The structure of the tetraploid sour cherry 'Schattenmorelle' (*Prunus cerasus* L.) genome reveals insights into its segmental allopolyploid nature

Thomas W. Wöhner[1], Ofere F. Emeriewen[1], Alexander H.J. Wittenberg[2], Koen Nijbroek[2], Rui Peng Wang[2], Evert-Jan Blom[2], Jens Keilwagen[4], Thomas Berner[4], Katharina J. Hoff[3], Lars Gabriel[3], Hannah Thierfeldt[3], Omar Almolla[5], Lorenzo Barchi[5], Mirko Schuster[1], Janne Lempe[1], Andreas Peil[1], Henryk Flachowsky[1]

[1]Julius Kühn Institute (JKI) – Federal Research Centre for Cultivated Plants, Institute for Breeding Research on Fruit Crops, Pillnitzer Platz 3a, D-01326, Dresden, Germany

[2]KeyGene N.V., P.O. Box 216, 6700 AE Wageningen, The Netherlands

[3]Institute of Mathematics and Computer Science, University of Greifswald, Walther-Rathenau-Str. 47, 17489 Greifswald, Germany

[4]Julius Kühn Institute (JKI) – Federal Research Centre for Cultivated Plants, Institute for Biosafety in Plant Biotechnology, Erwin-Baur-Str. 27, D-06484 Quedlinburg, Germany

[5]DISAFA – Plant genetics, University of Turin, Grugliasco (TO), 10095 Italy

[*]corresponding author: thomas.woehner@julius-kuehn.de

Keywords: genome assembly, *P. cerasus*, sour cherry, tetraploid

Abstract

Sour cherry (*Prunus cerasus* L.) is an economically important allotetraploid cherry species believed to have evolved in the Caspian Sea and Black Sea regions. How, when and where exactly the evolution of this species took place is unclear. It resulted from a hybridization of the tetraploid ground cherry (*Prunus fruticosa* Pall.) and an unreduced (2n) pollen of the diploid ancestor sweet cherry (*P. avium* L.). Some indications implement that the genome of sour cherry is segmental allopolyploid, but how it is structured and to what extent is unknown. To get an insight, the genome of the sour cherry cultivar 'Schattenmorelle' was sequenced at ~400x using Illumina

NovaSeq™ short-read and Oxford Nanopore long-read technologies (ONT R9.4.1 PromethION). Additionally, the transcriptome of 'Schattenmorelle' was sequenced using PacBio Sequel II SMRT cell sequencing at ~300x. The final assembly resulted in a ~629 Mbp long pseudomolecule reference genome, which could be separated into two subgenomes each split into eight chromosomes. Subgenome *Pce*$_S$_a which originates from *P. avium* has a length of 269 Mbp, whereas subgenome *Pce*$_S$_f which originates from *P. fruticosa* has a length of 299.5 Mbp. The length of unassembled contigs was 60 Mbp. The genome of the sour cherry shows a size-reduction compared to the genomes of its ancestral species. It also shows traces of homoeologous sequence exchanges throughout the genome. Comparative positional sequence and protein analyses provided evidence that the genome of sour cherry is segmental allotetraploid and that it has evolved in a very recent event in the past.

Background

Cherries include several species of the genus *Prunus*, which belong to the sub-family Spiraeoideae in the plant family Rosaceae (Potter et al. 2007). Although the fruits of several cherry species are used for consumption, only a few of them are grown and marketed on an economically significant scale (Quero-García et al. 2019). The two economically most important cherry species worldwide are the sweet cherry (*Prunus avium* L.) and the sour cherry (*Prunus cerasus* L.). Both species are thought to have originated in the Caspian Sea and Black Sea region (Quero-García et al. 2019).

Sour cherries commercial cultivation is mainly localized in Eastern and Central Europe, North America, and Central and Western Asia on an area of 217,960 ha. They are mainly cultivated for processing jams, juices, and whole fruits in preserved or dried forms. Furthermore, they are used in dairy products and baked goods. Their global production in 2021 was 1.51M tons of fruit and the cross production value in 2020 was 1.2 billion $US (https://www.fao.org/faostat/en/#data). The variability of morphological and fruit characteristics is very high in sour cherry, including fruit and juice colour, size, firmness, and fruit compounds (Schuster et al. 2017). Furthermore, tree growth varies from rather slender and upright to small and bushy types. Variation within ecotypes, differing e.g. in cold tolerance or growth habit, have been selected across Europe over time (Dirlewanger et al. 2007; Hancock 2008). However, just a small number of cultivars actually dominates the cultivation of sour cherry. ‘Schattenmorelle’ is the dominant cultivar (cv) in Middle Europe (Figure 1), whereas sour cherry production in the United States is still based on ‘Montmorency’ (Quero-Garcia et al. 2019).

'Schattenmorelle' was first described in France and today it is known in many countries with different names. In Poland, for example, it is called 'Łutovka' and in France 'Griotte du Nord' or 'Griotte Noir Tardive'.

The sour cherry is an allotetraploid with 2n=4x=32 chromosomes. It originated as a hybrid of an unreduced 2n pollen grain of *P. avium* (2n=2x=16) and a 1n egg cell of the tetraploid ground cherry *P. fruticosa* (2n=4x=32) (Kobel 1927; Olden and Nybom 1968). Evidence of hybridization events between sweet and ground cherries has already been found several times in areas where both species occur simultaneously (Macková et al. 2018, Hrotkó et al. 2020). The resulting hybrids were usually triploid

and were assigned to the secondary species *P.* ×*mohacsyana* Kárpáti. Natural occurrences of tetraploid sour cherries can be found in Eastern Turkey and the Caucasus region. There, they grow in forests and are used as wild forms for fruit production. The real area of origin is not known so far. Although *P. cerasus* can also be found in the wild in Europe, it is rather unlikely that those sour cherries are spontaneous hybrids. Since sour cherries are cultivated almost in many areas of the Northern hemisphere, they are often rather allochthonous individuals. The origin of the sour cherry thus seems to be based on a few hybridisation events. The results obtained by Olden and Nybom (1968) in experiments on the resynthesis of the species *P. cerasus* confirmed this hypothesis. The progeny from crosses between sweet and ground cherry showed the characteristic phenotype of the sour cherry. Studies based on chloroplast DNA markers strongly suggest that hybridisation between *P. avium* and *P. fruticosa* led to the emergence of *P. cerasus* at least twice (Dirlewanger et al. 2007). Furthermore, the hypothesis could also be confirmed by genomic *in situ* hybridisation (Schuster and Schreiber 2000) and transcriptome sequencing (Bird et al. 2022).

The sour cherry genome is presumed to have a genome size of 599 Mbp (Dirlewanger et al. 2007). It consists of the two subgenomes, each with eight chromosomes in the haploid set of chromosomes. One subgenome comes from the sweet cherry (*Pce*_a), while the other is from the ground cherry (*Pce*_f). However, the genome does not appear to be completely allopolyploid, since it has long been suspected that parts of the sour cherry genome might be segmental (Beaver and Lezzoni 1993; Olden and Nybom 1968; Raptopoulus 1941; Schuster and Wolfram 2005). The impact of hybridization and polyploidization between sweet and ground cherry has not been investigated so far, nor when this event originates. Cai et al. (2018) assumes that a mix of multi- and bivalent pairing led to imbalance segregation of chromosomes during meiosis in sour cherry, which in turn suggests that the sour cherry genome has not yet stabilized (Mason et al. 2020), or is still in the process of doing so. Recent advances in genome sequencing shed light into the complex structure and shape of polyploid genomes and their evolution (Zhang et al. 2021, Edger et al. 2019, Bertioli et al. 2019, Wu et al. 2021, Wang et al. 2019).

Here we report a high quality pseudo-chromosome-level genome assembly of the tetraploid sour cherry 'Schattenmorelle' (hereinafter referred to as Pce_S) generated

with a combination of Illumina NovaSeq short-read and Oxford Nanopore long-read sequencing technology. The sequences were scaffolded to chromosomes by Hi-C. In parallel, a full-length transcriptome of 'Schattenmorelle' was generated with the PacBio Sequel II SMRT cell long-read technology. Comparative sequence and amino acid analyses between data sets of *Prunus avium* cv 'Tieton' (hereinafter referred to as Pa_{T}) and *Prunus fruticosa* ecotype Hármashatárhegy (hereinafter referred to as Pf_{eH}) as representatives of the two ancestral species (Wang et al. 2019, Wöhner et el. 2021) and the two subgenomes of 'Schattenmorelle' Pce_{S}_a and Pce_{S}_f shed light into the evolution of sour cherry. Finally, possible HE within the sub-genomic structure of sour cherry were spotted, explaining the sour cherry genome segmental allopolyploidy.

Results

De novo assembly and scaffolding

A total of 68 Gb of paired-end Illumina sequencing data was obtained, corresponding to ~114x coverage of the estimated genome size of 599 Mbp. Using two PromethION flow cells, a total of 178 Gb was produced (~300x coverage). The longest ONT reads that together resulted in a 20x coverage were selected for assembly, having a minimum read length of 64,214 bp. Table S1 summarizes the properties of the 20-WGS-PCE.1.0 assembly after polishing. The *Prunus avium* and *Prunus fruticosa* contigs were then separated successfully by read mapping and contig selection that fit the hypothesis of 1 or more clear coverage peaks from the 20-WGS-PCE.1.0 assembly. The resulting two datasets, representing the subgenomes *Pce*$_S$_a and *Pce*$_S$_f, were purged and used for scaffolding using HI-C. After manual curation of both datasets, the final consensus genome assembly was scaffolded from 935 and 865 contigs of the *Pce*$_S$_a and *Pce*$_S$_f subgenomes, respectively. Eight clusters ideally representing the eight chromosomes were obtained for each subgenome (Figure S1). The final genome sequence is 628.5 Mbp long and consists of eight chromosomes for each subgenome (Figure 2). A total of 269 Mbp were assigned to subgenome *Pce*$_S$_a (N50 of 31.5 Mbp) and 299.5 Mbp (N50 of 39.4 Mbp) to *Pce*$_S$_f. Eighty-six and 134 unassembled contigs were unassigned to chromosomes for *Pce*$_S$_a (22.7 Mbp) and *Pce*$_S$_f (37.3 Mbp), respectively.

The longest scaffold from *Pce*$_S$_a is 52.8 Mbp and 53.5 Mbp from subgenome *Pce*$_S$_f (Table S2). Except for chromosome five, all scaffolds obtained from subgenome *Pce*$_S$_f are longer compared to the corresponding chromosome of subgenome *Pce*$_S$_a. The chloroplast sequence obtained was 158,178 bp and the mitochondrial sequence was 343,516 bp long (Figure S2).

Transcriptome sequencing, isoseq-analysis, structural and functional annotation

The total repeat content of the entire *Pce*$_S$ genome sequence was 49.7%. The total repeat content of subgenome *Pce*$_S$_a was 48.3% and of subgenome *Pce*$_S$_f 50.9%,

respectively (Table 1). The class I elements Gypsy comprised the largest fractions of repetitive elements in the *Pce*$_S$ genome sequence. A quantitative reduction between elements of this family was also detected in the *Pce*$_S$_f subgenome with a difference of 10.7%. Several elements could only be detected in one genotype of the two ancestral species. The TAD1 class I element only occurred in *Pf*$_{eH}$, while class II, order TIR - IS3EU, P, and Sola-3 were specifically detected in the genome of *Pa*$_T$. No element was found, which was only present in one of the two subgenomes of *Pce*$_S$. Several elements occurred in both subgenomes (class I, LINE – R1-LOA, RTE-X, SINE – tRNA-DEU- L2, class II, TIR – TcMar- Mariner, and DADA elements) but were not detected in *Pf*$_{eH}$ and *Pa*$_T$. The class I elements of the order LTR (ERV1, Pao) and Academ/-2 were only detected in one of the two genomes representing the ancestral species and in *Pce*$_S$_a and *Pce*$_S$_f. Iso-Seq results are summarized in Table S3. In total 248,218 high quality isoforms have been identified. Both, the high and the low quality isoforms have been used for genome annotation where each gene might be represented by multiple isoforms. A total of 107,508 transcripts (*Pce*$_S$_a: 53,497; *Pce*$_S$_f: 54,011) were predicted from the 60,123 gene models (*Pce*$_S$_a: 29,069; *Pce*$_S$_f: 31,054) obtained by structural annotation procedures (Table S4). Interproscan analysis detected 1,381,841 functional annotations (*Pce*$_S$_a: 649,310; *Pce*$_S$_f: 687,531) using 16 databases. Two third (71,870) of the transcripts were assigned with GO terms and 9,114 were found to be involved in annotated pathways.

Completeness and quality of the genome and transcriptome

BUSCO completeness of the *Pce*$_S$ genome was 99.0% (S: 16.7%, D: 82.3%, F: 0.4%, M: 0.6%, n: 1,614) respectively and comparable with *P. persica* 'Lovell' (99.3%) and *P. avium* 'Tieton' (98.3%, Figure S3). Completeness of subgenome *Pce*$_S$_a was higher (C: 89.4%, S: 84.8%, D: 4.6%, F: 1.5%, M: 9.1%, n: 1614) compared to subgenome *Pce*$_S$_f (C: 87.1%, S: 80.9%, D: 6.2%, F: 1.2%, M: 11.7%, n: 1,614). The calculated LAI index was 6.3 and low in comparison to other genomes (*Pp*$_L$: 17.6, *Pa*$_T$: 10.3, *Pf*$_{eH}$: 13.1). The LAI index for subgenome *Pce*$_S$_a was 7.1. The LAI index for subgenome *Pce*$_S$_f was 5.6 (Figure S4). The nucleotide heterozygosity rates were 94.9% for aaaa, 2.39 for aaab, 2.4 for aabb; 0.001 for aabc and 0.308 for abcd (Figure 3). The

comparison of genetic position and physical position of up to 1,856 markers of the five genetic sour cherry maps (Table S5a) showed a good co-linearity to the genome sequence (Figure S5). Busco evaluation on completeness of the annotated proteins resulted in 99.2% [C: 99.2% (S: 8.4%, D: 90.8%), F: 0.4%, M: 0.4%, n: 1,614]. The chloroplast sequence obtained contained 427 genes, 21 rRNAs and 136 tRNAs, whereas the mitochondrial sequence contained 188 genes, 3 rRNAs and 152 tRNAs (Figure S2). An ab-initio and homology-based gene prediction with 14 reference species was performed (IAA). Based on the homology prediction, thirty-four percent of the proteins showed the highest IAA towards *Prunus fruticosa* and 17.9% towards *P. avium*. Only 5.2% of the proteins showed no IAA to any of the used reference datasets used, which was due to ab-initio prediction. The data is summarized in Figure S6.

A comparison of transcripts of Pce_S and the annotation datasets of Pf_{eH} and Pa_T enabled a quantitative comparison of shared transcripts within the datasets (Table 2). A total number of 26,532 shared transcripts were found between the two subgenomes Pce_S_a and Pce_S_f and the genomes of Pf_{eH} and Pa_T. Thirty-eight percent of the *P. cerasus* proteins had a greater IAA to Pf_{eH}, whereas 54% showed a greater IAA to Pa_T. Eight percent showed an identical IAA to both ancestral species. A larger number of transcripts of both sour cherry subgenomes (Table 2) were assigned to the annotation data set of Pf_{eH}. A total of 13,425 transcripts from the Pce_S_a subgenome and 13,107 from the Pce_S_f subgenome were found in the genome sequences of Pf_{eH} and Pa_T. Seventy-five percent of the pool from the Pce_S_a subgenome showed a higher IAA to Pa_T and 17% to Pf_{eH}, while 59% from the pool originating from the Pce_S_f subgenome showed a higher IAA to Pf_{eH} and 32% to Pa_T.

Identification of syntenic regions and inversions

The sequences of the two subgenomes Pce_S_a and Pce_S_f and the genotypes Pa_T and Pf_{eH} of the two ancestral species *P. avium* and *P. fruticosa* were screened for duplicated regions using DAGchainer as previously published for peach (Verde et al. 2012). The seven major triplicated regions were found nearly one to one in *P. avium*

but not in *P. fruticosa,* which lacked regions 4 and 7 corresponding to Verde et al. (2012). *P. avium* and *P. fruticosa* seem to derive from the same paleohexaploid event like peach, but with a loss of the fourth and seventh paleoset of paralogs in *P. fruticosa.* The graphical analysis is summarized in figure S7).

Thirteen inversions were detected through positional co-linearity comparison between the two subgenomes using the molecular markers from the 9+6k SNP array (Figure S8). Five inversions were found between subgenome *Pce*$_S$_a and the genome sequence of *Pa*$_T$. Eleven inversions were found between sugenome *Pce*$_S$_f and *Pf*$_{eH}$ (Table S6). By comparing the position of amino acid sequences of orthologous proteins (synteny), we found 21 inversions when comparing *Pce*$_S$_f with *Pf*$_{eH}$. Only seven were found between *Pce*$_S$_a and *Pa*$_T$ and 16 between both subgenomes *Pce*$_S$_a and *Pce*$_S$_f (Figure S9).

Detection of de novo homoeologous exchanges

For the detection of de novo homoeologous exchanges we used three approaches by comparing inter- and intraspecific %-covered bases (genomic and transcriptomic) and %-IAA between proteins of *Pce*$_S$ to *Pa*$_T$ and *Pf*$_{eH}$ (Figure 4, Figure S10, S11). *Pce*$_S$ short-reads were mapped against *Pa*$_T$ and *Pf*$_{eH}$ and only species specific reads (*Pa*$_T$ and *Pf*$_{eH}$) were filtered into read-subsets. The obtained read-subsets were re-mapped against *Pce*$_S$_a and *Pce*$_S$_f and base coverage was calculated. A total of 1024 regions (100k window) were intraspecific %-covered bases from mapped reads (*Pce*$_S$_a to *Pa*$_T$, *Pce*$_S$_f to *Pf*$_{eH}$) was < than interspecific %-covered bases from mapped reads (*Pce*$_S$_a to *Pf*$_{eH}$, *Pce*$_S$_f to *Pa*$_T$) were discovered. In a second approach, translocations between the two subgenomes were localized by a short-read mapping analyses. Short-reads (RNAseq) from *P. cerasus, P. avium* and *P. fruticosa* obtained from Bird et al. (2022) were mapped on *Pce*$_S$. A total of 148 regions were intraspecific difference of %-covered bases from obtained RNAseq reads (*Pa* and *Pce*$_S$_a, *Pf* and *Pce*$_S$_f) > than interspecific difference of %-covered bases from obtained RNAseq reads (*Pf* and *Pce*$_S$_a, *Pa* and *Pce*$_S$_f) indicated homoeologous exchanges between the two subgenomes. Finally, 367 regions were the proportion of transcripts with intraspecific amino acid identity (*Pa*$_T$ and *Pce*$_S$_a, *Pf*$_{eH}$ and *Pce*$_S$_f) < than the proportion of

transcripts with interspecific amino acid identity (*Pf*$_{eH}$ and *Pce*$_{S}$_a , *Pa*$_{T}$ and *Pce*$_{S}$_f) were identified (Figure 4). Several regions were confirmed by calculating the 70% quantile of the IAA value within a window of 1 Mbp windows (Note S2). This confirms that there are transcripts in the *Pce*$_{S}$ genome whose IAA to the homoeologous representative genome (*Pf*$_{eH}$) is greater than to the homologous (*Pa*$_{T}$) representative. A total of 16 in *Pce*$_{S}$_a and 12 in *Pce*$_{S}$_f regions spanning 49 250k-windows were finally identified that match all three criteria indicating denovo homoelogous exchanges within the subgenomes (Figure 4, Figure S10). No evidence for an introgression of other *Prunus* species was found (Note S2, Note S3).

LTR dating and divergence of time estimation.

Left and right LTR identity of a subset of 2,385 (*Pce*$_{S}$_a), 3,028 (*Pce*$_{S}$_f), 3,130 (*Pa*$_{T}$) and 3,992 (*Pf*$_{eH}$) LTRs was analysed. The homologous genomes shared 200 (*Pce*$_{S}$_a versus *Pa*$_{T}$) and 100 (*Pce*$_{S}$_f versus *Pf*$_{eH}$) LTRs whereas 12 LTRs were shared by *Pce*$_{S}$_a versus *Pf*$_{eH}$ and *Pce*$_{S}$_f versus *Pa*$_{T}$. Only five common LTRs were found between *Pce*$_{S}$_a and *Pce*$_{S}$_f and 13 between *Pa*$_{T}$ and *Pf*$_{eH}$. A summary of the LTRs insertion time is shown in figure 5A.

The youngest shared LTRs between *Pce*$_{S}$_a and *Pa*$_{T}$ were calculated with 103,896.1 generations and between *Pce*$_{S}$_f and *Pf*$_{eH}$ with 97,402.6 generations. When comparing the homeologous chromosomes, the youngest shared LTRs between *Pa*$_{T}$ and *Pf*$_{eH}$ was calculated with 116,883.1 generations, between *Pce*$_{S}$_a and *Pce*$_{S}$_f with 194,805.2 generations. LTRs of *Pa*$_{T}$ were also found in subgenome *Pce*$_{S}$_f and calculated with 207,792.2 generations. LTRs of *Pf*$_{eH}$ were detected in subgenome of *Pce*$_{S}$_a and calculated with 149,350.6 generations. This indicates an exchange of LTRs between the two subgenomes.

A total of 834 single copy orthogroups among nine genomes were found and used for single protein alignments. Single alignments were concatenated and a final alignment with nine amino acid sequences representing each species with 419,586 amino acid positions was used for phylogenetic tree construction. Using the RelTime method, the

estimated divergence time between the genera *Malus* and *Prunus* was 50.4 Mya. The species groups *P. persica* and *P. mume* diverged from the *P. yedonensis*/*P.avium*/*P.fruticosa* group 11.6 Mya. Based on this model, the divergence of the two subgenomes of *P. cerasus* compared to the genome sequences of Pa_T and Pf_{eH} was estimated with 2.93 Mya and 5.5 Mya respectively (Figure 5B).

Discussion

The genome of the economically most important sour cherry 'Schattenmorelle' in Europe was sequenced using a combination of Oxford Nanopore R9.4.1 PromethION long read technology and Illumina NovaSeq™ short read technology. After assignment of the long-reads to the two subgenomes and Hi-C analysis, the final assembly was 629 Mbp.

This sequence was used to study structural changes present in the allotetraploid sour cherry genome after its emergence. Therefore, the sour cherry genome sequence was compared to the published genome sequences of *Prunus avium* 'Tieton' (Pa_T, Wang et al. 2019) and *Prunus fruticosa* ecotype Hármashatárhegy (Pf_{eH}, Wöhner et al. 2021a) representing genotypes of the two ancestral species. The size of the subgenome Pce_S originating from *P. avium* was 269 Mbp. A similar genome size (271 Mbp) is described for the *Prunus avium* 'Big Star' (Pinosio et al. 2020) and 'Sato Nishiki' (Shirarsawa et al. 2017). Larger differences were found to 'Regina' with 279 Mbp and 'Tieton' with 344 Mbp (Wang et al. 2020, Le Dantec et al. 2019). Differences were also found between the size of subgenome Pce_S_f (299 Mbp) and the genome of the ground cherry genotype *P. fruticosa* ecotype Hármashatárhegy (366 Mbp, Wöhner et la. 2021a). These differences indicate a reduction of the subgenome Pce_S_a by 0.49%-21%, whereas for subgenome Pce_S_f a reduction of 18.29% was found. The reduction in genome size for allotetraploid species in comparison to their ancestral genomes is common, as already reported for *Nicotiana tabacum* (1.9%-14.3%) and *Gossypium* species (Leitch et al., 2008; Hawkins et al., 2009; Renny-Byfield et al., 2011), with an overall downsizing rate for angiosperms calculated as 0%-30% (Zenil-Ferguson et al. 2016). Genome downsizing in repose to a genome hybridisation event can be explained with evolutionary advantages, which give these species with smaller

genomes a selection advantage in the long-term (Knight et al. 2005, Zenil-Ferguson et al. 2016).

Although downsizing of the *P. cerasus* subgenomes is most probable, enlargement and expansion of the genomes of ancestral species during evolution would be another possibility. However, an increase in genome size during the evolution of a species has only rarely been documented (Leitch et al. 2008, Jackobs et al. 2004, Kim et al. 2014). BUSCO analysis provides additional evidence for the reduction of genome size. Although the number of genes does not correlate with genome size in eukaryotes (Pierce 2012), differences between the ancestral genomes and sour cherry could be observed when looking at BUSCO-completeness. Considering both subgenomes and the genomes of Pa_{T} and Pf_{eH}, a completeness of > 96.4% was obtained. However, the completeness of the single subgenomes was only 89.4% for Pce_{S}_a and 87.1% for Pce_{S}_f. Structural differences between the *P. cerasus* subgenomes and the genomes of Pa_{T} and Pf_{eH} were also found by comparing the number of repetitive elements. While the content of repetitive elements differs by only 0.86% between Pf_{eH} and Pce_{S}_f, it is 17.8% between Pa_{T} and Pce_{S}_a. Whether this is a consequence of hybridisation remains speculative and would deserve further studies. An increase of class I elements Gypsy from six percent in the Pa_{T} genome to 7.3% in subgenome Pce_{S}_a indicates an expansion of this class following the formation of the sour cherry genome or a possible reduction of non-repetitive sequences in the corresponding subgenome resulting in a smaller genome size.

A comparison of syntenic regions showed a high degree of collinearity between Pce_{S}, Pa_{T} and Pf_{eH} genomes (Figure 2, Figure S7-9), with single inversions between the respective chromosome pairs. Using the genome of *P. persica,* seven triplicated regions were detected in Pce_{S}_a, confirming that these genomes descend from a palaeohexploid ancestor. However, the triplicated regions 4 and 7 in Pce_{S}_f were only detected in highly fragmented form or have been lost. Hao et al. (2022), who described a rapid loss of homoeologs immediately after polyploidy events, described similar finding.

Based on Ranallo-Benavidez (2020) the results from the k-mer analysis confirm that the genome of sour cherry can be considered as highly heterozygous and segmental allotetraploid. Furthermore, genomes of segmental allopolyploids may possess a mix of auto- and allopolyploid segments through duplication-deletion events as a result of homoeologous exchanges leading to either reciprocal translocations or homoeologous non-reciprocal translocations (Mason and Wendel, 2020). Whereas autotetraploids have an aaab > aabb rate, allotetraploids are considered to have aaab < aabb. The near identical rate between aaab and aabb in *Pce*$_S$ provides strong evidence that the sour cherry is a segmental mix of auto- and allotetraploidy.

Due to this assumption of segmental allotetraploidy, homoeologous recombination between homoeologous chromosomes is very likely. This is confirmed by the coverage and amino acid identity analyses. Homoeologous exchange events between the chromosomes of subgenome *Pce*$_S$_a and *Pce*$_S$_f were detected (Figure 4, Figure S10). These exchanges are not balanced but probably a product of a duplication/deletion event as described by Mason et al. (2020), generating the proposed mosaic of genomic regions representing one or the other subgenome.
Additional mapping of transcriptomes from six other *Prunus* species indicates that no major introgression from one of these species occurred. Using 14 reference species, 60,123 gene models were annotated. Almost the same number was assigned to the two *P. cerasus* subgenomes (Table 2). No evidence was found for large introgressions from any of the reference species (Note S2, S3). By comparing the amino acid identity of the proteins of *Pa*$_T$ and *Pf*$_{eH}$ with the respective sour cherry subgenome, the identified translocations via read mapping could be confirmed. The majority of the transcripts (51%) could be assigned to the genotypes *Pa*$_T$ and *Pf*$_{eH}$ of the two ancestral species *P. avium* and *P. fruticosa* (Figure S6). 5.2% of the transcripts could not be assigned to any of the reference species. Only < 1% of the transcripts could not be assigned to one of the ancestral species. They showed equivalent matches to both species and are probably a product of ab-initio prediction. A total of 49,698 proteins in subgenome *Pce*$_S$_a and 48,576 proteins in *Pce*$_S$_f shared only 13,435 and 13,107 proteins with *Pa*$_T$ and *Pf*$_{eH}$, respectively. A total of 75% of the proteins of subgenome *Pce*$_S$_a matched better to *Pa*$_T$ compared to *Pf*$_{eH}$, whereas only 59% of *Pce*$_S$_f mapped

373 better to *Pf*$_{eH}$ than to *Pa*$_{T}$ (Table 2). Subgenome *Pce*$_{S}$_a seems to be closer to *Pa*$_{T}$ than
374 *Pce*$_{S}$_f to *Pf*$_{eH}$. This was confirmed by an evolutionary approach that calculated the
375 separation of the subgenome *Pce*$_{S}$_f from *P. fruticosa* 5.5 Mya, while subgenome
376 *Pce*$_{S}$_a separated from *P. avium* 2.93 Mya (Figure 5B).
377 To validate these results, the insertion events of long terminal repeats between *P.*
378 *avium, P. fruticosa* and the subgenomes were calculated. Assuming a *Prunus*-specific
379 rate of $7.7*10^{-9}$ mutations per generation (Xie et al. 2016), LTRs of the same type with
380 the same insertion time were identified in the same positional order in the different
381 (sub)genomes. The most recent co-occurring LTRs between the genomes of *Pa*$_{T}$ and
382 *Pf*$_{eH}$ could be dated at 116,883.1 generations. Exact data on the duration of the
383 generation time of *Prunus* species in natural habitats do not exist. Although the
384 juvenile phase of many *Prunus* species is usually completed after 5 years (Besford et
385 al. 1996), it can be assumed that the times for a generation are considerably higher.
386 Many fruit species are hardly or not at all able to rejuvenate by seeds under natural
387 conditions (Coart et al. 2003), or they rejuvenate mainly by root suckers (Li et al. 2022).
388 Other studies on *Prunus* therefore assume a duration of 10 years per generation
389 (Wang et al. 2021), although even this seems rather too little.
390 Assuming that a generation change is to be expected after 10 to 60 years (Besford et
391 al. 1996), this would correspond to a time period of ~ 1 to 6 Mya. The youngest co-
392 occurring LTR could be estimated at 1.9 Mya. This suggests that *P. fruticosa* and *P.*
393 *avium* probably shared a gene pool between ~1 Mya and 2 Mya. It should be noted
394 that this estimate can vary greatly depending on the number of years per generation
395 used in the calculation (Figure 5A). Based on the results of the protein dating a
396 generation time of 30 years is more likely for *P. avium*. For *P. fruticosa*, which occurs
397 less frequently in natural habitats and reproduces mainly via root suckers, the
398 generation time seems to be somewhat longer at 55 to 60 years. Some LTRs present
399 in *Pa*$_{T}$, but absent in *Pf*$_{eH}$ were found in subgenome *Pce*$_{S}$_f only and vice versa. Other
400 class I elements (LTR - ERV1, Pao) and Academ/-2 specifically detected in one of the
401 two genotypes *Pa*$_{T}$ and *Pf*$_{eH}$ representing the two ancestral species of sour cherry and
402 in *Pce*$_{S}$_a and *Pce*$_{S}$_f, which indicates a transfer of these elements between the two
403 subgenomes following the formation of the allotetraploid *P. cerasus* genome. This is
404 a further indication for a segmental exchange between the two sour cherry

subgenomes. Mason et al. (2020) speculated that uni-directional homeologous exchange was observed in recent or synthetic allopolyploids. However, our results confirm this hypothesis by the evidence that sour cherry is a recent allopolyploid with autopolyploid segments derived from uni-directional homoeologous exchanges.

Conclusion

After sequencing of the genome of the sour cherry 'Schattenmorelle', the following can be concluded. The genome of sour cherry is segmental allotetrayploid. It consists of two subgenomes, one derived from the sweet cherry *P. avium* and one from the ground cherry *P. fruticosa*. DNA sequences have been repeatedly exchanged between the two subgenomes. At the same time, a reduction in genome size has taken place. Other *Prunus* species have not contributed to the evolution of this species. No evidence was found for introgressions in the sour cherry genome derived from *Prunus* species other than *P. avium* and *P. fruticosa*. Sour cherry is a very young *Prunus* species. The origin of this species is estimated about 1 mya earliest.

Material and Methods

Plant Material, DNA and RNA extraction, sequencing and iso-seq analysis

Snap frozen *Prunus cerasus* L. 'Schattenmorelle' (accession KIZC99-2, Figure 1, supplements 1.1) young leaf material was sent to KeyGene N.V. (Wageningen, The Netherlands). High molecular weight extracted DNA (Wöhner et al. 2021a) was used to generate 1D ligation (SQK-LSK109) libraries which were subsequently sequenced on two Oxford Nanopore Technologies (ONT) R9.4.1 PromethION flow cells. The same material was used to generate an Illumina PCR free paired-end library (insert size of ~550 bp) which was sequenced on a HiSeq 4000™ platform using 150bp and 125bp paired-end sequencing.

Snap frozen tissues from buds, flowers, leaves and fruits were collected and in the field and total RNA was extracted with Maxwell® RSC Plant RNA Kit (Promega). Two pools were generated and used for PacBio Iso-Seq library preparation (Procedure &

Checklist – Iso-Seq™ Express Template Preparation for Sequel® and Sequel II Systems, PN 101-763-800 Version 02). Each library pool was sequenced on a single 8M ZMW PacBio Sequel II SMRT cell (supplements 1.1). Obtained full-length reads with 5'-end primer, the 3'-end primer and the poly-A tail were filtered and these sequences were trimmed off. Transcripts containing (artificial) concatemers were completely discarded. Isoforms (consensus sequence) generated by full-length read clustering (based on sequence similarity), were finally polished with non-full-length reads using Arrow (SMRT Link v7.0.0, https://www.pacb.com/wp-content/uploads/SMRT_Tools_Reference_Guide_v600.pdf).

De novo assembly and scaffolding

The aligner Minimap2 (v2.16-r922, Li et al. 2018) and assembler Miniasm (v0.2-r137-dirty, Li et al. 2016) were used for raw data assembly generation. Racon (vv1.4.10, Vaser et al. 2017) and Pilon (v1.22, Walker et al. 2014) were used for base-quality improvement with raw ONT and Illumina read data. Chromosome-scale scaffolding was performed by Phase Genomics (Seattle, Washington, USA) with Proximo Hi-C (supplements 1.2). The resulting assembly was designated as 20-WGS-PCE_<Avium|Fruticosa>.2.0 _<Contig|Scaffold>.

Correctness, completeness and contiguity of the Prunus cerasus genome sequence

The BUSCO (Benchmark Universal Single-Copy Orthologs - Galaxy Version 4.1.4) software was used for quantitative and quality assessment of the genome assemblies based on near-universal single-copy orthologs. The long terminal repeat (LTR) assembly Index (LAI) Ou et al. (2018) was calculated with LTR_retriever 2.9.0 (https://github.com/oushujun/LTR_retriever) to evaluate the assembly continuity between the final genome sequence of *P. cerasus* 'Schattenmorelle' and *Prunus fruticosa* ecotype Hármashatárhegy (Wöhner et al. 2021a, Pf_1.0), *P. avium* 'Tieton' (Wang et al. 2020) and *P. persica* 'Lovell' (Verde et al. 2017) respectively. LTR_harvest (genometools 1.6.1 implementation) was used to obtain LTR-RT candidates. The genome size was also estimated by k-mer analysis (supplements 1.3) using Illumina short read data. The merged datasets were subsequently used to generate a histogram dataset representing the k-mers from all datasets. GenomeScope (Galaxy

Version 2.0, Ranallo-Benavidez 2020) was used to generate a histogram plot of k-mer frequency of different coverage depths using the tetraploid ploidy level (k-mer length 19). Marker sequences and genetic positions from five available genetic sour cherry maps (M172x25-F1, US-F1, 25x25-F1, Montx25-F1, RE-F1) and 14,644 SNP markers (9+6k array) were downloaded from the Genome Database for Rosaceae (GDR, https://www.rosaceae.org/). The marker sequences were mapped on the chromosome sequences using the mapping software bowtie2 (Galaxy Version 2.5.0+galaxy0, Langmead et al. 2012) implementation on the Galaxy server (https://usegalaxy.org) with standard settings.

Structural and functional annotation

For an inter species repeat comparison, a species-specific repeat library was generated with RepeatModeler open-1.0.11, and the genome was subsequently masked with RepeatMasker open-4.0.7. For structural genome annotation, another species-specific repeat library for PCE_1.0 was generated with RepeatModeler2 (Flynn et al., 2020) version 2.0.2, and the genome was subsequently masked with RepeatMasker 4.1.2. (Further details on the repeat masking software configuration are available in supplements 1.4.1.).

To generate extrinsic evidence for structural annotation of protein coding genes, short read RNA-Seq library SRR2290965 was aligned to the genome using HiSat2 version 2.1.0 (Kim et al., 2019). The output SAM file was converted to BAM format using SAMtools (Li et al, 2009). The resulting alignment file was further used by both BRAKER1 (Hoff et al., 2016; Hoff et al., 2019) and GeMoMa (Keilwagen et al. 2016). Further, a custom protocol was used for integrating long read RNA-Seq data into genome annotation (supplements 1.4.2). In short, protein coding genes were called in Cupcake transcripts using GeneMarkS-T (Tang et al., 2015) and these predictions were converted to hints for BRAKER1. In addition, intron coverage information from long read to genome spliced alignment with Minimap2 (Li et al. 2018) was provided to BRAKER1.

A combination of BRAKER1 (Hoff et al., 2016, Hoff et al., 2019), BRAKER2 (Bruna et al., 2021), TSEBRA (Gabriel et al., 2021), and GeMoMa (Keilwagen et al. 2016) was

used for the final annotation of protein coding genes. BRAKER pipelines use a combination of evidence-supported self-training GeneMark-ET/EP (Lomsadze et al., 2014; Bruna et al., 2020) (here version 4.68) to generate a training gene set for the gene prediction tool AUGUSTUS (Stanke et al., 2008; here version 3.3.2). BRAKER1 version 2.1.6 was here provided with BAM-files of from short and long read RNA-Seq to genome alignments, and with gene structure information derived from Cupcake transcripts using GeneMarkS-T. This generated a gene set that consists of ab initio and evidence supported predictions. A separate gene set was generated with BRAKER2, which uses protein to generate a gene set. We used the OrthoDB version 10 (Kriventseva, 2019) partion of plants in combination with the full protein sets of *Prunus fruticosa* (Wöhner et al., 2021b), *Prunus armeniaca* (GCA 903112645.1)*, Prunus avium* (GCF_002207925.1)*, Prunus dulcis* (GCF_902201215.1)*, Prunus mume* (GCF_000346735.1)*, Prunus persica* (GCF_000346465.2) as input for BRAKER2. Both the BRAKER1 and BRAKER2 AUGUSTUS gene sets were combined with a GeneMarkS-T derived gene set using TSEBRA (Gabriel et al., 2021) from the long_reads branch on GitHub with a custom configuration file (supplements 1.4.3.) incorporating evidence from BRAKER1 and BRAKER2.

GeMoMa was run on the genome assembly of Schattenmorelle using 14 reference species and experimental transcript evidence (supplements 1.4.4). GeMoMa gene predictions of each reference species were combined with TSEBRA predictions using the GeMoMa module GAF and subsequently, UTRs were predicted in a two-step process based on mapped Iso-seq and RNA-seq data using the GeMoMa module AnnotationFinalizer (supplements 1.4.5). First, UTRs were predicted based on Iso-seq data. Second, UTRs were predicted based on RNA-seq data for gene predictions without UTR prediction from the first step. An assembly hub for visualization of the *Prunus cerasus* genome with structural annotation was generated using MakeHub (Hoff, 2019; supplements 2). The functional annotation was performed with the Galaxy Europe implementation of InterProScan (Galaxy Version 5.59-91.0+galaxy3, Jones et al. 2014, Cock et al 2013, Zdobnov et al. 2001, Quevillon et al. 2005, Hunter et al. 2009). The chloroplast and mitochondria sequences were annotated with GeSeq (Tillich et al. 2017, supplements 1.4.5).

Identification of syntenic regions

Structural comparison of orthologous loci between the subgenomes *Pce*$_S$_a and *Pce*$_S$_f of *Prunus cerasus* and the two genotypes *Pa*$_T$ and *Pf*$_{eH}$ as representatives of the two genome donor species *P. avium* and *P. fruticosa* was calculated with the final annotations using SynMap2 (Haug-Baltzell et al. 2017) available at the CoGe platform (https://genomevolution.org/coge/). Analysis on triplication events were performed with standard settings and Last as Blast algorithm at a ratio coverage depth 3:3 in SynMap2 (Haug-Baltzell et al. 2017).

Identification of homoeologous exchange regions

Homoeologous exchanges were identified on the amino acid, transcript and genomic level.

Calculation of amino acid identity

Identity of amino acids (IAA) between all reference annotations homology-based gene prediction was calculated by GeMoMa using the default parameters. Subsequently, the *Pce*$_S$ genome was divided into 250k windows, and the percentage of proteins showing a higher IAA between *Pf*$_{eH}$ (Wöhner et al. 2021b) and *Pa*$_T$ (Wang et al. 2020) to the respective subgenome (*Pce*$_S$_a and *Pce*$_S$_f) was determined. The percentage of proteins in this window, which were more similar to *Pa*$_T$ was finally subtracted from the percentage of proteins which were more similar to *Pf*$_{eH}$. A proportion of transcripts with higher intraspecific amino acid identity (between *Pa*$_T$ and *Pce*$_S$_a or *Pf*$_{eH}$ and *Pce*$_S$_f) is expected compared to the proportion of transcripts with interspecific amino acid identity match (*Pf*$_{eH}$ and *Pce*$_S$_a or *Pa*$_T$ and *Pce*$_S$_f). Opposite cases, indicate potential translocations between the two subgenomes *Pce*$_S$_a and *Pce*$_S$_f and were plotted into a circus plot (Figure S11).

Read mapping and coverage analysis

RNAseq raw data published by Bird et al. (2022) was obtained from NCBI sequence read archive (SRA) for the following species: *P. cerasus* (SRX14816146,

SRX14816142, SRX14816138), *P. fruticosa* (SRX14816141), *P. avium* (SRX14816143), *P. canescens* (SRX14816137), *P. serrulata* (SRX14816136), *P. mahaleb* (SRX14816140), *P. pensylvanica* (SRX14816144), *P. maackii* (SRX14816139), *P. subhirtella* (SRX14816145). Reads were adapter- and quality trimmed using the software Trim Galore (version 0.6.3, parameters --quality 30 –length 50). Trimmed reads were mapped against the *P. cerasus* subgenomes Pce_S_a and Pce_S_f using STAR (version 2.7.8a, parameter --twopassMode Basic). The subsequent analysis was performed in accordance to Keilwagen et al. (2022). The Pce_S genome was divided into 250k windows. The percentage of covered bases using RNAseq data of *P. cerasus* (SRX14816146, SRX14816142, SRX14816138) was estimated at a depth of 1 for each window. The same was done with all other RNAseq data sets. The percentage of covered bases from *P. avium* (SRX14816143) was subtracted from the percentage of covered bases from *P. cerasus* (SRX14816146, SRX14816142, SRX14816138). The same was done using the reads of *P. fruticosa* (SRX14816141). For subgenome Pce_S_a it is expected that the intraspecific difference for transcripts of data set *P. avium* (SRX14816143) is lower (close to 0) than the interspecific difference for transcripts of data set *P. fruticosa* (SRX14816141) and vice versa. Opposite cases indicate potential homoeologous exchanges between the two subgenomes Pce_S_a and Pce_S_f and were plotted into a circos plot (Figure S11).

The nucleotide short reads from Pce_S were mapped against the genomes of the two ancestral species Pa_T and $Pf_{eH.}$ Subsequently, the mapped reads were filtered using samtools for mapped reads in proper pair (-f 3) and primary alignments and not supplementary alignment (-F 2304). Those reads were divided into four groups according to the following criteria: 1. unique match to Pa_T: 2. unique match to Pf_{eH}; 3. match to Pa_T and Pf_{eH}; 4. no match to Pa_T and Pf_{eH} (unique to Pce_S). The first two separated read sets were then re-mapped against the subgenomes Pce_S_a and Pce_S_f. The percentage of covered bases was calculated for a 100 k window. For the subgenomes of Pce_s, the percentage of intraspecific covered bases (Pce_S_a to Pa_T, Pce_S_f to Pf_{eH}) should be higher compared to the percentage of interspecific covered bases (Pce_S_a to Pf_{eH}, Pce_S_f to Pa_T). The opposite case indicates possible translocations and were plotted into a circus plot (Figure S11). Additionally, regions

of the Schattenmorelle genome assembly were determined that are uniquely covered by *Pa*$_T$ and *Pf*$_{eH}$ filtered read sets.

LTR insertion estimation

The difference (identity) of left and right LTR was calculated using the script EDTA_raw.pl from the software EDTA version 1.9 (https://github.com/oushujun/EDTA, Ou et al. 2019). As input files we used the genome sequences of *P. cerasus* (*Pce*$_S$_a and *Pce*$_S$_f), *Pa*$_T$ (NCBI BioProject acc. no. PRJNA596862), *Pf*$_{eH}$ (NCBI BioProject acc. no. PRJNA727075) and a curated library of representative transposable elements from *Viridiplantae* (https://www.girinst.org/repbase/). Because trees are not annual plants, the identity obtained from the resulting .pass.list file was used for the estimation of generation time after LTR insertion using the formula T=K/2μ (K is the divergence of the LTR = 1 – identity) assuming a *Prunus* specific mutation frequency of μ=7.7 * 10^{-9} (Xie et al. 2016) per generation.

Protein clustering, multiple sequence alignment and divergence of time estimation

The protein datasets from *Pce*$_S$*_a* and *Pce*$_S$*_f, Pa*$_T$*, Pf*$_{eH}$*, Pp* (*Prunus persica* Whole Genome Assembly v2.0, v2.0.a1)*, Pm* (*Prunus mume* Tortuosa Genome v1.0)*, Py* (*Prunus yedoensis* var. *nudiflora* Genome v1.0)*, Md* (*Malus* x *domestica* HFTH1 Whole Genome v1.0) and *At* (TAIR10.1, RefSeq GCF_000001735.4) from the annotation step were uploaded to Galaxy_Europe server as .fasta. The Proteinortho (Galaxy Version 6.0.32+galaxy0) was used to find orthologous proteins within the datasets. MAFFT (Galaxy Version 7.505+galaxy0) was used to align the obtained single copy orthogroups. The final alignments were merged with the Merge.files function (Galaxy Version 1.39.5.0). Finally, the alignments were concatenated into a super protein and the final sequences were aligned with MAFFT. A phyogenetic tree was reconstructed with RAxML (maximum likelihood based inference of large phylogenetic trees, Galaxy Version 8.2.4+galaxy3) and the obtained .nhx file was reformatted as .nwk file for further processing using CLC Mainworkbench (21.0.1, QIAGEN Aarhus A/S). Evolutionary analyses were conducted in MEGA X (Kumar et al. 2018). Estimation of pairwise divergence time was performed according to Shirasawa et al. (2019) with a

627 divergence time from the reference species peach and apple (34-67 Mya,
628 www.timetree.org). Specific parameters for the calculation are listed in supplements.
629
630

Data Availability

Data supporting the findings of this study are deposited into the Open Agrar repository (https://doi.org/10.5073/20230324-105730-0, Wöhner et al. (2023)) and on personal request to the corresponding author. An assembly hub for genome and annotation visualization is permanently hosted at http://bioinf.uni-greifswald.de/private-hubs/pcer/hub.txt .

Author contributions

TWW, OFE, HF wrote the manuscript. AHJW, KN, RPW and EJB performed DNA isolation, sequencing, genome assembly and scaffolding. KJH, JK, LG, HT and TB performed masking and annotation of the dataset. OA and LB performed the LTR-dating and TW did the corresponding analysis and results compilation. JK performed BUSCO analysis, read mapping and coverage analysis. TW the did the k-mer analysis, interproscan, synteny analysis, phylogenetic analysis and circus plots. TW and JL performed the Lai-Index analysis. HF, MS and AP conceived the study and made substantial contributions to its design, acquisition, analysis and interpretation of data. All authors contributed equally to the finalization of the manuscript.

Acknowledgments

We would acknowledge the Galaxy Europe, Galaxy USA server administration for support and provision of resources. Special thanks to Eric Lyon for support with Synmap2.

References

Arang Rhie (2020). Meryl. In GitHub repository. GitHub. https://github.com/marbl/meryl.

Arulsekar, S., and Parfitt, D.E. (1986). Isozyme Analysis Procedures for Stone Fruits, Almond, Grape, Walnut, Pistachio, and Fig. HortSci **21** (4): 928–933.

Beaver, J.A., and Iezzoni, A.F. (1993). Allozyme Inheritance in Tetraploid Sour Cherry (Prunus cerasus L.). jashs **118** (6): 873–877.

Bertioli, D.J., Jenkins, J., Clevenger, J., Dudchenko, O., Gao, D., Seijo, G., Leal-Bertioli, S.C.M., Ren, L., Farmer, A.D., and Pandey, M.K., et al. (2019). The genome sequence of segmental allotetraploid peanut Arachis hypogaea. Nat Genet **51** (5): 877–884.

Bird, K.A., Jacobs, M., Sebolt, A., Rhoades, K., Alger, E.I., Colle, M., Alekman, M.L., Bies, P.K., Cario, A.J., and Chigurupati, R.S., et al. (2022). Parental origins of the cultivated tetraploid sour cherry (Prunus cerasus L.). Plants People Planet **4** (5): 444–450.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30** (15): 2114–2120.

Brůna, T., Hoff, K.J., Lomsadze, A., Stanke, M., and Borodovsky, M. (2021). BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom Bioinform **3** (1): lqaa108.

Brůna, T., Lomsadze, A., and Borodovsky, M. (2020). GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. NAR Genom Bioinform **2** (2): lqaa026.

Cai, L., Stegmeir, T., Sebolt, A., Zheng, C., Bink, Marco C. A. M., and Iezzoni, A. (2018). Identification of bloom date QTLs and haplotype analysis in tetraploid sour cherry (Prunus cerasus). Tree Genetics & Genomes **14** (2): 1–11.

Chester, M., Gallagher, J.P., Symonds, V.V., Da Cruz Silva, A.V., Mavrodiev, E.V., Leitch, A.R., Soltis, P.S., and Soltis, D.E. (2012). Extensive chromosomal variation in a recently formed natural allopolyploid species, Tragopogon miscellus (Asteraceae). Proceedings of the National Academy of Sciences of the United States of America **109** (4): 1176–1181.

687 **Clarke, J.B., and Tobutt, K.R.** (2009). A standard set of accessions, microsatellites
688 and genotypes for harmonising the fingerprinting of cherry collections for the
689 ECPGR. Acta Horticulturae (814 (VOL 2)): 615–618.
690 **Clayton, J.W., and Tretiak, D.N.** (1972). Amine-Citrate Buffers for p H Control in
691 Starch Gel Electrophoresis. J. Fish. Res. Bd. Can. **29** (8): 1169–1172.
692 **Coart, E., Vekemans, X., Smulders, M.J.M., Wagner, I., van Huylenbroeck, J., van**
693 **Bockstaele, E., and Roldán-Ruiz, I.** (2003). Genetic variation in the endangered
694 wild apple (*Malus sylvestris* (L.) Mill.) in Belgium as revealed by amplified fragment
695 length polymorphism and microsatellite markers. Molecular Ecology **12** (4): 845–
696 857.
697 **Cock, P.J.A., Grüning, B.A., Paszkiewicz, K., and Pritchard, L.** (2013). Galaxy tools
698 and workflows for sequence analysis with applications in molecular plant
699 pathology. PeerJ **1**: e167.
700 **Dirlewanger, E., Claverie, J., Wünsch, A., and Iezzoni, A.F.** (2007). Cherry. In Fruits
701 and Nuts (Springer, Berlin, Heidelberg), pp. 103–118.
702 **Edger, P.P., Poorten, T.J., VanBuren, R., Hardigan, M.A., Colle, M., McKain, M.R.,**
703 **Smith, R.D., Teresi, S.J., Nelson, A.D.L., and Wai, C.M.**, et al. (2019). Origin and
704 evolution of the octoploid strawberry genome. Nat Genet **51** (3): 541–547.
705 **Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., and**
706 **Smit, A.F.** (2020). RepeatModeler2 for automated genomic discovery of
707 transposable element families. Proc. Natl. Acad. Sci. U.S.A. **117** (17): 9451–9457.
708 **Gabriel, L., Hoff, K.J., Brůna, T., Borodovsky, M., and Stanke, M.** (2021). TSEBRA:
709 transcript selector for BRAKER. BMC Bioinformatics **22** (1): 566.
710 **Hancock, J.F.** (2008). Temperate Fruit Crop Breeding: Germplasm to Genomics
711 (Dordrecht: Springer Science+Business Media B.V).
712 **Hao, Y., Fleming, J., Petterson, J., Lyons, E., Edger, P.P., Pires, J.C., Thorne, J.L.,**
713 **and Conant, G.C.** (2022). Convergent evolution of polyploid genomes from across
714 the eukaryotic tree of life. G3 (Bethesda, Md.) **12** (6).
715 **Haug-Baltzell, A., Stephens, S.A., Davey, S., Scheidegger, C.E., and Lyons, E.**
716 (2017). SynMap2 and SynMap3D: web-based whole-genome synteny browsers.
717 Bioinformatics **33** (14): 2197–2198.

718 **Hawkins, J.S., Proulx, S.R., Rapp, R.A., and Wendel, J.F.** (2009). Rapid DNA loss
719 as a counterbalance to genome expansion through retrotransposon proliferation
720 in plants. Proceedings of the National Academy of Sciences of the United States
721 of America **106** (42): 17811–17816.
722 **Höfer, M., Flachowsky, H., Schröpfer, S., and Peil, A.** (2021). Evaluation of Scab
723 and Mildew Resistance in the Gene Bank Collection of Apples in Dresden-Pillnitz.
724 Plants **10** (6): 1227.
725 **Hoff, K.J.** (2019). MakeHub: Fully automated generation of UCSC Genome Browser
726 Assembly Hubs (Cold Spring Harbor Laboratory).
727 **Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M.** (2016).
728 BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-
729 ET and AUGUSTUS. Bioinformatics **32** (5): 767–769.
730 **Hoff, K.J., Lomsadze, A., Borodovsky, M., and Stanke, M.** (2019). Whole-Genome
731 Annotation with BRAKER. In Gene Prediction (Humana, New York, NY), pp. 65–95.
732 **Hrotkó, K., Feng, Y., and Halász, J.** (2020). Spontaneous hybrids of Prunus fruticosa
733 Pall. in Hungary. Genet Resour Crop Evol **67** (2): 489–502.
734 **Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork,**
735 **P., Das, U., Daugherty, L., and Duquenne, L.**, et al. (2009). InterPro: the
736 integrative protein signature database. Nucleic Acids Res **37** (Database issue):
737 D211-5.
738 **Jakob, S.S., Meister, A., and Blattner, F.R.** (2004). The considerable genome size
739 variation of Hordeum species (poaceae) is linked to phylogeny, life form, ecology,
740 and speciation rates. Mol Biol Evol **21** (5): 860–869.
741 **Janick, J.** (2010). Horticultural Reviews, Volume 19 (New York, NY: John Wiley &
742 Sons).
743 **Jia, J., Xie, Y., Cheng, J., Kong, C., Wang, M., Gao, L., Zhao, F., Guo, J., Wang,**
744 **K., and Li, G.**, et al. (2021). Homology-mediated inter-chromosomal interactions
745 in hexaploid wheat lead to specific subgenome territories following
746 polyploidization and introgression. Genome Biol **22** (1): 26.
747 **Jo, Y., Chu, H., Cho, J.K., Choi, H., Lian, S., and Cho, W.K.** (2015). De novo
748 transcriptome assembly of a sour cherry cultivar, Schattenmorelle. Genomics Data
749 **6**: 271–272.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., and Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics **30** (9): 1236–1240.

Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. Computer applications in the biosciences CABIOS **8** (3): 275–282.

José Quero-García, Amy Iezzoni, Gregorio López-Ortega, Cameron Peace, Mathieu Fouché, Elisabeth Dirlewanger, and Mirko Schuster (2019). Advances and challenges in cherry breeding. In Achieving sustainable cultivation of temperate zone tree fruits and berries (Burleigh Dodds Science Publishing), pp. 55–88.

Keilwagen, J., Hartung, F., and Grau, J. (2019). GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. In Gene Prediction (Humana, New York, NY), pp. 161–177.

Keilwagen, J., Wenk, M., Erickson, J.L., Schattat, M.H., Grau, J., and Hartung, F. (2016). Using intron position conservation for homology-based gene prediction. Nucleic Acids Res **44** (9): e89.

Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol **37** (8): 907–915.

Kim, S., Park, M., Yeom, S.-I., Kim, Y.-M., Lee, J.M., Lee, H.-A., Seo, E., Choi, J., Cheong, K., and Kim, K.-T., et al. (2014). Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. Nat Genet **46** (3): 270–278.

Knight, C.A., and Beaulieu, J.M. (2008). Genome size scaling through phenotype space. Ann Bot **101** (6): 759–766.

Kobel, F. (1927). Zytologische Untersuchungen an Prunoideen und Pomoideen. Art. Institut Orell Füssli

Krebs, S.L., and Hancock, J.F. (1989). Tetrasomic inheritance of isoenzyme markers in the highbush blueberry, Vaccinium corymbosum L. Heredity **63** (1): 11–18.

Kriventseva, E.V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F.A., and Zdobnov, E.M. (2019). OrthoDB v10: sampling the diversity of animal, plant,

782 fungal, protist, bacterial and viral genomes for evolutionary and functional
783 annotations of orthologs. Nucleic Acids Res **47** (D1): D807-D811.
784 **Kumar S., Stecher G., Li M., Knyaz C., and Tamura K.** (2018). MEGA X: Molecular
785 Evolutionary Genetics Analysis across computing platforms. Molecular Biology
786 and Evolution **35**:1547-1549.
787 **Langmead, B., and Salzberg, S.L.** (2012). Fast gapped-read alignment with Bowtie
788 2. Nat Methods **9** (4): 357–359.
789 **Lansari, A., and Iezzoni, A.** (1990). A Preliminary Analysis of Self-incompatibility in
790 Sour Cherry. HortSci **25** (12): 1636–1638.
791 **Le Dantec, L., Girollet, N., Gouzy, J., Sallet, E., Fouche, M., Quero-Garcia, J., &**
792 **Dirlewanger, E.** (2019). An Improved Assembly of the Diploid 'Regina'Sweet
793 Cherry Genome. In Plant & Animal Genome Conference XXVII (p. PO1193).
794 **Leitch, I.J., Hanson, L., Lim, K.Y., Kovarik, A., Chase, M.W., Clarkson, J.J., and**
795 **Leitch, A.R.** (2008). The ups and downs of genome size evolution in polyploid
796 species of Nicotiana (Solanaceae). Ann Bot **101** (6): 805–814.
797 **Li, H.** (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy
798 long sequences. Bioinformatics **32** (14): 2103–2110.
799 **Li, H.** (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics
800 **34** (18): 3094–3100.
801 **Li, L., Chen, M., Zhang, X., and Jia, X.** (2022). Spatial Distribution Pattern of Root
802 Sprouts under the Canopy of *Malus sieversii* in a Typical River Valley on the
803 Northern Slopes of the Tianshan Mountain. Forests **13** (12): 2044.
804 **Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G.,**
805 **Abecasis, G., and Durbin, R.** (2009). The Sequence Alignment/Map format and
806 SAMtools. Bioinformatics **25** (16): 2078–2079.
807 **Lomsadze, A., Burns, P.D., and Borodovsky, M.** (2014). Integration of mapped
808 RNA-Seq reads into automatic training of eukaryotic gene finding algorithm.
809 Nucleic Acids Res **42** (15): e119.
810 **Macková, L., Vít, P., and Urfus, T.** (2018). Crop-to-wild hybridization in cherries-
811 Empirical evidence from Prunus fruticosa. Evolutionary Applications **11** (9): 1748–
812 1759.

Mason, A.S., and Wendel, J.F. (2020). Homoeologous Exchanges, Segmental Allopolyploidy, and Polyploid Genome Evolution. Front. Genet. **11**: 1014.

Multivariate analysis of a sour cherry germplasm c germplasm collection. J Amer Soc Hort Sci .

OLDÉN, E.J., and NYBOM, N. (1968). ON THE ORIGIN OF PRUNUS CERASUS L. Hereditas **59** (2-3): 327–345.

Ordidge, M., Litthauer, S., Venison, E., Blouin-Delmas, M., Fernandez-Fernandez, F., Höfer, M., Kägi, C., Kellerhals, M., Marchese, A., and Mariette, S. (2021). Towards a Joint International Database: Alignment of SSR Marker Data for European Collections of Cherry Germplasm. Plants **10** (6): 1243.

Ou, S., Chen, J., and Jiang, N. (2018). Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Research **46** (21): e126.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., Lugo, C.S.B., Elliott, T.A., Ware, D., and Peterson, T., et al. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol **20** (1): 275.

Pinosio, S., Marroni, F., Zuccolo, A., Vitulo, N., Mariette, S., Sonnante, G., Aravanopoulos, F.A., Ganopoulos, I., Palasciano, M., and Vidotto, M., et al. (2020). A draft genome of sweet cherry (Prunus avium L.) reveals genome-wide and local effects of domestication. The Plant Journal **103** (4): 1420–1432.

Pierce, B. A. (2012). Genetics: a conceptual approach. Macmillan.

Potter, D., Eriksson, T., Evans, R.C., Oh, S., Smedmark, J.E.E., Morgan, D.R., Kerr, M., Robertson, K.R., Arsenault, M., and Dickinson, T.A., et al. (2007). Phylogeny and classification of Rosaceae. Plant Syst. Evol. **266** (1-2): 5–43.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. (2005). InterProScan: protein domains identifier. Nucleic Acids Research **33** (Web Server issue): W116-20.

Ranallo-Benavidez, T.R., Jaron, K.S., and Schatz, M.C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun **11** (1): 1432.

Raptopoulos, T. (1941). Chromosomes and fertility of cherries and their hybrids. Journal of Genetics, 42(1), 91-114.

Renny-Byfield, S., Chester, M., Kovařík, A., Le Comber, S.C., Grandbastien, M.-A., Deloger, M., Nichols, R.A., Macas, J., Novák, P., and Chase, M.W., et al. (2011). Next generation sequencing reveals genome downsizing in allotetraploid Nicotiana tabacum, predominantly through the elimination of paternally derived repetitive DNAs. Mol Biol Evol **28** (10): 2843–2854.

Rutland, C.A., Hall, N.D., and McElroy, J.S. (2021). The Impact of Polyploidization on the Evolution of Weed Species: Historical Understanding and Current Limitations. Front. Agron. **3**: 5.

Santi, F., and Lemoine, M. (1990). Genetic markers for Prunus avium L. 2. Clonal identifications and discrimination from P cerasus and P cerasus × P avium. Ann. For. Sci. **47** (3): 219–227.

Schiessl, S.-V., Katche, E., Ihien, E., Chawla, H.S., and Mason, A.S. (2019). The role of genomic structural variation in the genetic improvement of polyploid crops. The Crop Journal **7** (2): 127–140.

Schuster, M., Apostol, J., Iezzoni, A., Jensen, M., and Milatović, D. (2017). Sour cherry varieties and improvement. In Cherries: botany, production and uses (Wallingford: CABI), pp. 95–116.

Schuster, M., and Schreibner, H. (2000). GENOME INVESTIGATION IN SOUR CHERRY, P. CERASUS L. Acta Hortic. (538): 375–379.

Schuster, M., and Wolfram, B. (2008). NEW SOUR CHERRY CULTIVARS FROM DRESDEN-PILLNITZ. Acta Hortic. (795): 83–86.

Shirasawa, K., Isuzugawa, K., Ikenaga, M., Saito, Y., Yamamoto, T., Hirakawa, H., and Isobe, S. (2017). The genome sequence of sweet cherry (Prunus avium) for use in genomics-assisted breeding. DNA Res **24** (5): 499–508.

Shirasawa, K., Esumi, T., Hirakawa, H., Tanaka, H., Itai, A., Ghelfi, A., Nagasaki, H., and Isobe, S. (2019). Phased genome sequence of an interspecific hybrid flowering cherry, 'Somei-Yoshino' (Cerasus × yedoensis). DNA Res **26** (5): 379–389.

Smit, A.F., Hubley, R., and Green, P. (2015). RepeatModeler Open-1.0. 2008–2015. Seattle, USA: Institute for Systems Biology. Available from: httpwww. repeatmasker. org, Last Accessed May 1., 2018.

Smit, A.F., Hubley, R., and Green, P. (2015). RepeatMasker Open-4.0. 2013–2015.

Soltis, D.E., Haufler, C.H., Darrow, D.C., and Gastony, G.J. (1983). Starch Gel Electrophoresis of Ferns: A Compilation of Grinding Buffers, Gel and Electrode Buffers, and Staining Schedules. American Fern Journal **73** (1): 9.

Soltis, D.E., and Rieseberg, L.H. (1986). AUTOPOLYPLOIDY IN TOLMIEA MENZIESII (SAXIFRAGACEAE): GENETIC INSIGHTS FROM ENZYME ELECTROPHORESIS. American Journal of Botany **73** (2): 310–318.

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics **24** (5): 637–644.

STEBBINS, G.L. (1947). Types of polyploids; their classification and significance. Advances in genetics **1**: 403–429 (Elsevier).

Tamura, K., Battistuzzi, F.U., Billing-Ross, P., Murillo, O., Filipski, A., and Kumar, S. (2012). Estimating divergence times in large molecular phylogenies. Proceedings of the National Academy of Sciences of the United States of America **109** (47): 19333–19338.

Tamura, K., Tao, Q., and Kumar, S. (2018). Theoretical Foundation of the RelTime Method for Estimating Divergence Times from Variable Evolutionary Rates. Mol Biol Evol **35** (7): 1770–1782.

Tang, S., Lomsadze, A., and Borodovsky, M. (2015). Identification of protein coding regions in RNA transcripts. Nucleic Acids Res **43** (12): e78.

Tao, Q., Tamura, K., Mello, B., and Kumar, S. (2020). Reliable Confidence Intervals for RelTime Estimates of Evolutionary Divergence Times. Mol Biol Evol **37** (1): 280–290.

The Galaxy platform for accessible, reproducible and collaborative biomedical analyses. (2022). update. Nucleic Acids Research, 2022, 50. Jg., Nr. W1, S. W345-W351.

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E.S., Fischer, A., Bock, R., and Greiner, S. (2017). GeSeq - versatile and accurate annotation of organelle genomes. Nucleic Acids Res **45** (W1): W6-W11.

van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. Nat Rev Genet **10** (10): 725–732.

Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. **27** (5): 737–746.

Verde, I., Jenkins, J., Dondini, L., Micali, S., Pagliarani, G., Vendramin, E., Paris, R., Aramini, V., Gazza, L., and Rossini, L., et al. (2017). The Peach v2.0 release: high-resolution linkage mapping and deep resequencing improve chromosome-scale assembly and contiguity. BMC Genomics **18** (1): 225.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., and Young, S.K., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLOS ONE **9** (11): e112963.

Wang, J., Liu, W., Zhu, D., Hong, P., Zhang, S., Xiao, S., Tan, Y., Chen, X., Xu, L., and Zong, X., et al. (2020). Chromosome-scale genome assembly of sweet cherry (Prunus avium L.) cv. Tieton obtained using long-read and Hi-C sequencing. Hortic Res **7** (1): 122.

Wang, X., Liu, S., Zuo, H., Zheng, W., Zhang, S., Huang, Y., Pingcuo, G., Ying, H., Zhao, F., and Li, Y., et al. (2021). Genomic basis of high-altitude adaptation in Tibetan *Prunus* fruit trees. Current biology CB **31** (17): 3848-3860.e8.

Wang, Z., Miao, H., Liu, J., Xu, B., Yao, X., Xu, C., Zhao, S., Fang, X., Jia, C., and Wang, J., et al. (2019). Musa balbisiana genome reveals subgenome evolution and functional divergence. Nat. Plants **5** (8): 810–821.

Wang, X., Morton, J.A., Pellicer, J., Leitch, I.J., and Leitch, A.R. (2021). Genome downsizing after polyploidy: mechanisms, rates and selection pressures. The Plant Journal **107** (4): 1003–1015.

Wöhner, T.W., Emeriewen, O.F., Wittenberg, A.H.J., Schneiders, H., Vrijenhoek, I., Halász, J., Hrotkó, K., Hoff, K.J., Gabriel, L., and Lempe, J., et al. (2021a). The draft chromosome-level genome assembly of tetraploid ground cherry (Prunus fruticosa Pall.) from long reads. Genomics **113** (6): 4173–4183.

Wöhner, T.W., Emeriewen, O.F., Wittenberg, A.H.J., Schneiders, H., Vrijenhoek, I., Halász, J., Hrotkó, K., Hoff, K.J., Gabriel, L., and Lempe, J., et al. (2021b). Supporting Materials for - The Draft Chromosome-level Genome Assembly of Tetraploid Ground Cherry (Prunus fruticosa Pall.) from Long Reads.

https://www.openagrar.de/receive/openagrar_mods_00070329 (2021) (accessed 1 June 2021)

Wöhner, T., Emeriewen, Ofere, F., Wittenberg, Alexander, H.J., Nijbroek, K., Wang, Rui, Peng, Blom, E.-J., Keilwagen, J., Berner, T., Hoff, Katharina, J., and Gabriel, L., et al. (2023). Data set: The structure of the tetraploid sour cherry 'Schattenmorelle' (Prunus cerasus L.) genome reveals insights into its segmental allopolyploid nature. https://doi.org/10.5073/20230324-105730-0

Xie, Z., Wang, L., Wang, L., Wang, Z., Lu, Z., Tian, D., Yang, S., and Hurst, L.D. (2016). Mutation rate analysis via parent-progeny sequencing of the perennial peach. I. A low rate in woody perennials and a higher mutagenicity in hybrids. Proceedings. Biological sciences **283** (1841).

Yan, M., Zhang, X., Zhao, X., and Yuan, Z. (2019). The complete mitochondrial genome sequence of sweet cherry (Prunus avium cv. 'summit'). Mitochondrial DNA Part B **4** (1): 1996–1997.

Zdobnov, E. M., and Apweiler, R. (2001). InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**(9), 847–848. https://doi.org/10.1093/bioinformatics/17.9.847

Zdobnov, E.M., and Apweiler R. (2001). R. InterProScan: protein domains identifier. Bioinformatics (Oxford, England) **17**: 847–848.

Zenil-Ferguson, R., Ponciano, J.M., and Burleigh, J.G. (2016). Evaluating the role of genome downsizing and size thresholds from genome size distributions in angiosperms. American Journal of Botany **103** (7): 1175–1186.

Zhang, Z., Fu, T., Liu, Z., Wang, X., Xun, H., Li, G., Ding, B., Dong, Y., Lin, X., and Sanguinet, K.A., et al. (2019). Extensive changes in gene expression and alternative splicing due to homoeologous exchange in rice segmental allopolyploids. Theor Appl Genet **132** (8): 2295–2308.

Figure legends

Figure 1 Morphology of *P. cerasus* L. 'Schattenmorelle'. (a) mature tree habitus, (b) leaves, (c) inflorescence, (d) fruits.

Figure 2 The genome of *P.cerasus* 'Schattenmorelle'. Circos plot of 16 pseudomolecules of the subgenomes of *Pce*$_S$_a and *Pce*$_S$_f. (a) chromosome length (Mb); (b) gene density in blocks of 250k: (c) distribution of repetitive sequences in blocks of 250k (d) Gypsy elements in blocks of 250k; (e) Copia elements in block of 250k; (f) GC content in blocks of 1 Mb. (g) The inner ring shows markers from the 6+9k SNP array located on both subgenomes.

Figure 3 GenomeScope (Galaxy Version 2.0) estimation of the *P. cerasus* genome size by k-mer counts obtained from the software Meryl (Galaxy Version 1.3+galaxy2). Both programs are integrated on the GalaxyServerEurope. The k-mer-peaks indicate that k-mers with a length of 19 bp occur in heterozygote (100x depth, 200x depth, 300x depth) and homozygote (400x depth) constitution within the genome. Coverage depth of individual k-mers is assigned as coverage.

Fig.4 Detected regions of homoeologous exchanges in the genome of *P. cerasus* 'Schattenmorelle'. Circos plot of 16 pseudomolecules of the subgenomes of *Pce*$_S$_a and *Pce*$_S$_f. (a) chromosome length (Mb); (b) 16 in *Pce*$_S$_a and 12 in *Pce*$_S$_f detected regions that match all three following analysis methods: (c) 1024 regions (100k window) were intraspecific %-covered bases from mapped reads (*Pce*$_S$_a to *Pa*$_T$, *Pce*$_S$_f to *Pf*$_{eH}$) was < than interspecific %-covered bases from mapped reads (*Pce*$_S$_a to *Pf*$_{eH}$, *Pce*$_S$_f to *Pa*$_T$); (d) 148 regions were intraspecific difference of %-covered bases from obtained RNAseq reads (*Pa* and *Pce*$_S$_a, *Pf* and *Pce*$_S$_f) > than interspecific difference of %-covered bases from obtained RNAseq reads (*Pf* and *Pce*$_S$_a, *Pa* and *Pce*$_S$_f); (e) 367 regions were the proportion of transcripts with intraspecific amino acid identity (*Pa*$_T$ and *Pce*$_S$_a, *Pf*$_{eH}$ and *Pce*$_S$_f) < than the proportion of transcripts with interspecific amino acid identity (*Pf*$_{eH}$ and *Pce*$_S$_a , *Pa*$_T$ and *Pce*$_S$_f).

Fig. 5 Investigation on the evolution of the genome of *P. cerasus* 'Schattenmorelle'. **A** Determination of insertion time from shared long terminal repeats (LTRs) in *P. cerasus* subgenome *avium* (*Pce*$_S$_a) and *P. cerasus* subgenome *fruticosa* (*Pce*$_S$_f) compared to *P. avium* 'Tieton' (*Pa*$_T$) and *P. fruticosa* ecotype Hármashatárhegy (*Pf*$_{eH}$)**.** **B** Estimation of divergence of time (Mya) of *P. cerasus* subgenomes *Pce*$_S$_a and *Pce*$_S$_f compared to the donor species *P. avium* (Pa) and *P. fruticosa* (Pf)*. Prunus yedonensis*

(Pyn); *Prunus avium* (Pa); *Prunus persica* (Pp); *Prunus mume* (Pm); *Malus domestica* (Md); Paleocene (PAL); Eocene (EOC); Oligocene (OLI); Miocene (MIO); Pliocene (PII); Pleistocene (PLEI).

Tables

Table 1 Characterization of repetitive sequences of *P. fruticosa* ecotype Hármashatárhegy (Pf_{eH}) compared to *P. avium* 'Tieton' (Pa_T), *P. persica* 'Lovell', and the two subgenomes of *P. cerasus* 'Schattenmorelle' Pce_S_a and Pce_S_f

Table 2 Comparison between the number of transcripts and %-IAA obtained from *P. fruticosa* ecotype Hármashatárhegy (Pf_{eH}) and *P. avium* cv 'Tieton' (Pa_T) representing the two ancestral species of *P. cerasus*

Supplemental information (SI)

Document S1 - Supplemental material and methods

Supplemental Figures

Figure S1 Hi-C heatmap post-scaffolding for the subgenomes Pce_S_a (left) and Pce_S_f (right) of *P. cerasus* cv 'Schattenmorelle'. The heatmap indicates the density of paired Hi-C reads which interact to each other in close proximity. High intense colour indicates high interaction.

Figure S2 The chloroplast (a) and mitochondrial (b) sequences of *P. cerasus* L. cv 'Schattenmorelle'.

Figure S3 Analysis of completeness of the *P. cerasus* cv. Schattenmorelle subgenomes *P. cerasus* cv 'Schattenmorelle' subgenome *avium* (Pce_S_a) and *P. cerasus* cv 'Schattenmorelle' subgenome *fruticosa* (Pce_S_f) and combined datasets compared to *P. avium* cv. 'Tieton' (Pa_T) and *P. persica* cv. Lovell (Pp_L) by mapping of a set of universal single-copy orthologs using BUSCO. The bar charts indicate complete **s**ingle copy (orange), complete **d**uplicated (gray), **f**ragmented (yellow) and **m**issing (blue) genes. For evaluation the embryophyta_odb10 BUSCO dataset (n=1614) was used. *P. cerasus* cv. Schattenmorelle show a 99 % completeness (S:

16.7 %, D: 82.3 %, F: 0.4 %, M: 0.6 %, n: 1614) which reaches the completeness of *P. avium* cv. 'Tieton' (C: 98.3 %, S: 95.6 %, D: 2.7 %, F: 0.5 %, M:1.5 %, n:1614) and *P. persica* 'Lovell' (C: 99.3 %, S: 97.5 %, D: 1.8%, F: 0.1 %, M: 0.6 %, n:1614).

Figure S4 Assessing the quality of repetitive sequences between the chromosome sequences of *P. avium* 'Tieton' (A), *P. fruticosa* ecotype Hármashatárhegy (B), *P. perisca* 'Lovell', and (C) *P. cerasus* subgenome *avium* and *P. cerasus* subgenome *fruticosa* using the LAI index. The genomes *P. cerasus* [this study] and *P. avium* were sequenced with ONT 9.4.1 and Illumina (Wang wet al. 2020), *P. fruticosa* with ONT 10.3 (Wöhner et al. 2021) and *P. persica* with Illumina and Sanger sequencing of fosmid and BAC clones (Verde et al. 2017).

Figure S5 Collinearity plots between the five published genetic maps of sour cherry (M172x25-F1, US-F1, 25x25-F1, Montx25-F1, RE-F1) and the *P. cerasus* cv 'Schattenmorelle' subgenome *avium* (*Pce*$_S$_a) and *fruticosa* (*Pce*$_S$_f).X-axis represents the genetic position of a marker in the genetic linkage map given in centi Morgan (cM). Y-axis represents the physical position of a marker sequence within the genome sequence of the respective subgenome given in Mega base pairs (Mbp).

Figure S6 Percentage of *P. cerasus* (*Pce*) proteins by IAA compared with 15 reference species. *P. fruticosa* ecotype Hármashatárhegy (*P*$_{feH}$), *P. avium* 'Tieton' (*Pa*$_T$), *P. yedonensis* (*Pyed*), *P. domestica* (*Pd*), *P. armeniaca* (*Par*), *P. persica* (*Pp*), *Pyrus communis* (*Pyrco*), *Populus trichocarpa* (*Poptri*), *Vitis vinifera* (*Vv*), *Arabidopsis thaliana* (*At*), *Malus domestica* (*Md*).

Figure S7 Synmap2 plots of self-self comparisons between (A) *Prunus persica* 'Lovell' (*Pp*$_L$) and *P. avium* 'Sato Nishiki' (*Pa*$_S$), (B) *P. fruticosa* 'Hármashatárhegy' (*Pf*$_{eH}$), (C) *P. cerasus_avium* 'Schattenmorelle' (*Pce*$_S$_a), (D) *P. cerasus_fruticosa* 'Schattenmorelle' (*Pce*$_S$_f) for the identification of triplicated regions (TR) 1-7.

Figure S8 Positional co-linearity comparison between the two subgenomes *P. cerasus_avium* 'Schattenmorelle' (*Pce*$_S$_a), *P. cerasus_fruticosa* 'Schattenmorelle'

(*Pce*$_S$_f) and *P. avium* 'Tieton' (*Pa*$_T$), *P. fruticosa* 'Hármashatárhegy' (*Pf*$_{eH}$), using the molecular markers from the 9+6k SNP array. The plots were generated using the R-software package chromoMap v0.4.1.

Figure S9 Synteny between (a) *P. cerasus* subgenome *_avium* (*Pce*$_S$_a) and *P. cerasus* subgenome *_fruticosa* (*Pce*$_S$_f), and (b) the subgenomes and the genotypes *Prunus avium* 'Tieton' (*Pa*$_T$) and *P. fruticosa* (*Pf*$_{eH}$) of the ancestral species *P. avium* and *P. fruticosa.* Blue arrows indicate positions where inversions occurred.

Figure S10 Detected regions of homoeologous exchanges in the genome of *P. cerasus* 'Schattenmorelle'. Circos plot of 16 single pseudomolecules 1 to 8 (see continuating plots) of the subgenomes of *Pce*$_S$_a and *Pce*$_S$_f. (a) chromosome length (Mb); (b) region in *Pce*$_S$_a and in *Pce*$_S$_f detected that match all three following analysis methods: (c) regions (100k window) were intraspecific %-covered bases from mapped reads (*Pce*$_S$_a to *Pa*$_T$, *Pce*$_S$_f to *Pf*$_{eH}$) was < than interspecific %-covered bases from mapped reads (*Pce*$_S$_a to *Pf*$_{eH}$, *Pce*$_S$_f to *Pa*$_T$); (d) regions were intraspecific difference of %-covered bases from obtained RNAseq reads (*Pa* and *Pce*$_S$_a, *Pf* and *Pce*$_S$_f) > than interspecific difference of %-covered bases from obtained RNAseq reads (*Pf* and *Pce*$_S$_a, *Pa* and *Pce*$_S$_f); (e) regions were the proportion of transcripts with intraspecific amino acid identity (*Pa*$_T$ and *Pce*$_S$_a, *Pf*$_{eH}$ and *Pce*$_S$_f) < than the proportion of transcripts with interspecific amino acid identity (*Pf*$_{eH}$ and *Pce*$_S$_a , *Pa*$_T$ and *Pce*$_S$_f).

Supplemental Tables

Table S1 Statistics of different assemblies for *P. cerasus* cv 'Schattenmorelle' (*Pce*$_S$) and the subgenomes *Pce*$_S$_a and *Pce*$_S$_f

Table S2 Pseudomolecule statistics for *Pce*$_S$

Table S3 Iso-Seq results

Table S4 Functional annotation results generated by interproscan using BRAKER & GeMoMa combination of ab-initio and homology-based structural gene annotation and statistics

Table S5 Mapping of marker sequences to *Prunus cerasus* 'Schattenmorelle' (*Pce*$_S$) genome

Table S6 Position of inversions within the subgenomes of *Pce*$_S$ indicated by collinearity of marker positions.

Supplemental Notes

Note S1 Access to the assembly hub for genome and annotation visualization at UCSC Genome Browser.

Note S2 Calculation of the 70% quantile from IAA.

Note S3 Calculation of %-covered bases obtained from RNAseq data.