wild hybridization in cherries—empirical evidence from *Prunus fruticosa*. *Evol Appl*. 2018;**11**:1748–59.
12. Iezzoni AF, Hancock AM. A comparison of pollen size in sweet and sour cherry. *HortScience*. 1984;**19**:560–2.
13. Mochalova OV. Siberian gene pool of steppe cherry polyploids (*Prunus fruticosa* pall.): cytomorphological estimation and prospects for breeding. *BIO Web Conf*. 2020;**24**:00056.
14. Sebolt AM, Iezzoni AF, Tsukamoto T. S-genotyping of cultivars and breeding selections of sour cherry (*Prunus cerasus* L.) in the Michigan State University sour cherry breeding program. *Acta Hortic*. 2017;31–40.
15. Wang D. *RFLP Mapping, QTL Identification, and Cytogenetic Analysis in Sour Cherry*. 1998.
16. Cai L, Stegmeir T, Sebolt A *et al*. Identification of bloom date QTLs and haplotype analysis in tetraploid sour cherry (*Prunus cerasus*). *Tree Genet Genomes*. 2018;**14**:1–11.
17. Wöhner TW, Emeriewen OF, Wittenberg AHJ *et al*. The draft chromosome-level genome assembly of tetraploid ground cherry (*Prunus fruticosa* pall.) from long reads. *Genomics*. 2021;**113**:4173–83.
18. Vanderzande S, Zheng P, Cai L *et al*. The cherry 6 + 9K SNP array: a cost-effective improvement to the cherry 6K SNP array for genetic studies. *Sci Rep*. 2020;**10**:1–14.
19. Kistner E, Kellner O, Andresen J *et al*. Vulnerability of specialty crops to short-term climatic variability and adaptation strategies in the Midwestern USA. *Clim Chang*. 2018;**146**:145–58.
20. Marino GP, Kaiser DP, Gu L *et al*. Reconstruction of false spring occurrences over the southeastern United States, 1901–2007: an increasing risk of spring freeze damage? *Environ Res Lett*. 2011;**6**:024015.

21. Unterberger C, Brunner L, Nabernegg S *et al.* Spring frost risk for regional apple production under a warmer climate. *PLoS One*. 2018;**13**:1–18.
22. Quero-Garciá J, Campoy JA, Barreneche T *et al.* Present and future of marker-assisted breeding in sweet and sour cherry. In: *Acta Horticulturae*. Vol 1. Acta Horticulturae: Yamagata, Japan, 2019,1–14.
23. Wang J, Liu W, Zhu D *et al.* Chromosome-scale genome assembly of sweet cherry (*Prunus avium* L.) cv. Tieton obtained using long-read and Hi-C sequencing. *Hortic Res*. 2020;**7**:1–11.
24. Rodriguez A, Sherman W, Scorza R *et al.* 'Evergreen' peach, its inheritance and dormant behavior. *J Am Soc Hortic*. 1994;**119**: 789–92.
25. Bielenberg DG, Wang Y, Fan S *et al.* A deletion affecting several gene candidates is present in the evergrowing peach mutant. *J Hered*. 2004;**95**:436–44.
26. Li Z, Reighard GL, Abbott AG *et al.* Dormancy-associated MADS genes from the EVG locus of peach [*Prunus persica* (L.) Batsch] have distinct seasonal and photoperiodic expression patterns. *J Exp Bot*. 2009;**60**:3521–30.
27. V da S F, Guitton B, Costes E *et al.* I want to (bud) break free: the potential role of DAM and SVP-like genes in regulating dormancy cycle in temperate fruit trees. *Front Plant Sci*. 2019;**9**: 1–17.
28. Fadón E, Fernandez E, Behn H *et al.* A conceptual framework for winter dormancy in deciduous trees. *Agron*. 2020;**10**:241.
29. Quesada-Traver C, Guerrero BI, Badenes ML *et al.* Structure and expression of bud *dormancy-associated MADS-box genes* (*DAM*) in european plum. *Front Plant Sci*. 2020;**11**:1288.
30. Goeckeritz C, Hollender CA. There is more to flowering than those *DAM* genes: the biology behind bloom in rosaceous fruit trees. *Curr Opin Plant Biol*. 2021;**59**:101995.
31. Calle A, Grimplet J, Dantec LL *et al.* Identification and characterization of DAMs mutations associated with early blooming in sweet cherry, and validation of DNA-based markers for selection. *Front Plant Sci*. 2021;**12**.
32. Ushijima K, Sassa H, Tao R *et al.* Cloning and characterization of cDNAs encoding S-RNases from almond (*Prunus dulcis*): primary structural features and sequence diversity of the S-RNases in Rosaceae. *Mol Gen Genet*. 1998;**260**:261–8.
33. Ikeda K, Igic B, Ushijima K *et al.* Primary structural features of the S haplotype-specific F-box protein, SFB, in *Prunus*. *Sex Plant Reprod*. 2004;**16**:235–43.
34. Hauck NR, Yamane H, Tao R *et al.* Accumulation of nonfunctional S-haplotypes results in the breakdown of gametophytic self-incompatibility in tetraploid *Prunus*. *Genetics*. 2006;**172**: 1191–8.
35. Nunes MDS, Santos RAM, Ferreira SM *et al.* Variability patterns and positively selected sites at the gametophytic self-incompatibility pollen *SFB* gene in a wild self-incompatible *Prunus spinosa* (Rosaceae) population. *New Phytol*. 2006;**172**: 577–87.
36. Tsukamoto T, Hauck NR, Tao R *et al.* Molecular characterization of three non-functional *S*-haplotypes in sour cherry (*Prunus cerasus*). *Plant Mol Biol*. 2006;**62**:371–83.
37. Tsukamoto T, Potter D, Tao R *et al.* Genetic and molecular characterization of three novel *S*-haplotypes in sour cherry (*Prunus cerasus* L.). *J Exp Bot*. 2008;**59**:3169–85.
38. Tao R, Iezzoni AF. The S-RNase-based gametophytic self-incompatibility system in *Prunus* exhibits distinct genetic and molecular features. *Sci Hortic*. 2010;**124**:423–33.
39. Tsukamoto T, Hauck NR, Tao R *et al.* Molecular and genetic analyses of four nonfunctional S haplotype variants derived from a common ancestral S haplotype identified in sour cherry (*Prunus cerasus* L.). *Genetics*. 2010;**184**:411–27.
40. Sassa H. Molecular mechanism of the S-RNase-based gametophytic self-incompatibility in fruit trees of Rosaceae. *Breed Sci*. 2016;**66**:116–21.
41. Matsumoto D, Tao R. Distinct self-recognition in the *Prunus* S-RNase-based gametophytic self-incompatibility system. *J Hortic*. 2016;**85**:289–305.
42. Schuster M. Self-incompatibility (*S*) genotypes of cultivated sweet cherries—an overview update. *OpenAgrar Repository*. 2020;**2020**:1–45.
43. Rhie A, Walenz BP, Koren S *et al.* Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. 2020;**21**:1–27.
44. Simão FA, Waterhouse RM, Ioannidis P *et al.* BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *J Bioinform*. 2015;**31**:3210–2.
45. Roach MJ, Schmidt S, Borneman AR. Purge Haplotigs: Synteny reduction for third-gen diploid genome assemblies. *Bioinform*. 2018;**19**:460.
46. Haug-Baltzell A, Stephens SA, Davey S *et al.* SynMap2 and SynMap3D: web-based whole-genome synteny browsers. *J Bioinform*. 2017;**33**:2197–8.
47. Verde I, Jenkins J, Dondini L *et al.* The peach v2.0 release: high-resolution linkage mapping and deep resequencing improve chromosome-scale assembly and contiguity. *Genomics*. 2017;**18**:225.
48. Dirlewanger E, Graziano E, Joobeur T *et al.* Comparative mapping and marker-assisted selection in Rosaceae fruit crops. *PNAS*. 2004;**101**:9891–6.
49. Dirlewanger E, Quero-García J, Dantec LL *et al.* Comparison of the genetic determinism of two key phenological traits, flowering and maturity dates, in three *Prunus* species: peach, apricot and sweet cherry. *J Hered*. 2012;**109**:280–92.
50. Aranzana MJ, Decroocq V, Dirlewanger E *et al. Prunus* genetics and applications after de novo genome sequencing: achievements and prospects. *Hortic Res*. 2019;**6**:58.
51. Lovell JT, MacQueen AH, Mamidi S *et al.* Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass. *Nature*. 2021;**590**:438–44.
52. Mitros T, Session AM, James BT *et al.* Genome biology of the paleotetraploid perennial biomass crop *Miscanthus*. *Nat Commun*. 2020;**11**:5442.
53. Jia K, Wang Z, Wang L *et al.* SUBPHASER: a robust allopolyploid subgenome phasing method based on subgenome-specific k-mers. *New Phytol*. 2022;**235**:801–9.
54. Gordon SP, Levy JJ, Vogel JP. PolyCRACKER, a robust method for the unsupervised partitioning of polyploid subgenomes by signatures of repetitive DNA evolution. *Genomics*. 2019;**20**:580.
55. Pinosio S, Marroni F, Zuccolo A *et al.* A draft genome of sweet cherry (*Prunus avium* L.) reveals genome-wide and local effects of domestication. *Plant J*. 2020;**103**:1420–32.
56. Gill N, Findley S, Walling JG *et al.* Molecular and chromosomal evidence for allopolyploidy in soybean. *Plant Physiol*. 2009;**151**: 1167–74.
57. Luo S, Mach J, Abramson B *et al.* The cotton centromere contains a Ty3-gypsy-like LTR retroelement. Dawson DS (ed.). *PLoS One* 2012;7, e35261,
58. Hartley G, O'Neill R. Centromere repeats: hidden gems of the genome. *Genes*. 2019;**10**.
59. Campbell MS, Holt C, Moore B *et al.* Genome annotation and curation using MAKER and MAKER-P. *Curr Protoc Bioinform*. 2014;**48**:4.11.1–39.

60. Bowman MJ, Pulman JA, Liu TL *et al.* A modified GC-specific MAKER gene annotation method reveals improved and novel gene predictions of high and low GC content in *Oryza sativa*. *Bioinformatics*. 2017;**18**:522.
61. Wang J, Childs K. deFusion—a tool to untangle MAKER fused genome annotation. 2018.
62. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun*. 2020;**11**:1432.
63. Yocca A. *Ka Ks Pipeline*. 2019.
64. Tang H, Wang X, Bowers JE *et al.* Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res*. 2008;**18**:1944–54.
65. The Uniprot Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 2021;**49**:D480–9.
66. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. 2019;**20**:238.
67. Camacho C, Coulouris G, Avagyan V *et al.* BLAST+: architecture and applications. *Bioinformatics*. 2009;**10**.
68. Dunn N, Unni D, Diesh C *et al.* Apollo: democratizing genome annotation. *PLoS Comput Biol*. 2019;**15**:e1006790.
69. Shi S, Li J, Sun J *et al.* Phylogeny and classification of *Prunus* sensu lato (Rosaceae). *J Integr Plant Biol*. 2013;**55**:1069–79.
70. Wöhner TW, Emeriewen OF, Wittenberg AHJ *et al.* The structure of the tetraploid sour cherry 'Schattenmorelle' (*Prunus cerasus* L.) genome reveals insights into its segmental allopolyploid nature. 2023;421.
71. Quero-García J, Iezzoni AF, Pulawska J *et al.*, eds. *Cherries: Botany, Production, and Uses*. Boston, MA: CABI International; 2017.
72. Deb SK, Edger PP, Pires JC *et al.* Patterns, mechanisms, and consequences of homoeologous exchange in allopolyploid angiosperms: a genomic and epigenomic perspective. *New Phytol*. 2023.
73. Bertioli DJ, Jenkins J, Clevenger J *et al.* The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. *Nat Genet*. 2019;**51**:877–84.
74. Akšić MF, Cerović R, Ercişli S *et al.* Microsporogenesis and meiotic abnormalities in different 'Oblačinska' sour cherry (*Prunus cerasus* L.) clones. *Flora*. 2016;**219**:25–34.
75. Iezzoni AF, Sebolt AM, Wang D. Sour cherry breeding program at Michigan State University. *Acta Hortic*. 2005;**667**:131–4.
76. Baek S, Choi K, Kim G *et al.* Draft genome sequence of wild *Prunus yedoensis* reveals massive inter-specific hybridization between sympatric flowering cherries. *Genome Biol*. 2018;**19**:127.
77. Hedrick UP. New York State Agricultural Experiment Station. In: *The Cherries of New York*. J.B. Lyon Company, State printers: Albany, NY, 1915.
78. Stegmeir T. *Discovery of a QTL for Cherry Leaf Spot Resistance and Validation in Tetraploid Sour Cherry of QTLs for Bloom Time and Fruit Quality Traits from Diploid Prunus Species*. 2013.
79. Ramsey J, Schemske DW. Neopolyploidy in flowering plants. *Annu Rev Ecol Syst*. 2002;**33**:589–639.
80. Tayalé A, Parisod C. Natural pathways to polyploidy in plants and consequences for genome reorganization. *Cytogenet Genome Res*. 2013;**140**:79–96.
81. Soares NR, Mollinari M, Oliveira GK *et al.* Meiosis in polyploids and implications for genetic mapping: a review. *Genes*. 2021;**12**:1517.
82. Iezzoni A, Schmidt H, Albertini A. Cherries (Prunus spp.). In: *Genetic Resources of Temperate Fruit and Nut Crops*. International Society for Horticultural Science: Wageningen, Netherlands, 1990,110–73.
83. Macková L, Vít P, Ďurišová Ľ *et al.* Hybridization success is largely limited to homoploid *Prunus* hybrids: a multidisciplinary approach. *Plant Syst Evol*. 2017;**303**:481–95.
84. Galassi A, Cappellini P, Miotto G. A descriptive model for peach fruit growth. *Adv Hort Sci*. 2000;**14**:19–22.
85. Yoo S, Gao Z, Cantini C *et al.* Fruit ripening in sour cherry: changes in expression of genes encoding expansins and other cell-wall-modifying enzymes. *J Am Soc Hortic*. 2003;**128**:16–22.
86. Gasic K, Hernandez A, Korban SS. RNA extraction from different apple tissues rich in polyphenols and polysaccharides for cDNA library construction. *Plant Mol Biol Rep*. 2004;**22**:437–8.
87. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *J Bioinform*. 2011;**27**:764–70.
88. Koren S, Walenz BP, Berlin K *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. 2017;**27**:722–36.
89. Walker BJ, Abeel T, Shea T *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. 2014;**9**:e112963.
90. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods*. 2012;**9**:357–9.
91. Wick RR, Schultz MB, Zobel J *et al.* Bandage: interactive visualization of de novo genome assemblies. *J Bioinform*. 2015;**31**:3350–2.
92. Durand NC, Shamim MS, Machol I *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst*. 2016;**3**:95–8.
93. Li H, Durbin R. Fast and accurate short read alignment with burrows–wheeler transform. *J Bioinform*. 2009;**25**:1754–60.
94. Dudchenko O, Batra SS, Omer AD *et al.* De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*. 2017;**356**:92–5.
95. Durand NC, Robinson JT, Shamim MS *et al.* Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst*. 2016;**3**:99–101.
96. Ou S, Chen J, Jiang N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res*. 2018;**46**:e126.
97. Ou S, Jiang N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob DNA*. 2019;**10**:48.
98. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *Bioinformatics*. 2008;**9**:18.
99. Ou S, Jiang N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol*. 2018;**176**:1410–22.
100. Ou S, Su W, Liao Y *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol*. 2019;**20**:275.
101. Jung S, Ficklin SP, Lee T *et al.* The genome database for Rosaceae (GDR): year 10 update. *Nucleic Acids Res*. 2014;**42**:D1237–44.
102. Jung S, Lee T, Cheng C-H *et al.* 15 years of GDR: new data and functionality in the genome database for Rosaceae. *Nucleic Acids Res*. 2019;**47**:D1137–45.
103. Tang H, Zhang X, Miao C *et al.* ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol*. 2015;**16**:3.
104. Berardini TZ, Reiser L, Li D *et al.* The arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. *Genesis*. 2015;**53**:474–85.
105. Campbell MS, Law M, Holt C *et al.* MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol*. 2014;**164**:513–24.

106. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *J Bioinform*. 2014;**30**:2114–20.
107. Dobin A, Davis CA, Schlesinger F *et al.* STAR: ultrafast universal RNA-seq aligner. *J Bioinform*. 2013;**29**:15–21.
108. Li H, Handsaker B, Wysoker A *et al.* The sequence alignment/map format and SAMtools. *J Bioinform*. 2009;**25**:2078–9.
109. Kovaka S, Zimin AV, Pertea GM *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol*. 2019;**20**:278.
110. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinformatics*. 2013;**14**:178–92.
111. Trapnell C, Roberts A, Goff L *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. *Nat Protoc*. 2012;**7**:562–78.
112. De Coster W, D'Hert S, Schultz DT *et al.* NanoPack: visualizing and processing long-read sequencing data. *J Bioinform*. 2018;**34**:2666–9.
113. Wick R, Volkening J, Loman N. *Porechop*. 2018.
114. Li H. Minimap2: pairwise alignment for nucleotide sequences. *J Bioinform*. 2018;**34**:3094–100.
115. Slater G, Birney E. Automated generation of heuristics for biological sequence comparison. *Bioinformatics*. 2005;**6**:31.
116. Stajich J. Turns EXONERATE GFF output into GFF for Gbrowse use—process_exonerate_gff3.P1. 2015.
117. Stanke M, Schöffmann O, Morgenstern B *et al.* Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *Bioinformatics*. 2006;**7**:62.
118. Mistry J, Chuguransky S, Williams L *et al.* Pfam: the protein families database in 2021. *Nucleic Acids Res*. 2021;**49**:D412–9.
119. Eddy SR. the HMMER development team. In: *HMMER: Biological Sequence Analysis Using Profile Hidden Markov Models*. 2020.
120. Johnson LS, Eddy SR, Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure. *Bioinformatics*. 2010;**11**.
121. Dainat J. AGAT: another GFF analysis toolkit to handle annotations in any GTF/GFF format. *Zenodo*. 431.
122. Thomas PD, Campbell MJ, Kejariwal A *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res*. 2003;**13**:2129–41.
123. Haft DH, Loftus BJ, Richardson DL *et al.* TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res*. 2001;**29**:41–3.
124. Quevillon E, Silventoinen V, Pillai S *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res*. 2005;**33**:W116–20.
125. Ashburner M, Ball CA, Blake JA *et al.* Gene ontology: tool for the unification of biology. *Nat Genet*. 2000;**25**:25–9.
126. Carbon S, Douglass E, Good BM *et al.* The gene ontology resource: enriching a GOld mine. *Nucleic Acids Res*. 2021;**49**:D325–34.
127. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;**32**:1792–7.
128. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*. 2006;**34**:W609–12.
129. Yang Z. PAML 4: a program package for phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007;**24**:1586–91.
130. Wickham H. *Ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.
131. Korf I. Gene finding in novel genomes. *Bioinformatics*. 2004;**5**:59.
132. Molinari NA, Petrov DA, Price HJ *et al.* Synteny and collinearity in plant genomes. *Science*. 2008;**320**:486–8.
133. R Core Team. *R: A Language and Environment for Statistical Computing*. 2022.
134. Kolde R. *Pretty Heatmaps*. 2019.
135. Sun X, Jiao C, Schwaninger H *et al.* Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nat Genet*. 2020;**52**:1423–32.
136. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;**30**:772–80.
137. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *J Bioinform*. 2009;**25**:1972–3.
138. Madeira F, Pearce M, Tivey ARN *et al.* Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res*. 2022;**50**:W276–9.
139. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *J Bioinform*. 2014;**30**:1312–3.
140. McKain M. *PhyDS: Phylogenetic iDentification of Subgenomes*. 2022.
141. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol*. 2016;**33**:1635–8.
142. Anand L, Rodriguez Lopez CM. ChromoMap: an R package for interactive visualization of multi-omics data and annotation of chromosomes. *Bioinformatics*. 2022;**23**:33.
143. Sanderson MJ. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *J Bioinform*. 2003;**19**:301–2.
144. Agarwala R, Barrett T, Beck J, *et al.* Database resources of the National Center for biotechnology information. *Nucleic Acids Res*. 2016;**44**:D7–19.