Running title: Phylogenomics and biogeography of the apple genus (*Malus*)

Phylogenomic analyses in the apple genus *Malus* s.l. reveal widespread hybridization and allopolyploidy driving the diversifications, with insights into the complex biogeographic history in the Northern Hemisphere

Bin-Bin Liu[1,2*], Chen Ren[3,4], Myounghai Kwak[5], Richard G.J. Hodel[2], Chao Xu[1], Jian He[6], Wen-Bin Zhou[7], Chien-Hsun Huang[8], Hong Ma[9], Guan-Ze Qian[10], De-Yuan Hong[1], Jun Wen[2*]

Correspondance: Bin-Bin Liu (liubinbin@ibcas.ac.cn) or Jun Wen (wenj@si.edu)

[1]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China

[2]Department of Botany, National Museum of Natural History, Smithsonian Institution, PO Box 37012, Washington, DC 20013-7012, USA

Full list of author information is available at the end of the article

Abstract: Phylogenomic evidence from an increasing number of studies has demonstrated that different data sets and analytical approaches often reconstruct strongly supported but conflicting relationships. In this study, hundreds of single-copy nuclear (SCN) genes (785) and complete plastomes (75) were used to infer the phylogenetic relationships and estimate the historical biogeography of the apple genus *Malus* sensu lato, an economically important lineage disjunctly distributed in the Northern Hemisphere involved in known and suspected hybridization and allopolyploidy events. The nuclear phylogeny recovered the monophyly of *Malus* s.l. (including *Docynia*); however, it was supported to be biphyletic in the plastid phylogeny. An ancient chloroplast capture event best explains the cytonuclear discordance that occurred in the Eocene in western North America. Our conflict analysis demonstrated that ILS, hybridization, and allopolyploidy could explain the widespread nuclear gene tree discordance. We detected one deep hybridization event (*Malus doumeri*) involving the ancestor of pome-bearing species and *Docynia delavayi*, and one recent hybridization event (*Malus coronaria*) between *M. sieversii* and a combined clade of *M. ioensis* and *M. angustifolia*. Furthermore, our historical biogeographic analysis combining living and fossil species supported a widespread East Asian-western North American origin of *Malus* s.l., followed by a series of extinction events in the Eocene in northern East Aisa and western North America. This study provides a valuable evolutionary framework for the breeding and crop improvement of apples and their close relatives.

Keywords: deep genome skimming; historical biogeography; genomic discordance; massive extinction; reticulate evolution; single-copy nuclear genes.

Introduction

The apple genus *Malus* Mill. (tribe Maleae, Rosaceae) is of great economic importance, with the domesticated apple (*M. domestica* (Suckow) Borkh.), various crabapples, as well as some important ornamentals (e.g., *M. halliana* Koehne, *M. hupehensis* (Pamp.) Rehder, and *M. micromalus* Makino)[1]. *Malus* comprises ca. 38-55 species disjunctly distributed across the temperate Northern Hemisphere[2,3], from East Asia (ca. 27 species)[2], Central Asia (ca. two species), Europe (three species)[4], and the Mediterranean (ca. two species)[4] to western (one species) and eastern North America (three species)[3]. Hybridization, polyploidy, and apomixis within the apple genus *Malus* have been reported to have occurred frequently in the wild and horticulture, and introgression has led to many taxa described in *Malus*[3,5]. These reticulated processes have significantly challenged the accurate phylogenetic inference of *Malus*; however, this genus also provides an ideal case for studying the phylogenomic discordance and its underlying potential causes. Based on morphology, many taxonomists have struggled to propose a reasonable classification system[5-15], and over-emphasis of one or a few morphological characters resulted in the controversial taxonomy at the sectional level, varying from two, five, six, to eight sections[15-17]. It has been challenging to elucidate the evolutionary relationships among *Malus* members based solely on morphological evidence. Robinson et al. (2001) first used molecular evidence (plastid *matK* and nuclear ribosomal internal transcribed spacer (nrITS) sequences) to explore the phylogenetic relationships of five sections sensu Rehder (1940). Subsequently, phylogenetic studies on *Malus* and its close relatives used more plastid and nrITS sequences[18], whole plastome and/or entire nuclear ribosomal DNA (nrDNA)[19-21], and transcriptomic dataset[22]. These studies provided significant insights into the major clades, but many questions remain due to the limited informative sites and/or taxon sampling (Fig. 1). The phylogenetic analysis in the framework of Maleae so far has shown strongly supported yet discordant nuclear and chloroplast topologies of *Malus* sect. *Chloromeles* (Decne.) Rehder and *M.* sect. *Eriolobus* lineage, which makes *Malus* s.l. either monophyletic (nuclear sequences: Fig. 1f-i) or biphyletic (whole plastome: Fig. 1e), suggesting possible hybridization and/or chloroplast capture events

in the origin of this clade[21]. However, Lo & Donoghue (2012) produced a monophyletic *Malus* s.l. based on 11 plastid regions as Robinson et al. (2001) did using *matK* regions (Fig. 1a, b). Generally, phylogenetic relationships within *Malus* s.l. are not fully resolved, particularly along the backbone of the phylogeny.

Phylogenomics has revolutionized plant systematics and evolution as well as the associated fields in the last decades, enabling the utilization of a large number of nuclear genes for testing phylogenetic hypotheses, especially untangling recalcitrant relationships using traditional molecular systematic approaches[23-25]. While data collection efforts have significantly increased overall phylogenetic support, these analyses are a double-edged sword by demonstrating that different data sets and analytical approaches often reconstruct strongly supported but conflicting relationships[26-28]. Underlying these conflicting topologies are strongly discordant gene trees[29], which may be due to several biological processes such as gene duplication, incomplete lineage sorting (ILS), and gene flow (e.g., hybridization, allopolyploidy, and introgression). High levels of gene tree discordance have occurred in the nuclear genome and in the plastome, with the latter often regarded as a single locus[30-34]. The evidence so far suggests that gene flow between lineages promoted the diversification of land plants[34,35]. Additionally, conflicts between nuclear and organellar genomes, often called cytonuclear discordance, complicate phylogenetic inference. Generally, cytonuclear conflict has been interpreted as a result of introgressive chloroplast capture, which has been found in various lineages of angiosperms[21,36-38]. Despite widespread introgression detected in angiosperms, current methods for phylogenetic inference often either assume no gene tree discordance (i.e., concatenated supermatrices) or only consider a coalescent model in which all discordance is attributed to ILS[39,40]. Recently, through calculating unique, conflicting, and concordant bipartitions (e.g., *phyparts*)[29] or quartet-based evaluation (e.g., Quartet-Sampling)[41], the highly supported but sometimes conflicting relationships can be quantified. Additionally, phylogenetic network analysis has allowed estimating species trees to account for ILS and introgression[42-44], and this approach can explore the mechanisms of the discordance.

Malus s.l. has a wide disjunct distribution in the major refugia of the Northern Hemisphere: East

Asia, the Mediterranean, western North America, and eastern North America, as well as Europe. This distribution pattern provides an ideal case study for exploring the evolution of the major patterns of biogeographic disjunctions in the Northern Hemisphere. Such disjunctions are generally considered to be remnants of a more continuously distributed, mixed mesophytic forest during the Tertiary[45,46]. Due to the geologic and climatic oscillations, the once more widely distributed flora was fragmented and became relict in four major refugia: East Asia, eastern North America, western North America, and Southwest Europe. Based on phylogenetic analyses of 47 chloroplast genomes, Nikiforova et al. (2013) suggested a North American origin of *Malus*, which was in accordance with the fossil evidence, because many fossils were recorded in the middle to late Eocene from western North America[47-53]. Contrasting this New World origin hypothesis, Jin (2014) proposed an alternative hypothesis of East Asian origin based on biogeographic analyses using complete plastome data. Jin (2014) concluded that this conflict may be due to the lack of sampling of key early diverged lineages, such as *Malus doumeri* (Bois) A. Chev., *M. florentina* C.K.Schneid., and *M. trilobata* C.K.Schneid. in Nikiforova et al. (2013)'s study. However, Jin (2014)'s misidentified *M. doumeri* sample (actually *Pseudocydonia sinensis* (Thouin) C.K.Schneid.[21]) may have also biased the phylogenetic inference and historical biogeographic estimation. These two conflicting hypotheses showcase the need for a robust phylogenetic framework with a comprehensive taxon sampling scheme to reconstruct the biogeographic history of *Malus*.

In this study, we intend to explore gene tree concordance and the phylogenetic relationships among major clades to test the potential ILS and hybridization events in the evolutionary history of *Malus* s.l. We further conduct biogeographic analyses to infer geographic origins and timing of possible hybrid clades. Due to their significant economic importance, genomes of apples and their close relatives have been sequenced, such as *Malus baccata* (L.) Borkh.[54], *M.* × *domestica*[55-59], *M. sieversii* (Ledeb.) M.Roem.[59], and *M. sylvestris* Mill.[59], which provided substantial genome resources for exploring single-copy nuclear (SCN) genes for phylogenetic analysis. Additionally, numerous genome resequencing, transcriptomes, and raw genome skimming data are available from the NCBI sequence read archive (SRA: https://www.ncbi.nlm.nih.gov/sra). These raw genomic data, coupled with deep genome

skimming data sensu Liu et al. (2021) generated for this study, provide an excellent opportunity to explore a robust phylogenetic framework within *Malus* s.l. using a large-scale phylogenomic analysis.

This study employs extensive genomic data, sampling from 77 individuals by integrating genome resequencing[57], transcriptomic[22], and deep genome skimming data[60] to assemble 797 SCN genes and whole plastomes. Specifically, we aim to: (1) reconstruct a robust phylogenetic backbone of the apple genus *Malus* s.l.; (2) explore gene tree conflicts and evaluate the potential causes; and (3) investigate broad-scale biogeographic relationships and ancestral range evolution.

Results

Single-copy nuclear genes and plastome assembly

Raw reads for the 27 newly sequenced deep genome skimming data are available from the NCBI Sequence Read Archive (SRA: BioProject: PRJNA759205 with the accession for each sample listed in Supplementary Table S1). The number of clean reads ranged from 33 (*Cotoneaster salicifolius* var. *henryanus* (C.K.Schneid.) T.T.Yu) to 103 (*Pourthiaea zhejiangensis* (P.L.Chiu) Iketani & H.Ohashi) million with the sequencing depths varying from 6.7× to 20.8×, assuming an estimated genome size of around 750 Mb based on *Malus domestica* genome[55].

We designed a set of 797 SCN genes from six genomes as mentioned below for this phylogenomic study on *Malus* and its close relatives. The number of genes recovered for each sample varied from 665 (83.4%: *Aronia melanocarpa* (Michx.) Elliott) to 797 (100%: 11 samples listed in Supplementary Table S2) (also referring to Fig. 2), and the number of genes after cleaning ranged from 568 (71.3%: *Malus doumeri*) to 785 (98.5%: *Malus baccata*) (Fig. 1 and Supplementary Table S2).

We successfully assembled 69 plastomes for this study, except *Aronia melanocarpa*, *Crataegus pinnatifida* Bunge, *Eriolobus trilobatus* (Labill. ex Poir.) M.Roem., *Pyrus pyrifolia* (Burm.f.) Nakai, and *Sorbus commixta* Hedl. due to the limited plastid reads in the raw data. All plastomes were submitted to GenBank with accessions listed in Supplementary Table S1, and the aligned plastid matrix was

deposited in Dryad Digital Repository (Data S4) https://doi.org/10.5061/dryad.2jm63xsq5.

Nuclear phylogenetic analysis and gene tree discordance

We obtained sequences of 785 genes with at least 47 samples and 273 bp for each gene (Data S1, available from Dryad Digital Repository https://doi.org/10.5061/dryad.2jm63xsq5). We kept 604 genes with more than 900 bp in aligned length for the downstream phylogenetic analysis. To test the effect of missing data for phylogenetic analysis, we generated three datasets including different numbers of samples for each clean gene, i.e., 50%-sample dataset (604 genes), 80%-sample dataset (589 genes), and all-sample dataset (66 genes), and all these three datasets have been deposited in Dryad Digital Repository (Data S2-S4) https://doi.org/10.5061/dryad.2jm63xsq5. These three concatenated matrices consisted of 1,193,313 bp, 1,042,020 bp, and 152,891 bp in aligned length. We estimated nine phylogenetic trees (Fig. 3 and Supplementary Figs. S1-S9), in which six of them resulted in the same topology. Henceforth, we used the RAxML tree (Fig. 3: hereafter, referred to as "nuclear phylogeny") estimated from the 80%-sample dataset for the following analysis. The nuclear phylogeny recovered *Malus* as paraphyletic with the genus *Docynia* Decne. embedded in *Malus*. *Malus* s.l. was delimited into three strongly supported major clades (BS = 100, 100, 100). Clade I included most of the species of *Malus* sect. *Malus* and *M.* sect. *Sorbomalus* except for a Mediterranean species, *M. florentina*. We sampled 27 individuals of clade I representing 11 species, disjunctly distributed between East Asia, Europe & Central Asia, and western North America. Clade II is composed of all the eastern North American (three species) and the Mediterranean species (two species). Clade III included two species, *M. doumeri*, previously delimited in *M.* sect. *Docyniopsis* C.K.Schneid. and *Docynia delavayi* C.K.Schneid.

Conflicted phylogenetic positions were detected in these nine phylogenetic trees (Fig. 3 and Supplementary Figs. S1-S9), in which the placements of *Malus coronaria* (L.) Mill. from eastern North America (Fig. 3: clade V) and the Asian *M. sikkimensis* (Wenz.) Koehne (Fig. 3: clade III) varied significantly among trees. All three species trees (Supplementary Figs. S3, S6, S9) estimated from

ASTRAL-III supported the sister relationship between *M. coronaria* and a large clade consisting of all species in clade I, II, III, and IV (defined in Fig. 3) (i.e., *M. coronaria* bipartition A: Table 1), while the remaining six ML trees (Supplementary Figs. S1, S2, S4, S5, S7, S8) estimated from IQ-TREE2 and RAxML recovered *M. coronaria* as sister to the clade composed of another two eastern North American species (*M. angustifolia* (Aiton) Michx. and *M. ioensis* (Alph. Wood) Britton) (i.e., *M. coronaria* bipartition B: Table 1). Likewise, the phylogenetic position of *M. sikkimensis* showed incongruence among trees (Supplementary Figs. S1-S9). One bipartition supported by seven out of nine trees was the sister relationship between *M. sikkimensis* and the combined clade, including *M. baccata* var. *xiaojinensis* (M.H.Cheng & N.G.Jiang) Ponomar., *M. orientalis* Uglitzk., *M. sylvestris*, *M. sieversii*, *M. hupehensis* (Pamp.) Rehder, *M. toringo* (Siebold) de Vriese, *M. baccata*, and *M. rockii* Rehder (i.e., *M. sikkimensis* bipartition A: Table 1), while the other bipartition recovered by only two trees was the sister relationship between *M. sikkimensis* and the clade composed of two species (*M. fusca* (Raf.) C.K.Schneid. and *M. kansuensis* (Batalin) C.K.Schneid.) (*M. sikkimensis* bipartition B: Table 1). An edge-based phylogenomic support test implemented in *phyckle*[28] on the 50%-sample dataset (600 genes all having *Gillenia stipulata* (Muhl. ex Willd.) Nutt. as outgroup) revealed that *M. coronaria* bipartition A was supported by more genes (445) than the *M. coronaria* bipartition B (155 genes). Meanwhile, the summed difference in log-likelihood scores supported the *M. coronaria* bipartition A (sum ΔlnL = 17028.36) over the bipartition B (sum ΔlnL = 4126.411). Although the extreme outlier genes were excluded (Table 1), both the number of genes (427:153 = 2.79 ≈ 3:1) and summed difference (14731.52:3876.897) supported the *M. coronaria* bipartition A. Similarly, two different bipartitions for the placements of *M. sikkimensis* were analyzed by *phyckle*. *Malus sikkimensis* bipartition B was supported by only 191 genes (sum ΔlnL = 3987.746) compared to the 409 genes of bipartition A (sum ΔlnL = 11675.01). Even with the outlier genes removed, *M. sikkimensis* bipartition B had more gene support (393, sum ΔlnL = 9168.462) than that of bipartition A (188, sum ΔlnL = 3647.953).

The conflict analysis from *phyparts* showed that significant gene tree discordance was detected among nuclear genes regarding the placement of three major clades, and minimal informative gene trees

supported each clade (Fig. 3 and Supplementary Fig. S11). In contrast, the QS result demonstrated that all these five nodes related to three major clades were confirmed with full support (1/-/1; i.e., all informative quartets support that lineage). Although Quartet Sampling (QS) confirmed the monophyly of *Malus* s.l. (node A) with full support, only 159 out of the 314 informative gene trees (50.6%) were concordant with this topology (ICA = 0.23). In contrast, only 56% of sampled informative quartets supported node B with only one alternative discordant topology (QS score = 0.56/0/1), but the *phyparts* result showed only 38 out of the 339 informative gene trees (11.2%; ICA = 0.05) supported this clade. Similarly, node E was supported by only 11% of informative quartets with a skewed frequency for alternative discordant topologies (QS score = 0.11/0.6/0.99), and the result of *phyparts* supported this clade with only 17 of the 387 informative gene trees (4.4%; ICA = 0.03). Node D was recovered in all nine trees (Supplementary Figs. S1-S9), while this was supported by only 12 of the 383 informative gene trees with counter-support from ICA (-0.05). Additionally, QS conflict analysis also showed the weak quartet support (0.32) with a skewed frequency for discordant topologies (0.59). In contrast, node E was recovered with full QS support (1/-/1) and 157 of the 282 informative gene trees (ICA = 0.29).

Coalescence simulation and phylogenetic network analysis

The nuclear gene coalescent simulation analysis showed the distinguished empirical and simulated gene to gene distance distribution (Fig. 4h), suggesting that ILS alone can not explain most observed gene tree heterogeneity[61]. Of the 27-taxa dataset at the tribe level, the plot of pseudo-loglikelihood scores (Fig. 4g) showed that the optimal number of hybridization events inferred in the SNaQ network analysis was one (Fig. 4g: -ploglik = 5218.4), suggesting the hybrid origin of *Malus doumeri* between *Docynia delavayi* ($\gamma = 0.794$) and the ancestor of pome related members (i.e., the formerly Maloideae; $\gamma = 0.206$). As h_{max} increased beyond one, the pseudo-loglikelihood increased slightly; therefore, we considered $h_{max} = 1$ to be optimal for this 27-taxa dataset.

The nuclear gene discordance analysis of the 14-taxa sampling dataset at the *Malus* level showed much overlap between the empirical and simulated distance distributions. This revealed that ILS alone

might explain the observed phylogenomic discordance. Meanwhile, the optimal hybridization event inferred from SNaQ network analysis was also one (Fig. 5c: -ploglik = 491.5), because the score of pseudo-loglikelihood levels off when h_{max} increased beyond one. *Malus coronaria* was 64.2% sister to a clade composed of *M. ioensis* and *M. angustifolia*, and 35.8% sister to one Central Asian species, *M. sieversii* (Fig. 5a).

We used the filtered HyDe results (i.e., $0 < \gamma < 1$) for detecting hybridization events. A total of 594 out of the 2448 hypotheses tested by HyDe showed significant evidence of a hybridization event (Supplementary Table S3), and nearly every species sampled in this study was involved in hybridization. The γ value for 350 of the 594 hypotheses was greater than 0.7 and less than 0.3, indicating ancient hybridization events, and only 244 γ values were close to 0.5 ($0.3 < \gamma < 0.7$), suggesting recent hybridization events. *Malus orientalis* has been involved in the most number of hybridization hypotheses (66), following with *M. coronaria* (54 ones) and *M. sikkimensis* (24 ones).

Due to the similar phylogenetic pattern between allopolyploidy and hybridization speciation, we summarized the chromosome count data for all species available from previous studies (Table 2, Fig. 3) for distinguishing the phylogenomic discordance from the two mechanisms. Generally, the chromosome distribution in *Malus* s.l. showed that allopolyploidy may have promoted the diversification of *Malus*. We did not find allopolyploidy in clade III. However, the various proportion of allopolyploidy cases was detected in clades I and II.

Plastid phylogenetic analysis

The final alignment from 80 plastid coding genes (CDSs) included 75 taxa and 80,799 bp in aligned length. All three phylogenetic trees from RAxML, IQ-TREE2, and ASTRAL-III recovered the same topology (Fig. 6 and Supplementary Figs. S12-S14). We presented the topology from RAxML herein and referred it to as the plastid phylogeny in the following context. Although the plastid result confirmed the three major clades in the nuclear phylogeny (Fig. 3), their relative phylogenetic position varied greatly (Fig. 6), and significant cytonuclear discordance showed between the plastid tree and the

nuclear phylogeny (Fig. 7). The monophyly of *Malus* s.l. did not recover in the plastid phylogeny, and the eastern North American and Mediterranean species (clade II) were supported to be sister to *Pourthiaea* Decne. (Fig. 6). Due to the limited informative sites for each plastid coding gene, the *phyparts* resulted in nearly completely grey pies for each focal node, i.e., no or very few genes supported this node (Fig. 6 and Supplementary Fig. S16), suggesting the limited utility of *phyparts* in shallow phylogenies and/or plastid genes. By contrast, the QS conflict analysis showed full support for the five focal nodes (1/-/1).

Dating and ancestral area reconstruction

The historical biogeographic analysis based on the SCN and plastid datasets using BEAST2 supported the East Asian origin of *Malus* s.l. The Northern Hemisphere disjunct distribution was through six (SCN data) or five (plastid data) dispersal events (Fig. 8). However, the phylogenetic dating analysis from SCN and the plastid dataset resulted in different age estimates. Generally, the overall age estimates in the nuclear data appeared to be older than those estimated from the plastid coding genes (Fig. 8 and Supplementary Figs. S17, S18). *Malus* s.l. originated from East Asia in the early Eocene, ca. 47.37 million years ago [Mya] in the SCN dating analysis, as compared to 42.16 Mya (95% highest posterior density (HPD) interval: 41.2-44.39 Mya: Supplementary Fig. S19) in the plastid analysis. The current disjunct distribution has been stabilized by the late Oligocene (26.42 Mya: 22.91-27.92 Mya: Fig. 8) based on the SCN result and by the early Miocene (15.14 Mya: 12.55-17.73 Mya: Supplementary Fig. S17) based on the plastid result. The eastern North American and Mediterranean clade (clade II: Fig. 3) was estimated to have originated from western North America in the middle Eocene [SCN: 43.58 Mya (41.2-44.39 Mya) and plastid coding genes: 41.2 Mya: 41.2-46.67 Mya], and then dispersed to the Mediterranean in the late Eocene [SCN: 39.61 Mya: 36.18-40.81 Mya; plastid coding genes: 35.77 Mya: 33.99-37.99 Mya) through North Atlantic Land Bridge (NALB). The SCN analysis showed that the eastern North American species originated from the extinct Western North American species in the late Eocene (34.97 Mya). In contrast, the plastid analysis resulted in the middle Miocene origin (14.18 Mya).

Furthermore, our ancestral area reconstruction analysis from the SCN data showed that some *Malus* members dispersed back to western North America in the early Miocene (20.98 Mya: Fig. 8), which may be due to the climate oscillations then. However, our plastid result did not present this back dispersal (Fig. 8). Similarly, the Europe & Central Asia clade originated from an East Asian ancestor in the late Oligocene (SCN: 26.42 Mya: 22.91-27.92 Mya) and the middle Miocene (plastid coding genes: 15.14 Mya: 12.55-17.73 Mya). Contrastingly, the western North American lineage (*Malus fusca*) was estimated to have originated from East Asia in the early Oligocene (31.46 Mya: 26.08-34.92 Mya: Fig. 8) and migrated through the BLB by the SCN data, while it originated from Europe & Central Asia in the late Miocene (8.6 Mya: 6.46-10.91 Mya: Supplementary Fig. S17) and likely migrated through the NALB based on the plastid coding gene data.

Discussion

This study integrated hundreds of SCN genes and plastomes to resolve the backbone of *Malus* s.l. The potential roles of ILS, hybridization, and allopolyploidy for the underlying phylogenetic discordance are herein evaluated. We also elucidate the biogeographic diversification patterns of the widespread disjunct distributions in the Northern Hemisphere. Our results support the paraphyly of the apple genus *Malus*, with *Docynia* nested within it, and recovered three strongly supported major clades within *Malus* s.l. except for the unstable phylogenetic placements of *M. coronaria* and *M. sikkimensis* in different trees. Furthermore, our coalescent simulation analysis demonstrated that ILS is not the sole or dominant cause of phylogenomic conflict, and other processes (e.g., hybridization and allopolyploidy) may have driven the discordance. The phylogenetic network analysis at the *Malus* and the Maleae levels supported the hybrid origin of *M. coronaria* and *M. doumeri*, and the further gene tree discordance analysis (e.g., *phyparts*, QS, *phyckle*, and HyDe) promoted the understanding of underlying conflict at some nodes. Below we will integrate several lines of evidence to discuss the potential causes of conflicts.

The phylogenetic backbone and an early chloroplast capture event in *Malus* s.l.

The taxonomic circumscription of *Malus* s.l. has been controversial due to the phylogenetic position of *Docynia*. With the highly distinctive numbers of ovules per locule (3-10 in *Docynia* vs. 2 in *Malus*), *Docynia* has been recognized as a separate genus by taxonomists historically[1,5,6,10,74-77]. The nuclear and plastid phylogeny from hundreds of SCN genes in our study supported the paraphyly of *Malus* s.s., with *Docynia* nested within it (Figs. 3, 4), and this was consistent with the phylogenetic inferences in several recent molecular studies[18,21,22] (Fig. 1). *Docynia delavayi* (clade VIII: Fig. 3) was the sister of *Malus doumeri* (clade VII: Fig. 3), a species formerly treated in the *Malus* sect. *Docyniopsis* or the genus *Docyniopsis* (C.K.Schneid.) Koidz.. Several shared characteristics supported a close relationship between these species, i.e., cone-shaped non-adnate part of the ovaries, fully connate carpels, incurved and persistent calyx, numerous scattered sclereids throughout the flesh, juvenile leaves deeply lobed, and similar flavonoid chemistry[1,5,78]. Robertson et al. (1991) proposed to merge *Docynia* into *Docyniopsis* (= *Malus* sect. *Docyniopsis*) based on these shared characters. Given the more ovules per lucule (more than 2), *Docynia* was treated in *Cydonia* Mill. by Roemer (1847) and Wenzig (1883), and this treatment was not supported by our strongly supported nuclear and plastid phylogeny. Furthermore, several characters may easily distinguish *Docynia* from *Cydonia*, such as ovaries partially adnate to hypanthium in *Docynia* vs. fully adnate in *Cydonia*, styles fused at the base in *Docynia* vs. free in *Cydonia*, and stamens ca. 40 in *Docynia* vs. 25 in *Cydonia*[1,5]. Hence, the above lines of evidence support a redefined circumscription of *Malus* s.l., by merging *Docynia* into *Malus*.

Our nuclear phylogeny recovered three major clades within *Malus* s.l.; however, cytonuclear discordance was detected for the placement of clade II, i.e., the combined eastern North American and Mediterranean species. Clade II was sister to clade III in the nuclear phylogeny (Fig. 3), while sister to another East Asian genus *Pourthiaea* in the plastid phylogeny (Figs. 4, 7). Several possible causes may explain the conflict between nuclear and plastid topologies, such as ILS, allopolyploidy, and hybridization[79]. Our coalescence simulation analysis showed that no simulated nuclear genes were

309 concordant to the plastid tree, the well-supported incongruence of the clade II between the nuclear and
310 plastid trees may not be explained by ILS (Fig. 7). Allopolyploidy could be excluded for explaining the
311 discordance (Fig. 3, Table 2). Although the eastern Northern American species have been involved in
312 diploidy, triploidy, and tetraploidy[3,65], the two Mediterranean species were consistently diploid[65,73]. It is
313 less likely that allopolyploidy has resulted in cytonuclear discordance.

314 Hybridization might be the underlying mechanism for explaining the conflicts between nuclear and
315 plastid topologies (Fig. 7), especially in the context of the chloroplast capture hypothesis, which has
316 been well illustrated by recent studies[21,38,80]. The chloroplast capture events have also been used to
317 explain the topological conflicts in the AMP (*Amelanchier-Malacomelels-Peraphyllum*) clade of
318 Maleae[21]. Furthermore, hybridization has played an essential role in the diversification of the apple tribe
319 (Maleae)[5]. Our ancestral area reconstruction showed that the clade II originated from western North
320 America in the middle Eocene (Fig. 8 and Supplementary Fig. S17). The shared plastomes between
321 clade II and *Pourthiaea* implicated that the ancestor of clade II most likely captured the plastome of the
322 ancestor of *Pourthiaea* in western North America, and the ancestor of *Pourthiaea* may have been widely
323 distributed in the boreotropical flora of the Northern Hemisphere in Eocene[45,81]. According to the
324 chloroplast capture scenario of Liu et al. (2020a), we hypothesize that the ancestor of clade II might
325 have been widely distributed in western North America, and served as the paternal parent of the hybrid.
326 The pollen provider (the ancestor of the clade II) hybridized with the ancestor of *Pourthiaea* (the ovule
327 provider), and formed the hybrid founder population. Subsequent backcrossings with the paternal parent
328 promoted that the ancestor of clade II captured the complete plastome of the ancestor of *Pourthiaea*.
329 Morphologically, clade II showed similarities to *Pourthiaea*, such as the densely scattered sclereids in
330 the flesh of fruits[5,82-84].

Ancient and recent events of hybridization and allopolyploidy drive the diversification of *Malus* s.l.

Rapid diversification and reticulate evolution pose significant challenges for phylogenetic inference of *Malus* s.l. This study resolved three strongly supported clades using hundreds of SCN genes. However, underlying gene tree conflicts for most nodes showed the potential hybridization and allopolyploidy events in the early diversification of *Malus* s.l.[5,85].

Our phylogenetic network analysis at the Maleae level confirmed one hybridization event, i.e., the hybrid origin of *Malus doumeri* between *Docynia delavayi* and the ancestor of the pome-related members of Maleae (Fig. 4), supporting an ancient hybridization event in the evolutionary history of Maleae. *Malus doumeri* is a diploid[63,64] (Table1, $2n = 34$); hence its origin did not involve an allopolyploidy event. Additionally, Our HyDe analysis detected 34 significant hybridization hypotheses that supported the hybrid origin of *M. doumeri*, and 25 out of 34 ($\gamma < 0.3$ and $\gamma > 0.7$) ones supported the ancient hybridization hypothesis.

Although the hybrid origin of *Malus coronaria* was estimated from the SNaQ analysis, the distribution of chromosome count data ($2n = 34, 51, 68$) indicated that allopolyploidy might have been involved in the speciation of *M. coronaria*, because the result of allopolyploidy event may have resembled that of hybridization[86]. In addition, our HyDe analysis showed that 27 of the 54 significant γ values supported the ancient hybridization events, and the same number of significant γ values (27) confirmed the recent hybridization events. However, an alternative hypothesis may explain this genetic admixture. With the wide naturalization of the European and the Central Asian species (*M. sylvestris*, *M. sieversii*, and *M. orientalis*) in eastern North America, the two individuals sampled in this study may represent a recent hybrid in the wild or horticulture. We need to test more samples of *M. coronaria* from its distributional range.

The phylogenetic conflicting positions of *Malus sikkimensis* among nine inferred topologies (Supplementary Figs. S1-S9) suggested its genetic heterogeneity, which also implicated by far more

gene trees in conflict with the species tree than in concord (6/370) in the *phyparts* analysis and the limited QS quartets (0.4/0.76/0.99) (Fig. 3 and Supplementary Fig. S10). We suggest that allopolyploidy might have resulted in the phylogenetic discordance of *Malus sikkimensis* (Fig. 3) based on the uneven gene trees supporting each bipartition (Table 1). Other lines of evidence also support this hypothesis. The chromosome count of *M. sikkimensis* varied from diploid to tetraploid (Fig. 3, Table 2). Additionally, an equal number of significant ancient and recent hybridization events (12 : 12: Supplementary Table S2) estimated by HyDe analysis showed that frequent hybridization might have played an important role in the diversification of *M. sikkimensis*, which might have introgressed with other species in ancient and recent times.

Although the sister relationship between *Malus baccata* var. *xiaojinensis* and clade I (Fig 2) has been fully supported in all nine nuclear topologies (Supplementary Figs. S1-S9) and the nearly full support of the QS analysis (Supplementary Fig. S10), only 23 out of the 319 informative gene trees confirmed this relationship in the *phyparts* analysis (Supplementary Fig. S11). A series of previous studies have demonstrated exclusive apomixis for *M. baccata* var. *xiaojinensis*[87]. They may have derived from the hybridization events between *M. kansuensis* and *M. toringoides*[88].

An East Asian-western North American origin of *Malus* s.l. and its subsequent extinctions in the Eocene

Our historical biogeographic analysis inferred East Asian + western North America as the most likely ancestral area of *Malus* s.l. (Fig. 8). The common ancestor is postulated to have occupied a widespread East Asian-western North American range, consistent with the rich fossil records of clade II and III recovered in northeast Asia and western North America from Eocene to Pliocene (Table 3). However, we did not find any fossil records of clade I from southern East Asia, which may be due to the underexplored fossil discovery for this region. This result disagrees with the North American origin or the East Asian origin, as proposed by Nikiforova et al. (2013) and Jin (2014), respectively. The

conflicted hypothesis on the *Malus* origin may be due to the absence of fossil records and the uneven sampling in the two prior analyses. Nikiforova et al. (2013)'s investigation did not include taxa of clade III (*Malus doumeri* and *Docynia delavayi*), and Jin (2014)'s study misidentified *Pseudocydonia sinensis* to be *Docynia delavayi*.

Our divergence time estimation suggested that three major clades of *Malus* diversified in the late Oligocene (43.58 Mya, 95% HPD interval: 41.2-44.39 Mya). Due to the decreased CO_2 principal forcing and long-term cooling trend from the early Eocene, the high latitude *Malus* (clade III) from northern East Asia migrated to southern East Asia; the ancient western North American populations dispersed to eastern North America and the Mediterranean region (clade II: Fig. 8). Subsequent extinctions occurred in the northern cold area from the Eocene to the Quaternary because a series of fossil species from different eras have been discovered in northern East Asia (*Malus kingiensis* in the Eocene, *M. parahupehensis* in the Miocene, and *M. obensis* in the Pliocene: Table 3) and western North America (*M. collardii*, *M. florissantensis*, and *M. pseudocredneria* in the Eocene, and *M. idahoensis* in the Miocene: Table 3). These extinction events in northern East Asia may be related to the cooling events in geologic times, such as the Miocene cooling and drying occurred at approximately 15-10 Mya[89-91] and the enhanced aridity at the middle latitudes of the Northern Hemisphere at about 8-7 Mya[92,93]. The living species of clade III show preferences to cool habitats in the high altitudes of southern East Asia, suggesting the northern East Asian origin of clade III, such as *Malus doumeri* at 700-2400 m and *Docynia delavayi* at 1000-3000 m^2. This distribution pattern has also been reported in many other angiosperm lineages, such as *Astilbe* Buch.-Ham. ex D.Don[94], *Meehania* Britton ex Small & Vail[95], *Mitchella* L.[96], *Parthenocissus* Planch.[97], and *Vitis* L.[98]. The extinctions of the early diverged *Malus* in western North America may be due to the increasing seasonality and drying spreading in the western Cordillera and cooling events into the Pleistocene[99-101]. The warm and moist environment in southern East Asia, eastern North America, and the Mediterranean promoted its survival in the refugia for *Malus* there, and the dispersal and vicariance events in the middle to late Eocene further facilitated its survival diversification across the Northern Hemisphere (Fig. 8).

Materials & Methods

Taxon sampling, library preparation, and deep genome skimming sequencing

Taxon sampling is designed to resolve the phylogenetic placements and relationships of major clades within *Malus* s.l. For the convenience of discussion, we followed the widely accepted taxonomic system proposed by Phipps et al. (1990), in which they recognized five sections. Due to the great economic importance of *Malus*, numerous hybrid cultivars have been used as crops and ornamentals, which have hybridized either within one section or between sections, and this made it difficult to infer the phylogenetic relationships between cultivated and wild species. Hence, we excluded the widely recognized artificial hybrid species in *Malus*, e.g., *M.* × *asiatica* Nakai, *M.* × *astracanica* hort. ex Dum.Cours., *M.* × *cerasifera* Spach, *M.* × *dawsoniana* Rehder, *M.* × *domestica* (= *M. pumila* Mill.), *M.* × *floribunda* Siebold ex Van Houtte, *M.* × *halliana*, *M.* × *micromalus*, *M.* × *heterophylla* Spach, *M.* × *magdeburgensis* Schoch ex Rehder, *M.* × *prunifolia* (Willd.) Borkh., *M.* × *sargentii* Rehder, *M.* × *scheideckeri* Späth ex Zabel, *M.* × *soulardii* (L.H.Bailey) Britton, *M.* × *spectabilis* (Aiton) Borkh., and *M.* × *sublobata* Rehder. We hence sampled 39 ingroup individuals representing 18 wild species (out of ca. 24[15]), representing all five sections and *Docynia* within *Malus* s.l. In addition, due to the potential biphyly of *Malus* based on complete plastomes[21,103], we sampled 38 outgroups across the tribes Maleae and Gillenieae to resolve the *Malus* phylogeny and identify possible events of cytonuclear discordance (Supplementary Table S1). We investigated 77 individuals in total, of which 27 species of deep genome skimming data were generated for this study (Supplementary Table S1).

Total genomic DNAs were extracted from silica-gel dried leaves or herbarium specimens using a modified CTAB (mCTAB) method[104] in the lab of the Institute of Botany, Chinese Academy of Science (IBCAS) in China. The libraries were prepared in the lab of Novogene, Beijing, China using NEBNext® Ultra™ II DNA Library Prep Kit, and then paired-end reads of 2 × 150 bp were generated on the NovoSeq 6000 Sequencing System (Novogene, Beijing; 5-10 G data for each sample: Supplementary

Table S1).

Single-copy nuclear marker development

The SCN marker development followed the pipeline in Liu et al. (2021). Briefly, the coding regions of *Malus domestica* (GenBank assembly accession: GCA_000148765.2) were first input into MarkerMiner v.1.0[105] to identify the putative single-copy genes. The resulting genes were then filtered by successively BLASTing[106-108] against six available genomes [*Malus baccata* (accession: GCA_006547085.1), *M. domestica*, *Pyrus betulifolia* Bunge (accession: GCA_007844245.1), *P. bretschneideri* Rehder (accession: GCA_000315295.1), *P. ussuriensis* Maxim. × *P. communis* L. (accession: GCA_008932095.1), and *P. pyrifolia* (Burm.f.) Nakai (accession: GCA_016587475.1)] in Geneious Prime[109], with the parameters settings in the Megablast program[110] as a maximum of 60 hits, a maximum E-value of 1×10^{-10}, a linear gap cost, a word size of 28, and scores of 1 for match and -2 for mismatch in alignments. We first excluded genes with mean coverage > 1.1 for alignments, which generally indicate potential paralogy of the genes and/or the presence of highly repeated elements in the sequences. The remaining alignments were further visually examined to exclude those genes receiving multiple hits with long overlapping but different sequences during the BLAST. It should be noted that the alignments with mean coverage between 1.0 and 1.1 were generally caused by the presence of tiny pieces of flanking intron sequences in the alignments. These fragments were still accepted as a SCN gene here. After filtering, the remaining genes were used as references in the following gene assembly. The baits *in silico* could be available from the Dryad Digital Repository (Data 1): https://doi.org/10.5061/dryad.2jm63xsq5.

Data processing and the assembly of single-copy nuclear genes

Read processing and assembly followed the pipeline in Liu et al. (2021). Generally, we used Trimmomatic v. 0.39[111] for quality trimming and adapter clipping with the parameters (ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:1:true LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15

MINLEN:36). Then, the results were quality-checked with FastQC v. 0.11.9[112]. The number of clean reads after trimming was also calculated here for comparison (Supplementary Table S1). We used the HybPiper pipeline v. 1.3.1[113] for targeting SCN genes with default settings; BWA v. 0.7.1[114] to align and distribute reads to target genes; SPAdes v. 3.15.0[115] with a coverage cutoff value of 8 to assemble reads to contigs; and Exonerate v. 2.2.0[116] to align assembled contigs to target sequences and determine exon-intron boundaries. Python and R scripts included in the HybPiper pipeline[113] were used to retrieve the recovered gene sequences, summarize and visualize the recovery efficiency.

Nuclear datasets construction and phylogenetic analysis

Sequences for each SCN were aligned in MAFFT v. 7.480[117] with options "--localpair --maxiterate 1000". Due to the variable sequencing coverage of each sample in this study, we employed three steps to remove the poorly aligned regions. We used trimAL v. 1.2[118] to trim the alignment of each SCN, in which all columns with gaps in more than 20% of the sequences or with a similarity score lower than 0.001 were removed. Given the low-quality assembly in some sequences, Spruceup[119] was used to discover, visualize, and remove outlier sequences in the concatenated multiple sequence alignments with the window size 50 and overlap 25. Because the Spruceup algorithm works better, the more data it has, we concatenated all the SCN gene alignments with AMAS v. 1.0[120] before running Spruceup. We also used AMAS v. 1.0[120] to split the processed/trimmed alignment back into single-locus alignments. The resulting alignments for each SCN were trimmed again using trimAL v. 1.2[118] with the same parameters described above. Thirdly, we excluded the sequences less than 250 bp in each alignment with our customized python script (exclude_short_sequences.py, which can be available from Dryad Digital Repository https://doi.org/10.5061/dryad.2jm63xsq5) for decreasing the effect of missing data, because the short sequences in each alignment have few informative sites for the following coalescent-based species tree inference. The resulting SCN genes were used to infer individual ML gene trees using RAxML 8.2.12[121] with a GTRGAMMA model and the option "-f a" and 200 BS replicates to assess clade support for each SCN. TreeShrink v. 1.3.9[122] was used for detecting abnormally long branches in

each tree with the default false positive error rate $\alpha = 0.05$ and per-species mode. The shrunk trees and sequences have been used for the following phylogenetic inference, and hereafter these resulted sequences were referred to as "clean genes".

We generated three different datasets to reconstruct the phylogeny to account for the effect of missing data in each SCN gene: (1) 50%-sample dataset: each SCN gene with at least 900 bp and more than 50% samples (≥ 39 individuals); (2) 80%-sample dataset: each SCN gene with at least 900 bp and more than 80% samples (≥ 62 individuals); (3) all-sample dataset: each SCN with at least 900 bp and more than 100% samples (77 individuals). These three datasets can be available from the Dryad Digital Repository (Data 2, 3, 4): https://doi.org/10.5061/dryad.2jm63xsq5. We used both concatenated and coalescent-based methods for phylogenetic inference of each dataset. We used PartitionFinder2[123,124] to estimate the best-fit partitioning schemes and/or nucleotide substitution models under the corrected Akaike information criterion (AICc) and linked branch lengths, as well as with rcluster[125] algorithm options for the nuclear dataset. The resulting partitioning schemes and evolutionary models were used for the following Maximum Likelihood (ML) tree using IQ-TREE2 v. 2.1.3[126] with 1000 SH-aLRT and the ultrafast bootstrap replicates and RAxML 8.2.12[121] with GTRGAMMA model for each partition and clade support assessed with 200 rapid bootstrap (BS) replicates. The shrunk trees from TreeShrink[122] were used to estimate a coalescent-based species tree with ASTRAL-III (Zhang et al., 2018) using local posterior probabilities (LPP)[127] to assess clade support. Each of the gene trees was rooted, and low support branches (≤ 10) were collapsed using Newick Utilities[128] or *phyx*[129] since collapsing gene tree nodes with BS support below a threshold value will help to improve accuracy[130]. In total, nine phylogenies were generated for topological comparison, and these nine trees are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.2jm63xsq5.

Detecting and visualizing nuclear gene tree discordance

To explore the discordance among gene trees, we employed *phyparts* v. 0.0.1[29] to calculate the conflicting/concordant bipartitions by comparing the nuclear gene trees against the ML tree inferred

from RAxML with a BS threshold of 50 (i.e., gene-tree branches/nodes with less than 50% BS were considered uninformative) for filtering out poorly supported branches, thus alleviating noise in the results of the conflict analysis[29]. We also used the internode certainty all (ICA) value that resulted from *phyparts* to quantify the degree of conflict on each node of a species tree given individual gene trees[131]. *Phyparts* results were visualized with *phypartspiecharts.py* (by Matt Johnson, available from https://github.com/mossmatters/MJPythonNotebooks/blob/master/phypartspiecharts.py). Furthermore, in order to distinguish lack of support from conflicting support in the species tree, we conducted Quartet Sampling (QS)[41] analysis with 100 replicates and the log-likelihood cutoff 2. The QS method subsamples quartets from the input tree and alignment to assess the confidence, consistency, and informativeness of internal tree relationships, and the reliability of each terminal branch, and then four values are given in this analysis: QC = Quartet Concordance, QD = Quartet Differential, QI = Quartet Informativeness, and QF = Quartet Fidelity. The QS result was visualized with plot_QC_ggtree.R (by ShuiyinLIU, available from https://github.com/ShuiyinLIU/QS_visualization). Both the *phyparts* and QS results can provide alternative evidence for evaluating the discordance between gene trees.

Comparing the nine topologies inferred above, we found two species (*Malus coronaria* and *M. kansuensis*) with conflicting phylogenetic positions among trees (referring to the result below). We used the "alternative relationship test" in *phyckle*[28] to investigate conflicting bipartitions and discover the gene trees supporting each bipartition. Considering the complex origin of *Malus coronaria*, we employed the two bipartitions supported by Phylonetworks analysis (see Fig. 4 & Table 1 in Results). Two or more user-specified alternative bipartitions could be used as a constraint to infer gene trees. Arbitrarily, we set the cutoff of ΔlnL > 100 as the outlier genes[132]. The resulting gene dataset supporting each bipartition was then used to estimate phylogenetic inference based on the concatenated (IQ-TREE2 and RAxML) and coalescent-based methods (ASTRAL-III) mentioned above. The resulting 12 trees could be available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.2jm63xsq5.

528 **Coalescence simulation and phylogenetic network estimation**

529 To measure the goodness-of-fit of the coalescent model with ILS explaining the gene tree
530 discordance adequately, we performed a coalescent simulation analysis following the methods in
531 previous studies[34,133-135]. Briefly, if the simulated gene trees based on the coalescent model correspond
532 well to the empirical gene trees, the gene tree discordance may be explained by ILS. Given the
533 calculation of gene tree distances, we subsampled 27 species representing each major clade in the
534 phylogenetic framework of the tribe Maleae. The function "sim.coaltree.sp" implemented in the R
535 package Phybase v. 1.5[136] has been used to simulate 10,000 gene trees with the multispecies coalescent
536 (MSC) model.

537 Hybrid Detector (HyDe) can be used to detect hybridization using phylogenetic invariants arising
538 under the coalescent model with hybridization[137]. We sampled 40 taxa, including all 39 *Malus* s.l.
539 individuals and one outgroup (*Pyrus communis*) to detect the possible hybridization events within *Malus*
540 s.l. This dataset can be available from the Dryad Digital Repository (Data 6):
541 https://doi.org/10.5061/dryad.2jm63xsq5. The γ denotes the inheritance probability of parent 1 (P1),
542 while 1 - γ would be the probability of the hybrid population being sister to parent 2 (P2). Generally,
543 significant γ values close to 0.5 indicate a recent hybridization event; significant γ values closer to 0 or 1
544 indicate an ancient hybridization event remained in the extant species. We herein set the γ threshold at
545 0.3 and 0.7, which followed convention[132].

546 To explore the possibility of reticulation as a cause of discordance in the apples and their allies, we
547 employed the Species Networks applying Quartets (SNaQ) method[43] implemented in the software
548 PhyloNetworks 0.14.0[44], which explicitly accommodates introgression/gene flow and ILS. Given the
549 computational limitation of PhyloNetworks, we used two datasets to test for hybridization events, i.e.,
550 27-taxa sampling at the tribe level of Maleae and 14-taxa sampling at the genus level of *Malus*. These
551 two datasets are available from the Dryad Digital Repository (Data 7 & 8):
552 https://doi.org/10.5061/dryad.2jm63xsq5. Considering that *Malus* members have been reported to have
553 hybridized with many genera in Maleae, e.g., *Aria* (Pers.) Host and *Torminaria* M.Roem.[5], we sampled

27 individuals, including five species representing each major clade in *Malus* s.l. and 22 outgroup species in Maleae. This taxon sampling scheme represents a reasonable compromise between taxonomic coverage and computational cost. The 27-taxa dataset construction followed the method described above (i.e., Nuclear dataset construction and phylogenetic analysis). The best trees generated from RAxML were used to estimate the quartet concordance factors (CFs), representing the proportion of genes supporting each possible relationship between each set of four species. The resulting CFs and the ASTRAL species tree were used as initial input to run SNaQ analysis ($h = 0$), and the resulting best network was used as starting topology to run the next h value ($h + 1$), and so on. We investigated h values ranging from 0 to 5 with 50 runs in each h for estimating the best phylogenetic network. Each run generated a pseudo-deviance score: a value for fitting the network to the data, and estimated the inheritance probabilities (i.e. the proportion of genes contributed by each parental population to a hybrid taxon) for each network. Similarly, we also sampled 14 species, including 13 *Malus* species and one outgroup (*Pyrus communis*), to test the hybridization events among *Malus* members. The method followed that of the 27-taxa sampling dataset mentioned above. The best network was visualized using Dendroscope v 3.7.4[138].

Plastome assembly, annotation, phylogenetic analysis, and cytonuclear discordance

A two-step strategy was used for obtaining high-quality chloroplast genomes. NOVOPlasty v. 4.3.1[139] was applied first to assemble the plastomes with high-quality raw data. Then we used the successive assembly approach[140], combining the reference-based and the *de novo* assembly methods to assemble the remaining low-quality samples. With the *de novo* assembly and a seed-and-extend algorithm, NOVOPlasty was the least laborious approach and resulted in accurate plastomes; however, this program needs sufficient high-quality raw reads without gaps to cover the whole plastome. The whole plastomes assembled from NOVOPlasty then could be used as references for assembling the remaining samples. The successive method provided an excellent approach to obtaining relatively accurate and nearly complete plastomes with or without gaps from lower-coverage raw data. Due to the

579 sensitivity of Bowtie2 v. 2.4.2[141] to the reference, this successive method needs a closely related
580 reference sequence with increased time and RAM requirements. Several recent studies have described
581 the procedure in detail[21,60,103,140,142,143]. All assembled plastomes have been submitted to GenBank with
582 the accession numbers listed in Supplementary Table S1.

583 The assembled plastid genomes from the low-coverage and high-coverage datasets were annotated
584 using PGA[144] with a closely related plastome (MN062004: *Malus ioensis*) downloaded from GenBank
585 as the reference, and the results of automated annotation were checked manually. The coding sequences
586 of plastomes were translated into proteins to manually check the start and stop codons in Geneious
587 Prime[109]. The custom annotations in the GenBank format were converted into the FASTA format, and
588 five-column feature tables file required by NCBI submission using GB2sequin[145].

589 Given the considerable variation among plastid introns at the level of Maleae, we extracted 80
590 coding genes (CDSs) using Geneious Prime[109], and these CDSs were aligned by MAFFT v. 7.475[117]
591 with default parameters, respectively. This dataset with 77 samples and 80 plastid coding genes is
592 available from the Dryad Digital Repository (Data 5): https://doi.org/10.5061/dryad.2jm63xsq5. The
593 best-fit partitioning schemes and/or nucleotide substitution models for each dataset were estimated using
594 PartitionFinder2[123,124], under the corrected Akaike information criterion (AICc) and linked branch
595 lengths, as well as with rcluster[125] algorithm options. The partitioning schemes and evolutionary model
596 for each subset were used for the downstream phylogenetic analysis. Like the nuclear analysis, we
597 estimated the ML tree by IQ-TREE2 v. 2.1.3[126] with 1000 SH-aLRT and the ultrafast bootstrap replicates
598 and RAxML 8.2.12[121] with GTRGAMMA model for each partition and clade support assessed with 200
599 rapid BS replicates. We also used ASTRAL-III[130] for estimating a coalescent-based species tree. These
600 three trees are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.2jm63xsq5.
601 Cytonuclear discordance at various levels was detected in *Malus* s.l. (Fig. 7); coalescent simulation was
602 also employed for evaluating the importance of ILS in explaining cytonuclear discordance[35,135]. We used
603 the R package Phybase v. 1.5 to simulate 10,000 gene trees based on the ASTRAL species tree from
604 SCN genes under MSC model. We used *Phyparts* v. 0.01[29] to explore conflicts among the simulated

gene trees and plastid tree, and the proportion of gene tree concordance compared to the plastid genome tree was used to qualify the discordance.

Dating and ancestral area reconstruction

We aim to estimate the age of divergence of the three major clades identified from hundreds of SCN genes and 80 plastid coding genes in *Malus* s.l. The fossils used in this study are listed in Table 3. *Malus obensis* and the clade *Malus doumeri* - *Docynia delavayi* were grouped together because of the fruit similarity. *Malus parahupehensis* was thought to be similar to the living species *M. hupehensis*; however, this fossil species showed more similarities to *M. doumeri* because of the dentate margin and parallel, craspedodromous venation[15]. Thus, we grouped *M. parahupehensis*, *M. doumeri, M. obensis*, and *Docynia delavayi* as monophyletic. The earliest fossil, *Malus kingiensis,* was found in the middle Eocene from the western Kamchatka peninsula and can not be assigned to any living lineages of *Malus* s.l. based on the morphology; thus, this fossil species likely represented the stem clade of *Malus doumeri* and *Docynia delavayi*. However, only one fossil, *Malus antiqua*, was found in Europe, and this fossil with deeply lobed leaves was similar to the Mediterranean species. Therefore, we grouped this fossil with the two Mediterranean living species (*M. florentina* and *Eriolobus trilobatus*). Five fossil species were described from North America, especially the abundant fossil records in western North America. More or less deeply lobed leaves characterized all these leaf fossils, significantly distinct from the living western North American species (*Malus fusca*). They showed more similarities to the eastern North American and Mediterranean species, and these species were thus grouped together. Additionally, leaf fossil of *Malus* or *Pyrus* from the Republic site, Washington[146] was used to constrain the divergence between *Malus* and *Pyrus* at 46-44 Mya.

We ran the dating analyses based on the SCN and plastid datasets to test the divergence time differences with significant cytonuclear discordance. Given the intensive computational burden of dating analysis using BEAST2, we employed a 19-taxa dataset with only one individual for each species in *Malus* s.l. and *Pyrus communis* as the outgroup, and this dataset is available from the Dryad Digital

Repository (Data 13): https://doi.org/10.5061/dryad.2jm63xsq5. The divergence time estimation was run under a GTR model with a gamma rate inferred from PartitionFinder2[123,124], an uncorrelated lognormal relaxed clock[147], and the fossilized birth-death model[148,149]. Markow Chain Monte Carlo (MCMC) chains were run for 100,000,000, sampling every 20,000 generations in five parallel jobs. We used the LogCombiner v1.10 to combine log and tree files from the five independent runs of BEAST. The MCMC trace file was analyzed in Tracer v1.7.1[150]. Maximum credibility trees were generated in TreeAnnotator v1.10, and FigTree v1.4.4 visualized the MCC tree.

To test the ancestral areas of three major clades of *Malus*, we conducted the ancestral area construction using BioGeoBEARS v. 1.1.1[151] implemented in RASP v. 4.2[152]. Geological evidence suggests that an aridity barrier existed from the western-most part of China to the eastern Asian coast from the Paleogene to the Miocene. It has been hypothesized to have acted as a climate barrier between these two regions[153]; thus, we subdivided East Asia into the northern and the southern areas[154-159]. Six biogeographic areas were defined across the distribution of *Malus* s.l.: (A), Southern East Asia; (B), Northern East Asia; (C), Europe and Central Asia; (D), Mediterranean; (E), eastern North America; (F), western North America. The MCC tree summarized by TreeAnnotator was used as input of RASP. The *maxarea* was set to six, i.e., the number of potential areas of a hypothetical ancestor was restricted to a maximum of six regions. The model with the highest AICc_wt value has been chosen as the best model.

Conclusions

We resolved the phylogenetic backbone of *Malus* s.l. using 785 nuclear loci (77 taxa) and 80 plastid coding genes (75 taxa). The nuclear phylogeny supported the monophyly of *Malus* s.l. (including *Docynia*) and three strongly supported major clades within the genus. However, widespread gene tree conflicts among nuclear gene trees indicated the complicated evolutionary history of *Malus*, and ILS, hybridization, and allopolyploidy have played an important role in the evolution of *Malus*, explaining this cytonuclear discordance. We detected a deep hybridization event involving *Malus doumeri* as a hybrid between the ancestor of pome-beared species and *Docynia delavayi*, and a recent hybridization

event (*M. coronaria*) between *M. sieversii* and a taxon of the clade of *M. ioensis* and *M. angustifolia*. However, our plastid result recovered the biphyly of *Malus* s.l., with the combined eastern North American and the Mediterranean clade sister to the East Asian genus *Pourthiaea*. The well-supported cytonuclear discordance could be best explained by the chloroplast capture event that occurred in western North America in the Eocene. The phylogenomic case study of the apple genus implicated that multiple methods accounting for ILS and gene flow can help untangle complex phylogenetic relationships among species, and concatenation method or methods only accounting for ILS (coalescent-based method) are biased and not appropriate for phylogenetic inferences in lineages with a highly complex evolutionary history. Our historical biogeographic analysis without fossil species in northern East Asia and western North America resulted in the East Asian origin, contrastingly that involving fossil and living species supported a widespread East Asian-western North American origin of *Malus* s.l., followed by subsequent extinction events in northern East Asia and western North America in the Eocene, and this indicated that integrating fossil and living species could promote the more accurate estimation of dating analysis, as well as ancestral area reconstruction[160]. The robust phylogenomic framework for the apple genus should provide an evolutionary basis for the breeding and crop improvement of apples and their close relatives. The study also represents an excellent case for utilizing the recently proposed deep genome skimming approach for obtaining nuclear and organelle genes for robust phylogenetic reconstructions.

Acknowledgments

All computational analyses were conducted on the Smithsonian Institution High Performance Computing Cluster (SI/HPC, “Hydra”: https://doi.org/10.25572/SIHPC). National Natural Science Foundation of China supports this research (Grant No.: 32000163 & 31620103902).

Author details

[1]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China. [2]Department of Botany, National Museum of Natural History, Smithsonian Institution, PO Box 37012, Washington, DC 20013-7012, USA. [3]Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, Guangdong, China. [4]Guangdong Provincial Key Laboratory of Applied Botany, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, Guangdong, China. [5]National Institute of Biological Resources, Incheon 22689, South Korea. [6]School of Ecology and Nature Conservation, Beijing Forestry University, Beijing, 100083 PR China. [7]Department of Plant and Microbial Biology, North Carolina State University, Raleigh, NC 27965, USA. [8]State Key Laboratory of Genetic Engineering and Collaborative Innovation Center of Genetics and Development, Ministry of Education Key Laboratory of Biodiversity and Ecological Engineering, Institute of Plant Biology, Center of Evolutionary Biology, School of Life Sciences, Fudan University, Shanghai 200433, China. [9]Department of Biology, Huck Institutes of the Life Sciences, Pennsylvania State University, 510D Mueller Laboratory, University Park, PA 16802 USA. [10]College of Life Sciences, Liaocheng University, Liaocheng 252059, China

Author contributions

B.B.L. designed and led the project. D.Y.H. and J.W. supervised the study. B.B.L. carried out the phylogenomic analyses and wrote the draft manuscript. C.X. performed the experiment of deep genome skimming. C.R., M.K., R.H., J.H., and W.B.Z. participated in the phylogenetic and biogeographic

697 analyses. G.Z.Q. provided suggestions on structuring the paper. C.H.H. and H.M. provided part of the
698 transcriptomic data. All the authors contributed to the writing and interpretation of the results, and
699 approved the final manuscript.

700 ## Data availability

701 Raw reads have been deposited in the NCBI Sequence Read Archive (BioProject PRJNA759205). The
702 customized scripts for SCN gene analysis, tree files, pre-and post-filtered alignments for all analyses are
703 available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.2jm63xsq5.

704 ## Conflict of interest

705 The authors declare no competing interests.

706 ## References

707 1. Kalkman, C. Rosaceae in The Families and Genera of Vascular Plants. Vol. VI. Flowering Plants.
708 Dicotyledons. Celastrales, Oxalidales, Rosales, Cornales, Ericales Vol. 6 (ed Kubitzki, K.) 343-386
709 (Springer, 2004).
710 2. Lu, L.D. *et al.* Rosaceae in Flora of China. Volume 9: Pittosporaceae through Connaraceae Vol. 9
711 (eds Wu, Z.Y., Raven, P.H. & Hong, D.Y.) 46-434 (Science Press & Missouri Botanical Garden
712 Press, 2003).
713 3. Phipps, J. Rosaceae in Flora of North America North of Mexico Magnoliophyta: Picramniaceae to
714 Rosaceae Vol. 9 (eds Flora of North America editorial Committee) 18-662 (Oxford University
715 Press, 2004).
716 4. Kurtto, A.K. Rosaceae (pro parte majore) in Euro+Med Plantbase - the information resource for
717 Euro-Mediterranean plant diversity (2009).

718 5. Robertson, K.R., Phipps, J.B., Rohrer, J.R. & Smith, P.G. A synopsis of genera in Maloideae
719 (Rosaceae). *Syst. Bot.* **16**, 376-394 (1991).
720 6. Koehne, B.A.E. Deutsche dendrologie (Verlag von Ferdinand Enke, 1893).
721 7. Beissner, L., Schelle, E. & Zabel, H. Handbuch der Laubholz-Benennung (Verlagsbuchhandlung
722 Paul Parey, 1903).
723 8. Schneider, C.K. Illustriertes Handbuch der Laubholzkunde (Verlag von Gustav Fisher, 1906).
724 9. Koidzumi, G. Contributiones ad floram Asiae orientalis. *Acta Phytotax. Geobot.* **3**, 146-162 (1934).
725 10. Rehder, A. Manual of cultivated trees and shrubs hardy in North America, exclusive of the
726 subtropical and warmer temperate regions, ed. 2. (Macmillan, 1940).
727 11. Huckins, C.A. A revision of the sections of the genus *Malus* Miller. Ph.D. 796 (Cornell University,
728 1972).
729 12. Phipps, J.B., Robertson, K.R., Smith, P.G. & Rohrer, J.R. A Checklist of the Subfamily Maloideae
730 (Rosaceae). *Canad. J. Bot.* **68**, 2209-2269 (1990).
731 13. Langenfeld, V.T. Apple tree systematics in Apple Tree: Morphological Evolution, Phylogeny,
732 Geography, and Systematics of the Genus (ed Langenfeld, V.T.) 119-195 (Zinatne, 1991).
733 14. Yunong, L. A primarily modern systematics of the genus *Malus* Mill. in the world. *J. Fruit Sci.* S1
734 (1996).
735 15. Qian, G.Z. The taxonomic study of the genus *Malus* Mill. Ph.D. (Nanjing Forestry University,
736 Nanjing, 2005).
737 16. Li, Y.N. A critical review of the species and the taxonomy of *Malus* Mill in the world. *J. Fruit Sci.*
738 **13**, 63-81 (1996).
739 17. Robinson, J.P., Harris, S.A. & Juniper, B.E. Taxonomy of the genus *Malus* Mill. (Rosaceae) with
740 emphasis on the cultivated apple, Malus domestica Borkh. *Pl. Syst. Evol.* **226**, 35-58 (2001).
741 18. Lo, E.Y.Y. & Donoghue, M.J. Expanded phylogenetic and dating analyses of the apples and their
742 relatives (Pyreae, Rosaceae). *Mol. Phylogenet. Evol.* **63**, 230-243 (2012).

19. Jin, G.H. Phylogenomics and biogeography of Malus Mill. in *Kunming Institute of Botany, Chinese Academy of Sciences* Master 129 (University of Chinese Academy of Sciences, Kunming, 2014).
20. Zhang, S.D. *et al.* Diversification of Rosaceae since the Late Cretaceous based on plastid phylogenomics. *New Phytol.* **214**, 1355-1367 (2017).
21. Liu, B.-B., Campbell, C.S., Hong, D.-Y. & Wen, J. Phylogenetic relationships and chloroplast capture in the *Amelanchier-Malacomeles-Peraphyllum* clade (Maleae, Rosaceae): evidence from chloroplast genome and nuclear ribosomal DNA data using genome skimming. *Mol. Phylogenet. Evol.* **147**, 106784 (2020).
22. Xiang, Y.Z. *et al.* Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Mol. Biol. Evol.* **34**, 262-281 (2017).
23. Wen, J. *et al.* Transcriptome sequences resolve deep relationships of the grape family. *Plos One* **8**, e74394 (2013).
24. Léveillé-Bourret, E., Starr, J.R., Ford, B.A., Lemmon, E.M. & Lemmon, A.R. Resolving rapid radiations within angiosperm families using anchored phylogenomics. *Syst. Biol.* **67**, 94-112 (2018).
25. Herrando-Moraira, S. *et al.* Nuclear and plastid DNA phylogeny of tribe Cardueae (Compositae) with Hyb-Seq data: A new subtribal classification and a temporal diversification framework. *Mol. Phylogenet. Evol.* **137**, 313-332 (2019).
26. Feuda, R. *et al.* Improved modeling of compositional heterogeneity supports sponges as sister to all other animals. *Curr. Biol.* **27**, 3864-3870.e4 (2017).
27. Walker, J.F., Brown, J.W. & Smith, S.A. Analyzing contentious relationships and outlier genes in phylogenomics. *Syst. Biol.* **67**, 916-924 (2018).
28. Smith, S.A., Walker-Hale, N., Walker, J.F. & Brown, J.W. Phylogenetic conflicts, combinability, and deep phylogenomics in plants. *Syst. Biol.* **69**, 579-592 (2020).

29. Smith, S.A., Moore, M.J., Brown, J.W. & Yang, Y. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evol. Biol.* **15**, 150 (2015).

30. Gonçalves, D.J.P., Simpson, B.B., Ortiz, E.M., Shimizu, G.H. & Jansen, R.K. Incongruence between gene trees and species trees and phylogenetic signal variation in plastid genes. *Mol. Phylogenet. Evol.* **138**, 219-232 (2019).

31. Gonçalves, D.J.P., Jansen, R.K., Ruhlman, T.A. & Mandel, J.R. Under the rug: Abandoning persistent misconceptions that obfuscate organelle evolution. *Mol. Phylogenet. Evol.* **151**, 106903 (2020).

32. Doyle, J.J. Defining coalescent genes: Theory meets practice in organelle phylogenomics. *Syst. Biol.* (2021).

33. Walker, J.F., Walker-Hale, N., Vargas, O.M., Larson, D.A. & Stull, G.W. Characterizing gene tree conflict in plastome-inferred phylogenies. *PeerJ* **7**, e7747 (2019).

34. Morales-Briones, D.F. *et al.* Disentangling sources of gene tree discordance in phylogenomic data sets: testing ancient hybridizations in Amaranthaceae s.l. *Syst. Biol.* (2020).

35. Rose, J.P., Toledo, C.A.P., Lemmon, E.M., Lemmon, A.R. & Sytsma, K.J. Out of sight, out of mind: widespread nuclear and plastid-nuclear discordance in the flowering plant genus *Polemonium* (Polemoniaceae) suggests widespread historical gene flow despite limited nuclear signal. *Syst. Biol.* **70**, 162-180 (2020).

36. Soltis, D.E., Soltis, P.S., Collier, T.G. & Edgerton, M.L. Chloroplast DNA variation within and among genera of the *Heuchera* group (Saxifragaceae): evidence for chloroplast transfer and paraphyly. *Am. J. Bot.* **78**, 1091-1112 (1991).

37. Yi, T.S., Jin, G.H. & Wen, J. Chloroplast capture and intra- and inter-continental biogeographic diversification in the Asian - New World disjunct plant genus *Osmorhiza* (Apiaceae). *Mol. Phylogenet. Evol.* (Article) **85**, 10-21 (2015).

792 38. Liu, X., Wang, Z.S., Shao, W.H., Ye, Z.Y. & Zhang, J.G. Phylogenetic and taxonomic status
793 analyses of the Abaso section from multiple nuclear genes and plastid fragments reveal new
794 insights into the North America origin of *Populus* (Salicaceae). *Front. Plant Sci.* **7**, 2022 (2017).
795 39. Chifman, J. & Kubatko, L. Quartet inference from SNP data under the coalescent model.
796 *Bioinformatics* **30**, 3317-3324 (2014).
797 40. Mirarab, S. & Warnow, T. ASTRAL-II: coalescent-based species tree estimation with many
798 hundreds of taxa and thousands of genes. *Bioinformatics* **31**, i44-i52 (2015).
799 41. Pease, J.B., Brown, J.W., Walker, J.F., Hinchliff, C.E. & Smith, S.A. Quartet Sampling
800 distinguishes lack of support from conflicting support in the green plant tree of life. *Am. J. Bot.*
801 **105**, 385-403 (2018).
802 42. Than, C., Ruths, D. & Nakhleh, L. PhyloNet: a software package for analyzing and reconstructing
803 reticulate evolutionary relationships. *BMC Bioinform.* **9**, 322 (2008).
804 43. Solís-Lemus, C. & Ané, C. Inferring phylogenetic networks with maximum pseudolikelihood under
805 incomplete lineage sorting. *PLoS Genet.* **12**, e1005896 (2016).
806 44. Solís-Lemus, C., Bastide, P. & Ané, C. PhyloNetworks: A Package for phylogenetic networks.
807 *Mol. Biol. Evol.* **34**, 3292-3298 (2017).
808 45. Wolfe, J.A. Some aspects of plant geography of the Northern Hemisphere during the late
809 Cretaceous and Tertiary. *Ann. Missouri Bot. Gard.* 264-279 (1975).
810 46. Wen, J., Ickert-Bond, S., Nie, Z.-L. & Li, R. Timing and modes of evolution of eastern Asian-
811 North American biogeographic disjunctions in seed plants. 252-269 (2010).
812 47. Wolfe, J.A. & Wehr, W. Rosaceous Chamaebatiaria-like foliage from the Paleogene of western
813 North America. *Aliso: A Journal of Systematic and Evolutionary Botany* **12**, 177-200 (1988).
814 48. Manchester, S.R. Fruits and seeds of the Middle Eocene nut beds flora, Clarno Formation, Oregon.
815 *Palaeontogr. Am.* **58**, 1-205 (1994).
816 49. Wehr, W. & Hopkins, D. The Eocene orchards and gardens of Republic, Washington. *Wash. geol.*
817 **22**, 27-34 (1994).

818 50. Axelrod, D.I. The Eocene Thunder Mountain flora of central Idaho (University of California Press,
819 1998).
820 51. Leopold, E. & Clay-Poole, S. Florissant leaf and pollen floras of Colorado compared: climatic
821 implications in Fossil flora and stratigraphy of the Florissant Formation, Colorado. Proceedings of
822 the Denver Museum of Nature and Science, series Vol. 4 17-70 (2001).
823 52. Wheeler, E.A. & Manchester, S.R. Woods of the middle Eocene nut beds flora, Clarno Formation,
824 Oregon, USA (Published for the International Association of Wood Anatomists at the Nationaal
825 Herbarium Nederland, 2002).
826 53. Campbell, C.S., Evans, R.C., Morgan, D.R., Dickinson, T.A. & Arsenault, M.P. Phylogeny of
827 subtribe Pyrinae (formerly the Maloideae, Rosaceae): Limited resolution of a complex evolutionary
828 history. *Pl. Syst. Evol.* **266**, 119-145 (2007).
829 54. Chen, X. *et al.* Sequencing of a wild apple (*Malus baccata*) Genome unravels the differences
830 between cultivated and wild apple species regarding disease resistance and cold tolerance. *G3*
831 *(Bethesda)* **9**, 2051-2060 (2019).
832 55. Velasco, R. *et al.* The genome of the domesticated apple (Malus x domestica Borkh.). *Nat. Genet.*
833 **42**, 833-9 http://dx.doi.org/10.1038/ng.654 (2010).
834 56. Daccord, N. *et al.* High-quality de novo assembly of the apple genome and methylome dynamics of
835 early fruit development. *Nat. Genet.* **49**, 1099-1106 (2017).
836 57. Duan, N. *et al.* Genome re-sequencing reveals the history of apple and supports a two-stage model
837 for fruit enlargement. *Nat. Commun.* **8**, (2017).
838 58. Zhang, L. *et al.* A high-quality apple genome assembly reveals the association of a retrotransposon
839 and red fruit colour. *Nat. Commun.* **10**, 1494 (2019).
840 59. Sun, X. *et al.* Phased diploid genome assemblies and pan-genomes provide insights into the genetic
841 history of apple domestication. *Nat. Genet.* **52**, 1423-1432 (2020).

842 60. Liu, B.-B. *et al.* Capturing single-copy nuclear genes, organellar genomes, and nuclear ribosomal
843 DNA from deep genome skimming data for plant phylogenetics: A case study in Vitaceae. *J. Syst.*
844 *Evol.* (2021).
845 61. Maureira-Butler, I.J., Pfeil, B.E., Muangprom, A., Osborn, T.C. & Doyle, J.J. The reticulate history
846 of *Medicago* (Fabaceae). *Syst. Biol.* **57**, 466-482 (2008).
847 62. Liang, G.-L. Observations of chromosomes of *Malus* species in China. *Acta Phytotax. Sin.* **25**, 437-
848 441 (1987).
849 63. Chen, R.Y. Chromosome atlas of Chinese fruit trees and their close wild relatives. *Chromosome*
850 *atlas of Chinese principal economic plants. Tomus I* (1993).
851 64. Liang, G.-L. & Li, X.-L. Chromosome studies of Chinese species of *Malus* Mill. *Acta Phytotax.*
852 *Sin.* **31**, 236-251 (1993).
853 65. Schuster, M. & Büttner, R. Chromosome numbers in the *Malus* wild species collection of the
854 genebank Dresden-Pillnitz. *Genet. Resour. Crop Evol.* **42**, 353-361
855 http://dx.doi.org/10.1007/bf02432139 (1995).
856 66. Liang, G.-L., Li, Y.-N. & Li, X.-L. Evolutionary study of the chromosomes at pollen mother cell
857 meiosis in *Malus*. *J. Southwest Agric. Univ.* **18**, 299-307 (1996).
858 67. Liang, G.-L. Comparative studies of karyotypes in Chinese species of *Malus*. *J. Southwest Agric.*
859 *Univ.* **1**, 104-117 (1986).
860 68. Pogan, E., Jankun, A. & Wcislo, H. Further studies in chromosome numbers of Polish angiosperms.
861 Part XXIV. *Acta Biologica Cracoviensia. Series Botanica* **33**, 26, 29-31, 35, 38 (1991).
862 69. Liang, G.-L. A preliminary study on G--banding patterns at metaphase chromosomes in Malus
863 baccata. *Hereditas* **12**, 4 (1990).
864 70. Baranec, T. & Murin, A. Karyogical analyses of some Korean woody plants. *Biologia* **58**, 797-804
865 (2003).
866 71. Probatova, N.S. *et al.* Chromosome numbers of vascular plants from nature reserves of the
867 Primorsky Territory and the Amur River basin. *Bot. Zhurn. S.S.S.R.* **91**, 1117-1134 (2006).

868 72. Liang, G.-L. Induction of Giemsa C-bands and analyses of banding petterns in *Malus sikkimensis*.
869 *J. Southwest Agric. Univ.* **19**, 95-97 (1997).
870 73. Höfer, M. & Meister, A. Genome size variation in *Malus* species. *J. Bot.* **2010**, 1-8 (2010).
871 74. Decaisne, M.J. Mémoire sur la Famille des Pomacées. *Nouvelles archives du Muséum d'histoire*
872 *naturelle* **10**, 45-192 (1874).
873 75. Focke, W.O. Rosaceae in Die Natürlichen Pflanzenfamilien nebst ihren Gattungen und wichtigeren
874 Arten, insbesondere den Nutzpflanzen, unter Mitwirkung zahlreicher hervorragender Fachgelehrten
875 begründet T.3 Abt.3 (eds Engler, A., Krause, K., Pilger, R.K.F. & Prantl, K.) (W. Engelmann,
876 1894).
877 76. Rehder, A. Bibliography of cultivated trees and shrubs hardy in the cooler temperate regions of the
878 northern hemisphere (Jamaica Plain, 1949).
879 77. Delectis Florae Reipublicae Popularis Sinicae Agendae Academiae Sinicae Edita Flora Reipublicae
880 Popularis Sinicae (Science Press, 1974).
881 78. WILLIAMS, A.H. Chemical evidence from the flavonoids relevant to the classification of Malus
882 species. *Bot. J. Linn. Soc.* **84**, 31-39 (1982).
883 79. Rieseberg, L.H. & Soltis, D. Phylogenetic consequences of cytoplasmic gene flow in plants. *Trends*
884 *Plant Sci.* **5**, 65-84 (1991).
885 80. Acosta, M.C. & Premoli, A.C. Evidence of chloroplast capture in South American Nothofagus
886 (subgenus Nothofagus, Nothofagaceae). *Mol. Phylogenet. Evol.* **54**, 235-242 (2010).
887 81. Wen, J. Evolution of eastern Asian and eastern North American disjunct distributions in flowering
888 plants. *Annu. Rev. Ecol. Evol. Syst.* **30**, 421-455 (1999).
889 82. Liu, B.-B. & Hong, D.-Y. A taxonomic revision of the *Pourthiaea villosa* complex (Rosaceae).
890 *Phytotaxa* **244**, 201-247 (2016).
891 83. Liu, B.-B. & Hong, D.-Y. Identity of *Pourthiaea podocarpifolia* (Rosaceae). *Phytotaxa* **269**, 221-
892 230 (2016).

84. Liu, B.-B. & Hong, D.-Y. A taxonomic revision of four complexes in the genus *Pourthiaea* (Rosaceae) (Magnolia Press, 2017).
85. Campbell, C.S., Greene, C.W. & Dickinson, T.A. Reproductive biology in subfam. Maloideae (Rosaceae). *Syst. Bot.* **16**, 333-349 (1991).
86. Wagner, N.D., He, L. & Hörandl, E. Phylogenomic relationships and evolution of polyploid *Salix* species revealed by RAD sequencing data. *Front. Plant Sci.* **11**, (2020).
87. Cheng, M.H., Yang, X.H. & Zeng, W.G. A preliminary report on investigation and study of *Malus* xiaojinensis — an apple stock resource. *J. Southwest Agric. Univ.* **3**, 38-43 (1984).
88. Shi, S.-Y. *et al.* Genetic diversity of Malus toringoides (Rehd.) Hughes based on AFLP. *Acta Hortic. Sin.* **33**, 381 (2006).
89. Douglas, R.G. & Woodruff, F. Deep sea benthic foraminifera in The Sea The Oceanic Lithosphere (ed Emiliani, C.) 1233-1327 (Wiley-Interscience, 1981).
90. Haq, B.U., Hardenbol, J. & Vail, P.R. Chronology of Fluctuating Sea Levels Since the Triassic. *Science* **235**, 1156-1167 (1987).
91. Zachos, J., Pagani, M., Sloan, L., Thomas, E. & Billups, K. Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* **292**, 686-693 (2001).
92. Savage, J.M. & Vial, J.L. The geographic distribution of frogs: patterns and predictions (University of Southern California Press. Allan Hancock Foundation, 1973).
93. An, Z., Kutzbach, J.E., Prell, W.L. & Porter, S.C. Evolution of Asian monsoons and phased uplift of the Himalaya–Tibetan plateau since Late Miocene times. *Nature* **411**, 62-66 (2001).
94. Zhu, W.-D., Nie, Z.-L., Wen, J. & Sun, H. Molecular phylogeny and biogeography of *Astilbe* (Saxifragaceae) in Asia and eastern North America. *Bot. J. Linn. Soc.* **171**, 377-394 (2013).
95. Deng, T. *et al.* Does the Arcto-Tertiary biogeographic hypothesis explain the disjunct distribution of northern hemisphere herbaceous plants? The case of *Meehania* (Lamiaceae). *Plos One* **10**, e0117171 (2015).

96. Huang, W.-P. *et al.* Molecular phylogenetics and biogeography of the eastern Asian-eastern North American disjunct Mitchella and its close relative *Damnacanthus* (Rubiaceae, Mitchelleae). *Bot. J. Linn. Soc.* **171**, 395-412 (2013).
97. Nie, Z.L. *et al.* Molecular phylogeny and biogeographic diversification of *Parthenocissus* (Vitaceae) disjunct between Asia and North America. *Am. J. Bot.* **97**, 1342-53 (2010).
98. Ma, Z.-Y. *et al.* Phylogenomics, biogeography, and adaptive radiation of grapes. *Mol. Phylogenet. Evol.* **129**, 258-267 (2018).
99. Leopold, E.B. & MacGinitie, H.D. Development and affinities of Tertiary floras in the Rocky Mountains. *Floristics and Paleoflorists of Asia and Eastern North America* (1972).
100. Wing, S.L. Tertiary vegetation of North America as a context for mammalian evolution in Evolution of Tertiary Mammals of North America, Volume 1: Terrestrial Carnivores, Ungulates, and Ungulatelike Mammals (eds Janis, C.M., Scott, K.N. & Jacobs, L.L.) 37-60 (Cambridge University, 1998).
101. Graham, A. The role of land bridges, ancient environments, and migrations in the assembly of the North American flora. *J. Syst. Evol.* **56**, 405-429 (2018).
102. He, H. *et al.* New 40Ar/39Ar dating results from the Shanwang Basin, eastern China: Constraints on the age of the Shanwang Formation and associated biota. *Phys. Earth Planet. Inter.* **187**, 66-75 (2011).
103. Liu, B.-B., Liu, G.-N., Hong, D.-Y. & Wen, J. *Eriobotrya* belongs to *Rhaphiolepis* (Maleae, Rosaceae): evidence from chloroplast genome and nuclear ribosomal DNA data. *Front. Plant Sci.* **10**, 1731 (2020).
104. Li, J., Wang, S., Yu, J., Wang, L. & Zhou, S. A modified CTAB protocol for plant DNA extraction. *Chinese Bulletin of Botany* **48**, 72 (2013).
105. Chamala, S. *et al.* MarkerMiner 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes. *Appl. Plant Sci.* **3**, 1400115 (2015).

943 106. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic Local Alignment Search
944 Tool. *J. Mol. Biol.* **215**, 403-410 (1990).
945 107. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search
946 programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997).
947 108. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421 (2009).
948 109. Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the
949 organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649 (2012).
950 110. Morgulis, A. *et al.* Database indexing for production MegaBLAST searches. *Bioinformatics* **24**,
951 1757-1764 (2008).
952 111. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence
953 data. *Bioinformatics* **30**, 2114-2120 (2014).
954 112. Andrews, S. FastQC: A quality control tool for high throughput sequence data. accessed August 1,
955 2021 (2018).
956 113. Johnson, M.G. *et al.* HybPiper: Extracting coding sequence and introns for phylogenetics from
957 high‐throughput sequencing reads using target enrichment. *Appl. Plant Sci.* **4**, 1600016
958 http://dx.doi.org/10.3732/apps.1600016 (2016).
959 114. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform.
960 *Bioinformatics* **25**, 1754-1760 (2009).
961 115. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell
962 sequencing. *J. Comput. Biol.* **19**, 455-477 (2012).
963 116. Slater, G.S.C. & Birney, E. Automated generation of heuristics for biological sequence comparison.
964 *BMC Bioinform.* **6**, 31 (2005).
965 117. Nakamura, T., Yamada, K.D., Tomii, K. & Katoh, K. Parallelization of MAFFT for large-scale
966 multiple sequence alignments. *Bioinformatics* **34**, 2490-2492 (2018).
967 118. Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T. trimAl: a tool for automated alignment
968 trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).

969 119. Borowiec, M.L. Spruceup: fast and flexible identification, visualization, and removal of outliers
970 from large multiple sequence alignments. *J. Open Source Softw.* **4**, 1635 (2019).
971 120. Borowiec, M.L. AMAS: a fast tool for alignment manipulation and computing of summary
972 statistics. *PeerJ* **4**, e1660 (2016).
973 121. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
974 phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).
975 122. Mai, U. & Mirarab, S. TreeShrink: fast and accurate detection of outlier long branches in
976 collections of phylogenetic trees. *BMC Genom.* **19**, (2018).
977 123. Stamatakis, A. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with
978 thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690 (2006).
979 124. Lanfear, R., Frandsen, P.B., Wright, A.M., Senfeld, T. & Calcott, B. PartitionFinder 2: new
980 methods for selecting partitioned models of evolution for molecular and morphological
981 phylogenetic analyses. *Mol. Biol. Evol.* **34**, 772-773 (2016).
982 125. Lanfear, R., Calcott, B., Kainer, D., Mayer, C. & Stamatakis, A. Selecting optimal partitioning
983 schemes for phylogenomic datasets. *BMC Evol. Biol.* **14**, 82 (2014).
984 126. Minh, B.Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic inference in the
985 genomic era. *Mol. Biol. Evol.* **37**, 1530-1534 (2020).
986 127. Sayyari, E. & Mirarab, S. Fast Coalescent-Based Computation of Local Branch Support from
987 Quartet Frequencies. *Mol. Biol. Evol.* **33**, 1654-1668 (2016).
988 128. Junier, T. & Zdobnov, E.M. The Newick utilities: high-throughput phylogenetic tree processing in
989 the UNIX shell. *Bioinformatics* **26**, 1669-1670 (2010).
990 129. Brown, J.W., Walker, J.F. & Smith, S.A. Phyx: phylogenetic tools for unix. *Bioinformatics* **33**,
991 1886-1888 (2017).
992 130. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree
993 reconstruction from partially resolved gene trees. *BMC Bioinform.* **19**, 153 (2018).

994 131. Salichos, L., Stamatakis, A. & Rokas, A. Novel information theory-based measures for quantifying
995 incongruence among phylogenetic trees. *Mol. Biol. Evol.* **31**, 1261-1271 (2014).
996 132. G. J. Hodel, R., Zimmer, E. & Wen, J. A phylogenomic approach resolves the backbone of *Prunus*
997 (Rosaceae) and identifies signals of hybridization and allopolyploidy. *Mol. Phylogenet. Evol.*
998 107118 (2021).
999 133. Wang, K. *et al.* Incomplete lineage sorting rather than hybridization explains the inconsistent
1000 phylogeny of the wisent. *Commun. Biol.* **1**, 169 (2018).
1001 134. Yang, Y. *et al.* Prickly waterlily and rigid hornwort genomes shed light on early angiosperm
1002 evolution. *Nat. Plants* **6**, 215-222 (2020).
1003 135. He, J. *et al.* A phylotranscriptome study using silica gel-dried leaf tissues produces an updated
1004 robust phylogeny of Ranunculaceae. (Cold Spring Harbor Laboratory, 2021).
1005 136. Liu, L. & Yu, L. Phybase: an R package for species tree analysis. *Bioinformatics* **26**, 962-963
1006 (2010).
1007 137. Blischak, P.D., Chifman, J., Wolfe, A.D. & Kubatko, L.S. HyDe: a Python package for genome-
1008 scale hybridization detection. *Syst. Biol.* **67**, 821-829 (2018).
1009 138. Huson, D.H. *et al.* Dendroscope: An interactive viewer for large phylogenetic trees. *BMC*
1010 *Bioinform.* **8**, 460 http://dx.doi.org/10.1186/1471-2105-8-460 (2007).
1011 139. Dierckxsens, N., Mardulyn, P. & Smits, G. NOVOPlasty: de novo assembly of organelle genomes
1012 from whole genome data. *Nucleic Acids Res.* **45**, e18-e18 (2016).
1013 140. Zhang, N., Wen, J. & Zimmer, E.A. Congruent deep relationships in the grape family (Vitaceae)
1014 based on sequences of chloroplast genomes and mitochondrial genes via genome skimming. *Plos*
1015 *One* **10**, e0144701 (2015).
1016 141. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357
1017 (2012).

142. Liu, B.-B. *et al.* Phylogenomic analyses of the *Photinia* complex support the recognition of a new genus *Phippsiomeles* and the resurrection of a redefined *Stranvaesia* in Maleae (Rosaceae). *J. Syst. Evol.* **57**, 678-694 (2019).

143. Wang, Y.-B. *et al.* Major clades and a revised classification of *Magnolia* and Magnoliaceae based on whole plastid genome sequences via genome skimming. *J. Syst. Evol.* **58**, 673-695 (2020).

144. Qu, X.-J., Moore, M.J., Li, D.-Z. & Yi, T.-S. PGA: a software package for rapid, accurate, and flexible batch annotation of plastomes. *Plant Methods* **15**, 50 (2019).

145. Lehwark, P. & Greiner, S. GB2sequin-A file converter preparing custom GenBank files for database submission. *Genomics* **111**, 759-761 (2019).

146. MacGinitie, H.D. The Eocene Green River flora of northwestern Colorado and northeastern Utah (University of California Press, 1969).

147. Drummond, A.J., Ho, S.Y.W., Phillips, M.J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, (2006).

148. Stadler, T. Sampling-through-time in birth–death trees. *J. Theor. Biol.* **267**, 396-404 (2010).

149. Heath, T.A., Huelsenbeck, J.P. & Stadler, T. The fossilized birth-death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences* **111**, E2957-E2966 (2014).

150. Rambaut, A., Drummond, A.J., Xie, D., Baele, G. & Suchard, M.A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901-904 (2018).

151. Matzke, N.J. BioGeoBEARS: BioGeography with Bayesian (and likelihood) Evolutionary Analysis with R Scripts. version 1.1.1. November 6, 2018 (GitHub, 2018).

152. Yu, Y., Harris, A.J., Blair, C. & He, X. RASP (Reconstruct Ancestral State in Phylogenies): a tool for historical biogeography. *Mol. Phylogenet. Evol.* **87**, 46-49 (2015).

153. Guo, Z.T. *et al.* A major reorganization of Asian climate by the early Miocene. *Clim. Past* **4**, 153-174 (2008).

154. Donoghue, M.J., Bell, C.D. & Li, J. Phylogenetic Patterns in Northern Hemisphere Plant Geography. *Int. J. Plant Sci.* **162**, S41-S52 (2001).

155. Manos, P.S. & Stanford, A.M. The historical biogeography of Fagaceae: Tracking the tertiary history of temperate and subtropical forests of the Northern Hemisphere. *Int. J. Plant Sci.* **162**, S77-S93 (2001).

156. Xiang, Q.Y. & Soltis, D.E. Dispersal-Vicariance Analyses of Intercontinental Disjuncts: Historical Biogeographical Implications for Angiosperms in the Northern Hemisphere. *Int. J. Plant Sci.* **162**, S29-S39 (2001).

157. Milne, R.I. & Abbott, R.J. The origin and evolution of Tertiary relict floras. *Adv. Bot. Res.* **38**, 281-314 (2002).

158. Milne, R.I. Northern Hemisphere plant disjunctions: a window on tertiary land bridges and climate change? *Ann. Bot.* **98**, 465-472 (2006).

159. Wang, J.J. *et al.* The Biogeographic South-North Divide of Polygonatum (Asparagaceae Tribe Polygonateae) within Eastern Asia and Its Recent Dispersals in the Northern Hemisphere. *Plos One* **11**, e0166134 (2016).

160. Zhou, W.B., Xiang, Q.Y. & Wen, J. Phylogenomics, biogeography, and evolution of morphology and ecological niche of the eastern Asian-eastern North American *Nyssa* (Nyssaceae). *J. Syst. Evol.* **58**, 571-603 (2020).

Table 1. The result of *phyckle* analysis of detecting the gene dataset supporting each bipartition for *Malus coronaria* and *M. sikkimensis*. The result was estimated from the 50%-sample dataset.

conflict	Bipartition	number of genes	sum lnL difference	number non-outlier genes	sum lnL difference: outlier genes removed
M. coronaria	*M. coronaria* bipartition A: (*M. coronaria*,(*M. angustifolia*,*M. ioensis*))|(all other taxa)	445	17028.36	427	14731.52
	M. coronaria bipartition B: (*M. coronaria*,*M. sieversii*)|(all other taxa)	155	4126.411	153	3876.897
M. sikkimensis	*M. sikkimensis* bipartition A: (*M. sikkimensis*,(*M. baccata* var. *xiaojinensis*,*M. orientalis*,*M. sylvestris*,*M. sieversii*,*M.hupehensis*,*M.toringo*,*M. baccata*,*M. rockii*)|(all other taxa)	191	3987.746	188	3647.953
	M. sikkimensis bipartition B: (*M. sikkimensis*,(*M. fusca*,*M. kansuensis*))|(all other taxa)	409	11675.01	393	9168.462

sum lnL difference: the sum of log-likelihood difference; outlier genes indicate the genes for their log-likelihood more than 100.

Table 2. The sporophytic chromosome count number of *Malus* sampled in this study, along with their citations. These 18 species were grouped by Phipps et al. (1991)'s taxonomic system and the clades identified in this study.

clades identified in this study	section sensu Phipps et al. (1990)	species	Sporophytic chromosome count
clade I	*Malus* sect. *Malus*	*M. toringo*	**34, 51**[62-66]
		M. sieversii	**34**[62-65,67]
		M. sylvestris	**34**[63,65,68]
		M. orientalis	-
		M. baccata	**34**[62-64,66,67,69-71]
		M. rockii	**34, 51**[62-64,67]
		M. hupehensis	**34, 51, 68**[62-67]
		M. baccata var. *xiaojinensis*	**68**[63,64]
		M. sikkimensis	**34, 68**[62-64,66,67,72]
	Malus sect. *Sorbomalus*	*M. fusca*	**34**[3,65]
		M. kansuensis	**34**[62,63,65,66]
		M. angustifolia	**34, 68**[3]
clade II	*Malus* sect. *Chloromeles*	*M. coronaria*	**34, 51, 68**[3,65]
		M. ioensis	**34, 51, 68**[3,65]
clade III	*Malus* sect. *Sorbomalus*	*M. florentina*	**34**[65,73]
	Malus sect. *Eriolobus*	*M. trilobata*	**34**[65,73]
	Malus sect. *Docyniopsis*	*M. doumeri*	**34**[63,64]
	Docynia	*D. delavayi*	-

1076 Table 3. Fossil records of *Malus*.

Age	Fossil	fossil organ	geographic origin	reference
Eocene (47.8-41.2 Mya)	*Malus collardii* Axelrod	leaves	North America (Thunder Mountain, Idaho, USA)	Axelrod, 1998
Eocene (47.8-41.2 Mya)	*Malus kingiensis* Budants	leaves	Eurasia (Rebro Cape, western Kamchatka peninsula, Kamchatka Territory, Russian Federation)	Budantsev, 2006
Eocene (37.71-33.9 Mya)	*Malus florissantensis* (Cockerell) MacGinitie	leaves	North America (Green River Formation, Florissant, Colorado, USA)	Cockerell, 1908; MacGinitie, 1953
Eocene (37.71-33.9 Mya)	*Malus pseudocredneria* (Cockerell) MacGinitie	leaves	North America (Green River Formation, Florissant, Colorado, USA)	Cockerell, 1908; MacGinitie, 1969
Miocene (23.03-14.18 Mya)	*Malus idahoensis* R.W.Br.	leaves	North America (G. W. Oliver coal mine, Salmon, Idaho, USA)	Brown, 1935
Miocene (17 Mya)*	*Malus parahupehensis* J.Hsu & R.W.Chaney	leaves	NE China (Shanwang, Shandong, China)	Hsu & Chaney, 1940
Pliocene (5.33-2.58 Mya)	*Malus obensis* M.G.Gorbunov	fruits	Eurasia (W Siberia)	Gorbunov, 1959
Pliocene (5.33-2.58 Mya)	*Malus antiqua* Doweld = *Malus pulcherrima* Givulescu *nom. illeg.*	leaves	Europe (Chiuzbaia, Județul Maramureș, Romania)	Doweld (2018)
Pleistocene (3.2-2.5 Mya)	*Malus pseudoangustifolia* E.W.Berry	leaves	North America (right bank of Neuse River 4,5 miles above Seven Springs, Wayne County, North Carolina, USA)	Berry, 1926

1077 * The timing of the sedimentary sequence of the Shanwang Formation was estimated from $^{40}Ar/^{39}Ar$ analysis of the basalts[102].

1078

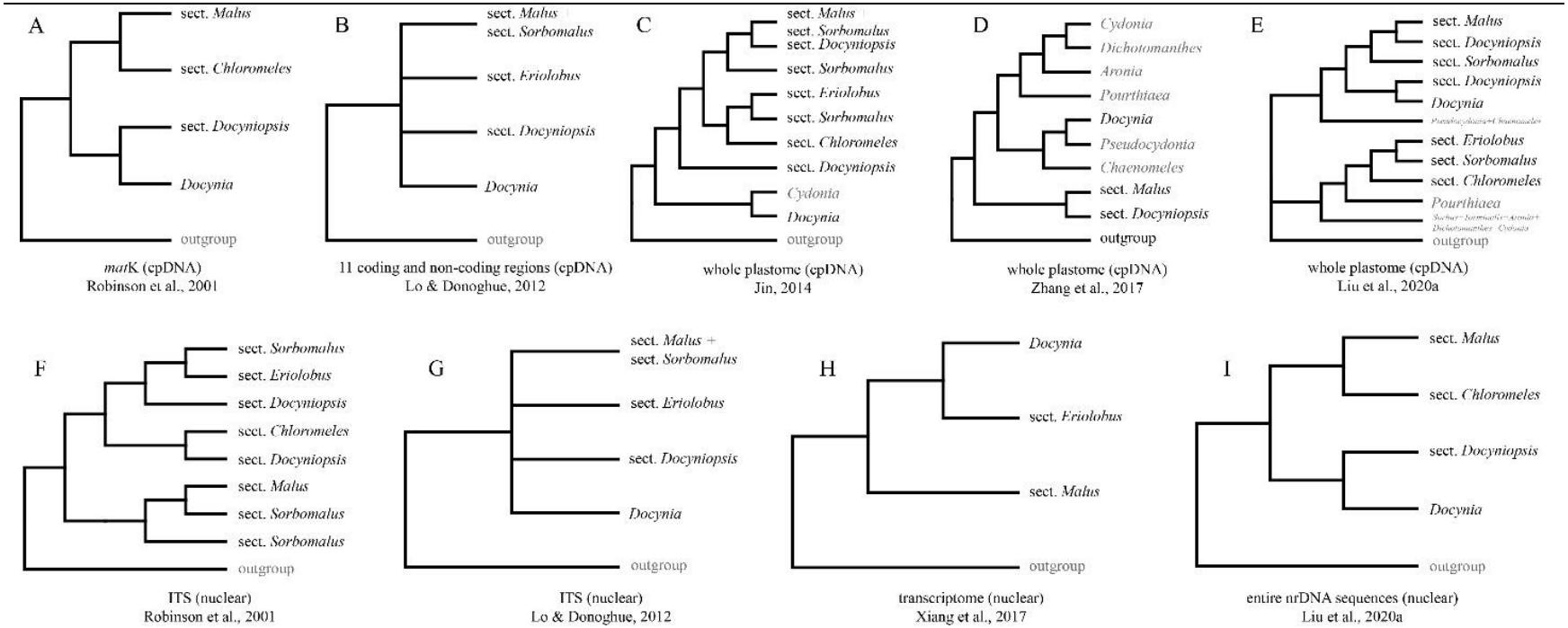

Fig. 1. Phylogenetic hypotheses among sections of *Malus* s.l. estimated by previous studies. **a**, plastid *matK* sequence[17]; **b**, 11 plastid coding and non-coding regions[18]; **c**, whole plastome[19]; **d**, whole plastome[20]; **e**, whole plastome[21]; **f**, nuclear ITS sequence[17]; **g**, nuclear ITS sequence[18]; **h**, transcriptome[22]; **i**, entire nrDNA sequences[21]. The sectional delimitation in *Malus* s.l. followed Phipps et al. (1990).

1079

1080

1081

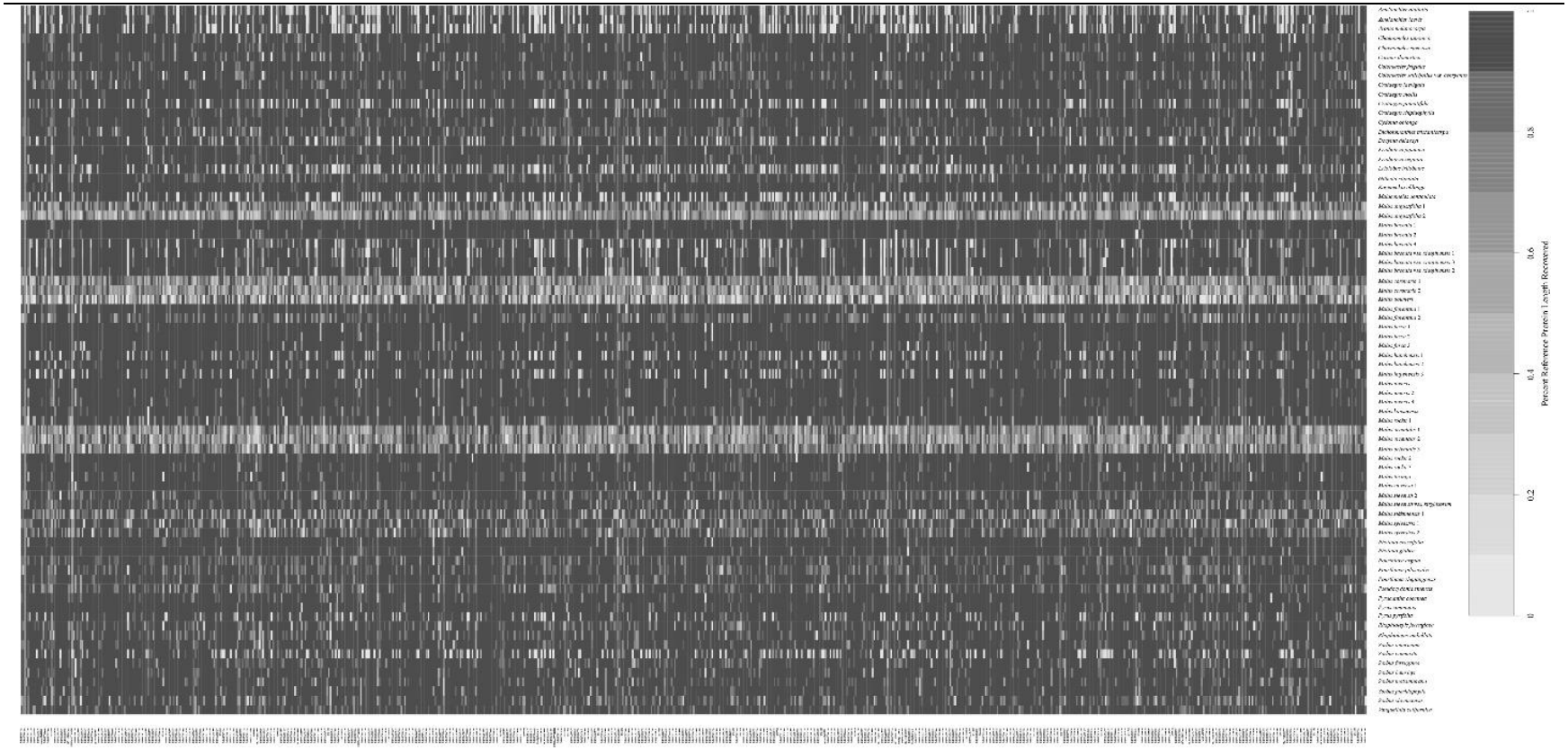

Fig. 2. Heat map showing recovery efficiency for 797 genes recovered by HybPiper. Each column is a gene, and each row is one sample. The shade of gray in the cell is determined by the length of sequence recovered by the pipeline, divided by the length of the reference gene (maximum of 1.0).

1082

1083

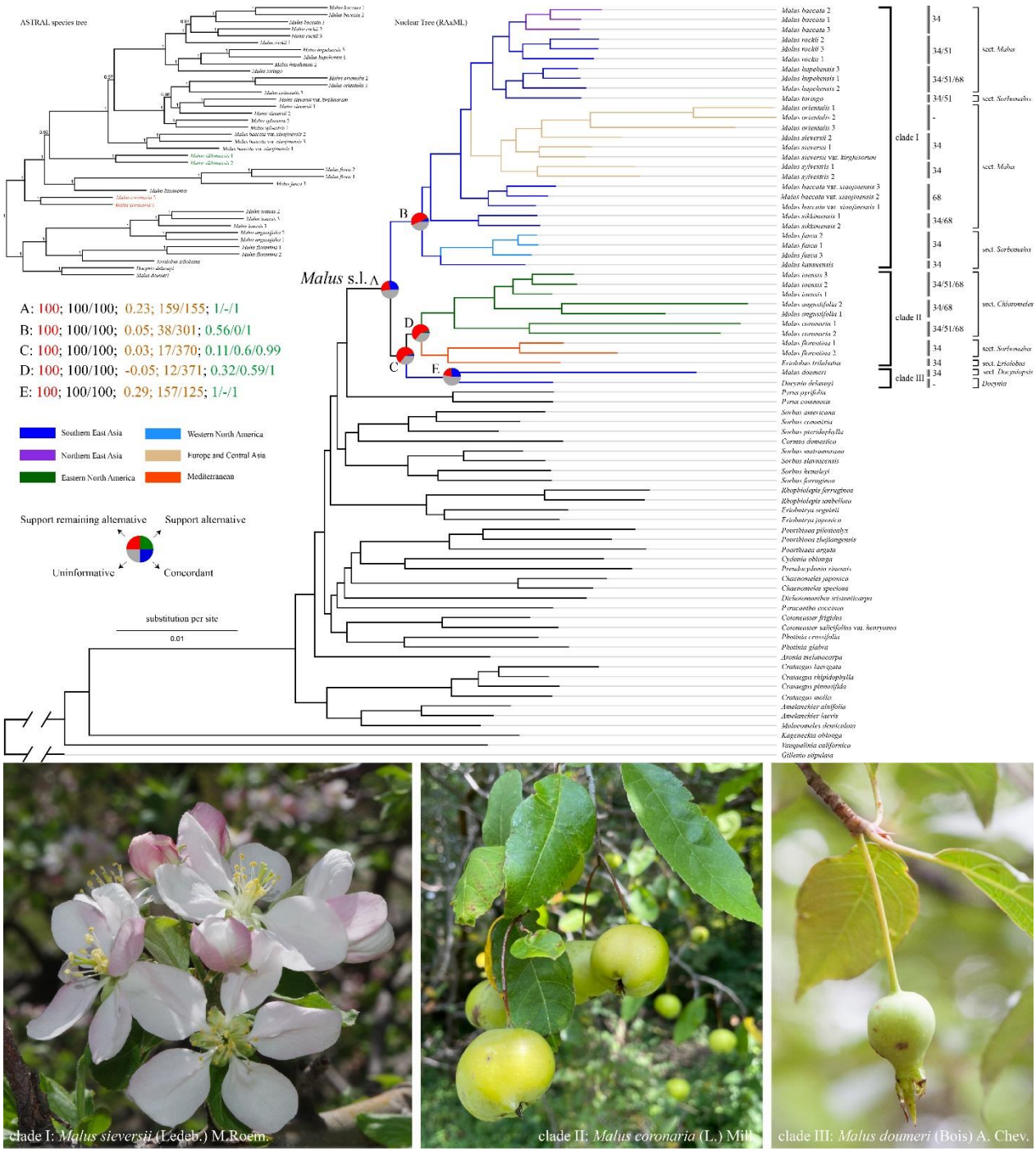

Fig. 3. Maximum likelihood phylogeny of *Malus* s.l. in the framework of Maleae inferred from RAxML analysis of the concatenated 80%-sample dataset. Pie charts on the focused five nodes (A, B, C, D, and E) present the proportion of gene trees that support that clade (blue), the proportion that support the main alternative bifurcation (green), the proportion that

support the remaining alternatives (red), and the proportion (conflict or support) that have < 50% bootstrap support (gray). All the other pie charts refer to Fig. S11 available from Dryad. The numbers (top left) indicate values associated with those nodes; they are bootstrap support values estimated from RAxML analysis (e.g., A: 100 labled by red; see Fig. S4 available on Dryad for all nodes BS), the SH-aLRT support and Ultrafast Bootstrap (UFBoot) support estimated from IQ-TREE2 (e.g., A: 100/100 labled by black; see Fig. S5 available on Dryad for all nodes support), the Internode Certainty All (ICA) score, the number of gene trees concordant/conflicting with that node in the species tree estimated from *phyparts* (e.g., 0.23; 159/155 labled by orange; see Fig. S11 available on Dryad for all values), and Quartet Concordance/Quartet Differential/Quartet Informativeness estimated from Quartet Samping analysis (e.g., 1/-/1 labled by green; see Fig. S10 available on Dryad for all scores). Branches are colored by their distribution, i.e., dark blue, Southern East Asia; purple, Northern East Asia; green, Eastern North America; light blue, Western North America; yellowish-brown, Europe and Central Asia; red, the Mediterranean. The sporophytic chromosome number is displayed to the right of each species label (see Table 2 for details). The Photo credits: clade I (*Malus sieversii*): Pan Li; clade II (*M. coronaria*): Richard G.J. Hodel; clade III (*M. doumeri*): Bin-Bin Liu.

1084

1085

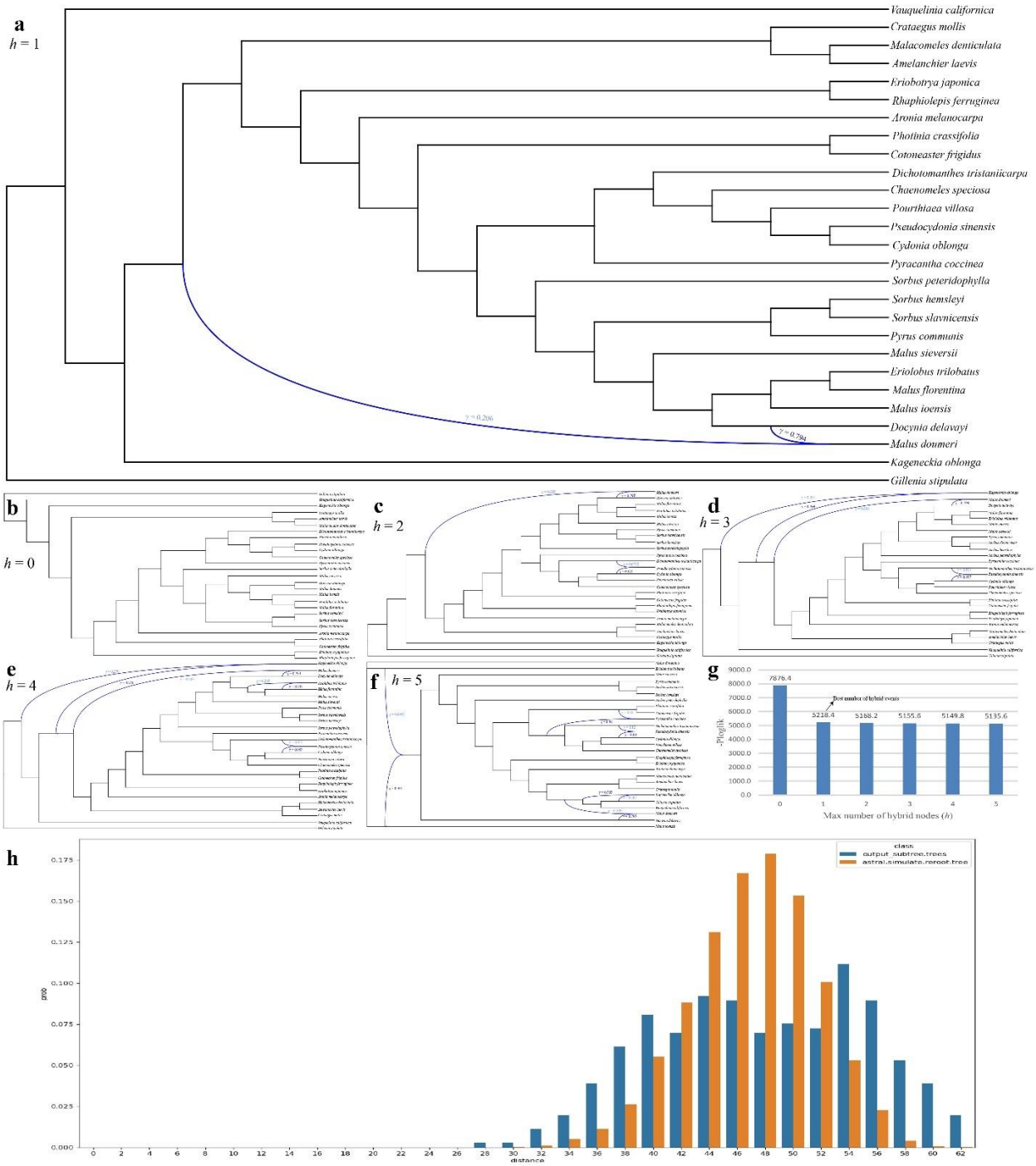

Fig. 4. Coalescent simulation and phylogenetic network analysis from the 27-taxa sampling at the tribe level of Maleae. **a-f**, Species networks inferred from SNaQ network analysis with 1 to 5 maximum number of reticulations. **g**, The pseudo-loglikelihood scores (-ploglik) in a bar chart indicated that *hmax* = 1 was the optimal network (A). **h**, Distribution of tree-to-tree distances between empirical gene trees and the ASTRAL species tree, compared to those from the coalescent simulation. Blue curved branches indicate the possible hybridization event. Dark blue and light blue numbers indicate the major and minor inheritance probabilities of hybrid nodes.

1086

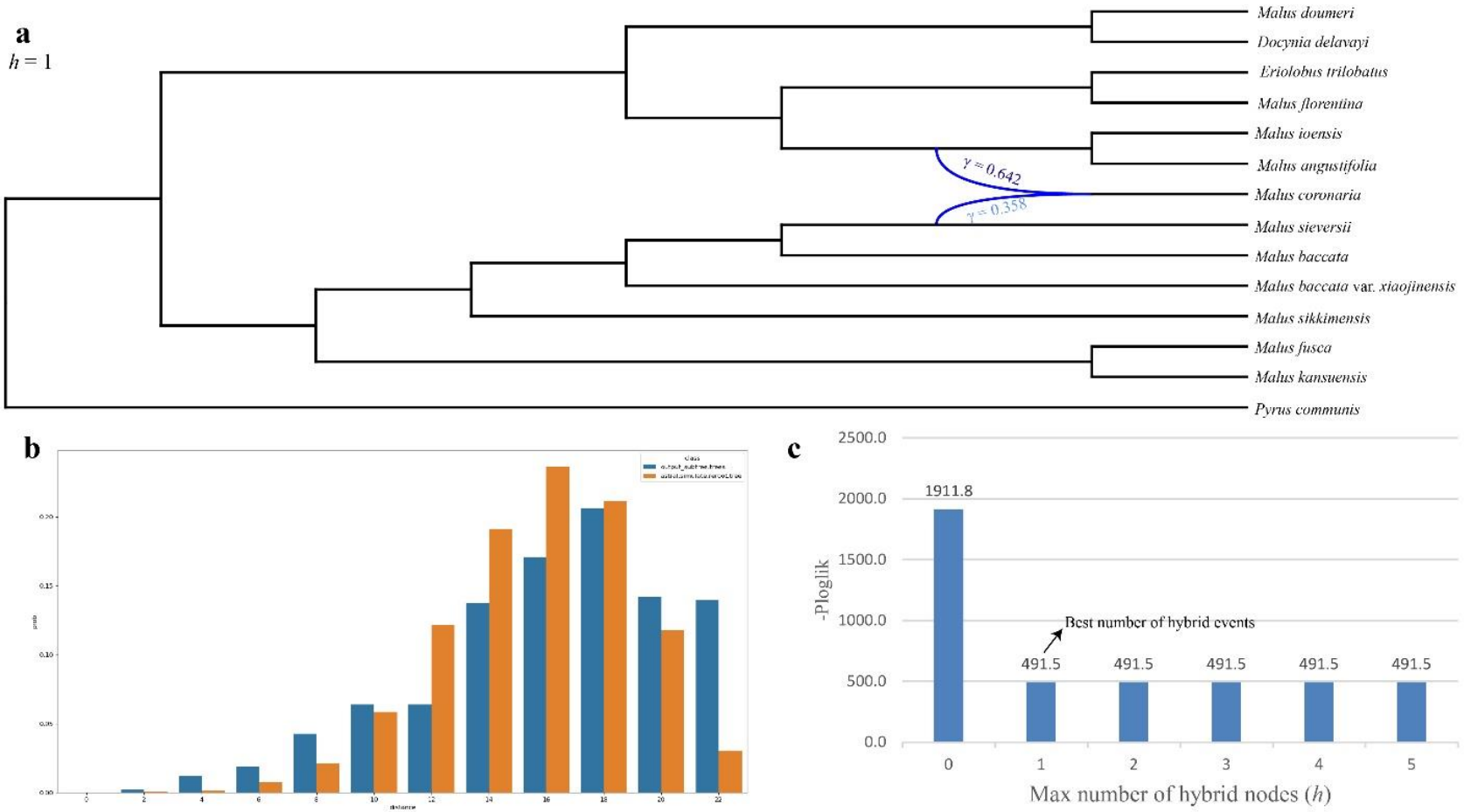

Fig. 5. Coalescent simulation and phylogenetic network analysis from the 14-taxa sampling at the genus level of *Malus*. **a**, Species networks inferred from SNaQ network analysis with *hmax* = 1 as the optimal network. **b**, Distribution of tree-to-tree distances between empirical gene trees and the ASTRAL species tree, compared to those from the coalescent simulation. **c**, The pseudo-loglikelihood scores (-ploglik) of one to five maximum number of reticulations. Blue curved branches indicate the possible hybridization event. Dark blue and light blue numbers indicate the major and minor inheritance probabilities of hybrid nodes.

1089

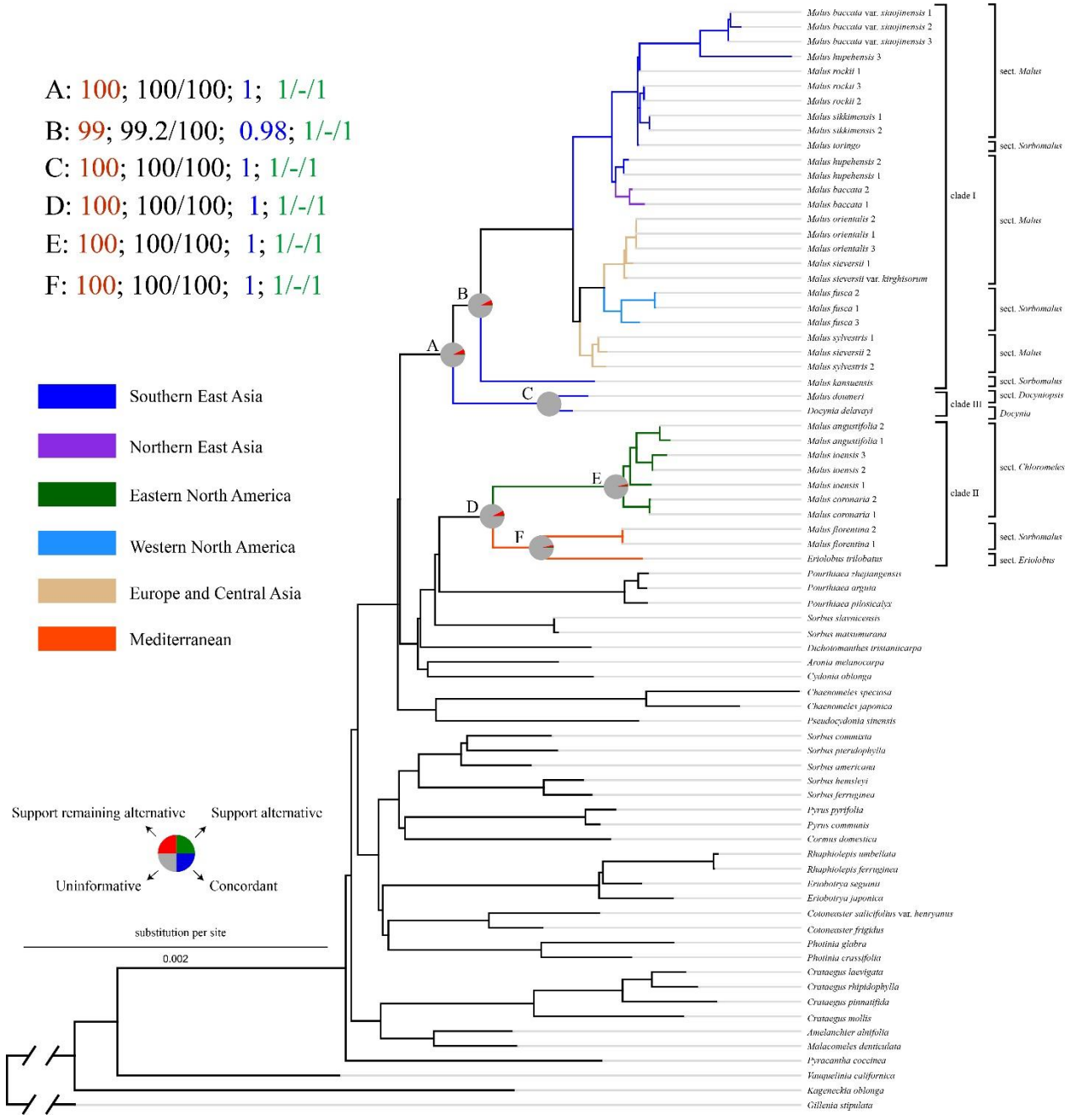

Fig. 6. Maximum likelihood phylogeny of *Malus* s.l. in the framework of Maleae inferred from RAxML analysis of 80 plastid coding regions. Pie charts on the focused five nodes (A, B, C, D, E, and F) present the proportion of gene trees that support that clade (blue), the proportion that support the main alternative bifurcation (green), the proportion that support the remaining alternatives (red), and the proportion (conflict or support) that have < 50%

bootstrap support (gray). All the other pie charts refer to Fig. S16, available from Dryad. The numbers (top left) indicate values associated with those nodes; they are bootstrap support values estimated from RAxML analysis (e.g., A: 100 labled by red; see Fig. S12 available on Dryad for all nodes BS), the SH-aLRT support and Ultrafast Bootstrap (UFBoot) support estimated from IQ-TREE2 (e.g., A: 100/100 labled by black; see Fig. S13 available on Dryad for all nodes support), the local posterior probability (LPP) estimated from ASTRAL-III (e.g., A: 1 labled by blue; see Fig. S14 available on Dryad for all LPP), and Quartet Concordance/Quartet Differential/Quartet Informativeness estimated from Quartet Samping analysis (e.g., 1/-/1 labled by green; see Fig. S15 available on Dryad for all scores). Branches are colored by their distribution, i.e., dark blue, Southern East Asia; purple, Northern East Asia; green, Eastern North America; light blue, Western North America; yellowish-brown, Europe and Central Asia; red, the Mediterranean.

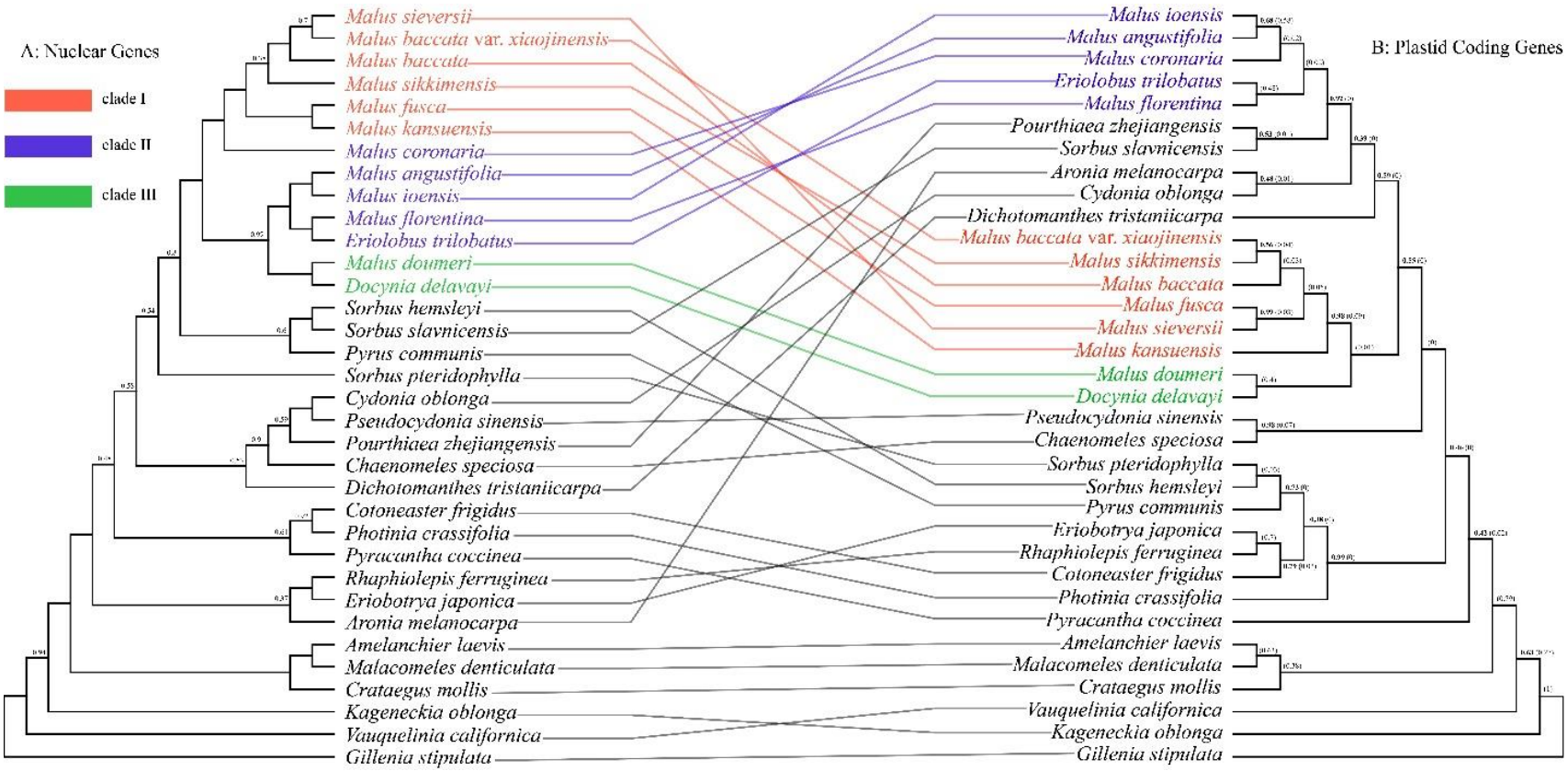

Fig. 7. Tanglegram illustrating the cytonuclear discordance. Left: ASTRAL species tree based on SCN genes; right: plastid tree estimated from 80 coding genes. All nodes have maximum support (LPP = 1) unless rooted. Numbers in the brackets in the plastid tree show the contribution of ILS to the conflicts between nuclear and plastid gene trees based on the multispecies coalescent model.

1099

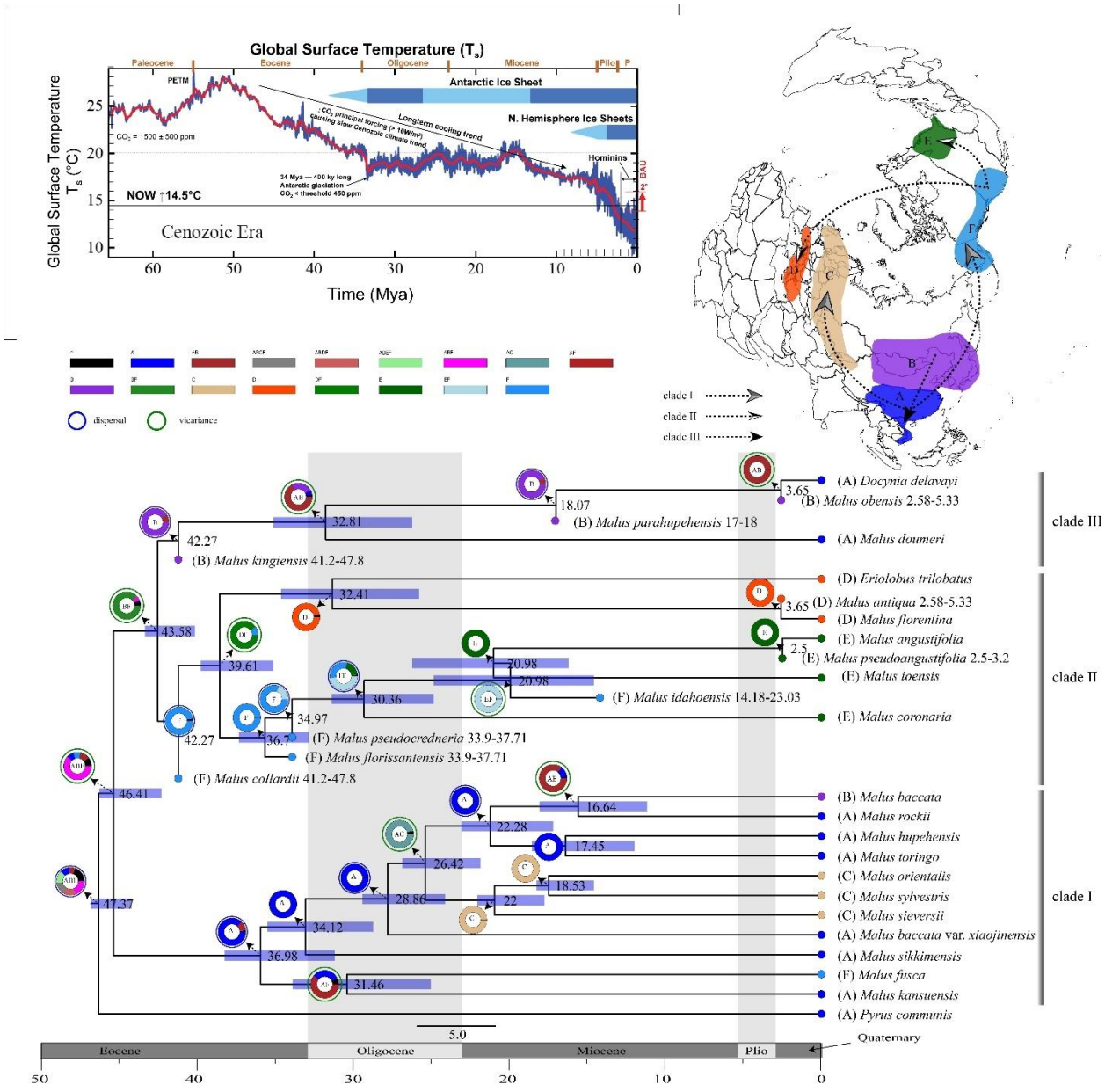

Fig. 8. Dated chronogram for the apple genus *Malus* inferred from BEAST with the Fossilized Birth-Death process based on the 19-taxa nuclear dataset. Also shown is the ancestral area reconstruction using BioGeoBEARS implemented in RASP, with the colored key identifying extant and possible ancestral ranges (upper-right map), (A), Southern East Asia; (B), Northern East Asia; (C), Europe and Central Asia; (D), Mediterranean; (E), Eastern North America; (F), Western North America. The upper-left chart displays a global surface temperature, indicating the major global climate trends in the Cenozoic Era (adapted from https://www.alpineanalytics.com/Climate/DeepTime/WebDownloadImages/CenozoicTsGlobal-7.5w.600ppi.png).

Supplementary files

Phylogenomic analyses in the apple genus *Malus* s.l. reveal widespread hybridization and allopolyploidy driving the diversifications, with insights into the complex biogeographic history in the Northern Hemisphere

Running title: Phylogenomics and biogeography of the apple genus (*Malus*)

Bin-Bin Liu[1,2*], Chen Ren[3,4], Myounghai Kwak[5], Richard G.J. Hodel[2], Chao Xu[1], Jian He[6], Wen-Bin Zhou[7], Chien-Hsun Huang[8], Hong Ma[9], Guan-Ze Qian[10], De-Yuan Hong[1], Jun Wen[2*]

[1]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China

[2]Department of Botany, National Museum of Natural History, Smithsonian Institution, PO Box 37012, Washington, DC 20013-7012, USA

[3]Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, Guangdong, China.

[4]Guangdong Provincial Key Laboratory of Applied Botany, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, Guangdong, China.

[5]National Institute of Biological Resources, Incheon 22689, South Korea.

[6]School of Ecology and Nature Conservation, Beijing Forestry University, Beijing, 100083 PR China.

[7]Department of Plant and Microbial Biology, North Carolina State University, Raleigh, NC 27965, USA.

[8]State Key Laboratory of Genetic Engineering and Collaborative Innovation Center of Genetics and

Development, Ministry of Education Key Laboratory of Biodiversity and Ecological Engineering, Institute of Plant Biology, Center of Evolutionary Biology, School of Life Sciences, Fudan University, Shanghai 200433, China.

[9]Department of Biology, Huck Institutes of the Life Sciences, Pennsylvania State University, 510D Mueller Laboratory, University Park, PA 16802 USA.

[10]College of Life Sciences, Liaocheng University, Liaocheng 252059, China

Correspondance: Bin-Bin Liu (liubinbin@ibcas.ac.cn) or Jun Wen (wenj@si.edu)

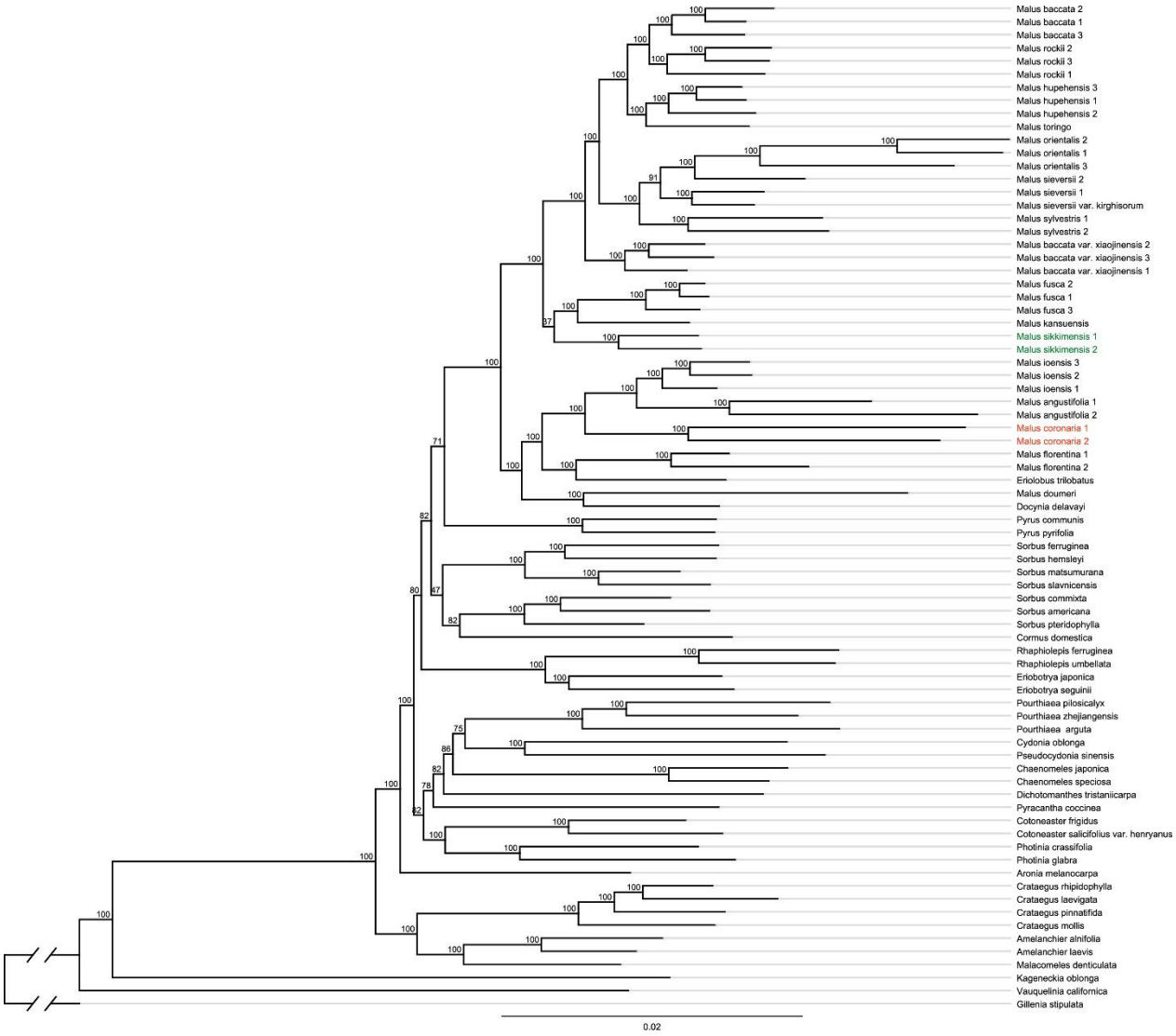

Fig. S1. Maximum likelihood phylogeny of *Malus* s.l. in the framework of Maleae inferred from RAxML analysis of the concatenated 50%-sample dataset. Numbers above the branches indicate the bootstrap support.

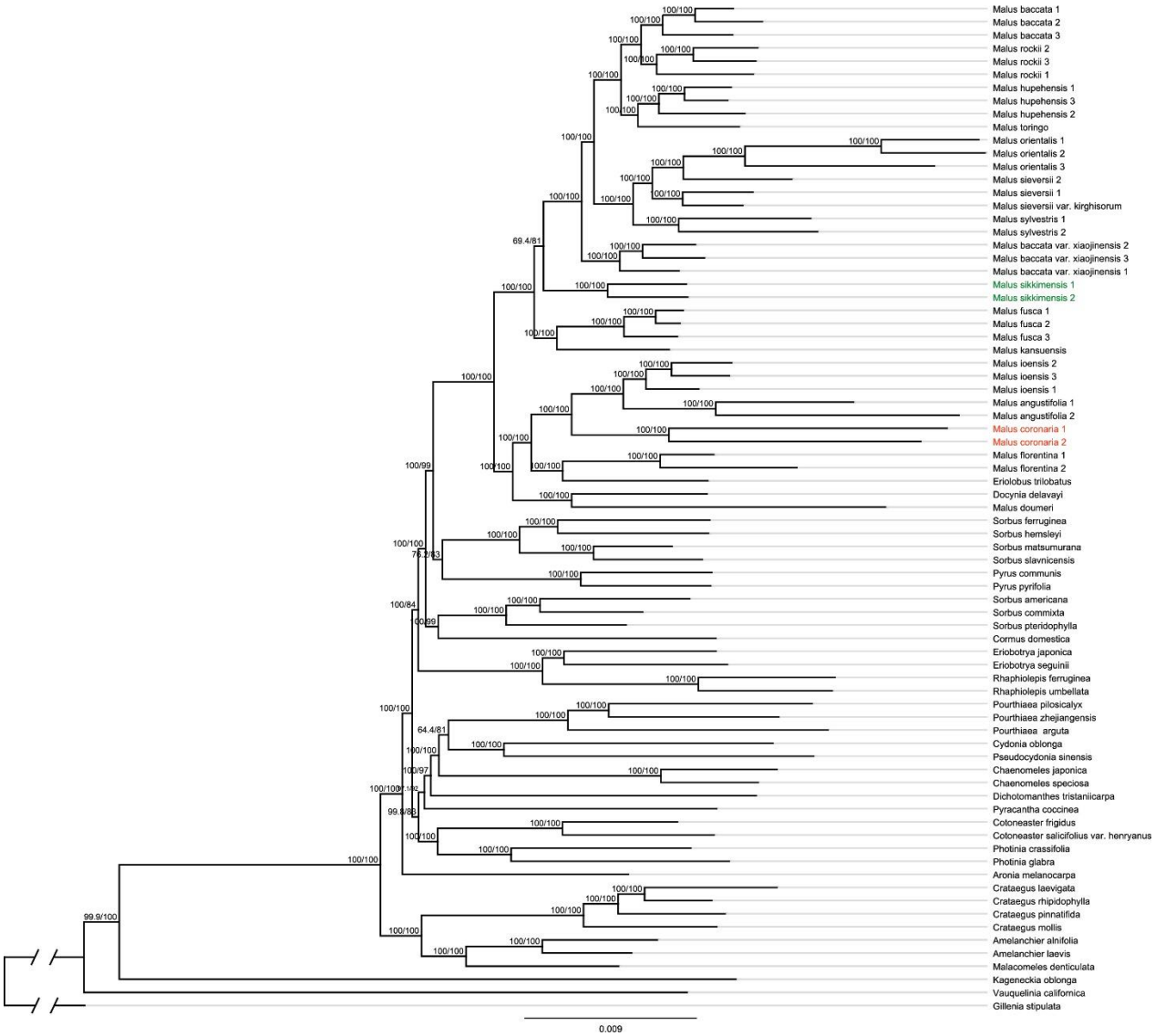

Fig. S2. Maximum likelihood phylogeny of *Malus* s.l. in the framework of Maleae inferred from IQ-TREE2 analysis of the concatenated 50%-sample dataset. Numbers above the branches indicate the SH-aLRT support and Ultrafast Bootstrap support.

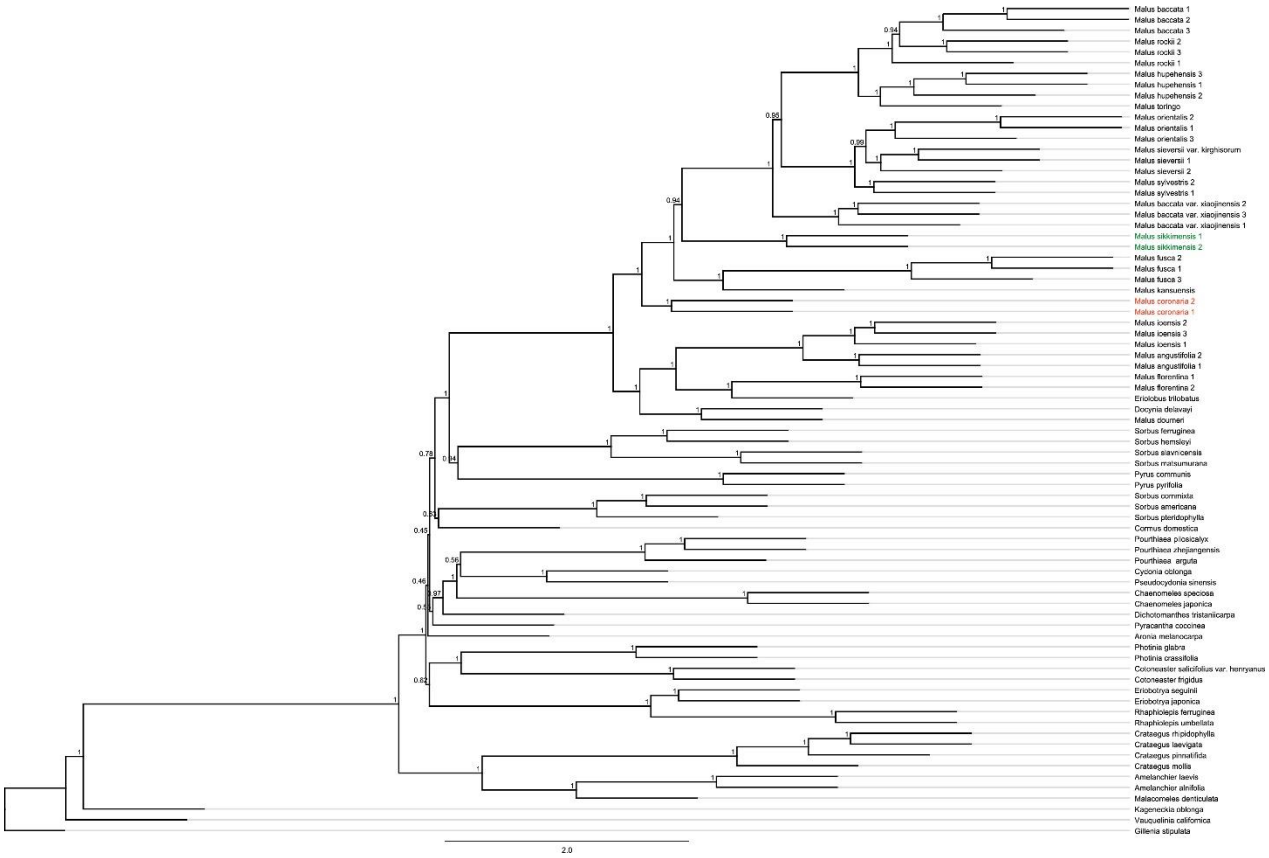

Fig. S3. Species tree of *Malus* s.l. in the framework of Maleae inferred from ASTRAL-III of the concatenated 50%-sample dataset. Numbers above the branches indicate the branch support values measuring the support for a local posterior possibility.

30

31

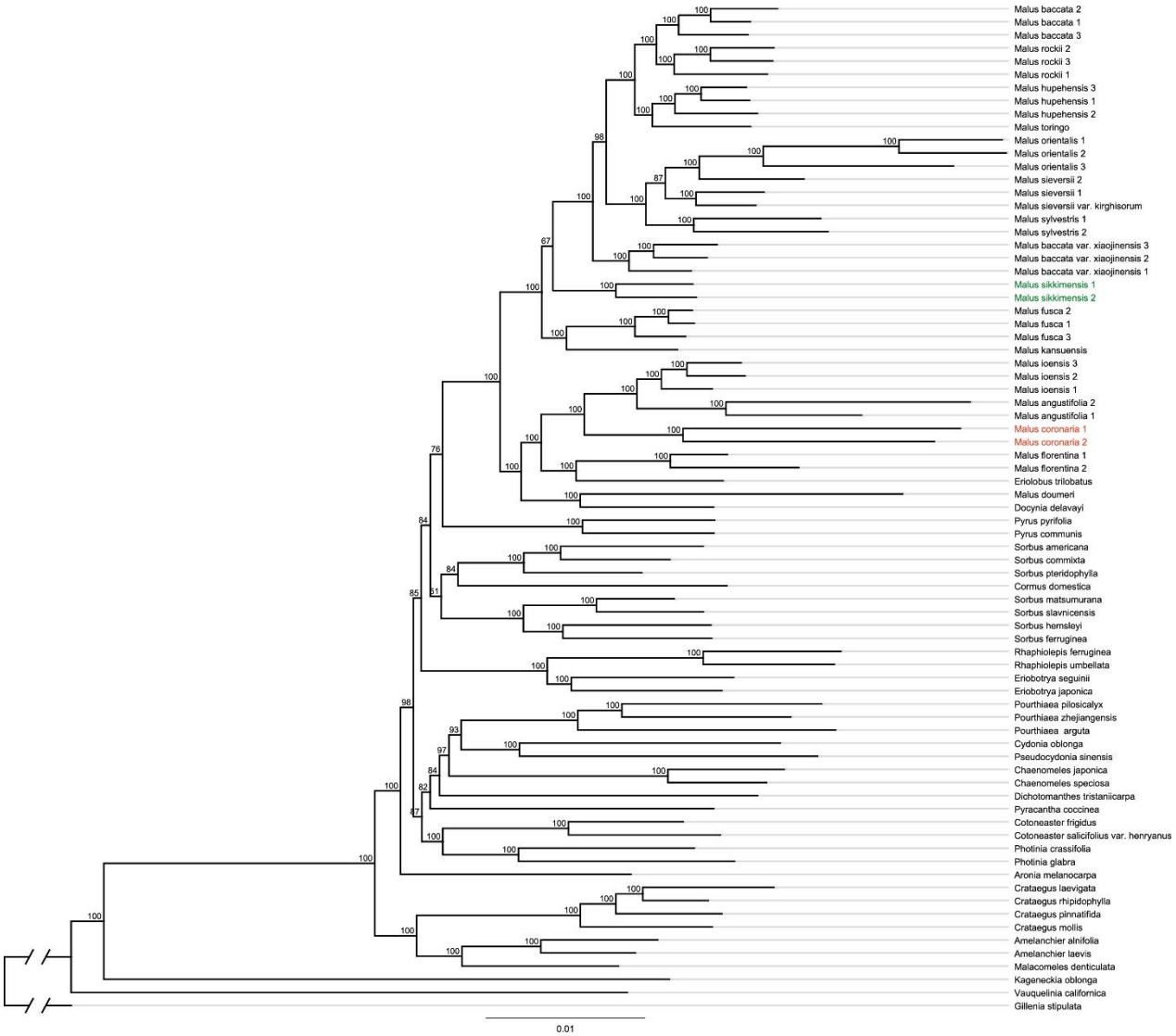

Fig. S4. Maximum likelihood phylogeny of *Malus* s.l. in the framework of Maleae inferred from RAxML analysis of the concatenated 80%-sample dataset. Numbers above the branches indicate the bootstrap support (BS).

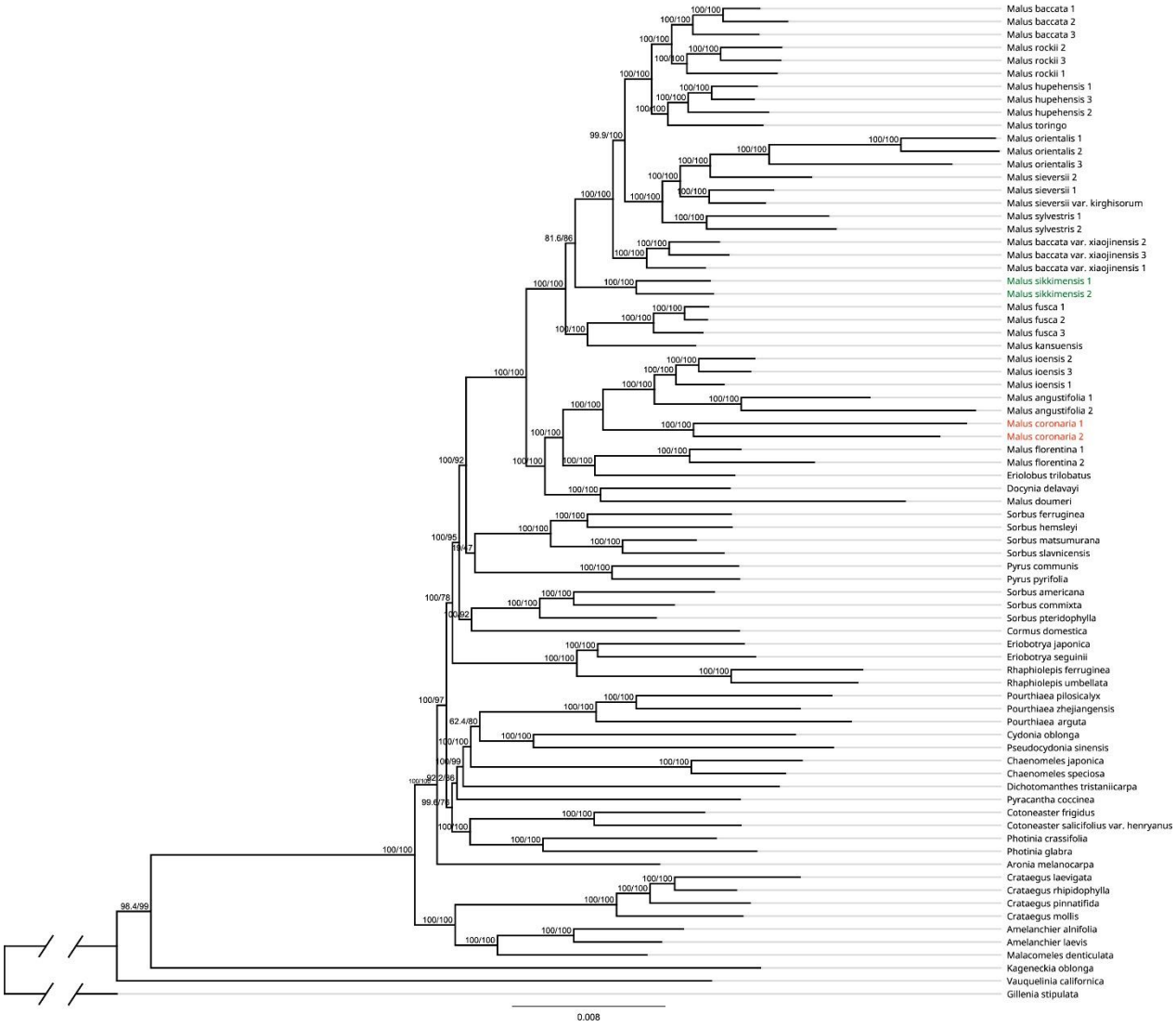

Fig. S5. Maximum likelihood phylogeny of *Malus* s.l. in the framework of Maleae inferred from IQ-TREE2 analysis of the concatenated 80%-sample dataset. Numbers above the branches indicate the SH-aLRT support and Ultrafast Bootstrap (UFBoot) support.

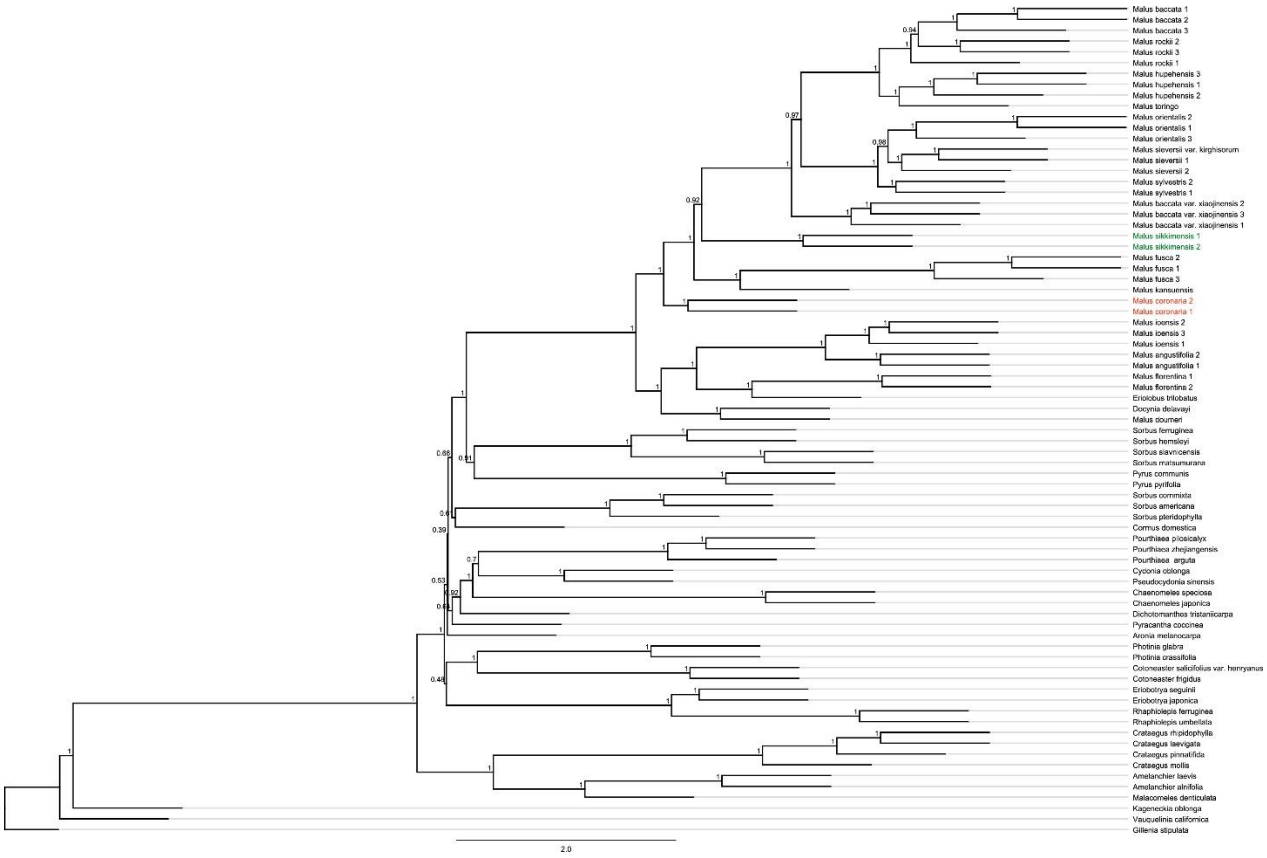

Fig. S6. Species tree of *Malus* s.l. in the framework of Maleae inferred from ASTRAL-III of the concatenated 50%-sample dataset. Numbers above the branches indicate the branch support values measuring the support for a local posterior possibility.

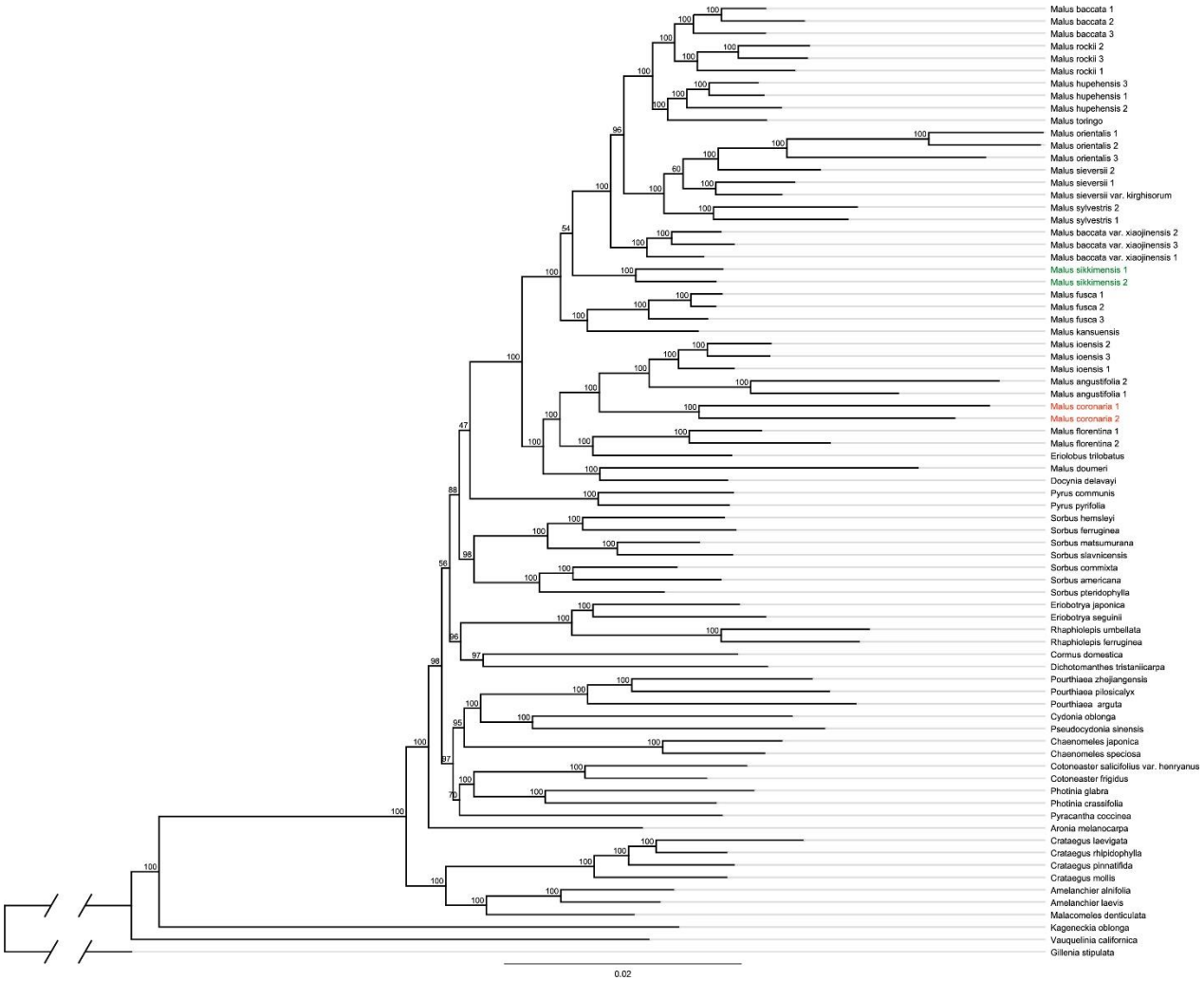

Fig. S7. Maximum likelihood phylogeny of *Malus* s.l. in the framework of Maleae inferred from RAxML analysis of the all-sample dataset. Numbers above the branches indicate the bootstrap support (BS).

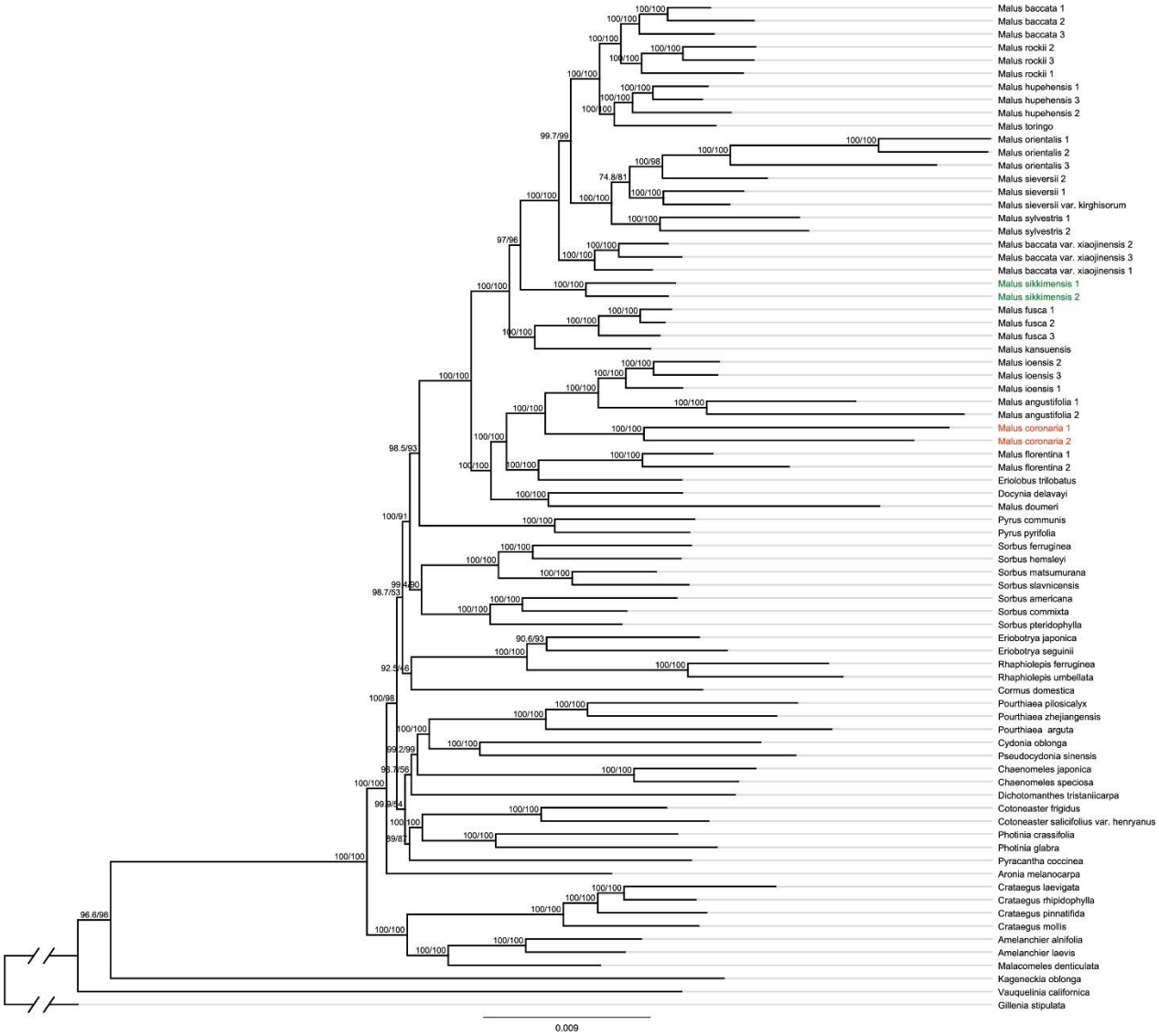

Fig. S8. Maximum likelihood phylogeny of *Malus* s.l. in the framework of Maleae inferred from IQ-TREE2 analysis of the all-sample dataset. Numbers above the branches indicate the SH-aLRT support and Ultrafast Bootstrap (UFBoot) support.

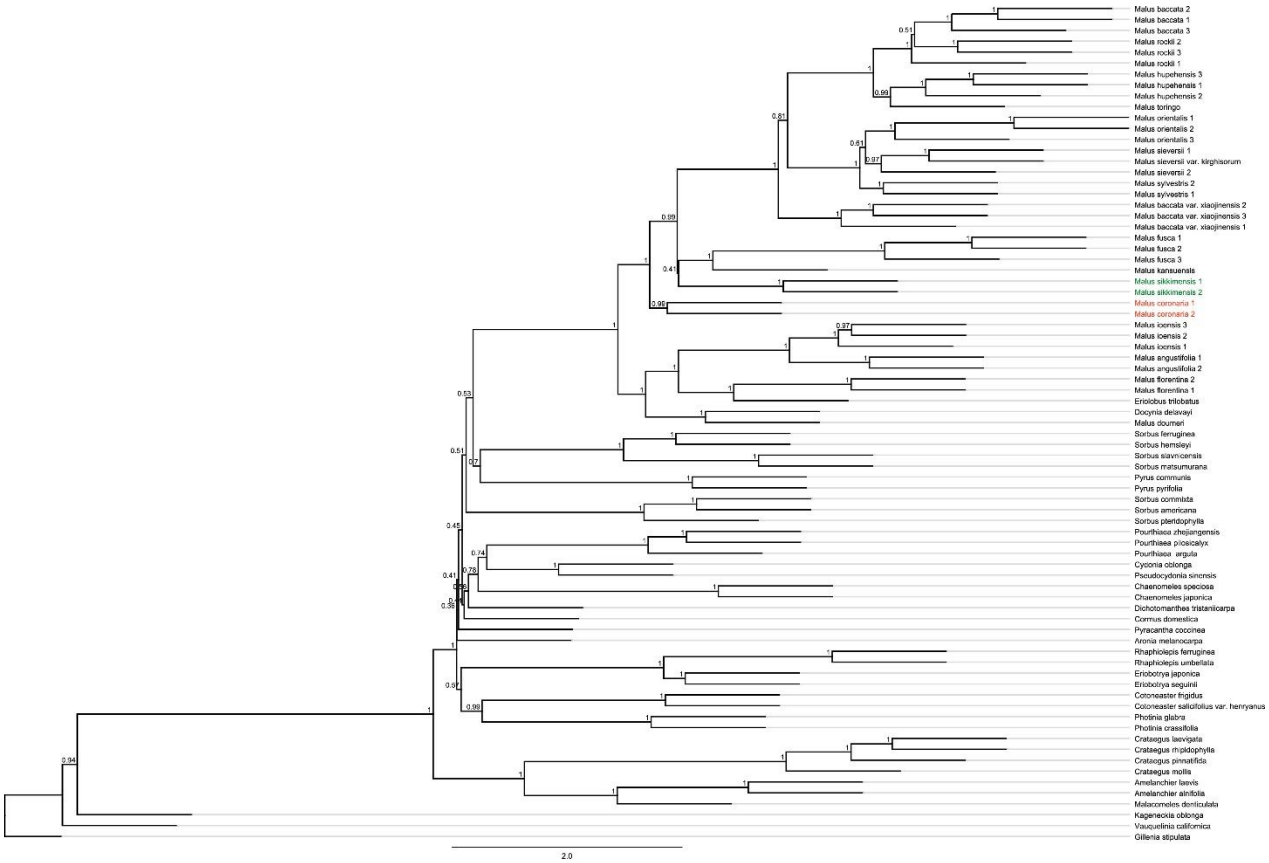

Fig. S9. Species tree of *Malus* s.l. in the framework of Maleae inferred from ASTRAL-III of the all-sample dataset. Numbers above the branches indicate the branch support values measuring the support for a local posterior possibility.

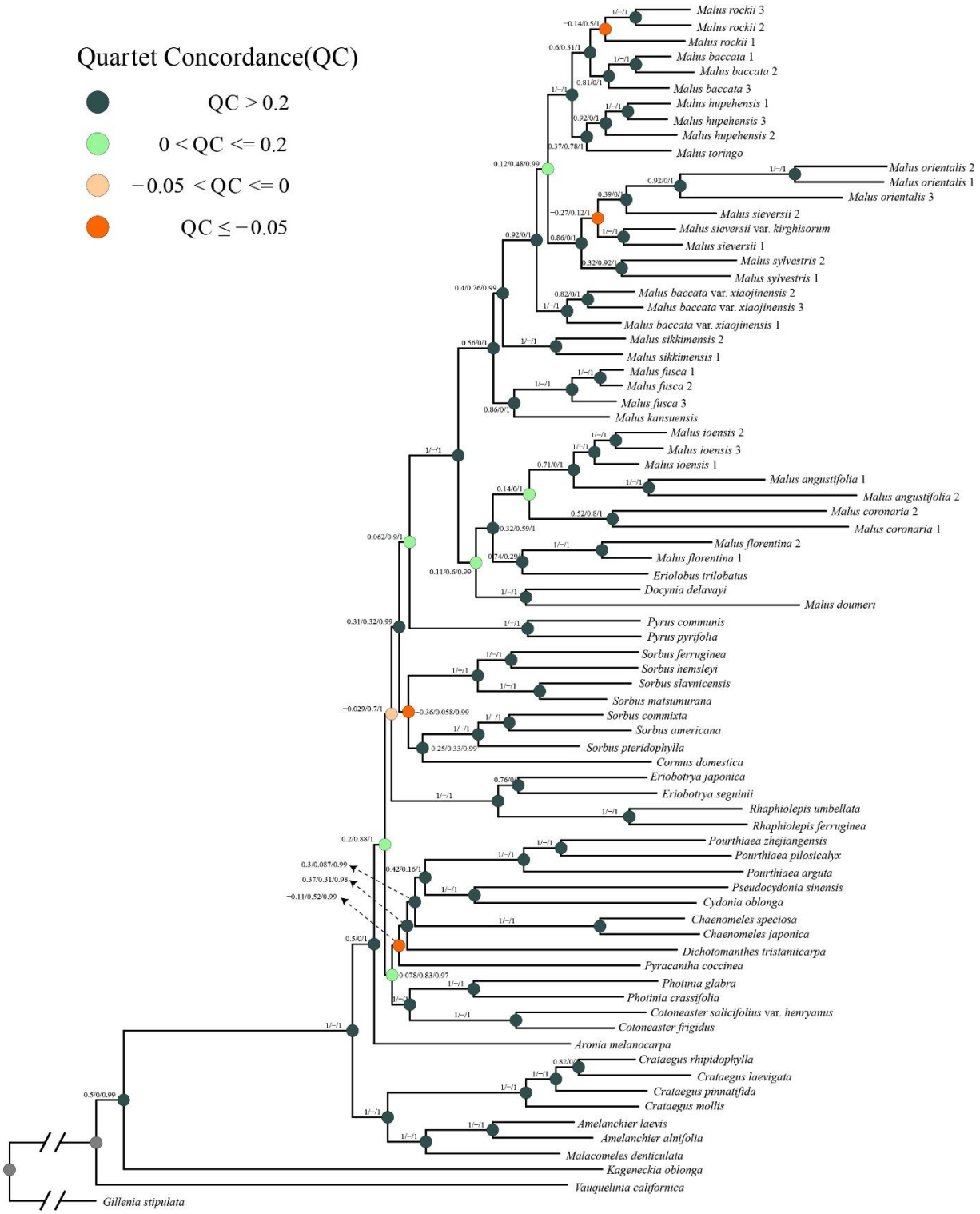

Fig. S10. Maximum likelihood phylogeny of *Malus* s.l. in the framework of Maleae inferred from RAxML analysis of the concatenated 80%-sample dataset. Quartet Samping Scores are shown on branches indicating Quartet Concordance/Quartet Differential/Quartet Informativeness. Quartet Concordance is color-coded according to the legend.

38

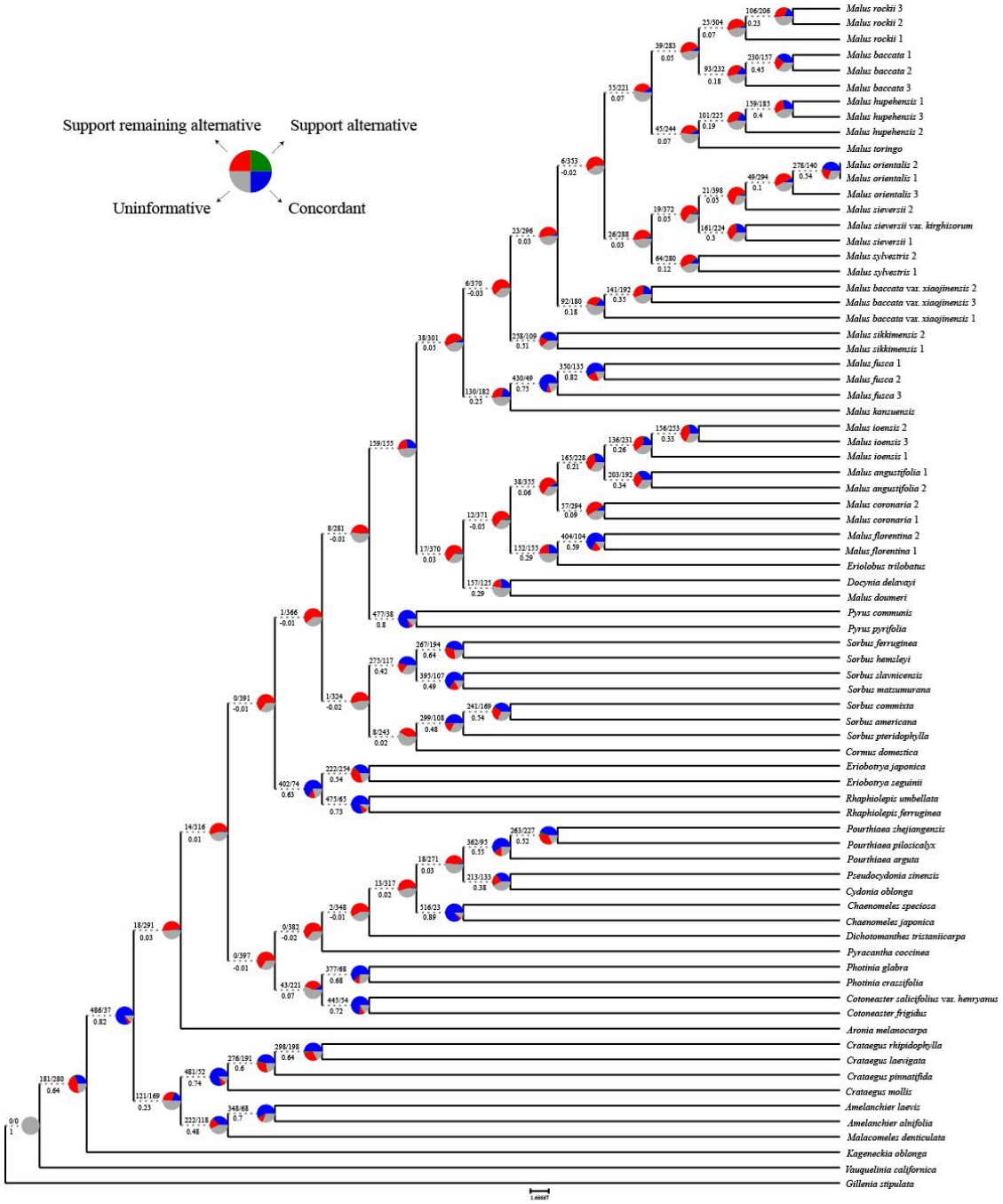

Fig. S11. Maximum likelihood cladogram of *Malus* s.l. inferred from RAxML analysis of the concatenated 80%-sample dataset. Numbers above branches indicate the number of gene trees concordant/conflicting with that node in the species tree. Numbers below the branches are the Internode Certainty All (ICA) score. Pie charts on nodes present the proportion of gene trees that support that clade (blue), the proportion that support the main alternative bifurcation (green), the proportion that support the remaining alternatives (red), and the proportion (conflict or support) that have < 50% bootstrap support (gray).

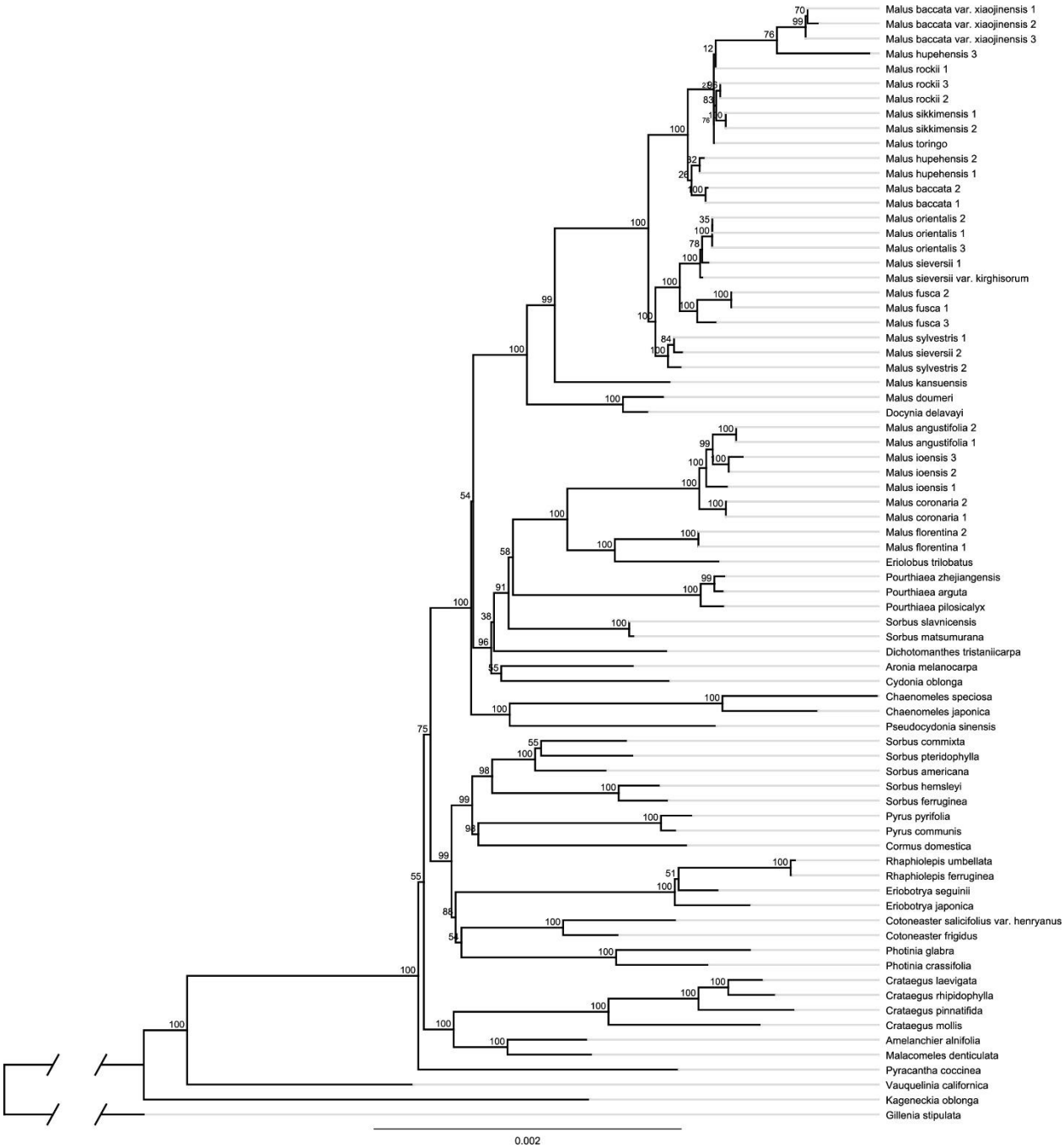

Fig. S12. Maximum likelihood phylogeny of *Malus* s.l. in the framework of Maleae inferred from RAxML analysis of the concatenated 80 plastid coding genes. Numbers above the branches indicate the bootstrap support.

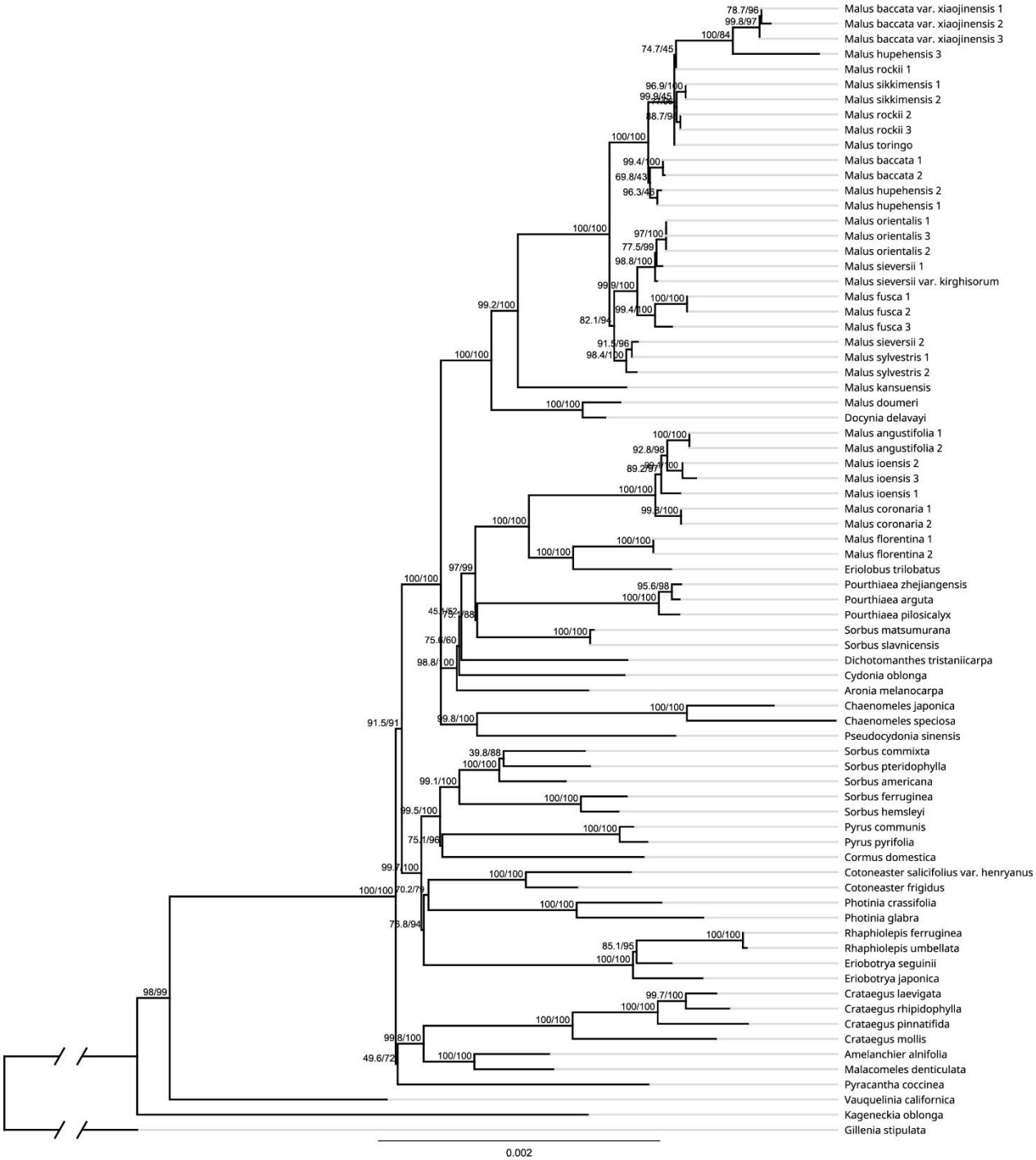

Fig. S13. Maximum likelihood phylogeny of *Malus* s.l. in the framework of Maleae inferred from IQ-TREE2 analysis of the concatenated 80 plastid coding genes. Numbers above the branches indicate the SH-aLRT support and Ultrafast Bootstrap support.

42

43

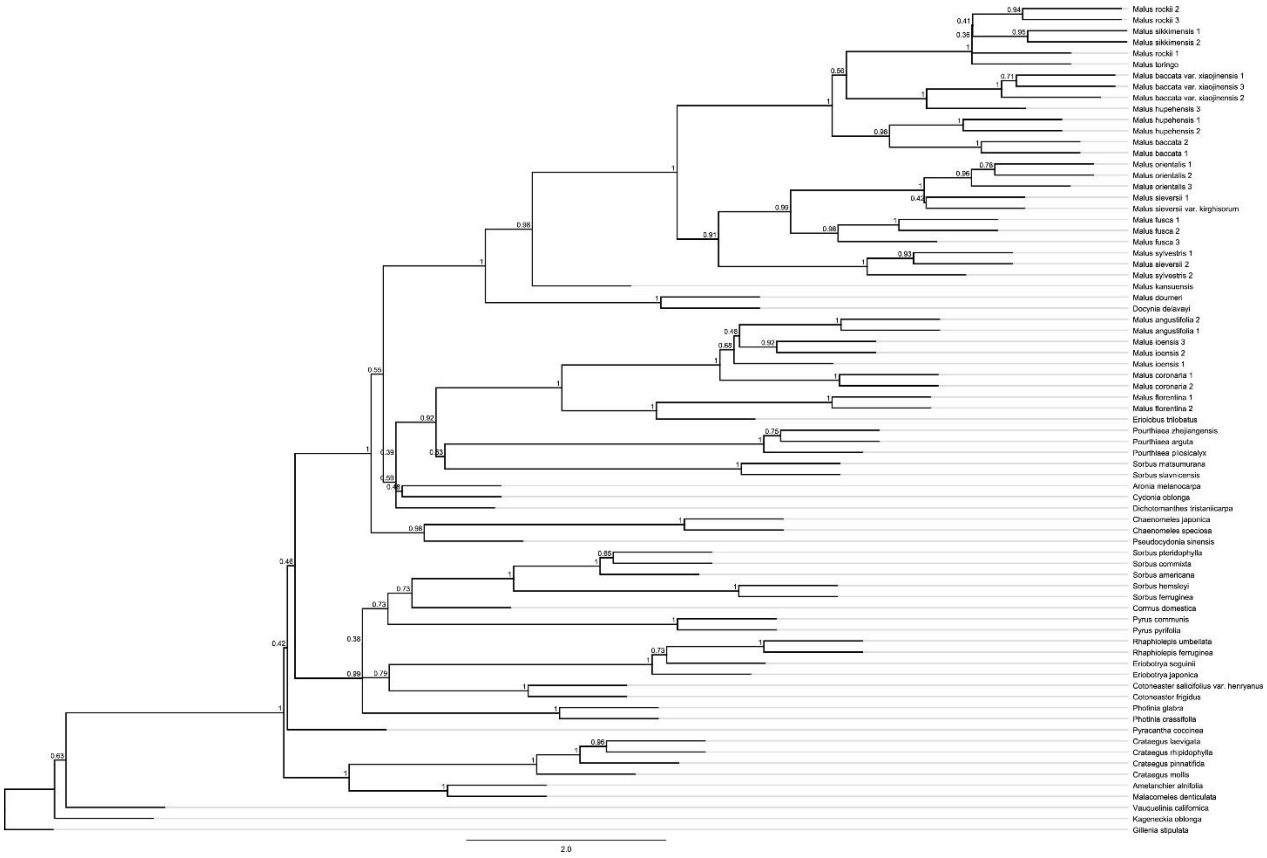

Fig. S14. Species tree of *Malus* s.l. in the framework of Maleae inferred from ASTRAL-III of the concatenated 80 plastid coding genes. Numbers above the branches indicate the branch support values measuring the support for a local posterior possibility.

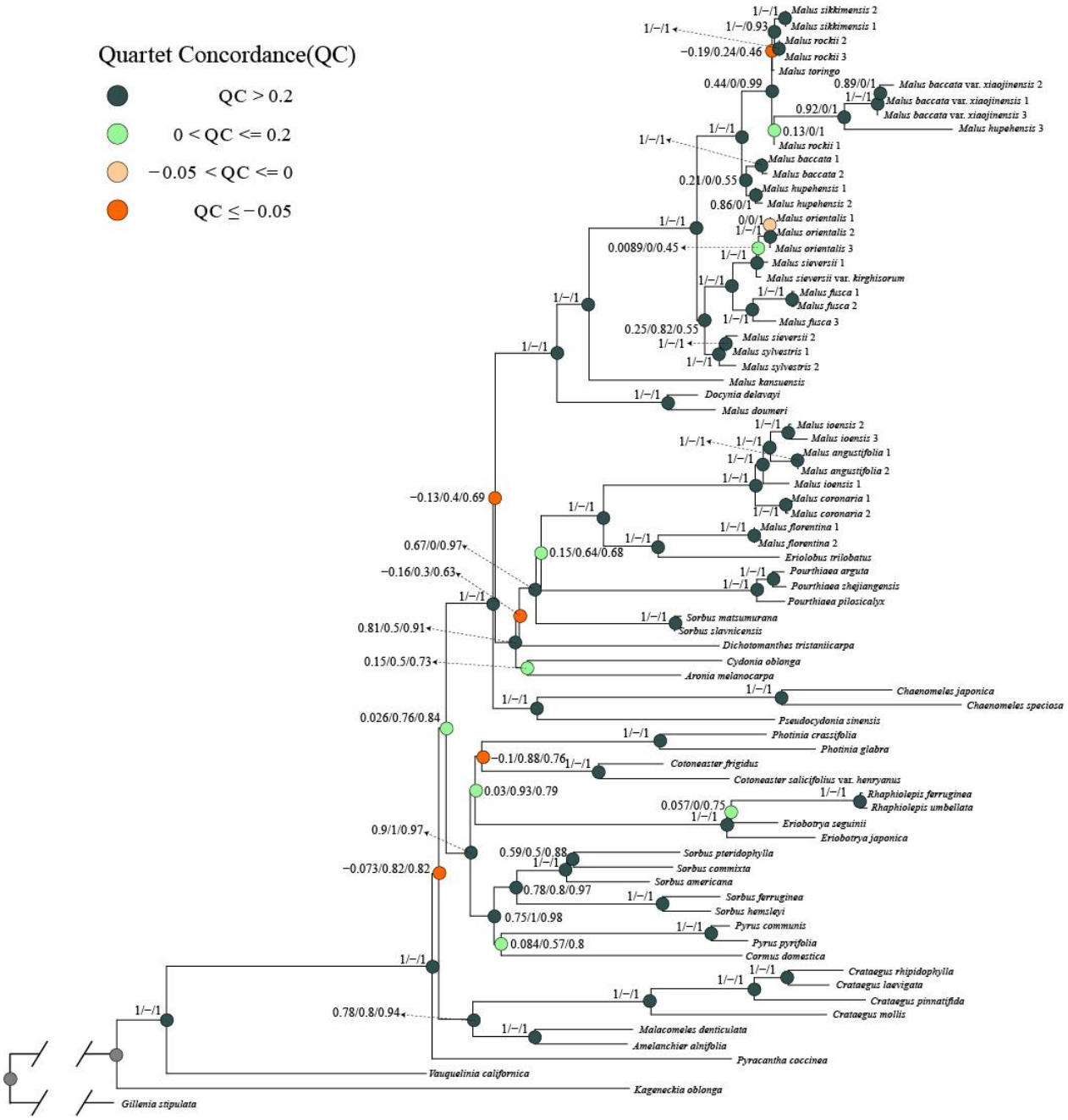

Fig. S15. Maximum likelihood phylogeny of *Malus* s.l. in the framework of Maleae inferred from RAxML analysis of the concatenated 80 plastid coding genes. Quartet Samping Scores are shown on branches indicating Quartet Concordance/Quartet Differential/Quartet Informativeness. Quartet Concordance is color-coded according to the legend.

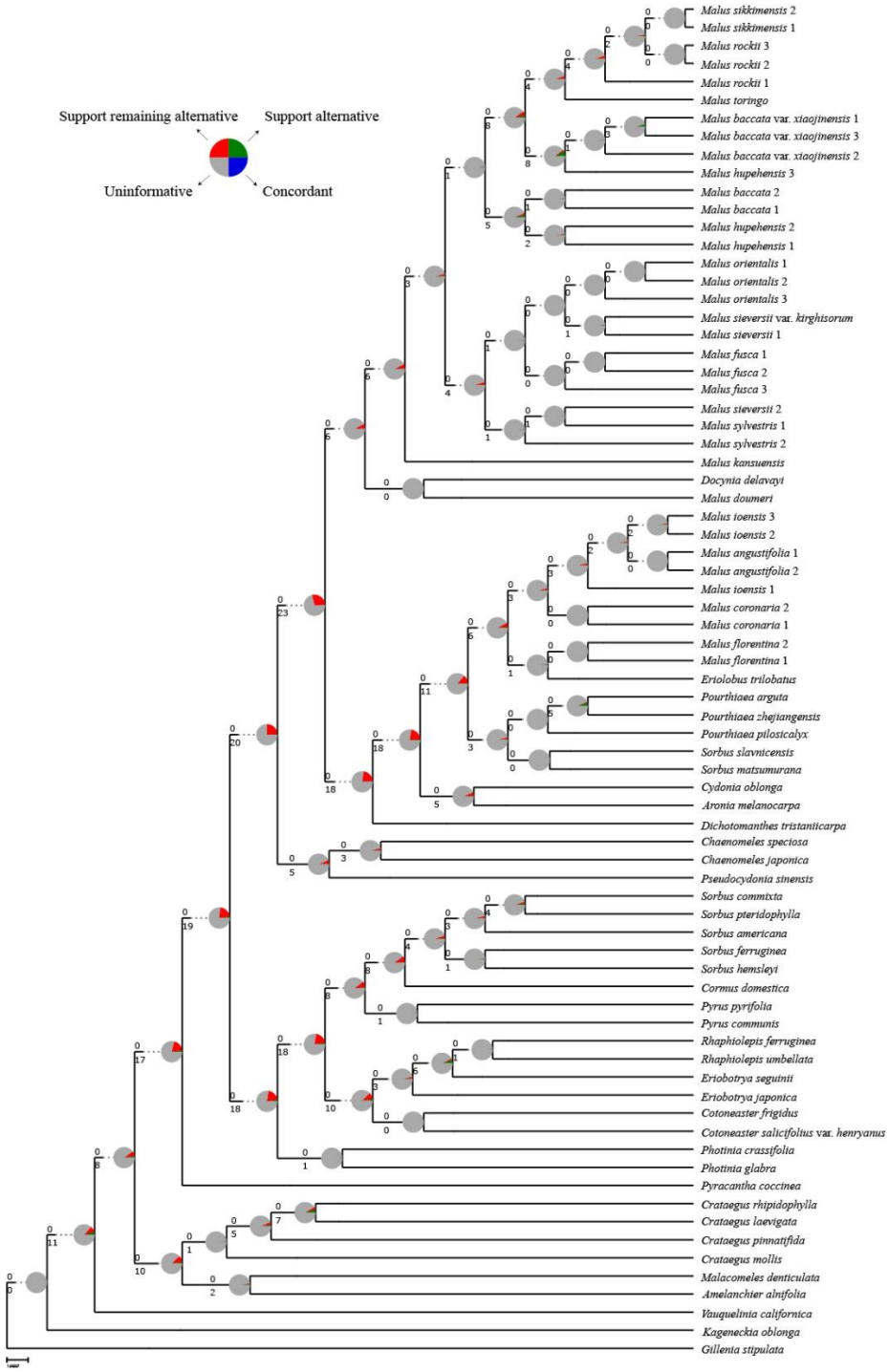

Fig. S16. Maximum likelihood cladogram of *Malus* s.l. inferred from RAxML analysis of the concatenated 80 plastid coding genes. Numbers above branches indicate the number of gene trees concordant/conflicting with that node in the species tree. Numbers below the branches are the Internode Certainty All (ICA) score. Pie charts on nodes present the proportion of gene trees that support that clade (blue), the proportion that support the main alternative bifurcation (green), the proportion that support the remaining alternatives (red), and the proportion (conflict or support) that have < 50% bootstrap support (gray).

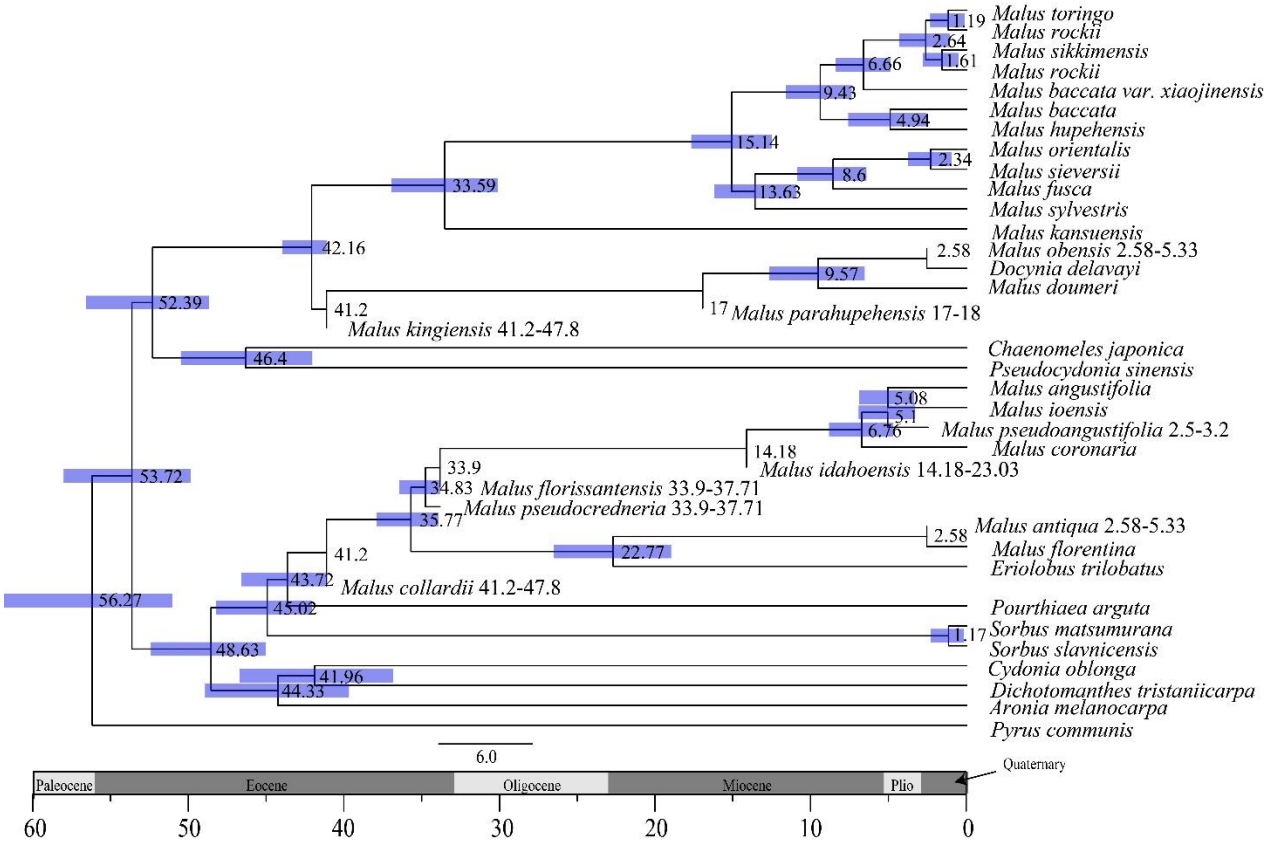

Fig. S17. Dated chronogram for the apple genus *Malus* inferred from BEAST with the Fossilized Birth-Death process based on the 80 plastid codinig genes dataset.

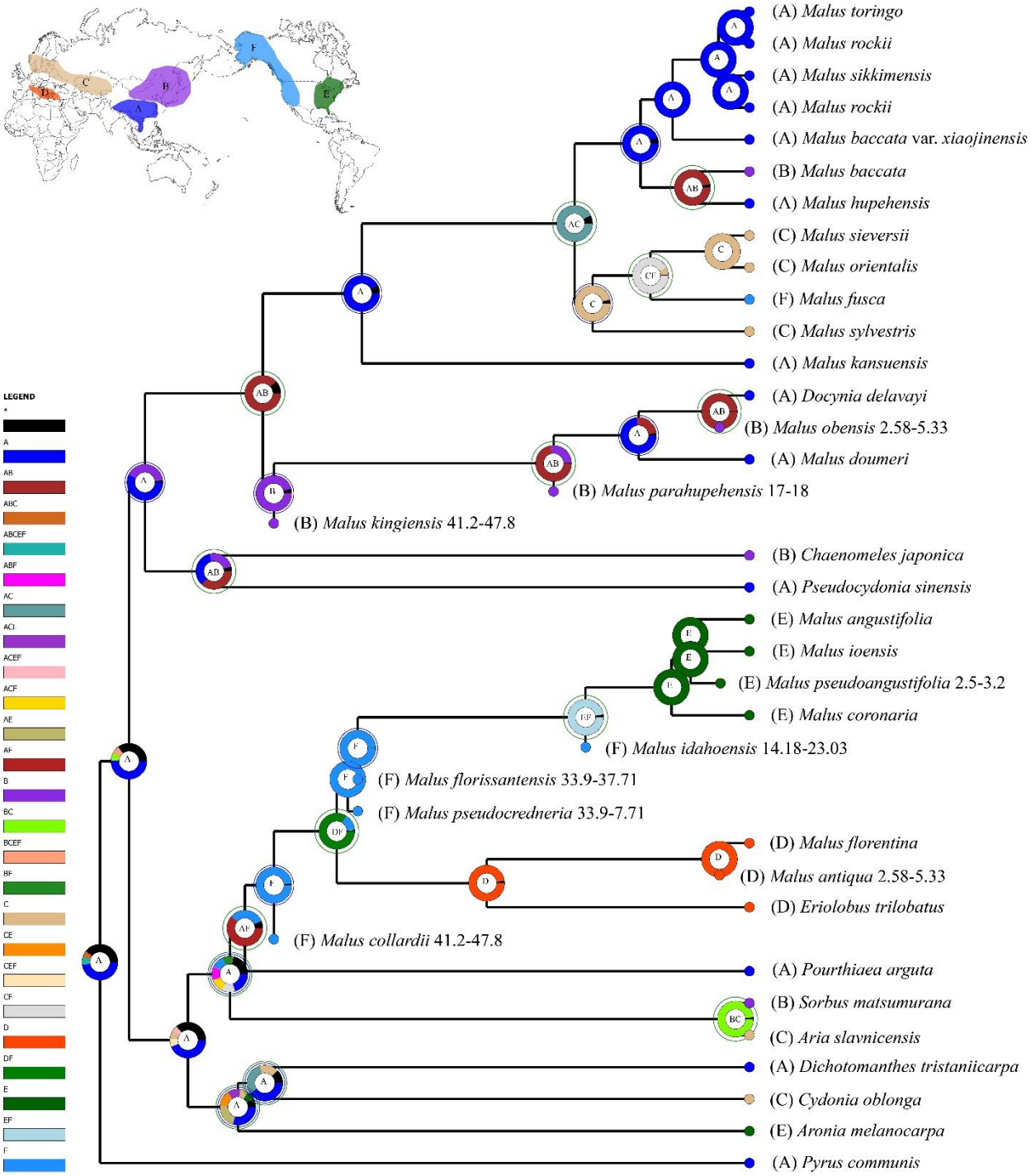

Fig. S18. The ancestral area reconstruction using BioGeoBEARS from the 80 plastid coding genes dataset, with the colored key identifying extant and possible ancestral ranges, (A), Southern East Asia; (B), Northern East Asia; (C), Europe and Central Asia; (D), Mediterranean; (E), Eastern North America; (F), Western North America.