A chromosome-scale assembly for tetraploid sour cherry (*Prunus cerasus* L.) 'Montmorency' identifies three distinct ancestral *Prunus* genomes

Charity Z. Goeckeritz[1] goeckeri@msu.edu, Kathleen E. Rhoades[1] rhoade24@msu.edu, Kevin L. Childs[2] kchilds@msu.edu, Amy F. Iezzoni[1] iezzoni@msu.edu, Robert VanBuren[1,*] vanbur31@msu.edu, Courtney A. Hollender[1,*] chollend@msu.edu

[1]Department of Horticulture, Michigan State University, 1066 Bogue St, East Lansing, MI 48824

[2]Department of Plant Biology, Michigan State University, 612 Wilson Road, East Lansing, MI 48824

[*]Co-corresponding authors

Abstract

Background

Sour cherry (*Prunus cerasus* L.) is a valuable fruit crop in the Rosaceae family and a hybrid between progenitors most closely related to extant *P. fruticosa* (ground cherry) and *P. avium* (sweet cherry). Sour cherry is an allotetraploid with few genomic resources, so a genome sequence would greatly facilitate the improvement of this crop. In *Prunus*, two known classes of genes are of particular importance to breeding strategies: the self-incompatibility loci (*S*-alleles), which determine compatible crosses and are critically important for successful fertilization and fruit set, and the Dormancy Associated MADS-box genes (DAMs), which strongly affect dormancy transitions and flowering time.

Results

Here we report a chromosome-scale genome assembly for sour cherry cultivar 'Montmorency', the predominant sour cherry cultivar grown in the U.S. We also generated a draft assembly of *P. fruticosa* to use alongside a published *P. avium* sequence for syntelog-based subgenome assignments for 'Montmorency'. Using hierarchal k-mer clustering and phylogenomics, we provide compelling evidence this allotetraploid 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). We therefore assigned the genome composition of 'Montmorency' to be AA'BB and show little to no recombination has occurred between progenitor subgenomes (A/A' and B). The *S*-alleles and DAMs in 'Montmorency' and *P. fruticosa* were manually annotated and demonstrated to support the three subgenome assignments. Lastly, the hybridization event that 'Montmorency' is descended from was

estimated to have occurred less than 1.61 million years ago, making sour cherry a relatively recent allotetraploid.

Conclusions

The genome of sour cherry cultivar Montmorency highlights the evolutionary complexity of the genus *Prunus*. These genomic resources will inform future breeding strategies for sour cherry, comparative genomics in the Rosaceae, and questions regarding neopolyploidy.

Keywords

allopolyploid, *Prunus*, neopolyploid, sour cherry, genome, k-mer clustering, breeding, subgenome, evolution, Rosaceae

Background

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). Sour cherry is an allotetraploid (2n=4x=32) and shown to be a hybrid between a tetraploid resembling *P. fruticosa* Pall. (ground cherry) and a diploid resembling *P. avium* L. (sweet cherry) (4–8). However, cytological and genetic data suggest that sour cherry could be considered a segmental allotetraploid (5,9,10). Trivalents and quadrivalents are common at meiosis, and although disomic inheritance is more common, tetrasomic inheritance has also been observed (5,9–11). 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 (4,7,11,12). 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 (7,13). 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 (6,7,14). To add to the polyploid complexity, it is unclear if *P. fruticosa* is an allo- or autotetraploid as to our knowledge, no rigorous study has been reported (13,15–17).

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 United States (18). Michigan, the number one producer of sour cherry in the U.S., lost more than 90 percent of its crop (18). Depending on the region, climate change is increasing the frequency of these events worldwide (18–20). 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* (21). 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 (9,11,21), since no public sour cherry genome sequence is available.

In this work, we constructed and annotated the first *P. cerasus* reference genome for the cultivar 'Montmorency,' a ~400-year-old French amarelle sour cherry selection of unknown origin but the most widely-grown cultivar in the United States. 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 (22), to assign subgenomes to the 'Montmorency' superscaffolds. 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 (*P. persica*) with major effects on dormancy transitions and flowering time in *Prunus* species (23–30). The second set includes the self-incompatibility *S*-allele genes that make up the *S*-haplotype, *S*-ribonuclease (*S*-RNase) and *S*-locus F-box (SFB) (31–41). The four *S*-haplotypes in 'Montmorency' have been previously characterized (38), but no *S*-haplotypes in *P. fruticosa* have been thoroughly described. 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 (**Supplementary Figure 1**). 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%) suggests multiple haplotypes were assembled. The Merqury plot (42) (**Supplementary Figure 2**) indicates high genome completeness as most k-mers in the Illumina dataset are found in the assembly. Additionally, k-mers from the Illumina reads are found in the assembly at the expected relative frequencies and there are 4 distinct peaks, indicating haplotypes are well-phased. A BUSCO assessment (43) 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.

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 8 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 (44) was used to set aside one of the two possible alleles of the *P. avium*-like sequences. 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. **Supplementary Figure 3** suggested 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 (2x, 3x, 4x) 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 (**Supplementary Figure 4**). These chromosomes were numbered according to their synteny with *Prunus persica* chromosomes 1 –

8 (**Figure 1a**) (45,46). 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 (22) 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 (47–49). When the scaffolded linkage groups were compared with a sour cherry genetic map of 545 unique markers (11), 78% mapped exactly once to each homoeolog. Furthermore, all markers showed nearly perfect linearity with the assembly (**Figure 2, Supplementary Figure 5**).

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* (50), *Nicotiana tobacum* (51), *Triticum aestivum* (51), *Eragrostis tef* (52) and *Panicum virgatum* (53). Using this strategy, we identified 840 25-base pair sequences (25-mers) with more than 10 copies on each 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, Supplementary Figure 6**). 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) (**Supplementary Figure 6**).

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 to 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 clearly 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 7 chromosomes. A total of 481 25-mers in 3 clusters differentiated the two chromosome sets (**Supplementary Figure 7;** 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, Supplementary Figure 8a**). Peak 25-mer densities colocalize with the lowest gene content and highest TE density, suggesting these 25-mer groups are most

198 abundant at or near putative centromeres. These estimated centromeric regions agree with a
199 previously-published *P. avium* genome (54). Group 2 and Group 3 densities, which distinguished
200 A/A' subgenomes from subgenome B in the 24-chromosome clustering analysis (**Figure 3a;**
201 **Supplementary Figure 6**) clearly exhibited contrasting peak 25-mer densities at approximate
202 centromeres (**Figure 4b, Supplementary Figure 8b**). Likewise, Group 5 and Group 6 25-mers,
203 which differentiated the A and A' subgenomes (**Supplementary Figure 7**), showed a similar
204 pattern when their densities were aligned to each of the homoeologous chromosome sets (**Figure**
205 **4c, Supplementary Figure 8c**). Since highly variable, repetitive satellite DNA is species-
206 specific and often associated with centromere and pericentromeric regions in eukaryotes (55–57),
207 these observations support the claim that A, A', and B represent distinct subgenomes.

208 As it is well-established that sour cherry is a tetraploid (5,10,13), we explored the relative dosage
209 of each 'Montmorency' subgenome by conducting a depth analysis of the Illumina reads against
210 the assembly. As suspected earlier, this analysis clearly indicated subgenome B had twice the
211 genome dosage of either subgenomes A or A' as regions along the length of subgenome B
212 frequently showed about twice the read depth of subgenomes A and A' **(Supplementary Figure**
213 **9**). Therefore, the genome structure of *P. cerasus* 'Montmorency' is AA'BB. A summary of
214 assembly and scaffolding statistics is given in **Table 1**.

215 *Annotation of the Prunus cerasus 'Montmorency' assembly*

216 For structural annotation of the gene space, RNA-sequencing of a variety of tissues and long-
217 read cDNA datasets were generated for 'Montmorency' and used as transcript evidence for de
218 novo gene annotation via MAKER (58,59). MAKER predicted a total of 92,783 protein-coding
219 genes in the full assembly of 'Montmorency' after filtering for gene predictions with known
220 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 (60). 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 (**Supplementary Figure 10**). 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 **Supplementary Table 1**, and **Supplementary Figure 11** 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 Prunus fruticosa draft genome

PacBio long-reads and Illumina short-reads were also generated for a *Prunus 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. Although the draft assembly is not chromosome-scale, the contigs are large enough for syntenic and gene orthology comparisons (**Supplementary Table 2,** see NG50); and the primary reason for generating this *P. fruticosa* resource was for the assignation of subgenome origins and divergence estimates between progenitors and A, A', and B in 'Montmorency'. 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. Additional assembly metrics can be found in **Supplementary Table 2**. 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; **Supplementary Figure 12**). Initially, efforts were made to alleviate these issues and reduce genome complexity by creating a representative haplotype with Purge Haplotigs (44). However, this method was quickly abandoned as it is unknown whether *P. fruticosa* is an allo- or autotetraploid, nor how divergent the alleles are. Given other metrics of the assembly are high and it is prudent to represent the allele diversity for accurate syntelog comparisons, we moved forward to include all contigs in structural annotation. Additional metrics of the draft assembly can be found in **Supplementary Table 2**.

Annotation of the Prunus 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 (61) were used as evidence for annotation of the *P. fruticosa* assembly with MAKER. 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 **Supplementary Table 3.** 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 (**Supplementary Figure 13).**

Progenitor assignments of the subgenomes in Prunus cerasus 'Montmorency'

267936 orthologs, or 93.1% of genes in all 7 'species' (*Malus* x *domestica*, *Prunus persica*, *P. avium*, *P. fruticosa*, 'Montmorency' subgenome A, subgenome A', subgenome B) were assigned to orthogroups using OrthoFinder v 2.5.4 (62). Out of all orthogroups, 12051 included at least one ortholog for every species. 336 of these were 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 10398 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% (10245/10370) of syntelog 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'.

DAM gene haplotypes identified in 'Montmorency' and Prunus 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* (26,29). 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 (63) revealed DAM gene candidates on 'Montmorency' chromosomes 1A, 1A', and 1B. Upon manual inspection and correction of these gene models using Apollo v 2.6.5 (64), three full haplotypes of DAM1 – DAM6 were found on chr1A, chr1A', and chr1B. 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 3 of the 9 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

311 regions were confirmed to be syntenic with DAM haplotypes found on 'Montmorency'
312 chromosomes 1A, 1A', and 1B (**Figure 6a**).

313 We next carried out a phylogenetic analysis of all 18 and 12 DAM genes in 'Montmorency' and
314 *P. fruticosa*, respectively, with manually annotated DAM genes from *P. avium* (**Figure 6b**) (30).
315 All gene numbers and NCBI identifiers associated with this analysis can be found in
316 **Supplementary Table 4**. All DAM1 – DAM6 genes formed well-supported monophyletic
317 clades (BSV >80%) consistent with their order in the tandem array (**Supplementary Figure 14**),
318 garnering further support that these genes were correctly identified. Furthermore, each
319 'Montmorency' DAM gene was sister to the syntelog of its respective progenitor. In other words,
320 all chr1A DAM genes were sister to *P. fruticosa* contig 8 DAM genes, chr1A' DAM genes were
321 sister to *P. fruticosa* contig 33 DAM genes, and chr1B DAM genes were sister to *P. avium* DAM
322 genes (**Figure 6b**). Interestingly, the gene downstream from the DAM haplotype for
323 'Montmorency' chr1A' (Pcer_010093) and *P. fruticosa* contig 33 (Pfrut_003731; note that two
324 gene models were created for this region) contain similar insertions of nearly 9 kb in the 6th
325 intron (99.87% identical), providing higher confidence these haplotypes have shared ancestry.
326 The size of this intron on chr1A, chr1B, and *P. fruticosa* contig 8 is approximately 767 bp.

327 *S-alleles identified in 'Montmorency' and Prunus fruticosa assemblies*

328 *S*-alleles, consisting of an RNase (*S*-RNase) and linked F-box protein (SFB), are responsible for
329 gametophytic self-incompatibility in *Prunus* (34). Four *S*-alleles identified in the 'Montmorency'
330 assembly on chromosomes 6A, 6A', and 6B, match the 'Montmorency' *S*-alleles previously
331 reported, i.e. S_6, S_{13m}, S_{35}, and S_{36a} (38,40). The linked SFBs and *S*-RNases comprising these four
332 *S*-allele haplotypes all have >99% identities to their published sequences (37,38,40). BLAST+
333 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 (40,41). 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* (37). 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 (39). 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 (40). 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 (40,65). 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' (**Supplementary Figure 15**); however, contig 1100 was too small (130 kb) to do a macrosyntenic analysis using MCScan. Of the sequences used as queries (**Supplementary Table 5**), 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 (**Supplementary Figure 16a**; note the amino acid fasta sequence reads through the stop codon to indicate conserved domains). The SFB protein on contig 53 has an open reading frame, expected protein domains, and an amino acid sequence identical to the *P. cerasus* SFB_{36b} sequence (**Supplementary Figure 16b**) (40). In summary, S_{36} variants have not been previously identified 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.

Prunus 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 (1–3,66) 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). The hybridization event 'Montmorency' is descended from would have occurred sometime after this estimate.

Because *P. fruticosa* is a verified tetraploid, the ortholog in each topology is an assembly collapse of four possible alleles. If *P. fruticosa* is an allotetraploid with two intact and divergent subgenomes, approximately half of the 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 *P. fruticosa* 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; **Supplementary Figure 17**). 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 – 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; **Supplementary Figure 17**). 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 (*Prunus cerasus*) cultivar Montmorency, which to our knowledge is the first genome sequence for sour cherry to be published. We chose to sequence 'Montmorency' as it is the most commonly grown cultivar in the United States. This chromosome-scale assembly is highly collinear with a published sour cherry genetic map (11), is syntenic with other *Prunus* species (22,46), and has a quality annotation with thorough manual curation. Additionally, we assembled and annotated a draft genome of *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 contributed by a *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 to 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). Our results suggest *P. fruticosa* is an allopolyploid hybrid of two distinct *Prunus* species. However, since large-scale evolutionary studies of *Prunus* rarely include this species (1–3), the extant relatives of its progenitor species are unknown and is a question for future research. 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 (40). An

423 interesting question moving forward is how the progenitor origins of the DAM genes may relate
424 to bloom time in diverse sour cherry accessions. Sour cherry exhibits a transgressive range in
425 bloom time: some genotypes bloom earlier or later than either progenitor (11). *P. fruticosa* is
426 native to colder northern latitudes in eastern Europe and flowers later than *P. avium*, which
427 originates from the Mediterranean region south of the Black Sea (67). One might expect to see
428 selective pressure on the DAM genes depending on whether a *P. cerasus* genotype distributes to
429 more northern or southern latitudes, tailoring bloom time to maximize reproductive potential.
430 However, this question remains largely unexplored.

431 Our subgenome assignments for 'Montmorency' also provide insights into the possible gametes
432 that formed 'Montmorency'. First, the 'Montmorency' subgenomes derived from the *P.*
433 *fruticosa*-like progenitor (A/A') do not show any evidence of recombination with the subgenome
434 derived from the *P. avium*-like progenitor (B). This lack of recombination supports the theory
435 that 'Montmorency' was formed by one gamete from the *P. fruticosa*-like ancestor and one
436 gamete from the *P. avium*-like ancestor. Second, since A and A' were readily distinguishable
437 according to gene and repetitive sequence differences (with the exclusion of chr8A and chr8A'),
438 this implies very few homoeologous exchanges (mixing of genomes) have occurred between
439 these more similar subgenomes. This finding once again suggests the *P. fruticosa*-like progenitor
440 was an allopolyploid and the gamete that gave rise to 'Montmorency' resulted from preferential
441 chromosome pairing of the two subgenomes. Chr8A and 8A' may be an exception; however, it is
442 unclear whether this is due to subgenome mixing, disproportionately high sequence similarity
443 among these chromosomes, assembly artifacts, or a combination of one or more scenarios. These
444 findings are consistent with the possibility that 'Montmorency' was formed from the fusion of a
445 reduced gamete from a *P. fruticosa*-like ancestor and an unreduced gamete from a *P. avium*-like

446 ancestor (8). In the case of 'Montmorency', chloroplast data identified the *P. fruticosa*-like
447 progenitor as the maternal parent (6).

448 It follows that the integrity of the three 'Montmorency' subgenomes is not likely to be
449 transmitted to the next generation. If the A and A' chromosomes preferentially pair, crossing
450 over will result in a patchwork of regions exchanged between these two homoeologous
451 chromosome sets. Additionally, cytological analysis of meiotic pairing for 'Montmorency'
452 shows lack of complete bivalent pairing: quadrivalents, trivalents and univalents occur (10,68).
453 Segregation data support primarily subgenome pairing (A with A' and B with B); however
454 genetic results are consistent with occasional homoeologous pairing (5,11). This suggests that
455 homoeologous exchanges could occur between the A and A' subgenome chromosomes and the B
456 subgenome chromosomes, further eroding the subgenome integrity in 'Montmorency' offspring.

457 Poor fertility supported by evidence of irregular meiosis and occasional tetrasomic inheritance
458 are prevalent in sour cherry germplasm (69). In general, the quantity of fruit set is vastly below
459 what the plant could support. Commercial cultivars such as 'Montmorency' are the exception as
460 this cultivar can achieve a "full crop" by setting approximately 30% of its fruit. Indeed, it is the
461 high fertility in 'Montmorency' that led it to be the predominant cultivar in the United States.
462 However, even in crosses between two productive cultivars, the low fertility in the offspring is
463 clear. For example, in one study (9), the mean fruit set and pollen germination for the German
464 sour cherry cultivar Schattenmorelle and the Hungarian cultivar Érdi Bötermö was 16.0% and
465 13.4%, respectively for fruit set, and 18.5% and 8.0% respectively for pollen germination. When
466 these two cultivars were crossed, the mean values for fruit set and pollen germination of the 86
467 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 that the allotetraploid event from which the 'Montmorency' lineage is descended occurred less than 1.61 million years ago. 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 (70). Thus, this sour cherry lineage may be considered younger than short-lived polyploids in terms of absolute number of generations. It must be noted, however, that the progenitor representatives used in dating analyses may greatly affect estimates. Additionally, since sour cherry is a species thought to have formed multiple times (6), the timing of hybridization for certain lineages would be expected to vary. Still, given these results and speculations, it would be reasonable to suggest sour cherry exemplifies the behavior of a neopolyploid actively undergoing the process of diploidization. Cytological and genetic data is consistent with a neopolyploid as irregular meiosis visualized as trivalents and quadrivalents is documented in several sour cherry genotypes (10,68), and genetic data indicates that although disomic inheritance is more common, tetrasomic inheritance has also been observed (9,11,71). 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 (72). 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 swiftly against these infertile genetic combinations and the polyploid would eventually 'diploidize' (72–74). 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 (13). 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 Crișana, respectively, favored fruit quality over fruit quantity, as these landrace cultivars have extremely low fruit set but highly desirable fruit quality (75). 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

514 scenario, a hybrid of a *P. fruticosa*-like and *P. avium*-like species (*Prunus* × *mohacsyana*) would
515 produce an unreduced gamete (3x) that either fertilized or was fertilized by a *P. avium* gamete
516 (1x) to form allotetraploid *P. cerasus*. Irrespective of these different evolutionary trajectories, it
517 is important to consider the 'Montmorency' subgenome structure reported herein, may be unique
518 to this genotype and therefore may not represent the subgenome structure of a range of sour
519 cherry accessions.

520 **Conclusions**

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

528 **Methods**

529 *Collection of materials*

530 Young leaves for gDNA libraries were collected fresh (for Hi-C, *P. cerasus* 'Montmorency'
531 only) or flash-frozen in liquid nitrogen and stored at -80°C until extraction (for both PacBio
532 SMRT sequencing and Illumina HiSeq) from a clone of 'Montmorency' and an accession of *P.*
533 *fruticosa* growing at Michigan State University's Clarksville Research Center in Clarksville,
534 Mich., in the spring of 2019. Tissues for RNAseq and Nanopore cDNA libraries were collected
535 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 **Supplementary Figure 18**. Fruit tissue was collected and staged based on the double-sigmoidal growth curve characteristic of *Prunus* sp. (76,77). Vegetative and floral meristems in typical floral positions were broadly characterized as pre-floral-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 (>30kb) was prepared and sequenced on six flow cells of a PacBio Sequel II machine for each species, producing 61.98 Gb of data (100X coverage of 621 Mb estimated genome size) for *P. cerasus* 'Montmorency' and 48.13 Gb of data (90.5X 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, USA). 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 (56X coverage) for *P. cerasus* 'Montmorency' and 40 Gbp of data was produced for *P. fruticosa* (75.2X 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

559 Proximo Library was created with a DpnII restriction enzyme. The Hi-C library was 150 bp
560 paired-end sequenced on a HiSeq4000 instrument at MSU's RTSF, producing 93.3 Gbp of data
561 or 150.5X physical coverage.

562 *RNA extraction, library preparation, and sequencing*

563 All RNA was extracted using a CTAB-based protocol (78). One Illumina TruSeq Stranded
564 mRNA library was prepared for each of the 2 – 3 biological replicates per tissue used for
565 RNAseq (**Supplementary Figure 18**, excl. c, d, g, j-l for 'Montmorency'). Young leaves, whole
566 flowers at balloon stage, whole fruits in stage I, whole fruits at the end of stage II, and whole
567 fruits in stage III, or RNA from 5 tissue-types total, were sequenced for *P. fruticosa* at MSU's
568 RTSF. 24 RNAseq libraries for 'Montmorency' and 14 RNAseq libraries for *P. fruticosa* were
569 150 bp paired-end sequenced on a HiSeq4000, producing between 25 and 36.6 million reads per
570 library. One library from 'Montmorency' (replicate of whole fruits at the end of stage II) and one
571 library from *P. fruticosa* (replicate of whole fruits at stage I) were deemed contaminants / low-
572 quality based on unusually poor alignment to the respective assemblies and were excluded from
573 downstream analyses. In addition to most of the tissue types used for RNAseq (excluding mature
574 fruit mesocarp and mature fruit exocarp), whole fruits at the beginning of phase II, vegetative
575 apices in early summer, transitioning apices in midsummer, mature leaves in late spring, mature
576 leaves in midsummer, and floral apices during organ development (confirmed via histological
577 observations – unpublished results) were included in the Nanopore cDNA-sequencing (12 tissue-
578 types total) for 'Montmorency' (**Supplementary Figure 18** c, d, g, j-l).

579 *Assembling the genome of Prunus cerasus 'Montmorency'*

Illumina gDNA reads' 25-mers were counted using Jellyfish v 2.2.10 (RRID:SCR_005491) (79) and the resulting histogram was visualized using GenomeScope v2.0 (RRID:SCR_017014) (80). Canu v 1.9 (RRID:SCR_015880) (81) was used to assemble the PacBio reads. Reads below 5kb 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 4 times with Pilon v1.23 (RRID:SCR_014731) (82). Reads were aligned to the assembly with Bowtie2 v 2.3.4.2 (RRID:SCR_016368) (83) until there was no further improvement in read alignment. The polished assembly was visualized with Bandage (RRID:SCR_022772) (84), a BUSCO analysis was done to assess gene space (RRID:SCR_015008) (43), and Merqury (RRID:SCR_022964) (42) was used to determine phasing quality and genome-completeness.

Scaffolding the Prunus 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, non-collapsed 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 (44) was used with a cutoff value of 99% alignment to set aside very similar alleles in the assembly prior to scaffolding. 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 this sequence. As a result, purged contigs greater than 400kb were added back into the assembly prior to scaffolding. After this size selection, we

603 verified we had removed mostly alternative alleles of the three assembled groups in two ways.
604 First, BUSCO analyses showed completeness to be very high (>90%) and duplication to be very
605 low (<3.0%) for each of the three pseudomolecule groups (**Table 1**). Second, a k-mer assessment
606 using Merqury indicated 25-mers in the remaining purged contigs were found only once and at
607 all relative multiplicities (1X - 4X) in the Illumina read dataset. If 25-mers at any multiplicity
608 had been present in the purged contigs more than once, this would suggest more than one
609 haplotype had been removed from the assembly (**Supplementary Figure 3**). After reducing the
610 complexity of the assembly as described above, Hi-C reads were aligned to the assembly using
611 BWA v 0.0.7.17 (RRID:SCR_010910) (85) within the Juicer pipeline (RRID:SCR_017226) (86).
612 The -S flag was set to exit the pipeline early after production of the merge_nodups.txt file. This
613 file was then used as input for 3D-DNA (RRID:SCR_017227) (87), which was run with "- -
614 editor-repeat-coverage 5" to prevent areas of the assembly with higher levels of coverage (due to
615 ploidy) being flagged as "junk." The output was a Hi-C matrix (.hic) that was manually edited in
616 JuiceBox Assembly Tools (RRID:SCR_021172) (88) to correct misassemblies. Following
617 manual editing, the new .hic and .asm files were used to create the chromosome-level .fasta file
618 with the script run-asm-pipeline-post-review.sh from the 3D-DNA suite of tools. The resulting
619 superscaffolds (chromosomes) were named according to a syntenic comparison with a peach
620 genome (45,46) and subgenome assignments based on 25-mer groups.

621 *Assessing repeat content and quality*

622 The LTR assembly index (LAI) (89) was determined by first identifying TEs with
623 LTR_FINDER_parallel (RRID:SCR_018969) (90) and LTRharvest (91), then combining the
624 output files and using it as input for LTR_retriever (92). The EDTA pipeline was used to
625 estimate repeat content and produce a custom repeat library (RRID:SCR_022063) (93).

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) (11) were downloaded from the Genomic Database for Rosaceae (GDR; https://www.rosaceae.org/) (94,95) and mapped to the 24 superscaffolds of the assembly using BLAST+ v 2.2.31(RRID:SCR_011820) (63). 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 (RRID:SCR_021171) (96).

Annotation of the genome of Prunus 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 (61,97). All data were processed to produce .gff3 files which were used as input for MAKER (RRID:SCR_005309) (59).

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 (RRID:SCR_011848) (98). Individual libraries, totaling 3.1 billion reads, were aligned to the 'Montmorency' genome assembly using default parameters in STAR v 2.7.3a (RRID:SCR_004463) (99). 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 (RRID:SCR_002105) (100). All alignments were merged, and a transcriptome assembly was

created using StringTie v 2.1.2 (RRID:SCR_016323) (101). 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 (RRID:SCR_011793) (102), and parameters were adjusted accordingly in StringTie v 2.1.2 ("-m 200 -t -c 3 -f 0.05 -g 50"). The final .gtf file was converted to .gff3 using the gffread function in Cufflinks v 2.2.1 (RRID:SCR_014597) (103) 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 (59).

Preparation of long-read RNA sequencing for MAKER

Nanopore reads were demultiplexed, trimmed, and filtered (reads <150bp were dropped) with Porechop v 0.2.4 (RRID:SCR_016967) (104) and NanoPack (105). 4.8 million reads were aligned at a rate of 89% to the 'Montmorency' genome assembly using minimap2 v 2.15 (RRID:SCR_018550) (106) with the following parameters: "-N 5 -ax splice -g2000 -G10k". 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 (RRID:SCR_002380) (61) and Arabidopsis protein sequences from TAIR10 (RRID:SCR_004618) (97) were downloaded in fasta format on 4/17/21 and 2/26/21, respectively. Sequences were aligned using Exonerate v 2.2.0 (RRID:SCR_016088) (107) and the 5 best matches for each alignment were kept in the following format: --ryo ">%qi length=%ql alnlen=%qal\n>%ti length=%tl alnlen=%tal\n". The

resulting .gff2 was converted to a .gff3 using the script process_exonerate_gff3.pl (108), and the features 'exonerate:protein2genome:local,' 'mRNA,' and 'CDS' were changed to 'protein2genome,' 'protein_match,' and 'match_part,' respectively, for compatibility with MAKER.

Running MAKER iteratively

MAKER was run similarly to (58,59), with the evidence detailed above and the custom repeat library created from the EDTA pipeline for masking (93). 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 (RRID:SCR_008417) (109). 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 (58).

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 (RRID:SCR_004726) (110) using hmmscan v 3.1b2 (RRID:SCR_005305) (111,112), and the predictions containing no known protein domains were removed. Second, defusion (60) was run to identify putatively-fused genes on the 24 chromosomes of the assembly (chr1[A, A', B] – chr8[A, A', B]). 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

692 defusion, we used an alternative method to identify candidates of the second class of gene
693 fusions. Putatively fused loci from the initial MAKER gene set were identified when two or
694 more distinct proteins-only gene predictions overlapped with gene predictions from the transcript
695 plus protein MAKER run. (see Supplementary Materials,
696 identify_fusion_candidates_w_PROTEIN_ONLY_datasets.bash). These candidate fusions were
697 checked alongside those identified by defusion in IGV and break points were manually added to
698 the .brk file as necessary (**Supplementary Figure 10**). The "defused" annotation was then
699 filtered again to remove predictions lacking a Pfam domain (60). Lastly, putative DAM genes, *S*-
700 RNases, and SFBs were manually annotated with Apollo v. 2.6.5 (RRID:SCR_001936) (64). The
701 old gene models were then removed with agat (113) and replaced with the corrected models to
702 produce the final annotation file.

703 *Assigning functions to genes*

704 Functional information was assigned to the .gff and protein and transcript .fasta files via a
705 BLAST+ v 2.9.0 (63) comparison of amino acid sequences to a Uniprot database (61) and
706 several accessory scripts within MAKER (maker_functional_gff, maker_functional_fasta) (114).
707 Moreover, Pfam (110), PANTHER (RRID:SCR_004869) (115), TIGRFAM
708 (RRID:SCR_005493) (116), InterProScan (RRID:SCR_005829) (117), and Gene Ontology (GO)
709 (RRID:SCR_002811) (118,119) database reference numbers or IDs were added to the .gff file by
710 scanning the amino acid sequences with InterProScan and using the MAKER accessory script
711 ipr_update_gff (114). Only hits with p-values < 1.0×10^{-10} were kept. Pfam, PANTHER, and
712 TIGRFAM hits are also provided as separate .csv files that include the gene ID and functional
713 descriptions.

714 *Assembly of a Prunus fruticosa draft genome*

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 (79) and the resulting histogram was visualized using GenomeScope v2.0 (80). The estimated heterozygosity for *P. fruticosa,* which is also an obligate-outcrossing tetraploid, was 3.956%. 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. Because it is unknown whether *P. fruticosa* tetraploidy resulted from genome-doubling or interspecific hybridization (auto- vs allopolyploidization), parameters in Canu v 1.9 (81) were set similarly to those used for *P. cerasus* 'Montmorency' so that multiple haplotypes could be assembled from the PacBio reads. 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 3 times with Pilon v1.23 (82). A BUSCO analysis (43) was done to assess gene space, and Merqury (42) was used to determine phasing quality and genome-completeness.

Annotation of a Prunus 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 (120) and AUGUSTUS (109) were trained on the .gff3 from the first MAKER run

738 before using them for the second run. The final predicted gene set was filtered to keep
739 predictions with known Pfam domains but lacking known TE domains. Due to limited resources,
740 no manual annotation was performed on the *P. fruticosa* contigs. Assigning gene function to *P.*
741 *fruticosa* gene predictions was done similarly as the *P. cerasus* 'Montmorency' annotation.

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

743 A synteny analysis was conducted between the 24 superscaffolds (chromosomes) of the *P.*
744 *cerasus* 'Montmorency' assembly and a *P. persica* genome (46) using the SynMap tool within
745 the Comparative Genomics Platform (CoGe) (45). Coding sequences (unmasked) of
746 'Montmorency' and *P. persica* were compared with default settings. Based on these synteny
747 results (**Figure 1A),** k-mer clustering, and phylogenomic comparisons of syntelogs as described
748 below, the 24 superscaffolds of the *P. cerasus* 'Montmorency' were named chr1[A, A', B] –
749 chr8[A, A', B].

750 *Syntenic comparison of the Prunus cerasus 'Montmorency' assembly with representative*
751 *progenitor genomes*

752 Macrosyntenic comparisons of 'Montmorency' with the *P. avium* 'Tieton' v2.0 genome, the *P.*
753 *fruticosa* draft genome, and itself were done using the MCScan package (RRID:SCR_017650)
754 (121) from JCVI (RRID:SCR_021641) (121,122). We first built .cds and .bed files for each
755 genome from the coding sequence and .gff files, respectively, then used command
756 "jcvi.compara.catalog ortholog" to generate a list of syntenic blocks between either *P. avium* and
757 'Montmorency', *P. fruticosa* and 'Montmorency', or between 'Montmorency' and itself. All
758 karyotype figures were constructed with the "jcvi.graphics.karyotype" command.

759 *k-mer clustering*

We used Jellyfish 2.2.10 (79) 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 (RRID:SCR_001905) (123). Further filtering was done if the 25-mers were not 1) present at 2x 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, and the package 'pheatmap' to create heat maps (RRID:SCR_016418). 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 (RRID:SCR_014601) (124).

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 (83). SAMtools v 1.9 (RRID:SCR_002105) (100) was then used to sort and calculate depth at every position. 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 (**Supplementary Figure 9**). Plots were created with R v. 4.2.1 (123) and ggplot2 (124).

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,' (125), *P. persica* 'Lovell' (46), *P. avium* 'Tieton' (22), *P. fruticosa* from the present study, *P. cerasus* 'Montmorency' subA, subA', and subB were either downloaded from GDR or generated as described above. Orthogroups (groups of orthologous genes) were identified between these 7 "species" using OrthoFinder v. 2.5.4 (RRID:SCR_017118) (62). Multiple sequence alignments (MSAs) of orthogroups were done within OrthoFinder using MAFFT v. 7.480 (RRID:SCR_011811) (126) with no alignment trimming (-z). Only orthogroups including all 7 species were used for downstream analysis. Protein sequence alignments were converted to nucleotide alignments using PAL2NAL v. 14.1 (127), and raw cds alignments were trimmed with trimAl v. 1.4.1 (RRID:SCR_017334) (128) using flag -automated1. Alignments before and after trimming were visualized with MView (129) to ensure high quality of the resulting alignments. A phylogenetic tree for each orthogroup was created with RAxML-NG v 1.0.0 (RRID:SCR_022066) (130) using the gamma + GTR model, 500 bootstrap replications, and an apple ortholog outgroup. A two-column list of 'Montmorency' orthologs was used as input for PhyDS (131) to identify gene-gene sister relationships to either a *P. fruticosa* or *P. avium* ortholog with bootstrap values (BSV) of at least 80%. PhyDS requires a 2-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 trees with a phylogenetic tree viewer (132) 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%)

805 using basic Unix commands (see Supplementary Materials). At the same time, syntenic gene
806 pairs between A, A', and B versus each representative progenitor (a total of 6 comparisons) were
807 identified using the default settings of the python version of MCScan (121). The orthologous
808 relationships identified with Orthofinder that fit the above criteria were interjoined with the
809 syntenic orthologous relationships identified by MCScan using R version 4.2.1 (123). These
810 high-confidence syntenic orthologs were mapped back to the 'Montmorency' assembly and
811 labeled as either '*P. avium*-like' or '*P. fruticosa*-like.' The R package chromoMap (133) was
812 used to visualize these results.

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

815 Separating A, A', and B for ortholog analyses was done to ensure more single-copy orthologous
816 groups could be identified between all "species" and used for phylogenomic dating. RAxML-NG
817 was used to create phylogenetic trees for single-copy orthogroups as described above. Based on
818 current knowledge of Rosaceae phylogenetics (1–3,66), the topolog(ies) for most accurately
819 estimating the divergence of 'Montmorency' subgenomes and its representative progenitors from
820 their most recent common ancestor (MRCA) should place apple as the super outgroup and peach
821 as sister to the cherry lineage. Additionally, to calculate divergence time without error from
822 possible homoeologous recombination, only orthogroups where 'Montmorency' homoeologs
823 from each subgenome were predominantly sister to one progenitor over the other (*P. avium* or *P.*
824 *fruticosa*) were included in the r8s analysis. This assumed that any previous homoeologous
825 recombination taking place between the subgenomes did not replace >50% of the original
826 sequence contributed by the progenitor. As a result, in any given tree, one homoeolog would not
827 be able to pair with *P. avium* or *P. fruticosa* (i.e., there are 2 representative progenitors but 3

828 'Montmorency' subgenomes). Thus, we were prepared to calculate node ages of multiple
829 topologies to obtain MRCA divergence estimates for each subgenome and its most closely-
830 related progenitor. This required each single-copy orthologous gene tree (n=336) to be manually
831 inspected and the frequencies of each topology noted. The two most frequent topologies
832 (**Supplementary Figure 17**) made it clear which progenitor subgenome A, A', and B was most
833 related to: subgenome B orthologs were almost always sister to *P. avium* while A and A'
834 orthologs were nearly equally likely to be sister to *P. fruticosa*. Orthogroup sequence alignments
835 showing the two most frequent topologies were separately concatenated for all "species" (n=7)
836 regardless of bootstrap support. Phylogenetic trees for the two concatenated alignments were
837 created with RAxML-NG (130) as described above. Bootstrap replications (500 per topology)
838 were used to calculate node age estimates and confidence intervals with r8s
839 (RRID:SCR_021161) (134) and R v. 4.2.1 (123). Based on Xiang et al. 2017, the *Malus*/*Prunus*
840 node was fixed at 95 million years ago (mya) and the peach/cherry node was constrained to a
841 minimum age of 10 mya for both analyses. The smoothing parameter was set to 1, rate was set to
842 gamma, and divergence time was set to penalized likelihood with the TN algorithm. 128633 sites
843 were used to determine divergence times of the topology in **Supplementary Figure 17A** and
844 78,183 sites were used for the topology in **Supplementary Figure 17B**.

845 *Identification of the DAM genes and S-alleles*

846 DAM (Dormancy-Associated MADS-box) genes were identified in 'Montmorency' and the *P.*
847 *fruticosa* contigs with BLAST+ v. 2.9.0 (63), using *P. persica* DAM1 – DAM6 coding
848 sequences from NCBI's GenBank (RRID:SCR_002760) (135) as query and genomic sequence
849 or transcripts from the MAKER pipeline as the target. The sequence IDs of the *P. persica* DAMs
850 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 7 different contigs, some containing only partial haplotypes of the expected 6 tandem genes with occasional erroneous fusions. Only genomic regions containing full haplotypes (6 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 (63) and the following complete cds sequences from NCBI's GenBank (135) 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. cerasusSFB$_{36b}$, P.cerasusSFB$_{36b3}$, P.cerasusSFB$_{36b2}$, P.aviumSFB$_{13}$, P.cerasusSFB$_{35}$, and P.cerasusSFB$_{36a}$ (**Supplementary Table 5**). 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

Protein sequences for *Arabidopsis thaliana* SEP3 and the six *P. avium* DAM genes were downloaded from NCBI **(Supplementary Table 4)**. Protein alignments were done with MUSCLE/3.8.31 (RRID:SCR_011812) (136) using default settings and the phylogenetic tree was constructed using RAxML-NG/1.0.0 (130). 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 (RRID:SCR_006472) **(Supplementary Table 5)**. *P. fruticosa S*-alleles were aligned to their 'Montmorency' counterparts with MUSCLE/3.8.31 and alignment figures were made using the R package gggenes (**Supplementary Figure 16**) (124).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

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. Scripts and example files associated with genome assembly, annotation, subgenome assignment analyses, and r8s estimates can be found at https://github.com/goeckeritz/Montmorency_genome. Scripts for k-mer hierarchal clustering, synteny, DAM gene phylogenetic analyses, and *S*-allele alignments can be found at https://github.com/KEBRhoades/Montmorency_genome.

Competing interests

The authors declare that they have no competing interests.

Funding

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

CAH, AFI, and RV conceptualized the experiments. CZG performed genome assembly, annotation, subgenome assignment using orthologs, and divergence time estimate analyses. KER performed the k-mer hierarchal clustering, synteny, DAM gene phylogenetic analyses, and *S*-allele alignments. KLC provided expertise and assistance with annotation and contributed code. CZG, KER, and AFI wrote the manuscript. All authors assisted with editing the manuscript.

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 thank Dr. Pat Edger for his advice on all phylogenomic analyses.

References

1. Chin SW, Shaw J, Haberle R, Wen J, Potter D. 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, Wen J, Li S, Yi T, 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(2):262–81.

3. G. J. Hodel R, 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;59:327–59.

5. Beaver JA, Iezzoni AF. Allozyme inheritance in tetraploid sour cherry (*Prunus cerasus* L.). J Am Soc Hortic. 1993;118(6):873–7.

6. Brettin TS, Karle R, Crowe EL, Iezzoni AF. Chloroplast inheritance and DNA variation in sweet, sour, and ground cherry. J Hered. 2000;91(1):75–9.

7. Iezzoni AF. Acquiring cherry germplasm from Central and Eastern Europe. HortScience. 2005;40(2):304–8.

8. Bird KA, Jacobs M, Sebolt A, Rhoades K, Alger EI, Colle M, et al. Parental origins of the cultivated tetraploid sour cherry (*Prunus cerasus* L.). Plants People Planet. 2022;1–7.

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. Wang D. RFLP mapping, QTL Identification, and cytogenetic analysis in sour cherry. [East Lansing, Michigan]: Michigan State University; 1998.

11. Cai L, Stegmeir T, Sebolt A, Zheng C, Bink MCAM, Iezzoni A. Identification of bloom date QTLs and haplotype analysis in tetraploid sour cherry (*Prunus cerasus*). Tree Genet Genomes. 2018;14(22):1–11.

12. Barać G, Ognjanov V, Vidaković DO, Dorić D, Ljubojević M, Dulić J, et al. Genetic diversity and population structure of European ground cherry (*Prunus fruticosa* Pall.) using SSR markers. Sci Hortic. 2017;224:374–83.

13. Macková L, Vít P, Urfus T. Crop-to-wild hybridization in cherries — Empirical evidence from *Prunus fruticosa*. Evol Appl. 2018;11(9):1748–59.

14. Iezzoni AF, Hancock AM. A comparison of pollen size in sweet and sour cherry. HortScience. 1984;19(4):560–2.

15. Mochalova OV. Siberian gene pool of steppe cherry polyploids (*Prunus fruticosa* Pall.): cytomorphological estimation and prospects for breeding. BIO Web Conf [Internet]. 2020;24(00056). Available from: https://www.bio-conferences.org/10.1051/bioconf/20202400056

16. Vanderzande S, Zheng P, Cai L, Barac G, Gasic K, Main D, 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(7613):1–14.

17. Wöhner TW, Emeriewen OF, Wittenberg AHJ, Schneiders H, Vrijenhoek I, Halász J, et al. The draft chromosome-level genome assembly of tetraploid ground cherry (*Prunus fruticosa* Pall.) from long reads. Genomics. 2021;113:4173–83.

18. Kistner E, Kellner O, Andresen J, Todey D, Morton LW. Vulnerability of specialty crops to short-term climatic variability and adaptation strategies in the Midwestern USA. Climatic Change. 2018;146:145–58.

19. Unterberger C, Brunner L, Nabernegg S, Steininger W, Steiner AK, Stabentheiner E, et al. Spring frost risk for regional apple production under a warmer climate. PLoS ONE. 2018;13(7):1–18.

20. Marino GP, Kaiser DP, Gu L, Ricciuto DM. 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. Quero-Garcıá J, Campoy JA, Barreneche T, Le Dantec L, Wenden B, Fouché M, et al. Present and future of marker-assisted breeding in sweet and sour cherry. In: Acta Horticulturae. Yamagata, Japan: Acta Horticulturae; 2019. p. 1–14. (VIII; vol. 1).

22. Wang J, Liu W, Zhu D, Hong P, Zhang S, Xiao S, 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(122):1–11.

23. Falavigna V da S, Guitton B, Costes E, Andrés F. 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;10(January):1–17.

24. Fadón E, Fernandez E, Behn H, Luedeling E. A conceptual framework for winter dormancy in deciduous trees. Agron. 2020;10(2):241.

975 25. Goeckeritz C, Hollender CA. There is more to flowering than those DAM genes: the
976 biology behind bloom in rosaceous fruit trees. Curr Opin Plant Biol. 2021;59(101995):1–
977 10.

978 26. Bielenberg DG, Wang Y, Fan S, Reighard GL, Scorza R, Abbott AG. A deletion affecting
979 several gene candidates is present in the evergrowing peach mutant. J Hered.
980 2004;95(5):436–44.

981 27. Rodriguez A, Sherman W, Scorza R, Wisniewski M, Okie W. 'Evergreen' peach, its
982 inheritance and dormant behavior. J Am Soc Hortic. 1994;119(4):789–92.

983 28. Li Z, Reighard GL, Abbott AG, Bielenberg DG. Dormancy-associated MADS genes from
984 the EVG locus of peach [*Prunus persica* (L.) Batsch] have distinct seasonal and
985 photoperiodic expression patterns. Journal of Experimental Botany. 2009;60(12):3521–30.

986 29. Quesada-Traver C, Guerrero BI, Badenes ML, Rodrigo J, Ríos G, Lloret A. Structure and
987 expression of bud Dormancy-Associated MADS-Box Genes (DAM) in European plum.
988 Front Plant Sci. 2020;10(7613).

989 30. Calle A, Grimplet J, Dantec LL, Wünsch A. Identification and characterization of DAMs
990 mutations associated with early blooming in sweet cherry, and validation of DNA-based
991 markers for selection. Front Plant Sci. 2021;12(621491).

992 31. Tao R, Iezzoni AF. The S-RNase-based gametophytic self-incompatibility system in
993 *Prunus* exhibits distinct genetic and molecular features. Sci Hortic. 2010;124:423–33.

994 32. Sassa H. Molecular mechanism of the S-RNase-based gametophytic self-incompatibility in
995 fruit trees of Rosaceae. Breed Sci. 2016;66:116–21.

996 33. Ikeda K, Igic B, Ushijima K, Yamane H, Hauck NR, Nakano R, et al. Primary structural
997 features of the *S* haplotype-specific F-box protein, SFB, in *Prunus*. Sex Plant Reprod. 2004
998 Jan 1;16:235–43.

999 34. Matsumoto D, Tao R. Distinct self-recognition in the *Prunus S*-RNase-based gametophytic
1000 self-incompatibility system. J Hortic. 2016;85(4):289–305.

1001 35. Ushijima K, Sassa H, Tao R, Yamane H, Dandekar AM, Gradziel TM, et al. Cloning and
1002 characterization of cDNAs encoding *S*-RNases from almond (*Prunus dulcis*): primary
1003 structural features and sequence diversity of the *S*-RNases in Rosaceae. Mol Gen Genet.
1004 1998 Nov;260:261–8.

1005 36. Nunes MDS, Santos RAM, Ferreira SM, Vieira J, Vieira CP. Variability patterns and
1006 positively selected sites at the gametophytic self-incompatibility pollen *SFB* gene in a wild
1007 self-incompatible *Prunus spinosa* (Rosaceae) population. New Phytol. 2006 Nov;172:577–
1008 87.

1009 37. Schuster M. Self-incompatibility (*S*) genotypes of cultivated sweet cherries – An overview
1010 update 2020. OpenAgrar Repository. 2020. p. 1–45.

1011 38. Tsukamoto T, Potter D, Tao R, Vieira CP, Vieira J, Iezzoni AF, et al. Genetic and
1012 molecular characterization of three novel *S*-haplotypes in sour cherry (*Prunus cerasus* L.). J
1013 Exp Bot. 2008;59(11):3169–85.

1014 39. Tsukamoto T, Hauck NR, Tao R, Jiang N, Iezzoni AF. Molecular characterization of three
1015 non-functional *S*-haplotypes in sour cherry (*Prunus cerasus*). Plant Mol Biol. 2006 Sep
1016 25;62(3):371–83.

1017 40. Tsukamoto T, Hauck NR, Tao R, Jiang N, Iezzoni AF. Molecular and genetic analyses of
1018 four nonfunctional *S* haplotype variants derived from a common ancestral S haplotype
1019 identified in sour cherry (*Prunus cerasus* L.). Genetics. 2010;184(2):411–27.

1020 41. Hauck NR, Yamane H, Tao R, Iezzoni AF. Accumulation of nonfunctional *S*-haplotypes
1021 results in the breakdown of gametophytic self-incompatibility in tetraploid *Prunus*.
1022 Genetics. 2006;172(2):1191–8.

1023 42. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: Reference-free quality,
1024 completeness, and phasing assessment for genome assemblies. Genome Biol.
1025 2020;21(245):1–27.

1026 43. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO:
1027 Assessing genome assembly and annotation completeness with single-copy orthologs. J
1028 Bioinform. 2015;31(19):3210–2.

1029 44. Roach MJ, Schmidt S, Borneman AR. Purge Haplotigs: Synteny reduction for third-gen
1030 diploid genome assemblies. BMC Bioinform. 2018;19(460).

1031 45. Haug-Baltzell A, Stephens SA, Davey S, Scheidegger CE, Lyons E. SynMap2 and
1032 SynMap3D: Web-based whole-genome synteny browsers. J Bioinform. 2017;33(14):2197–
1033 8.

1034 46. Verde I, Jenkins J, Dondini L, Micali S, Pagliarani G, Vendramin E, et al. The Peach v2.0
1035 release: high-resolution linkage mapping and deep resequencing improve chromosome-
1036 scale assembly and contiguity. BMC Genom [Internet]. 2017;18(225). Available from:
1037 http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3606-9

1038 47. Dirlewanger E, Graziano E, Joobeur T, Garriga-Caldere F, Cosson P, Howad W, et al.
1039 Comparative mapping and marker-assisted selection in Rosaceae fruit crops. PNAS.
1040 2004;101(26):9891–6.

1041 48. Dirlewanger E, Quero-garc J, Dantec LL, Lambert P, Ruiz D, Dondini L, et al. Comparison
1042 of the genetic determinism of two key phenological traits, flowering and maturity dates, in
1043 three *Prunus* species: peach, apricot and sweet cherry. J Hered. 2012;109:280–92.

1044 49. Aranzana MJ, Decroocq V, Dirlewanger E, Eduardo I, Gao ZS, Gasic K, et al. *Prunus*
1045 genetics and applications after de novo genome sequencing: achievements and prospects.
1046 Hortic Res [Internet]. 2019;6(58). Available from: http://dx.doi.org/10.1038/s41438-019-
1047 0140-8

1048 50. Mitros T, Session AM, James BT, Wu GA, Belaffif MB, Clark LV, et al. Genome biology
1049 of the paleotetraploid perennial biomass crop *Miscanthus*. Nat Commun. 2020;11(5442).

1050 51. Gordon SP, Levy JJ, Vogel JP. PolyCRACKER, a robust method for the unsupervised
1051 partitioning of polyploid subgenomes by signatures of repetitive DNA evolution. BMC
1052 Genom. 2019;20(580).

1053 52. VanBuren R, Wai CM, Wang X, Pardo J, Yocca AE, Wang H, et al. Exceptional
1054 subgenome stability and functional divergence in the allotetraploid Ethiopian cereal teff.

Nat Commun [Internet]. 2020;11(884). Available from: http://dx.doi.org/10.1038/s41467-020-14724-z

53. Lovell JT, MacQueen AH, Mamidi S, Bonnette J, Jenkins J, Napier JD, et al. Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass. Nature. 2021;590:438–44.

54. Pinosio S, Marroni F, Zuccolo A, Vitulo N, Mariette S, Sonnante G, 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.

55. Hartley G, O'Neill R. Centromere Repeats: Hidden gems of the genome. Genes [Internet]. 2019 [cited 2022 Nov 22];10(233). Available from: https://www.mdpi.com/2073-4425/10/3/223

56. Gill N, Findley S, Walling JG, Hans C, Ma J, Doyle J, et al. Molecular and chromosomal evidence for allopolyploidy in soybean. Plant Physiol. 2009 Nov 3;151(3):1167–74.

57. Luo S, Mach J, Abramson B, Ramirez R, Schurr R, Barone P, et al. The cotton centromere contains a Ty3-gypsy-like LTR retroelement. Dawson DS, editor. PLoS One [Internet]. 2012 [cited 2022 Dec 5];7(4). Available from: https://dx.plos.org/10.1371/journal.pone.0035261

58. Bowman MJ, Pulman JA, Liu TL, Childs KL. A modified GC-specific MAKER gene annotation method reveals improved and novel gene predictions of high and low GC content in *Oryza sativa*. BMC Bioinform. 2017;18(522).

59. Campbell MS, Law M, Holt C, Stein JC, Moghe GD, Hufnagel DE, et al. MAKER-P: A tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. 2014;164(2):513–24.

60. Wang J, Childs K. defusion - a tool to untangle MAKER fused genome annotation [Internet]. 2018. Available from: https://github.com/wjidea/defusion

61. The Uniprot Consortium. UniProt: The universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49(D1):D480–9.

62. Emms DM, Kelly S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(238).

63. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: Architecture and applications. BMC Bioinform. 2009;10(421).

64. Dunn N, Unni D, Diesh C, Munoz-Torres M, Harris NL, Yao E, et al. Apollo: Democratizing genome annotation. PLoS Comput Biol. 2019;15.

65. 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;(1161):31–40.

66. Shi S, Li J, Sun J, Yu J, Zhou S. Phylogeny and classification of *Prunus sensu lato* (Rosaceae). J Integr Plant Biol. 2013;55(11):1069–79.

67. Quero-García J, Iezzoni AF, Pulawska J, Lang G, editors. Cherries: Botany, Production, and Uses. Boston, MA: CABI International; 2017.

68. Akšić MF, Cerović R, Ercişli S, Jensen M. Microsporogenesis and meiotic abnormalities in different 'Oblačinska' sour cherry (*Prunus cerasus* L.) clones. Flora. 2016;219:25–34.

69. Iezzoni AF, Sebolt AM, Wang D. Sour cherry breeding program at Michigan State University. Acta Hortic. 2005;667:131–4.

70. Hedrick UP, New York State Agricultural Experiment Station. The Cherries of New York [Internet]. Albany, NY: J.B. Lyon Company, State printers; 1915 [cited 2022 Dec 5]. Available from: http://www.biodiversitylibrary.org/bibliography/54515

71. 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. [East Lansing, Michigan]: Michigan State University; 2013.

72. Ramsey J, Schemske DW. Neopolyploidy in flowering plants. Annu Rev Ecol Syst. 2002;33:589–639.

73. Soares NR, Mollinari M, Oliveira GK, Pereira GS, Vieira MLC. Meiosis in polyploids and implications for genetic mapping: A review. Genes [Internet]. 2021 [cited 2022 Nov 23];12(1517). Available from: https://www.mdpi.com/2073-4425/12/10/1517

74. Tayalé A, Parisod C. Natural pathways to polyploidy in plants and consequences for genome reorganization. Cytogenet Genome Res. 2013;140:79–96.

75. Iezzoni A, Schmidt H, Albertini A. Cherries (*Prunus* spp.). In: Genetic resources of temperate fruit and nut crops. Wageningen, Netherlands: International Society for Horticultural Science; 1990. p. 110–73.

1115 76. Yoo S dong, Gao Z, Cantini C, Loescher WH, Nocker SV. Fruit ripening in sour cherry:
1116 Changes in expression of genes encoding expansins and other cell-wall-modifying
1117 enzymes. J Am Soc Hortic. 2003;128(1):16–22.

1118 77. Galassi A, Cappellini P, Miotto G. A descriptive model for peach fruit growth. Adv Hort
1119 Sci. 2000;14:19–22.

1120 78. Gasic K, Hernandez A, Korban SS. RNA extraction from different apple tissues rich in
1121 polyphenols and polysaccharides for cDNA library construction. Plant Mol Biol Rep.
1122 2004;22:437–8.

1123 79. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of
1124 occurrences of k-mers. J Bioinform. 2011;27(6):764–70.

1125 80. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for
1126 reference-free profiling of polyploid genomes. Nat Commun [Internet]. 2020;11(1432).
1127 Available from: http://dx.doi.org/10.1038/s41467-020-14998-3

1128 81. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: Scalable and
1129 accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome
1130 Res. 2017;27:722–36.

1131 82. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An
1132 integrated tool for comprehensive microbial variant detection and genome assembly
1133 improvement. PLoS One. 2014;9(11):e112963.

1134 83. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods.
1135 2012;9:357–9.

1136 84. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: Interactive visualization of de novo
1137 genome assemblies. J Bioinform. 2015;31(20):3350–2.

1138 85. Li H, Durbin R. Fast and accurate short read alignment with Burrows – Wheeler transform.
1139 J Bioinform. 2009;25(14):1754–60.

1140 86. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, et al. Juicer
1141 provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst.
1142 2016;3(1):95–8.

1143 87. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Lander ES, Aiden AP, et al. De novo
1144 assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds.
1145 Science. 2017;356(6333):92–5.

1146 88. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al. Juicebox
1147 provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst.
1148 2016;3(1):99–101.

1149 89. Ou S, Chen J, Jiang N. Assessing genome assembly quality using the LTR Assembly Index
1150 (LAI). Nucleic Acids Res. 2018;46(21):e126.

1151 90. Ou S, Jiang N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid
1152 identification of long terminal repeat retrotransposons. Mob DNA. 2019;10(1):48.

1153 91. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de
1154 novo detection of LTR retrotransposons. BMC Bioinform [Internet]. 2008 [cited 2022 Sep
1155 30];9(18). Available from:
1156 https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-18

1157 92. Ou S, Jiang N. LTR_retriever: A highly accurate and sensitive program for identification of
1158 long terminal repeat retrotransposons. Plant Physiology. 2018;176(2):1410–22.

1159 93. Ou S, Su W, Liao Y, Chougule K, Agda JR, Hellinga AJ, et al. Benchmarking transposable
1160 element annotation methods for creation of a streamlined, comprehensive pipeline. Genome
1161 Biol. 2019;20(275).

1162 94. Jung S, Ficklin SP, Lee T, Cheng CH, Blenda A, Zheng P, et al. The Genome Database for
1163 Rosaceae (GDR): Year 10 update. Nucleic Acids Res. 2014;42(237-D1244):1237–44.

1164 95. Jung S, Lee T, Cheng CH, Buble K, Zheng P, Yu J, et al. 15 years of GDR: New data and
1165 functionality in the Genome Database for Rosaceae. Nucleic Acids Res. 2019 Jan
1166 8;47(D1):D1137–45.

1167 96. Tang H, Zhang X, Miao C, Zhang J, Ming R, Schnable JC, et al. ALLMAPS: robust
1168 scaffold ordering based on multiple maps. Genome Biol. 2015;16(3).

1169 97. Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, et al. The arabidopsis
1170 information resource: Making and mining the “gold standard” annotated reference plant
1171 genome. Genesis. 2015;53(8):474–85.

98. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. J Bioinform. 2014;30(15):2114–20.

99. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. J Bioinform. 2013;29(1):15–21.

100. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. J Bioinform. 2009;25(16):2078–9.

101. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 2019;20:278.

102. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief Bioinformatics. 2013;14(2):178–92.

103. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7(3):562–78.

104. Wick R, Volkening J, Loman N. Porechop [Internet]. 2018. Available from: https://github.com/rrwick/Porechop

105. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: Visualizing and processing long-read sequencing data. J Bioinform. 2018;34(15):2666–9.

106. Li H. Minimap2: Pairwise alignment for nucleotide sequences. J Bioinform. 2018;34(18):3094–100.

107. Slater G, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinform [Internet]. 2005 [cited 2022 Dec 7];6(31). Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-6-31

108. Stajich J. Turns EXONERATE gff output into GFF for Gbrowse use - process_exonerate_gff3.pl [Internet]. University of California, Riverside, CA; 2015. Available from: https://github.com/hyphaltip/genome-scripts/blob/master/data_format/process_exonerate_gff3.pl.

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

110. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021;49(D1):D412–9.

111. Eddy SR, the HMMER development team. HMMER: Biological sequence analysis using profile hidden Markov models [Internet]. Chevy Chase, Maryland, US: Howard Hughes Medical Institute; 2020. Available from: http://hmmer.org

112. Johnson LS, Eddy SR, Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinform [Internet]. 2020 [cited 2022 Dec 7];11(431). Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-431

113. Dainat J. AGAT: Another GFF Analysis Toolkit to handle annotations in any GTF/GFF format. Zenodo. p. https://www.doi.org/10.5281/zenodo.3552717.

114. Campbell MS, Holt C, Moore B, Yandell M. Genome annotation and curation using MAKER and MAKER-P. Curr Protoc Bioinform. 2014;48(December):4.11.1-4.11.39.

115. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. PANTHER: A library of protein families and subfamilies indexed by function. Genome Res. 2003;13:2129–41.

116. Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, et al. TIGRFAMs: A protein family resource for the functional identification of proteins. Nucleic Acids Res. 2001;29(1):41–3.

117. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan: Protein domains identifier. Nucleic Acids Res. 2005;33(suppl_2):W116–20.

118. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: Tool for the unification of biology. Nat Genet. 2000;25:25–9.

119. Carbon S, Douglass E, Good BM, Unni DR, Harris NL, Mungall CJ, et al. The Gene Ontology resource: Enriching a GOld mine. Nucleic Acids Res. 2021;49(D1):D325–34.

120. Korf I. Gene finding in novel genomes. BMC Bioinform. 2004;5(59).

121. Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Res. 2008;18:1944–54.

122. Molinari NA, Petrov DA, Price HJ, Smith JD, Gold JR, Vassiliadis C, et al. Synteny and collinearity in plant genomes. Science. 2008;320:486–9.

123. R Core Team. R: A language and environment for statistical computing. [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2022. Available from: https://www.R-project.org/

124. Wickam H. ggplot2: Elegant Graphics for Data Analysis [Internet]. New York: Springer-Verlag; Available from: https://ggplot2.tidyverse.org

125. Sun X, Jiao C, Schwaninger H, Chao CT, Ma Y, Duan N, et al. Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. Nat Genet. 2020;52:1423–32.

126. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.

127. Suyama M, Torrents D, Bork P. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34(suppl_2):W609–12.

128. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. J Bioinform. 2009;25(15):1972–3.

129. Madeira F, Pearce M, Tivey ARN, Basutkar P, Lee J, Edbali O, et al. Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Res. 2022 Jul 5;50(W1):W276–9.

130. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. J Bioinform. 2014;30(9):1312–3.

131. McKain M. PhyDS: Phylogenetic iDentification of Subgenomes [Internet]. 2022. Available from: https://github.com/mrmckain/PhyDS

132. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol. 2016;33(6):1635–8.

133. Anand L, Rodriguez Lopez CM. ChromoMap: An R package for interactive visualization of multi-omics data and annotation of chromosomes. J Bioinform [Internet]. 2022;23(33). Available from: https://doi.org/10.1186/s12859-021-04556-z

134. Sanderson MJ. r8s: Inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. J Bioinform. 2003;19(2):301–2.

135. Agarwala, R. et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2016;44(D1):D7–19.

136. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.

Figure Captions

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 8 chromosomes of *P. persica* 'Lovell' v2.0 (46). 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 (22). Each gray line represents a syntenic block between the genomes. Small rearrangements between *P. avium* and each of the Montmorency subgenomes are evident.

Figure 2: Linearity comparison of linkage group 1 and a published sour cherry genetic map (11). 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. Post-filtering, 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 (96). Other chromosome sets (2 – 8) are shown in Supplementary Figure 5.

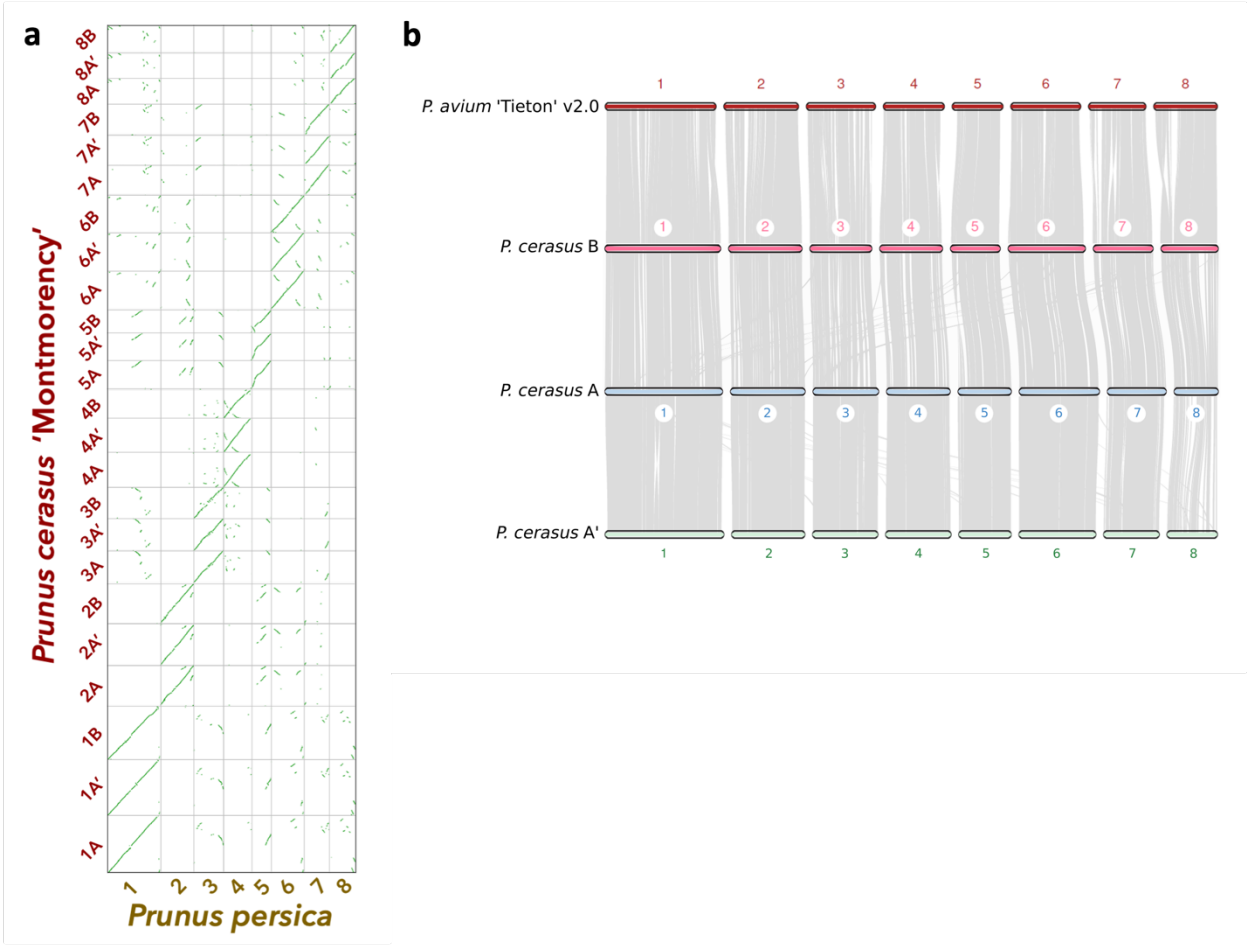
Figure 3: Chr8A and chr8A' affect the kmer groupings that differentiate subgenomes. a) Hierarchal clustering of all 24 chromosomes based on 25-mer abundance (present 10 times or more on every chromosome and at least twice as abundant on one homoeolog compared to its sisters). The star indicates chr8A and chr8A'. A corresponding heat map is shown in Supplementary Figure 6. 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 an assembly artifact (collapse) due to 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 group 8 chromosomes.

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 Supplementary Fig. 6) 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 (Fig. 3a). c) Group 5 and Group 6 25-mer densities (From Supplementary Fig. 7) 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 Supplementary Figure 8.

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 8 chromosome sets and colored by the progenitor they are most likely derived from. Window size for each tick mark ranges between 130 – 133 kilobases and is automatically optimized in chromoMap (reference 133) based on the largest chromosome's size. Mbp = Megabase pairs.

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* (23). *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.

1337 **Figures**



1338

1339 **Figure 1**

1340

1341

1342

1343

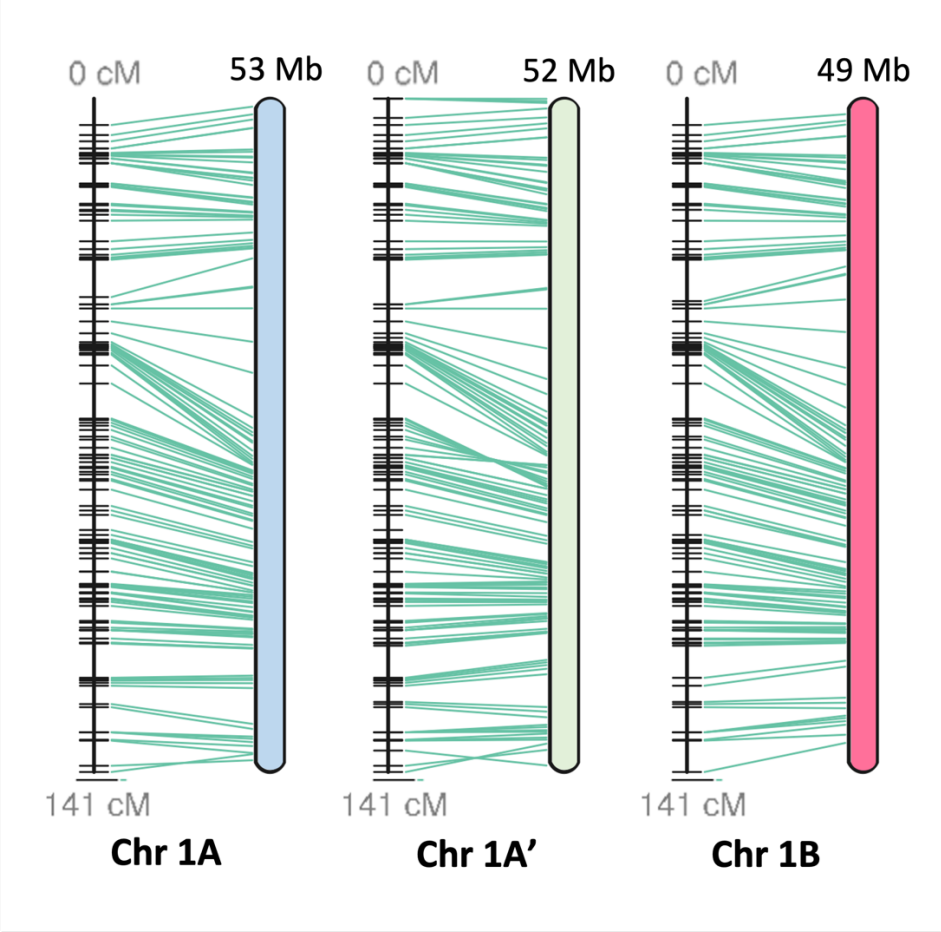


Figure 2

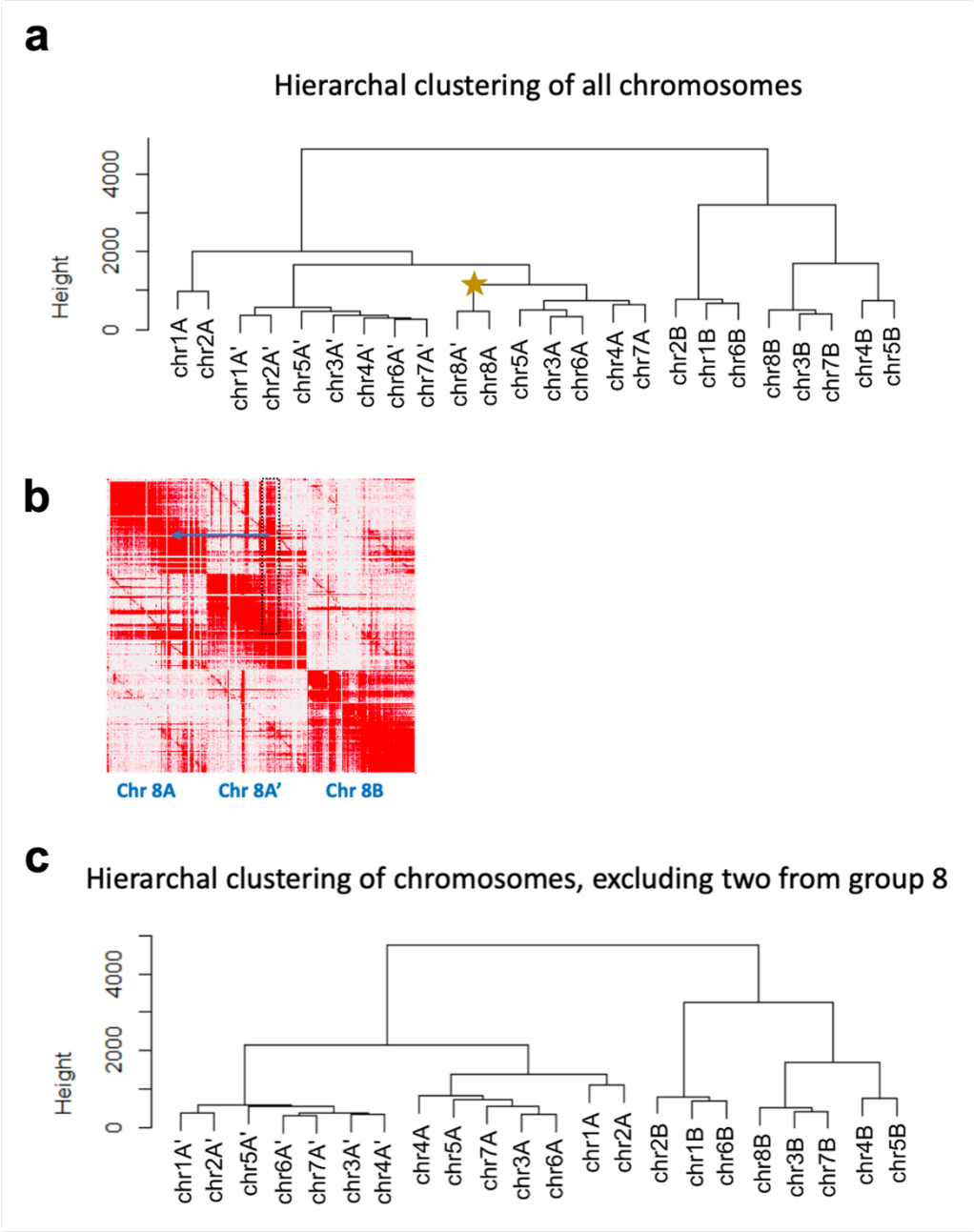


Figure 3

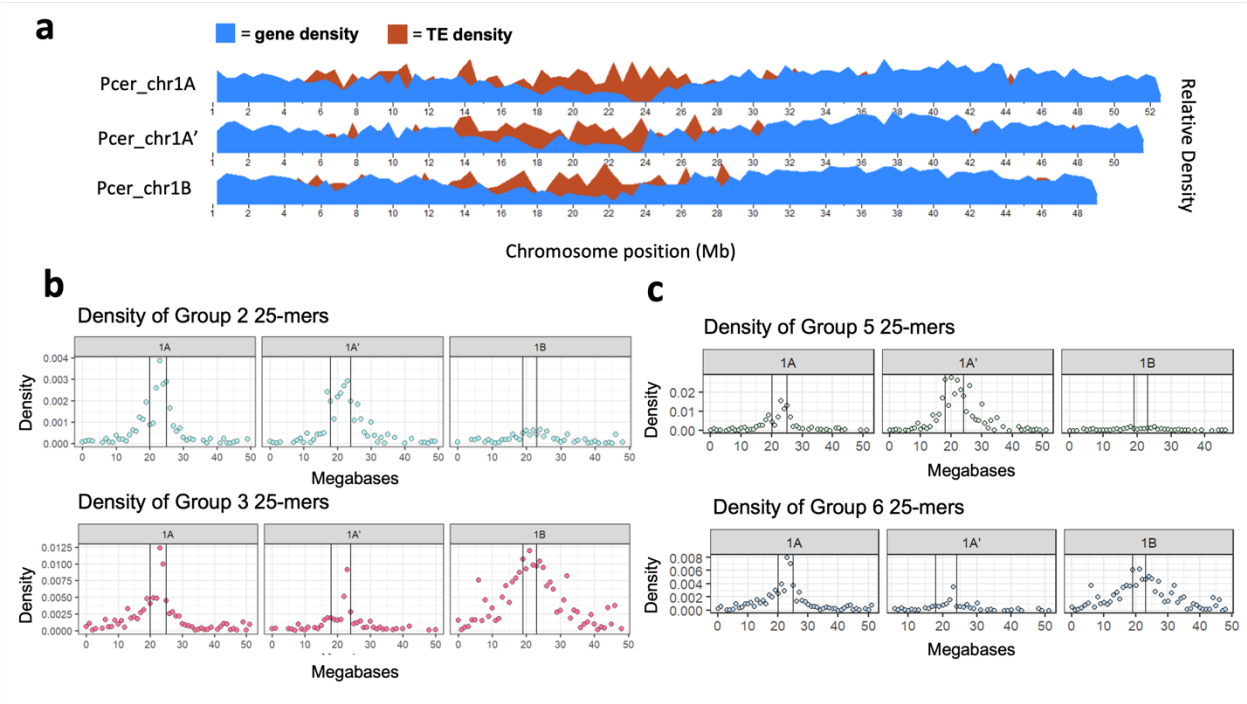


Figure 4

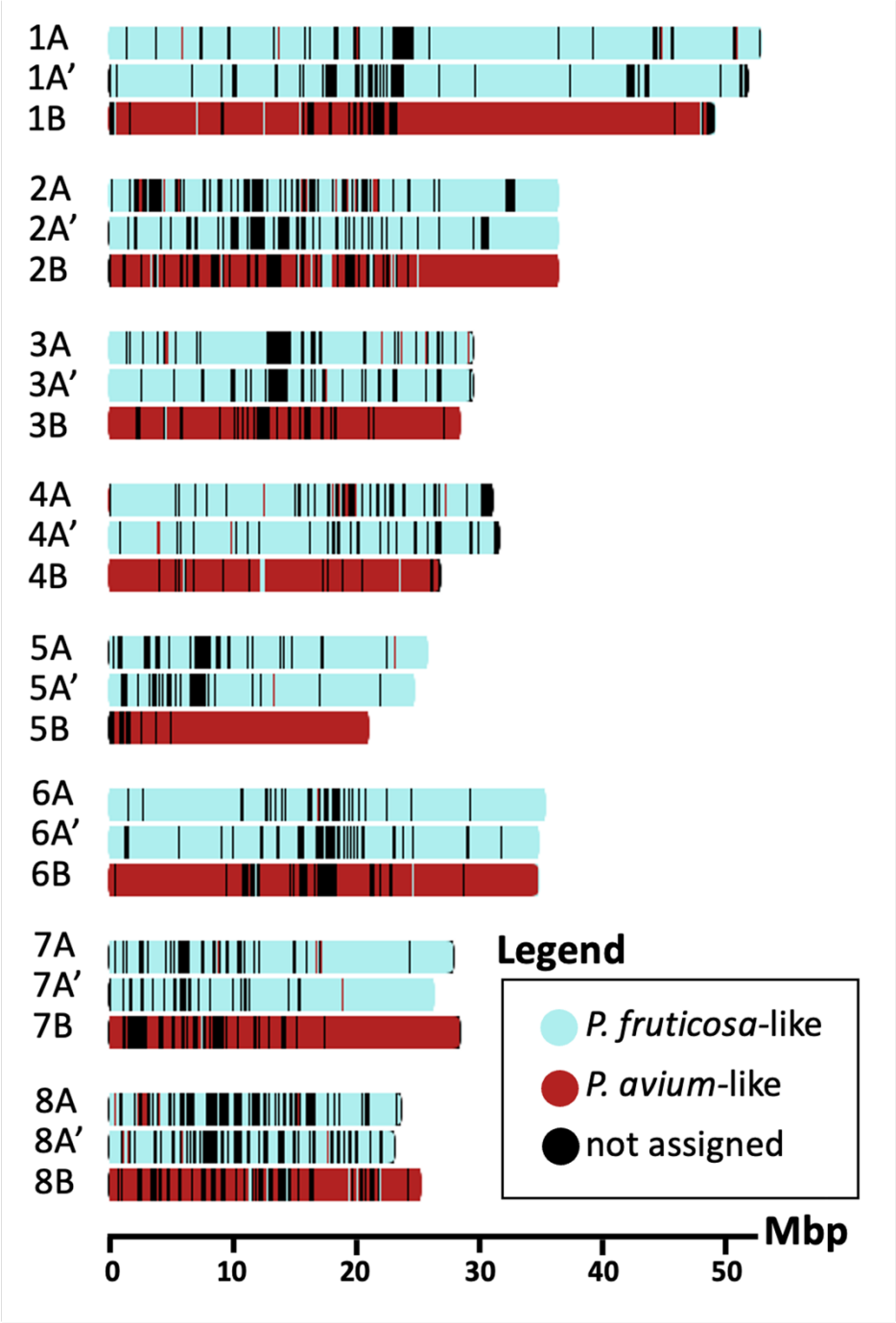
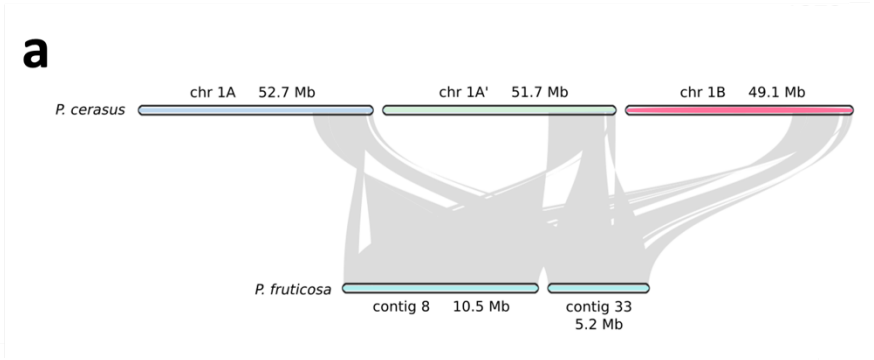
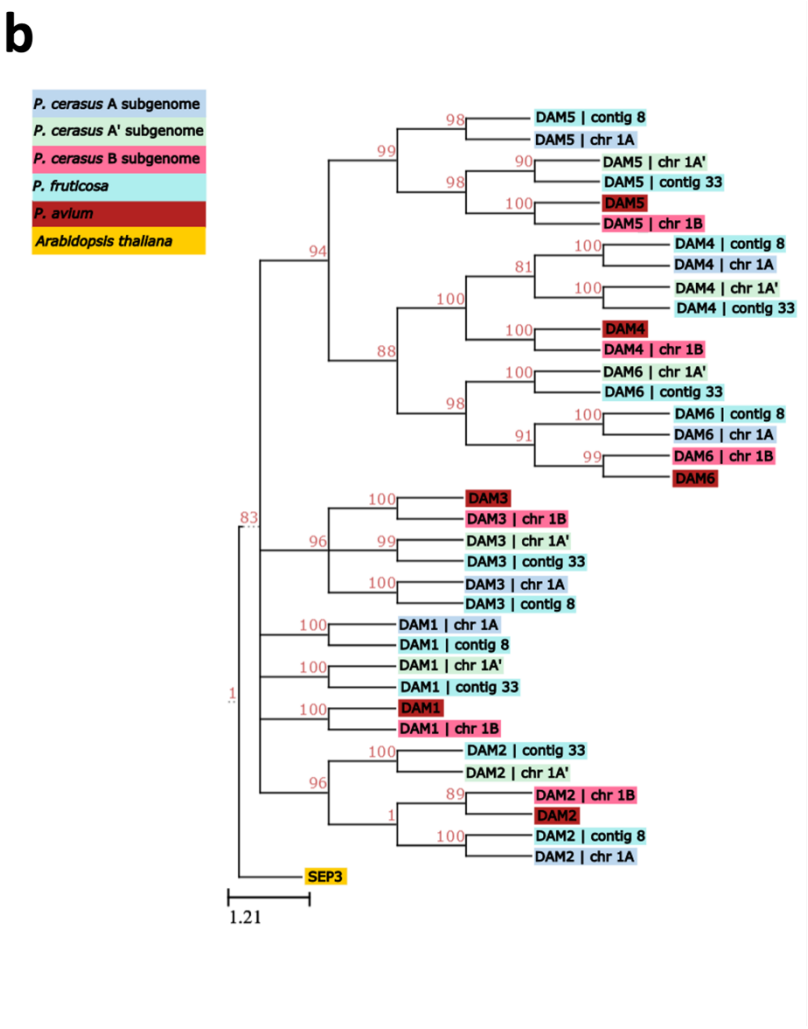


Figure 5





1387 **Figure 6**

1388

1389

Table Captions

Table 1: Summary of the 'Montmorency' assembly metrics. 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 (43); LTR = Long Terminal Repeat; TIR = Terminal Inverted Repeat; LAI = LTR Assembly Index (89).

Tables

Table 1

***P. cerasus* 'Montmorency' assembly metrics**					
Estimated haploid genome size (kmer analysis; k = 25)					621 Mb
Estimated Heterozygosity (total)					4.90%
class	aaab	aabb	aabc	abcd	
	2.430%	2.060%	0.001%	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	TIR	Helitron	Total	
	35.6%	11.6%	1.3%	48.5%	
LAI	Full assembly (incl. unanchored)				14.74
	Scaffolded assembly				17.09