RESEARCH

Complete chloroplast genomes of *Sorbus* sensu stricto (Rosaceae): comparative analyses and phylogenetic relationships

Chenqian Tang[1], Xin Chen[1*], Yunfei Deng[2], Liyang Geng[1], Jianhui Ma[1] and Xueyan Wei[1]

Abstract

Background: *Sorbus* sensu stricto (*Sorbus s.s.*) is a genus with important economical values because of its beautiful leaves, and flowers and especially the colorful fruits. It belongs to the tribe Maleae of the family Rosaceae, and comprises about 90 species mainly distributed in China. There is on-going dispute about its infrageneric classification and species delimitation as the species are morphologically similar. With the aim of shedding light on the circumscription of taxa within the genus, phylogenetic analyses were performed using 29 *Sorbus s.s.* chloroplast (cp) genomes (16 newly sequenced) representing two subgenera and eight sections.

Results: The 16 cp genomes newly sequenced range between 159,646 bp and 160,178 bp in length. All the samples examined and 22 taxa re-annotated in *Sorbus* sensu lato (*Sorbus s.l.*) contain 113 unique genes with 19 of these duplicated in the inverted repeat (IR). Six hypervariable regions including *trnR-atpA*, *petN-psbM*, *rpl32-trnL*, *trnH-psbA*, *trnT-trnL* and *ndhC-trnV* were screened and 44–53 SSRs and 14–31 dispersed repeats were identified as potential molecular markers. Phylogenetic analyses under ML/BI indicated that *Sorbus s.l.* is polyphyletic, but *Sorbus s.s.* and the other five segregate genera, *Aria*, *Chamaemespilus*, *Cormus*, *Micromeles* and *Torminalis* are monophyletic. Two major clades and four sub-clades resolved with full-support within *Sorbus s.s.* are not consistent with the existing infrageneric classification. Two subgenera, subg. *Sorbus* and subg. *Albocarmesinae* are supported as monophyletic when *S. tianschanica* is transferred to subg. *Albocarmesinae* from subg. *Sorbus* and *S. hupehensis* var. *paucijuga* transferred to subg. *Sorbus* from subg. *Albocarmesinae*, respectively. The current classification at sectional level is not supported by analysis of cp genome phylogeny.

Conclusion: Phylogenomic analyses of the cp genomes are useful for inferring phylogenetic relationships in *Sorbus s.s.* Though genome structure is highly conserved in the genus, hypervariable regions and repeat sequences used are the most promising molecule makers for population genetics, species delimitation and phylogenetic studies.

Keywords: *Sorbus* sensu stricto, Chloroplast genome, Repeats, Codon usage, Sequence divergence, Phylogeny

Introduction

The genus *Sorbus* L. (Maleae, Rosaceae), when established by Linnaeus [1], included only two pinnately leaved species, *S. aucuparia* L. and *S. domestica* L. The simple leaved species in *Sorbus* sensu lato (*Sorbus s.l.*) known to Linnaeus were assigned to other genera in the tribe Maleae. The taxonomy of *Sorbus* has historically been controversial. Taxonomists either adopted a broad definition [2–6] or segregated it into six small

*Correspondence: chenxinzhou@njfu.edu.cn

[1] Co-Innovation Center for Sustainable Forestry in Southern China, College of Biology and the Environment, Nanjing Forestry University, Nanjing 210037, Jiangsu, China
Full list of author information is available at the end of the article

Tang *et al. BMC Plant Biology*   (2022) 22:495

Page 2 of 16

genera, i.e., *Aria* (Pers.) Host, *Chamaemespilus* Medik., *Cormus* Spach, *Micromeles* Decne., *Sorbus* sensu stricto (*Sorbus s.s.*) and *Torminalis* Medik., with varied delimitation [1, 7–12]. Evidence from morphological [11, 13] and molecular analyses [14–19] suggested that *Sorbus s.l.* is polyphyletic and can be divided into five or six separate evolutionary lineages. Accordingly, *Sorbus s.l.* has been divided into five or six genera, and the genus *Sorbus s.s.* is restricted to species with pinnately compound leaves and small fruits [12].
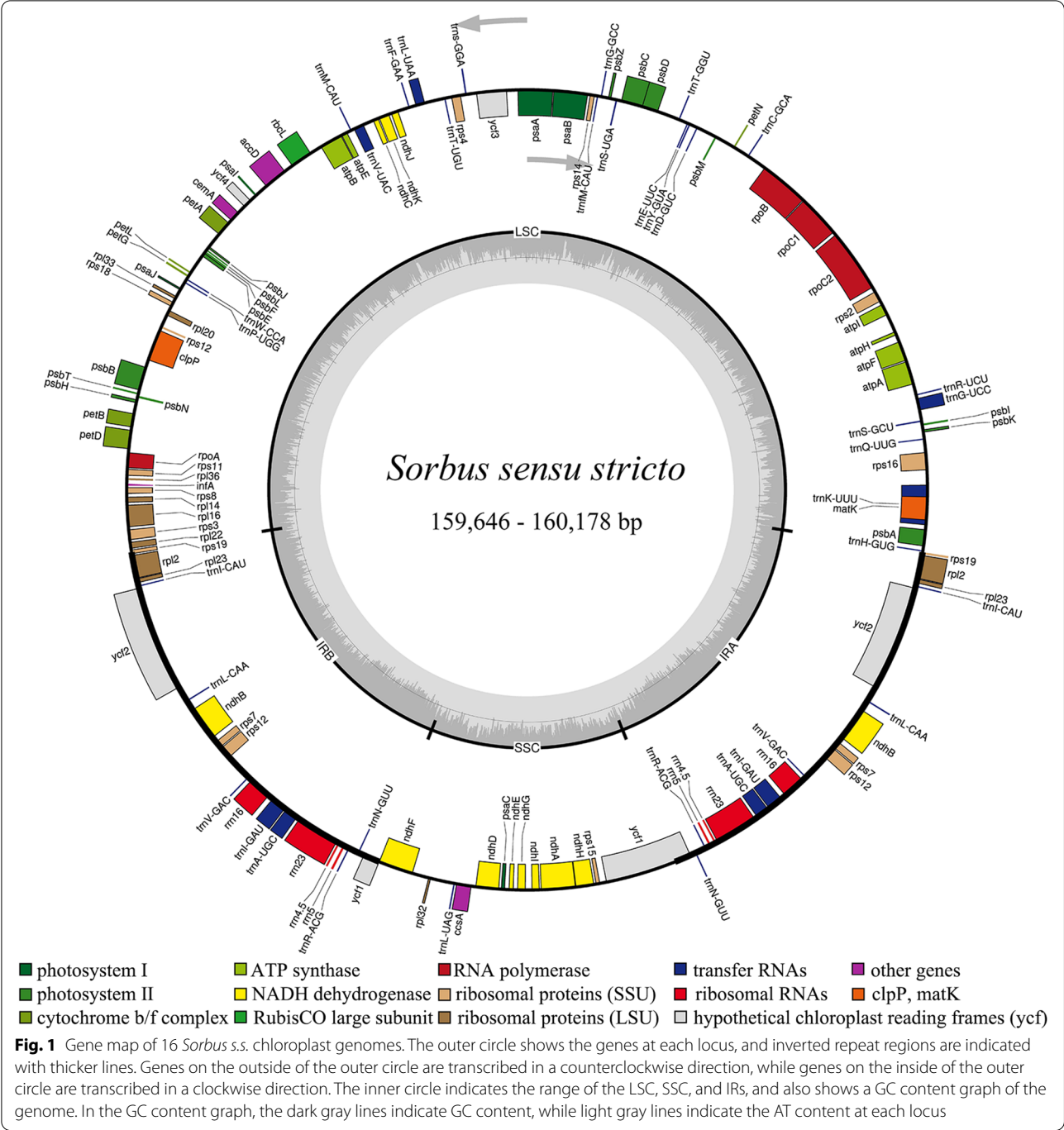
Currently, *Sorbus s.s.* consists of about 90 species with more than 60 native to China [5, 6, 12]. The genus is distributed in the temperate regions of the Northern Hemisphere with the greatest diversity in the mountains of south-western China, adjacent areas of Upper Burma and the Eastern Himalaya [12]. *Sorbus s.s.* species have great horticultural potential for their autumn leaf color and late summer and autumn fruit displays which range in colour from scarlet and deep crimson to orange, pink, yellow and pure white. However, relationships within the genus are still unresolved due to interspecific hybridization, apomictic polyploidy and the limited phylogenetic research data available, so intrageneric classifications proposed by previous taxonomists need to be tested. In the twentieth century, the broad definition of the *Sorbus* was adopted by most authors and the genus *Sorbus s.s.* was usually treated as a subgenus or a section within *Sorbus s.l.* Koehne [20] classified subg. *Aucuparia* (equivalent to *Sorbus s.s.*) into five unnamed groups because it was "impossible to divide the genus into well characterized sections". Yü and Kuan [21] separated sect. *Sorbus* (equivalent to *Sorbus s.s.*) into eight series based on morphological characters such as the presence of trichomes on the buds, the number and shape of the leaflets and the fruit color. Gabrielian [4] argued that some series proposed by Yü and Kuan [21] included distantly related taxa and assigned species of sect. *Sorbus* in Western Asia and the Himalayas to nine subsections based on comparative morphological and anatomical data. Phipps et al. [5] divided subg. *Sorbus* (equivalent to *Sorbus s.s.*) into two sections, nine series and five informal groups based on morphological characters such as the number and size of leaflets, free or united carpels and fruit color. The only recent revision of the genus S*orbus s.s.* was by published McAllister [12], who divided the genus into two subgenera and 11 sections based mainly on morphological characters such as the color of hairs on the buds, the number, size and shape of leaflets and the color of the fruits, combined with ploidy levels, breeding systems and geographical distribution. Both infrageneric classification and taxonomic inconsistency in species delimitation remain a challenge in the genus. The identities of *S. rehderiana* Koehne and *S. koehneana* C.K.Schneid. are examples

discussed here. *S. hypoglauca* (Cardot) Hand.-Mazz. was treated as a synonym of *S. rehderiana* by Yü and Lu [3] and Lu and Spongberg [6], *S. unguiculata* Koehne was reduced to the synonymy of *S. koehneana* by McAllister [12], while both of them were recognized as distinct species by McAllister [12] and Phipps et al. [5].

Previous molecular studies mainly focused on the phylogeny of the tribe Maleae, while few concentrated specifically on the infrageneric relationships within *Sorbus s.s.* Despite previous efforts to elucidate infrageneric relationships within the genus, relationships between the subgenera and sections have remained uncertain. Phylogenetic analyses using chloroplast markers [16–18, 22] or chloroplast (cp) genomes [19] supported the monophyly of the genus but did not support any existing infrageneric classifications. However, conflicting results were noted in nuclear DNA phylogenies. Based on ITS, Wang and Zhang [23] suggested that *Sorbus s.l.* was monophyletic, but that *Sorbus s.s* and the infrageneric groups were not monophyletic based on ITS phylogeny. Contrary to Wang and Zhang [23], based on ITS, Li et al. [24] supported the monophyly of *Sorbus s.s.* and the other four segregated segregated genera from *Sorbus s.l.*, *i.e.*, *Aria* (including *Micromeles*), *Chamaemespilus*, *Cormus* and *Torminalis*.

Chloroplast genomes of most vascular plants range from 120 to 160 kb, and their cp genomes have a conserved quadripartite structure composed of two copies of an inverted repeat (IR) which divides the remainder of the genomes into one large and one small single copy regions (LSC and SSC) [25]. Chloroplast genomes are frequently used in systematics for the simplicity of the circular structure, predominantly clonal inheritance along the maternal line, as well as being highly variable even at low taxonomic levels [26]. Knowledge of the organization and evolution of cp genomes in *Sorbus s.s.*, *Sorbus s.l.* and tribe Maleae has been expanding rapidly because of the fast growing number of completely sequenced genomes available. Currently, the cp genomes of more than 100 species in the tribe Maleae including 22 species of *Sorbus s.l.* have been reported and are available for use (https://www.ncbi.nlm.nih.gov).

Thus, in genus *Sorbus s.s.*, relationships among subgenera and sections remained uncertain. In this study, cp genomes of 15 *Sorbus s.s.* samples and an unidentified sample were newly sequenced and compared with other 22 other samples of *Sorbus s.l.* and 26 samples of other genera from tribe Maleae. The aims were: (1) to determine the structure of cp genomes in the 16 *Sorbus s.s.* samples; (2) to compare the structural variation, investigate and screen mutational hotspots, examine variations of simple sequence repeats (SSRs) and dispersed repeat sequences, and to calculate the nucleotide diversity in *Sorbus s.s.* cp genomes for future population genetic,
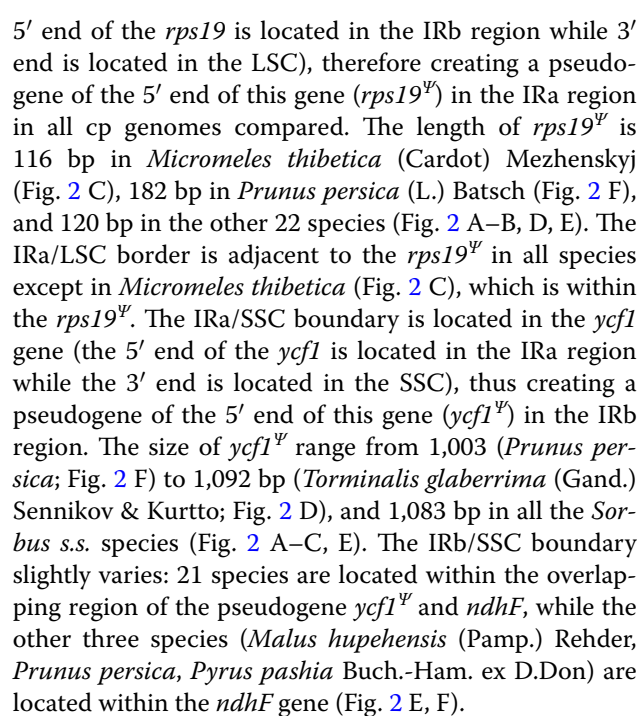
Fig. 1 Gene map of 16 *Sorbus s.s.* chloroplast genomes. The outer circle shows the genes at each locus, and inverted repeat regions are indicated with thicker lines. Genes on the outside of the outer circle are transcribed in a counterclockwise direction, while genes on the inside of the outer circle are transcribed in a clockwise direction. The inner circle indicates the range of the LSC, SSC, and IRs, and also shows a GC content graph of the genome. In the GC content graph, the dark gray lines indicate GC content, while light gray lines indicate the AT content at each locus

species delimitation and phylogenetic studies; (3) to reconstruct phylogenetic relationships among species in *Sorbus s.s.* and *Sorbus s.l.*

Results

Organization and features of the chloroplast genomes

The chloroplast genomes of the 15 species and the unidentified sample of *Sorbus s.s.* exhibit similar structure and organization (Table S1, Fig. 1). The size of the cp genomes of the 16 *Sorbus s.s.* samples range from 159,646 bp in *S. wilsoniana* C.K.Schneid. to 160,178 bp in *S. hypoglauca*. All the 16 cp genomes consist of a large single-copy (LSC) with lengths between 87,612 bp in *S. sargentiana* Koehne and 88,125 bp in *S. hypoglauca*, a small single-copy (SSC) with lengths between 19,219 bp in *Sorbus* sp. and 19,359 bp in *S.*

Fig. 2 Comparisons of the LSC, SSC and IRs boundaries (**A–D**) within *Sorbus s.l.* and (**E** and **F**) among three other Rosaceae cp genomes. Genes shown below are transcribed forward and those shown above the lines are transcribed in reverse. Minimum and maximum sizes for the regions and structures of each chloroplast type that compose the borders are indicated in base pairs (bp)

tianschanica Rupr.; and a pair of inverted repeats (IRs) with lengths between 26,378 bp (*S. aestivalis* Koehne and the other nine taxa) and 26,405 bp (*S. amabilis* Cheng ex T.T.Yu; Table S1). The total GC content is nearly similar, 36.5% for five samples and 36.6% for the other 11 samples (Table S1).

All the 16 cp genomes assembled here encode 113 unique genes (79 protein-coding genes, 30 tRNA genes and four rRNA genes), and 19 of these are duplicated in the IR, giving a total of 132 genes (Tables S1, S2 and Fig. 1). Eighteen genes contain one (*atpF, ndhA, ndhB, petB, petD, rpl2, rpl16, rpoC1, rps12, rps16, trnA-UGC, trnG-UCC, trnI-GAU, trnK-UUU, trnL-UAA* and *trnV-UAC*) or two (*clpP* and *ycf3*) introns, and six of these are the tRNA genes (Table S2, Fig. 1). The cp genomes consist of 56.5 or 56.6% coding regions (49.1 or 49.2% protein coding genes and 7.4% RNA genes) and 43.4 or 43.5% non-coding regions, including both intergenic spacers and introns (Table S1).

The boundaries between the IR and LSC/SSC regions of the 16 *Sorbus s.s.* cp genomes and the eight species in other genera in Rosaceae were compared (Fig. 2). The IRb/LSC boundary is located within the *rps19* gene (the

5′ end of the *rps19* is located in the IRb region while 3′ end is located in the LSC), therefore creating a pseudogene of the 5′ end of this gene (*rps19$^{\Psi}$*) in the IRa region in all cp genomes compared. The length of *rps19$^{\Psi}$* is 116 bp in *Micromeles thibetica* (Cardot) Mezhenskyj (Fig. 2 C), 182 bp in *Prunus persica* (L.) Batsch (Fig. 2 F), and 120 bp in the other 22 species (Fig. 2 A–B, D, E). The IRa/LSC border is adjacent to the *rps19$^{\Psi}$* in all species except in *Micromeles thibetica* (Fig. 2 C), which is within the *rps19$^{\Psi}$*. The IRa/SSC boundary is located in the *ycf1* gene (the 5′ end of the *ycf1* is located in the IRa region while the 3′ end is located in the SSC), thus creating a pseudogene of the 5′ end of this gene (*ycf1$^{\Psi}$*) in the IRb region. The size of *ycf1$^{\Psi}$* range from 1,003 (*Prunus persica*; Fig. 2 F) to 1,092 bp (*Torminalis glaberrima* (Gand.) Sennikov & Kurtto; Fig. 2 D), and 1,083 bp in all the *Sorbus s.s.* species (Fig. 2 A–C, E). The IRb/SSC boundary slightly varies: 21 species are located within the overlapping region of the pseudogene *ycf1$^{\Psi}$* and *ndhF*, while the other three species (*Malus hupehensis* (Pamp.) Rehder, *Prunus persica, Pyrus pashia* Buch.-Ham. ex D.Don) are located within the *ndhF* gene (Fig. 2 E, F).

Tang *et al. BMC Plant Biology*   (2022) 22:495

Page 5 of 16

Codon preference analysis

According to the codon usage analysis, the total sequence sizes of the protein coding genes are 78,570–78,588 bp in the 16 *Sorbus s.s.* genomes; 26,190–26,196 codons were encoded (Table S3). Leucine encoded with the highest number of codons ranging from 2,753 to 2,757, followed by isoleucine, with the number of codons encoded between 2,255 and 2,260. Cysteine is the least (297 or 298). The relative synonymous codon usage (RSCU) values vary a little among the 16 *Sorbus s.s.* sequences. Thirty codons are used frequently with RSCU > 1 and 32 codons used less frequently with RSCU < 1. UUA shows a preference in all the 16 cp genomes. The frequency of use for the start codons AUG and UGG, encoding methionine and tryptophan, show no bias (RSCU = 1). Codons with A (32.1%) or U (38.2%) in the third position are all 70.3%, thus the codon usage is biased towards A or U at the third codon position.

Repeated sequences analysis

The total number of SSRs in 16 *Sorbus s.s.* genomes ranges from 44 to 53 (Fig. 3 A–C; Table S4). The most abundant SSRs are A or T nucleotide repeats, which account for 88.2 to 96.3% (Table S4) of the total. The most common SSRs are mononucleotides, which range from 29 to 38, followed by tetranucleotides ranging from 5 to 7, and pentanucleotides ranging from 2 to 5. Dinucleotides are all four in the examined samples except for five in *S. tianschanica*. Trinucleotides were discovered only in three species: *S. filipes* Hand.-Mazz., *S. hypoglauca* and *S. rutilans* McAll. There are three hexanucleotides in *S. cibagouensis* H.Peng & Z.J.Yin, two in *S. helenae* Koehne, one in *S. aestivalis*, *S. albopilosa* T.T.Yu & L.T.Lu, *S. amabilis* and *S. rehderiana*, and none in the other 11 samples. SSRs are mainly distributed in the intergenic regions (76.6–89.4%), with much lower quantities distributed in the intron regions (10.6–21.3%) and exon regions (0–2.1%; Fig. 3B). Furthermore, SSRs are found mainly in LSC regions (78.4–89.4%), and are significantly lower in the SSC (6.4–17.6%) and IR (3.8–8%) regions (Fig. 3 C).

The REPuter screening discovered 14 to 31 dispersed repeats of 25 bp or longer among the 16 *Sorbus s.s.* cp genomes examined (Fig. 3 D–E). *Sorbus rehderiana* has the largest number of repeats with 31 and *S. cibagouensis* has the fewest with only 14. The majority of the repeats (19.1–44.4%) in all cp genomes range between 25 and 29 bp. The longest repeats is 123 bp and was only found in *S. foliolosa* Spach. Six taxa, *S. albopilosa*, *S. cibagouensis*, *S. helenae*, *S. pteridophyslla* Hand.-Mazz., *S. tianschanica* and *Sorbus* sp., have a maximum size of 44 bp. Only four taxa, *S. foliolosa*, *S. hypoglauca*, *S. rehderiana* and *S. ursina* S.Schauer, have repeats with a size

larger than 60 bp (Table S5, Fig. 3 E). Among these, forward repeats (6–21) were the most common, followed by palindromic repeats (6–11) and reverse repeats (1–3, Fig. 3 D).

Comparative analysis of chloroplast genomes

Comparative cp genome analysis reveals that noncoding regions are generally more divergent than coding regions and LSC/SSC regions are more divergent than IR regions (Fig. 4). The highest levels of divergence were found in 17 intergenic regions, 15 in the LSC regions, namely *trnH-psbA*, *trnK-rps16*, *trnG-trnR*, *trnR-atpA*, *atpF-atpH*, *atpH-atpI*, *trnC-petN*, *petN-psbM*, *trnT-psbD*, *psbZ-trnG*, *trnT-trnL*, *ndhC-trnV*, *trnM-atpE*, *accD-psaI* and *rps8-rpl14*; and two in the SSC regions, namely *ndhF-rpl32*, *rpl32-trnL*. Apart from these regions, two intron regions: *clpP* and *rpl16* also show high sequence variation.
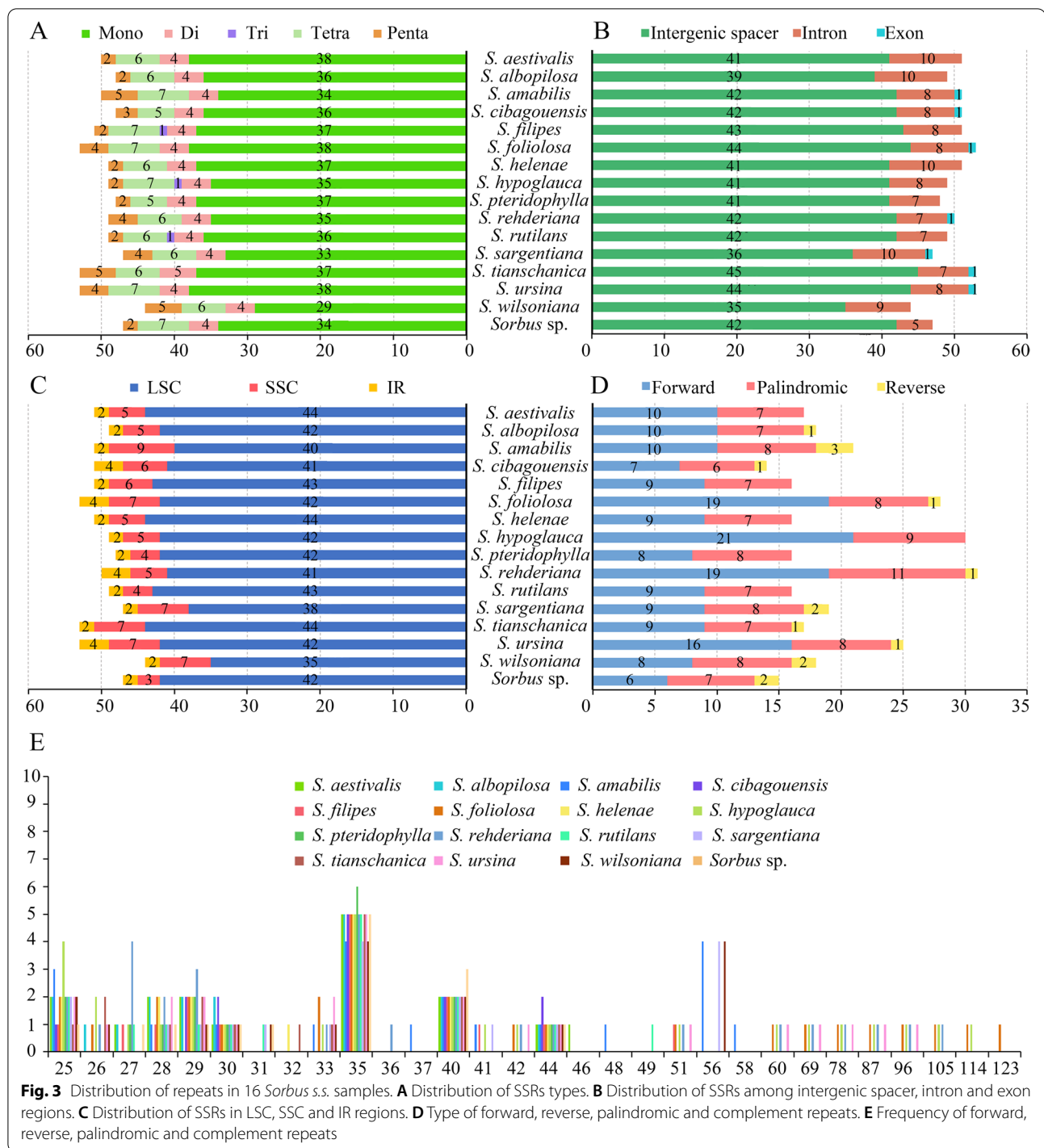
To elucidate levels of diversity at the sequence level, the nucleotide diversity (Pi) values were calculated. The Pi values range from 0 to 0.00975, with mean value of 0.00098 (Fig. 5, Table S6). The SSC region shows the highest nucleotide diversity (Pi = 0.00173), while the lowest Pi is in the IR boundary regions (Pi = 0.00016). Meanwhile, six hypervariable sites with Pi between 0.005 and 0.01 were screened, which are *trnR-atpA* (Pi = 0.00975), *petN-psbM* (Pi = 0.00932), *rpl32-trnL* (Pi = 0.00753), *trnH-psbA* (Pi = 0.00636), *trnT-trnL* (Pi = 0.00642) and *ndhC-trnV* (Pi = 0.00616).

Phylogenetic analysis

The ML and BI analyses of cp genomes result in highly congruent topologies. There are only slight differences in support values between the phylogenetic trees. Therefore, only the ML topology is shown here with the ML/BI support values added at each node (Fig. 6).
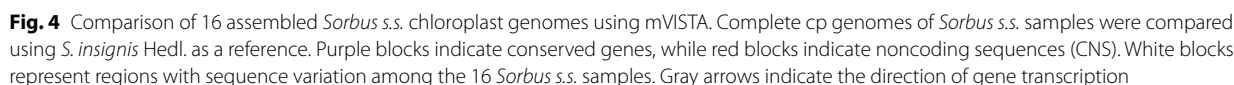
Our analyses confirmed that *Sorbus s.l.* is polyphyletic and the six segregate genera, *i.e.*, *Aria*, *Chamaemespilus*, *Cormus*, *Miromeles*, *Sorbus s.s.* and *Torminalis*, are monophyletic. *Aria*, *Chamaemespilus* and *Torminalis* are resolved in one branch near the base of the cp genome phylogeny together with *Malus trilobata* C.K. Schneid., *Aronia arbutifolia* (L.) Pers. and *Cydonia oblonga* Mill. *Miromeles* is sister to *Sorbus s.s.* and nested in one branch with *Cormus* and *Pyrus* L.

Within the monophyletic genus *Sorbus s.s.*, two major clades are resolved. Clade I comprises two fully supported subclades (A and B). Subclade A is consistent with subg. *Albocarmesinae* McAll. Nevertheless, three sections, sect. *Hypoglaucae* McAll., sect. *Insignes* (T.T. Yu) McAll. and sect. *Multijugae* (T.T.Yu) McAll. within subg. *Albocarmesinae*, are not monophyletic. Subclade B contains two samples representing *S. tianschanica*

Tang *et al. BMC Plant Biology*   (2022) 22:495

Page 6 of 16

Fig. 3 Distribution of repeats in 16 *Sorbus s.s.* samples. **A** Distribution of SSRs types. **B** Distribution of SSRs among intergenic spacer, intron and exon regions. **C** Distribution of SSRs in LSC, SSC and IR regions. **D** Type of forward, reverse, palindromic and complement repeats. **E** Frequency of forward, reverse, palindromic and complement repeats

which belongs to subg. *Sorbus* sect. *Tianshanicae* (Kom. ex T.T.Yu) McAll., although it is resolved with full-support on a branch with subg. *Albocarmesinae* with full-support. Clade II contains two fully supported subclades (C and D) and is sister to the rest of the genus. Subclade C includes five taxa belonging to

three different sections from two subgenera. *Sorbus aucuparia* in sect. *Sorbus* McAll. (subg. *Sorbus*) and *S. hupehensis* var. *paucijuga* in sect. *Discolores* (T.T.Yu) McAll. (subg. *Albocarmesinae*), are nested within sect. *Commixtae* McAll. (subg. *Sorbus*). Subclade D contains two species in subg. *Sorbus* sect. *Wilsonianae* McAll.

Tang *et al. BMC Plant Biology* (2022) 22:495

Page 7 of 16

Fig. 4 Comparison of 16 assembled *Sorbus s.s.* chloroplast genomes using mVISTA. Complete cp genomes of *Sorbus s.s.* samples were compared using *S. insignis* Hedl. as a reference. Purple blocks indicate conserved genes, while red blocks indicate noncoding sequences (CNS). White blocks represent regions with sequence variation among the 16 *Sorbus s.s.* samples. Gray arrows indicate the direction of gene transcription

Discussion

Gene, structure and the potential molecular markers

Sorbus s.s. can be easily identified by the pinnate leaves and colorful fruits with persistent sepals, stamens and styles. Understanding of the taxonomy and phylogenetic relationships in *Sorbus s.s.* has been particularly difficult because of the widespread occurrence of polyploidy associated with gametophytic apomixis [27–29]. In the present study, 29 *Sorbus s.s.* cp genomes (16 newly sequenced and 13 previously reported) representing 23 species, one variety and two unidentified taxa from both subgenera and eight out of the 11 sections were compared in all cp genomes in *Sorbus s.l.* to clarify phylogenetic relationships and resolve taxonomic uncertainties.

The structure, gene order and GC content are highly conserved and nearly similar in *Sorbus s.s.* samples analyzed here, and are identical to other cp genomes in angiosperms [30–35]. The size of the 29 cp genomes varied from 159,632 (*S. ulleungensi*s Chin S. Chang; NC037022) to 160,178 bp (*S. hypoglauca*). The *Sorbus s.s.* cp genomes sequenced here all contained 113 unique genes with the total GC content being 36.5 or 36.6% (Table S1). However, the absence of one to six of the following genes: *infA, psaC, psbL, rpl16, rrn4.5, rrn5, rrn16, rrn23, trnG-GCC, trnG-UCC, trnI-CAU* and *trnS-GGA,* were reported in 22 species in *Sorbus s.l.* (https://www.ncbi.nlm.nih.gov/, Table S8). Some species were found to contain different numbers of

Tang *et al. BMC Plant Biology* (2022) 22:495

Page 8 of 16
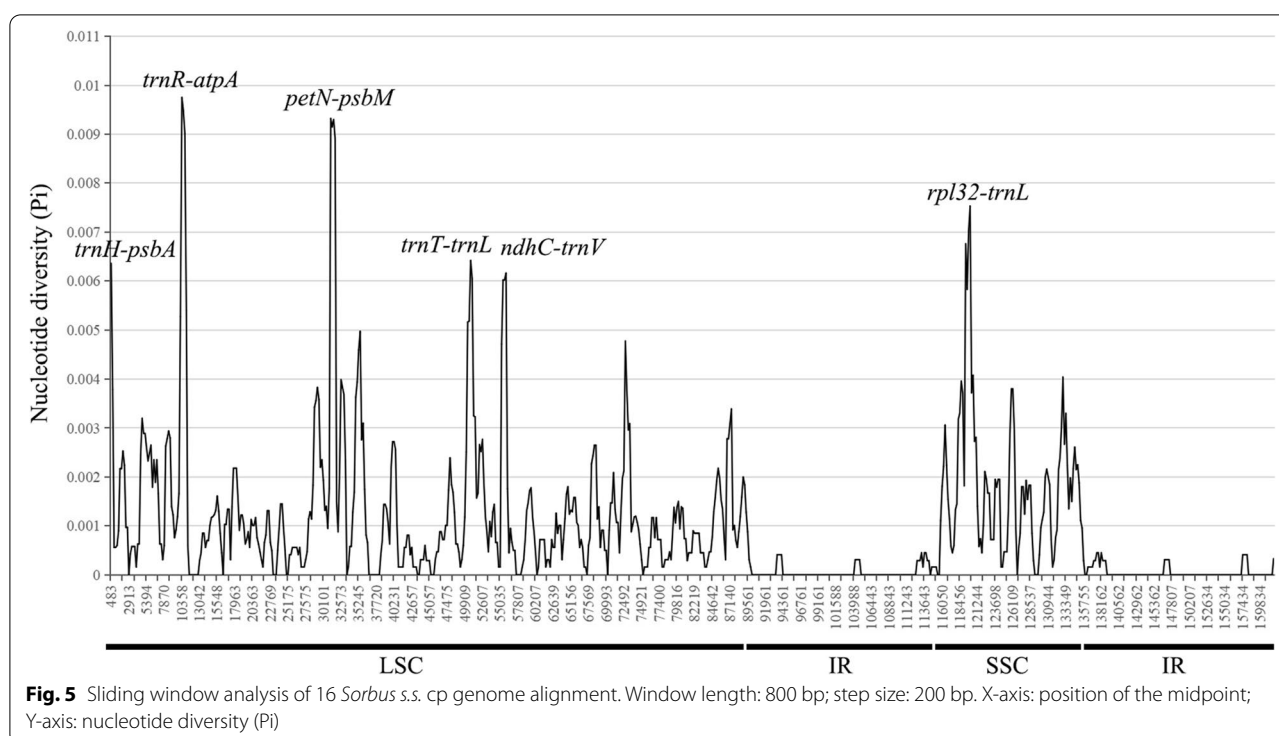
Fig. 5 Sliding window analysis of 16 *Sorbus s.s.* cp genome alignment. Window length: 800 bp; step size: 200 bp. X-axis: position of the midpoint; Y-axis: nucleotide diversity (Pi)

genes in different individuals, for examples, *S. amabilis* (MT357029) and *S. helene* (KY419924) were reported to contain 109 and 111 genes respectively, but both annotated 113 genes in the samples examined here. To eliminate the influence of annotation software and references used, the 22 samples were all re-annotated using Plastid Genome Annotator (PGA) program [36] and Geneious v.9.0.2 [37] with *S. insignis* (NC051947) and *Malus hupehensis* (NC040170) as references. Unexpectedly, no gene loss was found and all the 22 sequences re-annotated contained 113 genes which were identical to the samples examined in this study (Table S8).

Genome composition and natural selection are the two major factors affecting codon usage bias [38, 39]. The total number of 64 codons present across the *Sorbus s.s.* cp genomes encoding 20 amino acids (AAs) and codon usage is biased towards A or U at the third codon position, which is in consistent with other higher plants [40–43].

The contraction and expansion of IR regions are useful in evolutionary studies in some taxa [44]. However, the IR/SC boundaries are conserved in *Sorbus s.s.* and in most species of *Sorbus s.l.* All species compared in this study with the IRb/LSC boundary were located within the *rps19* gene and creating a pseudogene (*rps19*$^\Psi$) in the IRa, the IRa/SSC boundary located in the *ycf1* gene and creating a pseudogene (*ycf1*$^\Psi$) in the IRb region.

SSRs are useful markers to assess the organization of genomes and diversity at the species and population levels [45–47] and to analyse phylogenetic relationships in plants [48]. In this study, the number of SSRs found within *Sorbus s.s.* genomes ranges from 44 to 53, similar to SSRs previously documented in the genus [49, 50]. Consistent with the previous reports in other *Sorbus s.s.* species, mononucleotides are the most common SSRs and the largest amount of SSRs is located in the intergenic regions. SSRs are especially useful in establishing the amount of genetic diversity within and between populations [51] and in investigating the parentage of polyploid in *Sorbus s.s.* [52]. Dispersed repetitive sequences represent a major component of genomes and play a major role in genomic rearrangement and sequence variation [53, 54]. *Sorbus s.s.* species contain a substantial number of dispersed repeats and show marked difference in number ranging from 58 to 130 with a majority of the repeats ranging from between 20 and 25 bp.

Despite the high levels of gene conservation observed, 17 intergenic regions and two introns genes were identified as highly divergent in *Sorbus s.s.* (Figs. 4 and 5). Among them, some were shown in previous studies to be highly variable and of high phylogenetic utility, such as *trnK-rps16*, *atpH-atpI*, *trnT-psbD*, *ndhC-trnV*, *ndhF-rpl32* and *rpl32-trnL* [34, 55]. Consistent with the diverse patterns found in most angiosperms [56–58], sequence divergence in non-coding regions is higher than that in
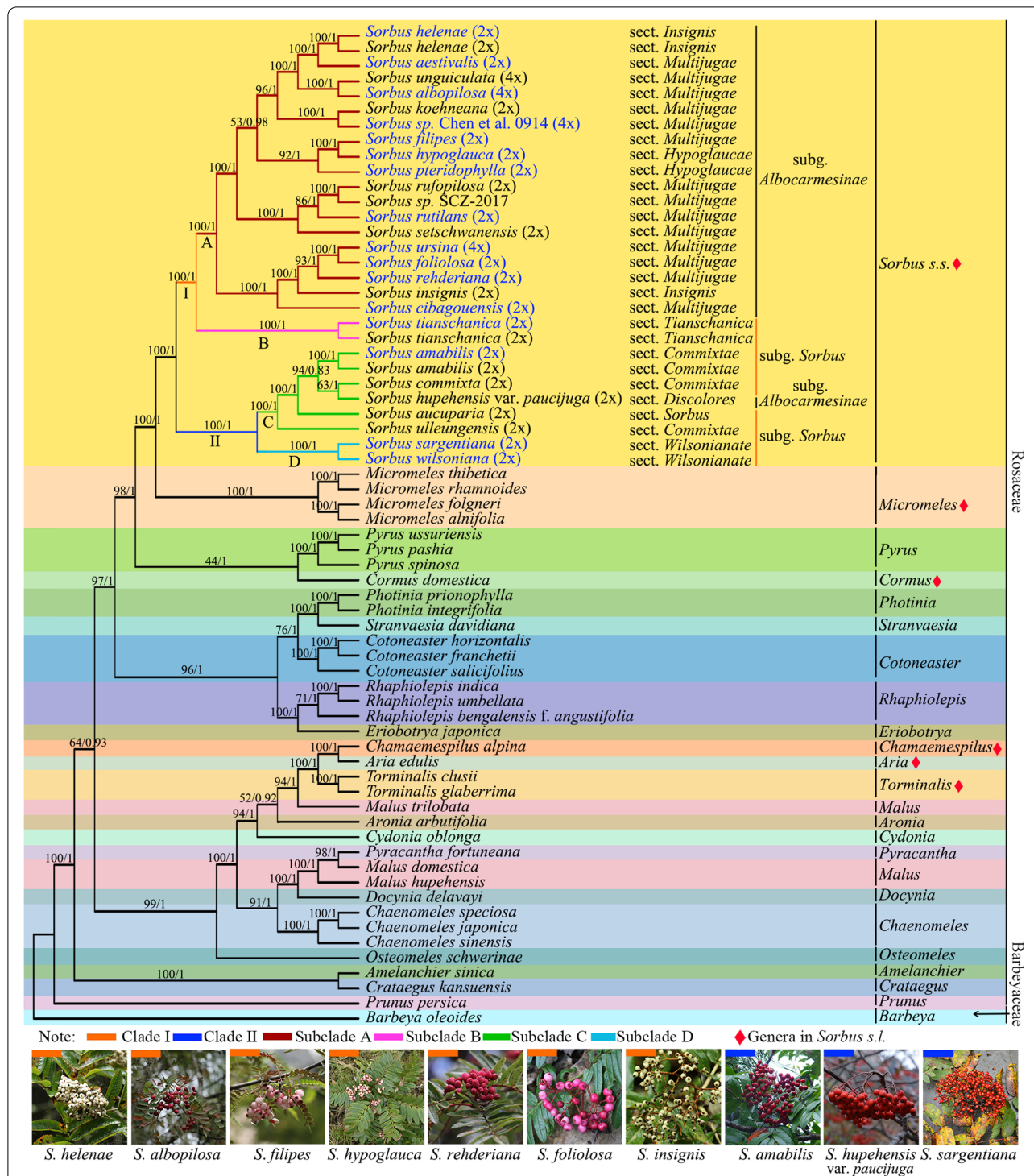
Fig. 6 Phylogenetic tree based on complete cp genomes resulting from the maximum likelihood (ML) analysis and Bayesian inference (BI) analysis. Bootstrap values in ML analysis and posterior probabilities in BI analysis are listed at nodes. Names of taxa newly sequenced in *Sorbus s.s.* are in blue

coding regions. Variable chloroplast sequences have been widely used for plant phylogeny reconstruction [58, 59] and for species identification [60, 61]. However, among the chloroplast sequences which have most frequently been used in phylogeny reconstruction of tribe Maleae, such as *trnH-psbA*, *trnL-trnF* and *trnG-trnS*, etc. [15, 17,

Tang *et al. BMC Plant Biology*   (2022) 22:495

Page 10 of 16

18], only one intergenic region, *trnH-psbA* (Pi = 0.00339; ranked 5) and one intron *rpl16* (Pi = 0.00339; ranked 15) show high variability in *Sorbus s.s.* The only one chloroplast marker, *rps16-trnK*, which was applied in the phylogeny and historical biogeography analysis of *Sorbus s.s* [22]. with Pi value (0.0032) ranked 16. Furthermore, the intergenic region *trnR-atpA* shows the highest Pi value (0.00975) in all *Sorbus s.s.* genomes, and is also hypervariable within the genomes of other species in Rosaceae genomes [18, 33, 62]. Rapid evolutionary rates of the six hypervariable regions, *trnR-atpA*, *petN-psbM*, *rpl32-trnL*, *trnT-trnL*, *trnH-psbA* and *ndhC-trnV*, have led to some topology confusion (Figs. S1,S2,S3,S4,S5,S6, and S7). Howerver, the phylogenetic analysis based on *trnR-atpA* and *trnT-trnL* resulted in similar topology as the complete cp genomes (Figs. S1, and S4), *rpl32-trnL* and *trnT-trnL* could well differentiated the two subgenera in *Sorbus s.s.* (Figs. S3, and S4). Meanwhile, *rpl32-trnL*, *trnT-trnL* and the combination performed well in species identification which grouped the two individuals of same species included in the studied together (Figs. S3, and S4). Thus, the new highly variable sequences generated in this study, especially the six hypervariable regions, *trnR-atpA*, *petN-psbM*, *rpl32-trnL*, *trnT-trnL*, *trnH-psbA* and *ndhC-trnV*, might be the most promising potential molecule makers in phylogeny reconstruction and DNA barcode identification for *Sorbus s.s.* plants.

Phylogenetic analysis

Chloroplast genomes are effective in inferring phylogenetic relationships at various taxonomic levels for the conservatism and uniparental heritance [63–65]. In this phylogenetic analysis using cp genomes, the monophyly and the infrageneric classification of *Sorbus s.s.* were investigated, as well as its relationship with other genera in Maleae. The status of *S. hypoglauca* and *S. unguiculata* was also re-evaluated.

In congruence with previous molecular phylogenetic studies [15, 17–19] and morphological research [11, 13], the generic circumscription of *Sorbus s.l.* is not supported by the phylogenetic analyses of this study. Six monophyletic lineages correspond to the six genera segregated from *Sorbus s.l.*, *Aria*, *Chamaemespilus*, *Cormus*, *Miromeles*, *Sorbus s.s.* and *Torminalis*, and are all well supported. However, the delimitations of three genera with simple leaves, *i.e.*, *Aria*, *Chamaemespilus* and *Micromeles*, was controversial. *Aria* was usually accepted in a broad sense in earlier morphological studies to include *Chamaemespilus* and *Micromeles* [11, 66, 67] or in molecular studies to include only *Micromeles* [14, 15, 24]. Our analyses indicated that Asiatic species formerly included in *Aria* with a persistent calyx are nested within *Micromeles* which forms the sister group of *Sorbus s.s.*,

but are only distantly related to *Aria edulis*, the type species of *Aria*. Therefore, it is proposed to treat *Micromeles* as an independent genus. All Asiatic simple leaved species formerly included in *Aria* have been transferred to *Micromeles* by Mezhenska et al. [68]. In our study, *Chamaemespilus* forms a sister group to *Aria*. The relationship between them needs to be further investigated.

The systematics of *Sorbus s.s.* have been discussed in terms of morphological [4, 5, 12, 20, 21] and molecular [17, 22–24] results. The topologies of the phylogenetic trees obtained here are congruent with that reported by Li et al. [22] using four nuclear markers (LEAFY-2, GBSSI-1, SBEI and WD) and one chloroplast marker (*rps16-trnK*). Corresponding to Li et al. [22], the two monophyletic clades resolved are largely congruent with the two subgenera, subg. *Albocarmesinae* and subg. *Sorbus*, as defined by McAllister [12]. However, the sections defined by McAllister [12] and infrageneric classification proposed by Koehne [20], Yü & Kuan [21], Gabrielian [4] and Phipps et al. [5], are not supported.

Species in clade I have white to crimson flowers, pinkish-red, white to pink or crimson fruits which will gradually become almost pure white with only the occasional crimson or pink fleck when ripe. Two monophyletic subclades, namely subclade A and subclade B are resolved in this clade. Subclade A consists of 16 species in subg. *Albocarmesinae* and two unidentified samples that are morphologically similar to species in this subgenus. Two species, *S. helenae* and *S. insignis* of sect. *Insignes* and two species of sect. *Hypoglaucae*, *S. hypoglauca* and *S. pteridophylla,* are nested within sect. *Multijug*ae. Thus, the three sections in subclade A, sect. *Hypoglaucae*, sect. *Insignes* and sect. *Multijug*ae are not monophyletic. McAllister [12] distinguished the three sections by the color of the hairs on buds and young shoots, whether petiole bases were sheathing or not and carpel apices free or fused, together with the ploidy levels. However, the two fully supported groups in subclade A lack of a consistent morphological synapomophy. Species in subclade A are sexual diploids or apomictic tetraploids. Four taxa, *S. albopilosa* (2C = 2.624 ± 0.047 pg), *S. unguiculata* (2C = 2.783 ± 0.103 pg), *S. ursina* (2C = 2.681 ± 0.028 pg) and the unidentified *Sorbus* sp. Chen et al. 0914 (2C = 2.765 ± 0.248 pg) are tetraploids (Chen et al. unpublished), the other 13 species are diploids [12, 69, 70]. The ploidy level of *Sorbu*s sp. SCZ-2017 is unknown. The tetraploids species, *S. albopilosa* and *S. unguiculata,* are clustered together and form a fully supported group with the diploid species *S. helenae* and *S. aestivalis*, *S. ursina* is grouped with the diploid species *S. foliolosa* and *Sorbus* sp. Chen et al. 0914 is grouped with diploid *S. koehneana*. However, the origin of tetraploid taxa and the relationship with the closely related diploid ones

Tang *et al. BMC Plant Biology*   (2022) 22:495

Page 11 of 16

need further study. Subclade B contains two samples of *S. tianschanica,* a species that was formerly included in sect. *Tianshanicae* under subg. *Sorbus* by McAllister [12] and it is also a sexual diploid. In accordance with previous publications [22], *S. tianschanica* is sister to the sampled species of subg. *Albocarmesinae* and it suggests that the species may be misplaced. *Sorbus tianschanica* can be distinguished from all other species of subg. *Sorbus* by its "very glossy twigs" [12]. Furthermore, McAllister [12] noted that sect. *Tianshanicae* has fruits with a distinctive pinkish-red color unknown in subg. *Sorbus*, and he thought that it might indicate some relationship with species in subg. *Albocarmesinae*. Our results suggest the transfer of *S. tianschanica* to subg. *Albocarmesinae* from subg. *Sorbus*. However, more samples from other species of sect. *Tianshanicae* need to be sequenced to confirm its placement at sectional level.

Species in clade II could be easily distinguished from species in clade I by their white flowers and orange-red to bright red fruits which lack any trace of white or crimson [12, 24, 71]. All species in clade II are sexual diploids. In clade II, two subclades, subclades C and D, are fully supported. Morphologically, species in subclades C have much small inflorescences and relatively larger fruits compared to species in subclades D. Subclade C contains five taxa formerly assigned to three sections, sect. *Commixtae*, sect. *Sorbus* and sect. *Discolores*. Two taxa, *Sorbus aucuparia* of sect. *Sorbus* and *S. hupehensis* var. *paucijuga* of sect. *Discolores*, are nested within sect. *Commixtae*. Morphologically, *S. aucuparia* can be easily distinguished from the other two species in sect. *Sorbus*, *S. esserteauiana* Koehne and *S. scalaris* Koehne by its smaller stipules while the other two have larger, persistent stipules; *S. hupehensis* var. *paucijuga* is more closely related to *S. amabilis* and *S. commixta* in having white flowers and small red fruits rather than to *S. hupehensis* C.K. Schneid. which has white fruits. Therefore, *S. aucuparia* and *S. hupehensis* var. *paucijuga* might merit transfer to sect. *Commixtae*. Subclade D includes two species, *S. sargentiana* and *S. wilsoniana* of sect. *Wilsonianae* in subg. *Sorbus*, and it is the only one section whose monophyly is supported in the present study.

Taxonomic inconsistencies in species delimitation also remain a challenge in the genus *Sorbus s.s. S. hypoglauca* (Cardot) Hand.-Mazz. was treated as a synonym of *S. rehderiana* by Yü and Lu [3], and Lu and Spongberg [6], but it was reinstated as a distinct species by McAllister [12]. In the present study, *S. hypoglauca* is sister to *S. filipes* but not *S. rehderiana*. *Sorbus hypoglauca* differs from both *S. filipes* and *S. rehderiana* in having large persistent stipules. Therefore, there is good support for the treatment of *S. hypoglauca* as a distinct species. *S. unguiculata* Koehne was reduced to synonymy with *S. koehneana* by McAllister [12], but was treated as a distinct species by Phipps et al. [5]. In our study, *S. unguiculata* is not clustered with *S. koehneana*, but formed a sister group to *S. albopilosa*. Morphologically, *S. unguiculata* differs from *S. koehneana* by the much more numerous of leaflets, and from *S. albopilosa* by the white, not red fruits. Therefore, *S. unguiculata* might merit treatment as a distinct species.

Conclusion

Complete chloroplast genomes of 29 samples in *Sorbus s.s.,* including 16 newly sequenced samples, representing both two subgenera and eight sections were compared. Although genome structure, organization and gene content are highly conserved in the genus, differences in number and distribution of repeat sequences and the six hypervariable regions could be used for molecular systematic, phylogeographic, and population genetic studies.

Sorbus s.s. and the other five genera segregated from *Sorbus s.l.* (*i.e.,* *Aria*, *Chamaemespilus*, *Cormus*, *Miromeles* and *Torminalis*) are strongly supported as monophyletic, while *Sorbus s.l.* is confirmed to be polyphyletic. The two subgenera of *Sorbus s.s.*, subg. *Sorbus* and subg. *Albocarmesinae* as defined by McAllister [12] are monophyletic when *S. tianschanica* is transferred to subg. *Albocarmesinae* and *S. hupehensis* var. *paucijuga* to subg. *Sorbus*. Nevertheless, except for sect. *Wilsonianae*, the seven sections in the genus *Sorbus s.s* as defined by McAllister [12] are not supported. To fully resolve relationships in *Sorbus s.s.*, more cp genomes need be sequenced and phylogenetic analysis with cp genome and nrDNA data combined with morphological comparisons are necessary.

Methods

Sampling, DNA extraction and sequencing

Leaf samples representing 15 *Sorbus s.s.* species and an unidentified sample were collected in the field between 2015 and 2018 from Anhui, Hubei, Sichuan, Xinjiang, Xizang and Yunnan Provinces in China. Fresh leaves were rapidly dried using silica gel for further DNA extraction. Voucher specimens were deposited in the Herbarium of Nanjing Forestry University (NF) and collection information is listed in Table 1.

Total DNA was extracted following the CTAB protocol [72]. DNA was quantified through fluorometry using Qubit Fluorometer or microplate reader and visualized in a 1% agarose-gel electrophoresis for quality check. The extracted genomic DNA was subjected to random degradation by Covaris, and then fragments with a size of 270 bp were selected by using AxyPrep Mag PCR clean up Kit. The selected fragments were amplified after suffering

Tang *et al. BMC Plant Biology* (2022) 22:495

Page 12 of 16

Table 1 Basic information of 16 *Sorbus s.s.* samples

Taxa	Voucher	Clean bases (G)	Clean reads	Average coverage (×)	Genbank accession number
S. aestivalis	CHINA. Sichuan, 29°39′42.46″N, 102°57′03.20″E, 2382 m, Xin Chen and Zhongren Xiong 0750 (NF)	2.35	7,200,013	46.6	ON049656
S. albopilosa	CHINA. Xizang, 29°49′22.06″N, 94°44′21.38″E, 3148 m, Xin Chen, Xiaochen Zhang and Zhongren Xiong 1030 (NF)	2.33	7,150,804	27.95	ON049662
S. amabilis	CHINA. Anhui, 30°08′09.60″N, 118°09′31.68″E, 1771 m, Xin Chen and Jiabao Li 1423 (NF)	4.52	13,851,710	62.15	ON049665
S. cibagouensis	CHINA. Xizang, 29°49′16.49″N, 95°42′40.63″E, 3272 m, Xin Chen, Xiaochen Zhang and Zhongren Xiong 0964 (NF)	2.90	8,895,466	61.75	ON049660
S. filipes	CHINA. Xizang, 29°48′23.40″N, 95°41′53.27″E; 3598 m, Xin Chen, Xiaochen Zhang and Zhongren Xiong 0988 (NF)	4.32	13,335,748	72.05	ON049661
S. foliolosa	CHINA. Xizang, 29°29′24.94″N, 94°55′14.81″E, 3650 m, Xin Chen 0620 (NF)	3.53	10,877,503	87.1	ON049652
S. helenae	CHINA. Sichuan, 29°31′26.41″N,103°20′12.88″E, 3014 m, Xin Chen, Xiaoyan Wang and Chunhui Wang 1937 (NF)	2.60	8,333,965	205.2	ON049667
S. hypoglauca	CHINA. Yunnan, 25°41′14.47″N, 100°06′32.61″E, 3211 m, Xin Chen, Xiaochen Zhang and Qi qi 1237 (NF)	2.36	7,244,523	12.8	ON049664
S. pteridophylla	CHINA. Xizang, 29°43′32.18″N, 95°38′41.52″E, 3224 m, Xin Chen 0572 (NF)	2.32	7,128,300	45.45	ON049651
S. rehderiana	CHINA. Sichuan, 30°01′21.08″N, 100°49′38.46″E, 4305 m, Xin Chen 0556 (NF)	2.34	7,192,370	32.95	ON049650
S. rutilans	CHINA. Sichuan, 29°59′20.01″N, 102°01′96.00″E, 3060 m, Xin Chen and Zhongren Xiong 0739 (NF)	2.90	8,905,432	155.35	ON049654
S. sargentiana	CHINA. Sichuan, 29°39′43.21″N, 102°57′03.97″E, 2387 m, Xin Chen and Zhongren Xiong 0749 (NF)	4.92	15,204,888	28.85	ON049655
S. tianschanica	CHINA. Xinjiang, 43°52′52.76″N, 88°07′44.44″E, 1927 m, Wenhao Fan 1761 (NF)	2.58	8,291,047	238.9	ON049666
S. ursina	CHINA. Sichuan, 30°53′35.86″N, 102°58′37.91″E, 3411 m, Xin Chen and Zhongren Xiong 0664 (NF)	3.22	9,861,892	97.2	ON049653
S. wilsoniana	CHINA. Hubei, 30°01′34.84″N, 109°45′42.59″E, 1856 m, Xin Chen and Yun Chen 0784 (NF)	3.45	10,636,280	58.75	ON049657
Sorbus sp.	CHINA. Xizang, 29°31′27.88″N, 96°46′43.05″E, 4146 m, Xin Chen, Xiaochen Zhang and Zhongren Xiong 0914 (NF)	2.91	8,920,370	110.75	ON049659

from end repair, the addition of polyA tail and adaptor ligation. The processed fragments were heat denatured to single strand after purification. The single strands were circularized, and single strand circle DNA was obtained as the final library. The final library was sequenced by Illumina HiSeq 4000 platform at BGI (Shenzhen, China) to generate raw deta. The generated raw sequencing data was filtered using the program SOAPnuke [73] with default parameters to remove adapters, low-quality reads with quality value ≤ 20, to finally obtain clean reads. The quantity and quality of clean reads sequenced for each *Sorbus s.s.* sample was analyzed with FastQC v.0.11.9 [74], and the details were provided in Table 1.

Genome assembly and annotation
The high-quality reads were used for de novo assembly to reconstruct *Sorbus s.s.* chloroplast genomes using GetOrganelle v.1.7.5.3 [75] with the reference cp genome sequence of *Torminalis glaberrima* (NC033975) with wordsize of 103 and K-mer sizes of 105 or 127 (the K-mer sizes were 105 in *S. helenae and S. tianschanica*, and were 127 in other 14 samples) and the coverage depth of the assembled genomes are provided in Table 1. Bandage software [76] was used to map all reads to the assembled cp genome sequence for visualization processing and obtaining accurate cp genomes. Complete cp genomes were annotated using the PGA program [35] with *S. insignis* as a reference, then, manually verified and corrected by comparison with five sequences in the same tribe Maleae, *Aria edulis* M. Roem. (NC045418), *Malus hupehensis*, *Micromeles thibetica* (MK920287), *Pyrus pashia* (NC034909) and *Torminalis glaberrima* using Geneious v.9.0.2 [36]. The cp genome maps were created using Organellar Genome DRAW (https://chlorobox.

Tang *et al. BMC Plant Biology* (2022) 22:495

Page 13 of 16

mpimp-golm.mpg.de/OGDraw.html). The complete cp genome sequences and gene annotation of the 16 newly assembled *Sorbus s.s.* samples were submitted to NCBI database (https://www.ncbi.nlm.nih.gov) under the accession numbers listed in Table 1. Meanwhile, all the 22 cp genomes in *Sorbus s.l.* (13 in *Sorbus s.s.*) reported previously were re-annotated.

Genome structure and codon usage analyses

The structure, size, gene content and GC content of cp genomes were identified using Geneious v.9.0.2. LSC, SSC, IRa and IRb region were plotted with boundary positions being compared using IRscope online software (https://irscope.shinyapps.io/irapp/) [77]. All CDSs were extracted using Geneious v.9.0.2. The amount of codon and RSCU ratio was calculated using CodonW v.1.4.2 software (http://codonw.sourceforge.net/) with default parameters.

Repeats analyses

SSRs were identified using the MISA online software (https://webblast.ipk-gatersleben.de/misa/) with the minimum repeat parameters set as 12, 6, 4, 3, 3, 3 repeat units for mono-, di-, tri- tetra-, penta-, and hexanucleotide SSRs, respectively. Online REPuter software (https://bibiserv.cebitec.uni-bielefeld.de/reputer) was used to identify and locate forward, palindromic, reverse and complement sequences with minimum repeat size of 20 bp, maximum repeat sequences number of 200 and the E-value below 1e-5.

Comparative analyses of chloroplast genomes

To identify variable regions and intra-generic variations within *Sorbus s.s.*, the alignment was visualized using online mVISTA (https://genome.lbl.gov/vista/index.shtml) in Shuffle-LAGAN mode, with the annotated cp genome of *S. insignis* as a reference. The 16 *Sorbus s.s.* cp genomes sequences were aligned in MAFFT [78]. The alignment was used to calculate the Pi value within *Sorbus s.s.* cp genomes. The sliding window analysis was performed in DnaSP v.5 [79] with step size of 200 bp and window length of 800 bp.

Phylogenetic analysis

The complete cp genome sequences of the 16 newly sequenced *Sorbus s.s.* with the other 48 cp genomes of tribe Maleae, one cp genomes of Amygdaleae and the outgroup (*Barbeya oleoides*) (Table S7) were aligned with the program MAFFT and any alignment issues were manually modified in Geneious v.9.0.2. Phylogenetic analyses were performed using both maximum likelihood (ML) and Bayesian inference (BI) methods based on the 63

complete cp genomes. ML analyses were implemented in RAxML v.8.0.0 [80] with GTR+GAMMA model. The best likelihood tree was obtained from 100 starting trees using rapid bootstrap analyses with 1000 bootstrap replicates. Multiparametric bootstrapping analyses with 1000 replicates was conducted to obtain the bootstrap for each node. BI analyses were conducted using MrBayes v.3.2.2 [81]. The best-fit nucleotide substitution model for BI analysis were inferred from Modeltest v.3.7 [82] and PAUP v.4.0 [83]. The Markov chain Monte Carlo (MCMC) analysis was run for 6,000,000 generations, and the trees were sampled every 1,000 generations with the initial 25% discarded as a burn-in fraction. Meanwhile, six hypervariable regions (*trnR-atpA*, *petN-psbM*, *rpl32-trnL*, *trnT-trnL*, *trnH-psbA* and *ndhC-trnV*) and the combination were uesd to reconstruct ML tree with 1000 bootstrap replicates, and compared with the tree based on complete cp genomes for the DNA barcode analysis. The resulting trees by ML and BI methods were rooted with *Barbeya oleoides* and visualized with FigTree v.1.4.3 [84].

Abbreviations

AAs: Amino acids; BI: Bayesian inference; CNS: Non-coding sequence; Cp: Chloroplast; IRs: Inverted repeats; LSC: Large single-copy region; MCMC: Markov chain Monte Carlo; ML: Maximum likelihood; nrDNA: Nuclear ribosome DNA; Pi: Nucleotide diversity; RSCU: Relative synonymous codon usage values; SSC: Small single-copy region; SSRs: Simple sequence repeats.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-022-03858-5.

Tang *et al. BMC Plant Biology*   (2022) 22:495

Page 14 of 16

Bootstrap value at nodes. **Fig. S6** Phylogenetic tree base on *ndhC-trnV* region resulting from ML analysis with Bootstrap value at nodes. **Fig. S7** phylogenetic tree base on 6 regions (*ndhC-trnV* + *petN-psbM*+ *rpl32-trnL* + *trnH-psbA* + *trnT-atpA*+ *trnT-trnL*) resulting from ML analysis with Bootstrap value at nodes.

Availability of data and materials

All 16 newly sequenced sequences in this study are available from the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm. nih.gov; accession numbers are ON049650–ON049657, ON049659–ON049662 and ON049664–ON049667; see Table 1). Information for other cp genomes used for phylogenetic analysis download from NCBI (https://www.ncbi.nlm. nih.gov) can be found in Additional Table 7: Table S7. Voucher specimens are deposited in the Herbarium of Nanjing Forestry University (NF) and collection information was listed in Table 1.

Declarations

Ethics approval and consent to participate

No specific permissions were required for the collection of plant material in this study. The field work and molecular experiments were carried out in compliance with the relevant laws of China. All specimens were identifed by Xin Chen.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Co-Innovation Center for Sustainable Forestry in Southern China, College of Biology and the Environment, Nanjing Forestry University, Nanjing 210037, Jiangsu, China. [2]Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China.

References

1. Linnaeus C. Species Plantarum. 1753. p. 1–477. https://www.biodiversi tylibrary.org/page/358496.
2. Hedlund T. Monographie der Gattung *Sorbus*. Kongliga Svenska Vetenskaps Akademiens Handlingar. 1901;35:1–147.
3. Yü TT, Lu LT. *Spiraea*, *Dichotomanthes*, *Cotoneaster*, *Sorbus*, *Chaenomeles*. In: Yü TT, editor. Flora Reipublicae Popularis Sinicae, vol. 36. Beijing: Science Press; 1974. p. 1–443. http://www.iplant.cn/info/Sorbus?t=z. (In Chinese)
4. Gabrielian E. The genus *Sorbus* L. in Western Asia and the Himalayas. Yerevan: Armenian Acadenmy of Sciences; 1978. p. 1–264.
5. Phipps JB, Robertson KR, Smith PG, Rohrer JR. A checklist of the subfamily Maloideae (Rosaceae). Can J Bot. 1990;68:2209–69. https://doi.org/10. 1139/b90-288.
6. Lu LT, Spongberg SA. *Sorbus* L. In: Wu ZY, Raven PH, Hong DY, editors. Flora of China. vol. 9. Beijing: Science Press; St. Louis: Missouri Botanical Garden Press; 2003. p. 144–170. http://foc.eflora.cn/content.aspx?Taxon Id=130718.
7. Roemer MJ. Familiarum naturalium regni vegetabilis synopses monographicae III. Rosiflorae. Amygdalacearum et Pomacearum. Weimar: Landes-Industrie-Comptoir; 1847.
8. Decaisne J. Mémoirs sur le famille des Pomacées. Nouv Arch Mus Hist Nat. 1874;10:113–92.
9. Die KE, der Pomaceen G. Wissenschaftliche beilage zum programm des falk realgymnasiums zu Berlin. Berlin: Verlagsbuchhandlung Hermann Heyfelder; 1890.
10. Koehne E. Die Gattungen der Pomaceen. Garten flora. 1891;40:4–7, 35–38, 59–61.
11. Robertson KR, Phipps JB, Rohrer JR, Smith PG. A synopsis of genera in Maloideae (Rosaceae). Syst Bot. 1991;16(2):376–94. https://doi.org/10. 2307/2419287.
12. McAllister H. The genus *Sorbus* Mountain Ash and other Rowans. London: Royal Botanical Gardens; 2005. p. 1–252.
13. Phipps JB, Robertson KR, Rohrer JR, Smith PG. Origins and evolution of subfam. Maloideae (Rosaceae). Syst Bot. 1991;16(2):303–32.
14. Campbell CS, Donoghue MJ, Baldwin BG, Wojciechowski MF. Phylogenetic relationships in Maloideae (Rosaceae): evidence from sequences of the internal transcribed spacers of nuclear ribosomal DNA and its congruence with morphology. Amer J Bot. 1995;82(7):903–18. https://doi. org/10.2307/2445977.
15. Campbell CS, Evans RC, Morgan DR, Dickinson TA, Arsenault MP. Phylogeny of subtribe Pyrinae (formerly the Maloideae, Rosaceae): limited resolution of a complex evolutionary history. Pl Syst Evol. 2007;266(1–2):119–45. https://doi.org/10.1007/s00606-007-0545-y.
16. Potter D, Eriksson T, Evans RC, Oh S, Smedmark JEE, Morgan DR, et al. Phylogeny and classification of Rosaceae. Pl Syst Evol. 2007;266(1–2):5–43. https://doi.org/10.1007/s00606-007-0539-9.
17. Lo EYY, Donoghue MJ. Expanded phylogenetic and dating analyses of the apples and their relatives (Pyreae, Rosaceae). Molec Phylogen Evol. 2012;63(2):230–43. https://doi.org/10.1016/j.ympev.2011.10.005.
18. Sun JH, Shi S, Li JL, Yu J, Wang L, Yang XY, et al. Phylogeny of Maleae (Rosaceae) based on multiple chloroplast regions: implications to genera circumscription. BioMed Res Int. 2018;2018:1–10. https://doi.org/10.1155/ 2018/7627191.
19. Ulaszewski B, Jankowska-Wróblewska S, Swiło K, Burczyk J. Phylogeny of Maleae (Rosaceae) based on complete chloroplast genomes supports the distinction of *Aria, Chamaemespilus* and *Torminalis* as separate genera, different from *Sorbus* sp. Plants. 2021;10(11):2534. https://doi.org/10. 3390/plants10112534.
20. Koehne E. Plantae Wilsonianae: an enumeration of the woody plants collected in western China for the Arnold Arboretum of Harvard University during the years 1907, 1908, and 1910. Cambridge: Cambridge University Press; 1913. p. 1–661. http://www.biodiversitylibrary.org/bibliography/191.
21. Yü TT, Kuan KJ. Taxa nova Rosacearum Sinicarum (I). Acta Phytotax Sin. 1963;8(3):202–234. https://www.plantsystematics.com/CN/abstract/abstr act1327.shtml. (In Chinese)
22. Li M, Tetsuo OT, Gao YD, Xu B, Zhu ZM, Ju WB, et al. Molecular phylogenetics and historical biogeography of *Sorbus sensu stricto* (Rosaceae). Molec Phylogen Evol. 2017;111:76–86. https://doi.org/10.1016/j.ympev. 2017.03.018.
23. Wang GX, Zhang ML. A molecular phylogeny of *Sorbus* (Rosaceae) based on ITS sequence. Acta Hort Sin. 2011;38(12):2387–94. https://doi.org/10. 16420/j.issn.0513-353x.2011.12.001. (In Chinese)
24. Li QY, Guo W, Liao WB, Macklin JA, Li JH. Generic limits of Pyrinae: insights from nuclear ribosomal DNA sequences. Bot Stud. 2012;53:151–64.

Tang *et al. BMC Plant Biology*   (2022) 22:495

Page 15 of 16

25. Smith DR, Keeling PJ. Mitochondrial and plastid genome architecture: reoccurring themes, but significant differences at the extremes. PNAS. 2015;112:10177–84. https://doi.org/10.1073/pnas.1422049112.

26. Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol. 2011;76:273–97. https://doi.org/10.1007/s11103-011-9762-4.

27. Campbell CS, Dickinson TA. Apomixis, patterns of morphological variation, and species concept in subfam. Maloideae (Rosaceae). Syst Bot. 1990;15(1):124–35.

28. Ludwig S, Robertson A, Rich TCG, Djordjević M, Cerović R, Houston L, et al. Breeding systems, hybridization and continuing evolution in avon gorge *Sorbus*. Ann Bot. 2013;111(4):563–75. https://doi.org/10.1093/aob/mct013.

29. Robertson A, Rich TCG, Allen AM, Houston L, Roberts C, Bridle JR, et al. Hybridization and polyploidy as drivers of continuing evolution and speciation in Sorbus. Molec Ecol. 2010;19:1675–90. https://doi.org/10.1111/j.1365-294X.2010.04585.x.

30. Initiative TAG. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature. 2000;408(6814):796–815. https://doi.org/10.1038/35048692.

31. Xing SC, Liu CJH. Progress in chloroplast genome analysis. Prog Biochem Biophys. 2008;35(1):21–8.

32. Jeon JH, Kim SC. Comparative analysis of the complete chloroplast genome sequences of three closely related East-Asian wild roses (*Rosa* sect. *Synstylae*; Rosaceae). Genes. 2019;10(1):23.

33. Sun JH, Wang YH, Liu YL, Xu C, Yuan QJ, Guo LP, et al. Evolutionary and phylogenetic aspects of the chloroplast genome of *Chaenomeles* species. Sci Rep. 2020;10:11466. https://doi.org/10.1038/s41598-020-67943-1.

34. Cho MS, Kim JH, Yamada T, Maki M, Kim SC. Plastome characterization and comparative analyses of wild crabapples (*Malus baccata* and *M. toringo*): insights into infraspecific plastome variation and phylogenetic relationships. Tree Genet Genomes. 2021;17:41.

35. Yan JW, Li JH, Yu L, Bai WF, Nie DL, Xiong Y, Wu SZ. Comparative chloroplast genomes of *Prunus* subgenus *Cerasus* (Rosaceae): insights into sequence variations and phylogenetic relationships. Tree Genet Genomes. 2021;17:50. https://doi.org/10.1007/s11295-021-01533-8.

36. Qu X, Moore M, Li D, Yi T. PGA: a software package for rapid, accurateand flexible batch annotation of 13 plastomes. Plant Methods. 2019;15(1):1–12. https://doi.org/10.1186/s13007-019-0435-7.

37. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28(12):1647–9. https://doi.org/10.1093/bioinformatics/bts199.

38. Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol. 1985;2(1):13–34. https://doi.org/10.1093/oxfordjournals.molbev.a040335.

39. Bernardi G, Bernardi G. Compositional constraints and genome evolution. J Mol Evol. 1986;24:1–11. https://doi.org/10.1007/BF02099946.

40. Sablok G, Nayak KC, Vazquez F, Tatarinova TV. Synonymous codon usage, GC_3, and evolutionary patterns across plastomes of three pooid model species: emerging grass genome models for monocots. Mol Biotechnol. 2011;49(2):116–28. https://doi.org/10.1007/s12033-011-9383-9.

41. Lee SR, Kim K, Lee BY, Lim CE. Complete chloroplast genomes of all six *Hosta* species occurring in Korea: molecular structures, comparative, and phylogenetic analyses. BMC Genomics. 2019;20:833. https://doi.org/10.1186/s12864-019-6215-y.

42. Ren T, Li ZX, Xie DF, Gui LJ, Peng C, Wen J, et al. Plastomes of eight *Ligusticum* species: characterization, genome evolution, and phylogenetic relationships. BMC Plant Biol. 2020;20:519. https://doi.org/10.1186/s12870-020-02696-7.

43. Chi XF, Zhang FQ, Dong Q, Chen SL. Insights into comparative genomics, codon usage bias, and phylogenetic relationship of species from Biebersteiniaceae and Nitrariaceae based on complete chloroplast genomes. Plants. 2020;9:1605. https://doi.org/10.3390/plants9111605.

44. Zhu AD, Guo WH, Gupta S, Fan WS, Mower JP. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. New Phytol. 2016;209:1747–56. https://doi.org/10.1111/nph.13743.

45. Yamamoto T. DNA markers and molecular breeding in pear and other Rosaceae fruit trees. Hort J. 2021;90(1):1–13. https://doi.org/10.2503/hortj.UTD-R014.

46. Eken BU, Kirdok E, Velioglu E, Ciftci YO. Assessment of genetic variation of natural populations of wild cherry (*Prunus avium* L.) via SSR markers. Turk J of Bot. 2022;46(1):14–25.

47. Khan G, Zhang FQ, Gao QB, Fu PC, Zhang Y, Chen SL. Spiroides shrubs on Qinghai-Tibetan Plateau: multilocus phylogeography and palaeodistributional reconstruction of *Spiraea alpina* and *S. Mongolica* (Rosaceae). Mol Phylogenet Evol. 2018;123:137–48.

48. Olmstead RG, Palmer JD. Chloroplast DNA systematics: a review of methods and data analysis. Amer J Bot. 1994;81(9):1205–24. https://doi.org/10.2307/2445483.

49. Wang Q, Niu Z, Li JB, Zhu KL, Chen X. The complete chloroplast genome sequence of the Chinese endemic species *Sorbus setschwanensis* (Rosaceae) and its phylogenetic analysis. Nordic J Bot. 2020;38(2): e02532. https://doi.org/10.1111/njb.02532.

50. Tang CQ, Qiu ZX, Tan C, Qian YM, Chen X. *Sorbus koehneana* (Rosaceae): its complete chloroplast genome and phylogenetic relationship with *S. unguiculata*. Acta Hort Sin. 2022;49(3):641–54. https://doi.org/10.16420/j.issn.0513-353x.2021-0040. (In Chinese)

51. Raspé O, Saumitou-Laprade P, Cuguen J, Jacquemart AL. Chloroplast DNA haplotype variation and population differentiation in *Sorbus aucuparia* L. (Rosaceae: Maloideae). Molec Ecol. 2000;9(8):1113–22.

52. Chester M, Cowan RS, Fay MF, Rich TCG. Parentage of endemic *Sorbus* L. (Rosaceae) species in the British Isles: evidence from plastid DNA. Bot J Linn Soc. 2007;154(3):291–304.

53. Borsch T, Quandt D. Mutational dynamics and phylogenetic utility of noncoding chloroplast DNA. Plant Syst Evol. 2009;282:169–99. https://doi.org/10.1007/s00606-009-0210-8.

54. Nie XJ, Lv SZ, Zhang YX, Du XH, Wang L, Biradar SS, et al. Complete chloroplast genome sequence of a major invasive species, crofton weed (*Ageratina adenophora*). PLoS ONE. 2012;7(5): e36869. https://doi.org/10.1371/journal.pone.0036869.

55. Shaw J, Lickey EB, Schilling EE, Small RL. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. Amer J Bot. 2007;94(3):275–88. https://doi.org/10.3732/ajb.94.3.275.

56. Huang H, Shi C, Liu Y, Mao SY, Gao LZ. Thirteen *Camellia* chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. BMC Evol Biol. 2014;14:151. https://doi.org/10.1186/1471-2148-14-151.

57. Barrett CF, Baker WJ, Comer JR, Conran JG, Lahmeyer SC, Leebens-Mack JH, et al. Plastid genomes reveal support for deep phylogenetic relationships and extensive rate variation among palms and other commelinid monocots. New Phytol. 2015;209:855–70. https://doi.org/10.1111/nph.13617.

58. Kartonegoro A, Veranso-Libalah MC, Kadereit G, Frenger A, Penneys DS, Mota de Oliveira S, et al. Molecular phylogenetics of the *Dissochaeta alliance* (Melastomataceae): Redefining tribe Dissochaeteae. Taxon. 2021;70(4):793–825. doi:https://doi.org/10.1002/tax.12508.

59. Mapaya RJ, Cron GV. A phylogeny of *Emilia* (Senecioneae, Asteraceae) – implications for generic and sectional circumscription. Taxon. 2020;70(1):127–38. https://doi.org/10.1002/tax.12417.

60. Kress WJ. Plant DNA barcodes: applications today and in the future. J Syst Evol. 2017;55(4):291–307. https://doi.org/10.1111/jse.12254.

61. Le DT, Zhang YQ, Xu Y, Guo LX, Ruan ZP, Burgess KS, Ge XJ. The utility of DNA barcodes to confirm the identification of palm collections in botanical gardens. PLoS ONE. 2020;15(7): e0235569. https://doi.org/10.1371/journal.pone.0235569.

62. Korotkova N, Nauheimer L, Ter-Voskanyan H, Allgaier M, Borsch T. Variability among the most rapidly evolving plastid genomic regions is lineage-Specific: implications of pairwise genome comparisons in *Pyrus* (Rosaceae) and other angiosperms for marker choice. PLoS ONE. 2014;9(11): e112998. https://doi.org/10.1371/journal.pone.0112998.

63. Rokas A, Holland PWH. Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol. 2000;15(11):454–9. https://doi.org/10.1016/s0169-5347(00)01967-4.

64. Li X, Yang Y, Henry RJ, Rossetto M, Wang Y, Chen S. Plant DNA barcoding: from gene to genome. Biol Rev. 2014;90(1):157–66. https://doi.org/10.1111/brv.12104.

Tang *et al. BMC Plant Biology*   (2022) 22:495

Page 16 of 16

65. Daniell H, Lin CS, Yu M, Chang WJ. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol. 2016;17:134. https://doi.org/10.1186/s13059-016-1004-2.

66. Aldasoro JJ, Aedo C, Navarro C, Garmendia FM. The genus *Sorbus* (Maloideae, Rosaceae) in Europe and in North Africa: morphological analysis and systematics. Syst Bot. 1998;23(2):189–212. https://doi.org/10.2307/2419588.

67. Robertson KR, Phipps JB, Rohrer JR. Summary of Leaves in the Genera of Maloideae (Rosaceae). AAnn Missouri Bot Gard. 1992;79(1):81–94. http://www.jstor.org/stable/2399811.

68. Mezhenska LO, Mezhenskyj VM, Yakubenko BY. NULESU Collections of fruit and ornamental plants. Lira-K, Kiev, **К**ОЛЕ**К**ЦІЯ НУБІ**П** У**К**РАЇ**Н**И **П**ЛО**Д**ОВИ**Х** І **ДЕК**ОРАТИ**В**НИ**Х** РОСЛИ**Н**; 2018. p. 1–107.

69. Xi LL, Li JB, Zhu KL, Qi Q, Chen X. Variation in genome size and stomatal traits among three *Sorbus* species. Pl Sci J. 2020;38(1):32–8. https://doi.org/10.11913/PSJ.2095-0837.2020.10032. (In Chinese)

70. Li JB, Zhu KL, Wang Q, Chen X. Genome size variation and karyotype diversity in eight taxa of *Sorbus sensu stricto* (Rosaceae) from China. Cytogenet Genome Res. 2021;15(2):137–48. https://doi.org/10.3897/compcytogen.v15.i2.58278.

71. Chang CS, Gil HY. *Sorbus ulleungensis*, a new endemic species on Ulleung Island. Korea Harvard Pap Bot. 2014;19(2):247–55. https://doi.org/10.3100/hpib.v19iss2.2014.n11.

72. Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull. 1987;19:11–5. https://doi.org/10.12691/ajmr-3-1-7.

73. Chen YX, Chen YS, Shi CM, Huang ZB, Zhang Y, Li S, et al. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. GigaScience. 2018;7(1):1–6. https://doi.org/10.1093/gigascience/gix120.

74. Simon A, FastQC. A quality control tool for high throughput sequence data. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

75. Jin JJ, Yu WB, Yang JB, Song Y, dePamphilis CW, Yi TS, et al. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. 2020;21(1):241. https://doi.org/10.1186/s13059-020-02154-5.

76. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics. 2015;31(20):3350–2. https://doi.org/10.1093/bioinformatics/btv383.

77. Amiryousefi A, Hyvönen J, Poczai P. IRscope: an online program to visualize the junction sites of chloroplast genomes. Bioinformatics. 2018;34(17):3030–1. https://doi.org/10.1093/bioinformatics/bty220.

78. Katoh K, Standley DM. MAFFT Multiple sequence alignment software version 7: improvements in performance and usability. Molec Biol Evol. 2013;30(4):772–80. https://doi.org/10.1093/molbev/mst010.

79. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009;25(11):1451–2. https://doi.org/10.1093/bioinformatics/btp187.

80. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3. https://doi.org/10.1093/bioinformatics/btu033.

81. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012;61(3):539–542. doi:https://doi.org/10.1093/sysbio/sys029.

82. Posada D, Crandall KA. MODELTEST: testing the model of DNA substitution. Bioinformatics. 1998;14(9):817–8. https://doi.org/10.1093/bioinformatics/14.9.817.

83. Matthews LJ, Rosenberger AL. Taxon combinations, parsimony analysis (PAUP*) and the taxonomy of the yellow-tailed woolly monkey. Lagothrix flavicauda Am J Phys Anthropol. 2008;137:245–55. https://doi.org/10.1002/ajpa.20859.

84. Rambaut A. FigTree, a Graphical Viewer of Phylogenetic Trees. Edinburgh: Institute of Evolutionary Biology University of Edinburgh; 2007. http://tree.bio.ed.ac.uk/software/figtree.

Publisher's Note