bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Molecular and morphological analyses clarify species delimitation and reveal a new *Betula* species in section *Costatae*

Luwei Wang[1,2#], Junyi Ding[1,2#], James S. Borrell[3], Hugh A. McAllister[4], Feifei Wang[1,2], Lu Liu[1,2], Nian Wang[1,2,5*]

[1]State Forestry and Grassland Administration Key Laboratory of Silviculture in downstream areas of the Yellow River, College of Forestry, Shandong Agricultural University, Tai'an 271018, China.

[2]Mountain Tai Forest Ecosystem Research Station of State Forestry and Grassland Administration, College of Forestry, Shandong Agricultural University, Tai'an 271018, China.

[3]Natural Capital and Plant Health Department, Royal Botanic Gardens Kew, Richmond, Surrey TW9 3AB, UK.

[4]School of Life Sciences, Biosciences Building, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK.

[5]State Key Laboratory of Crop Biology, Shandong Agricultural University, Tai'an 271018, Shandong province, China

Running title: species delimitation and a new *Betula* species in section *Costatae*

[#]contributed equally to this work

[*]For correspondence.

Email: nian.wang@sdau.edu.cn

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Background and Aims Delineating closely related and morphologically similar species with overlapping ranges can be difficult. Here, we use section *Costatae* (genus *Betula*) as a model to resolve species and subspecies boundaries in four morphologically similar trees: *Betula ashburneri*, *Betula costata*, *Betula ermanii* and *Betula utilis* (including ssp. *utilis*, and diploid and tetraploid races of ssp. *albosinensis*).

Methods We genotyped 298 individuals (20-80 per species) from 38 populations at 15 microsatellite markers and a subset of 34 individuals from 21 populations using restriction-site associated DNA sequencing (RAD-seq). Morphometric analysis was conducted to characterise leaf variation for a subset of 89 individuals.

Key Results Molecular analyses and leaf morphology found little differentiation between *B. ashburneri*, diploid *B. utilis* ssp. *albosinensis* and some samples of *B. utilis* ssp. *utilis* suggesting that these should be treated as a single species. By contrast, tetraploid *Betula utilis* ssp. *albosinensis* was divided into two groups with group I genetically similar to *B. utilis* ssp. *utilis* based on SNPs and group II, a very distinct cluster, which we propose as a new species, namely, *Betula buggsii*. Phylogenomic analysis based on 2,285,620 SNPs show a well-supported monophyletic clade of *B. buggsii*, forming a sister with a well-supported clade of *B. ashburneri*, diploid *B. albosinensis* and some samples of *B. utilis* ssp. *utilis*. Morphologically, *Betula buggsii* is characterised by elongated lenticels and a distinct pattern of bark peeling. *Betula buggsii* is geographically restricted to the Qinling-Daba Mountains.

Conclusions Our study reveals six genetically distinguishable species: *B. ashburneri*,

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

B. buggsii, *B. costata*, *B. utilis* ssp. *utilis*, *B. utilis* ssp. *albosinensis* and *B. ermanii*. Our research demonstrates an integrative approach in delimitating species using morphological and genetic samples from their nearly entire distributions. Analyses based on subsets of species' distributions may lead to erroneous species or subspecies delineation.

Keywords: birch, cryptic species, microsatellite markers, polyploidy, RAD-seq, species delineation

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

INTRODUCTION

Species delineation based on morphology may be confounded by intra-specific variation among populations and limited differentiation between closely-related species (Whittall et al., 2004; Leliaert et al., 2009; Wang et al., 2014b; Lissambou et al., 2019). Where species co-occur, this may be further exacerbated by introgression and hybridisation (Bacon et al., 2012; Andújar et al., 2014), or may be morphologically impossible due to cryptic speciation (Bickford et al., 2007; Fišer et al., 2018). Despite advances in phylogenetic methods, this has meant that many species rich genera have remained unresolved, hindering our understanding of species ecology and evolution as well as limiting our ability to deliver effective conservation management.

Betula L. (Betulaceae) is such a genus with many taxonomic issues. The genus consists of approximately 65 species and subspecies (Ashburner & McAllister, 2016) with some spanning a very broad latitudinal and longitudinal range, such as *B. platyphylla*, ranging from Europe to eastern Asia and from the Himalayas to Siberia. Species such as *B. michauxii* and *B. nana*, are morphologically convergent but comparatively distantly-related (Wang et al., 2016; Wang et al., 2020), having evolved independently in North America and Scotland respectively. Analysis of *Betula* taxonomy is complicated by these broad ranges, frequent inter-specific hybridisation (Anamthawat-Jónsson & Tómasson, 1999; Wang et al., 2014a; Zohren et al., 2016; Tsuda et al., 2017), polyploidy and considerable morphological variation (Wang et al., 2014b; Ashburner & McAllister, 2016).

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Where ranges overlap, introgression appears frequent between species of the same ploidy level (Nagamitsu et al., 2006; Ashburner & McAllister, 2016) and even differing ploidy levels (Anamthawat-Jónsson & Thórsson, 2003; Wang et al., 2014a; Zohren et al., 2016; Tsuda et al., 2017). *Betula* species from different subgenera appear able to hybridise readily, such as hybridisation between *B. alleghanensis* and *B. papyrifiera* (Thomson et al., 2015). Polyploidy is also common within *Betula*, accounting for nearly 60% of the described taxa, ranging from diploid to dodecaploidy, with cytotypes observed for some species, such as *B. chinensis* (6x and 8x) (Ashburner & McAllister, 2016).

In this study, we use section *Costatae* as a model in which to demonstrate combined morphological and genetic methods to resolve these taxonomic issues (Table 1). Section *Costatae* includes the diploids *B. ashburneri* and *B. costata*, with *B. ashburneri* discovered from south-east Tibet and reported to have distributions in north-west Yunnan and western Sichuan (McAllister & Rushforth, 2011) and with *B. costata* distributed in northern and northeastern China, Japan and Russian Far East (Ashburner & McAllister, 2016). Section *Costatae* also includes two tetraploids: *B. utilis* (subdivided into ssp. *utilis* and ssp. *albosinensis*) occurring from the Himalayas to north China without clear geographical and morphological intra-specific boundaries and *B. ermanii* from northeastern China, Japan and Russian Far East. Several varieties of these tetraploid species have also been named based on a limited number of herbarium specimens, such as *B. utilis* var. *prattii*, *B. albosinensis* var. *septentrionalis* and *B. ermanii* var. *lanata* (Ashburner & McAllister, 2016), though their taxonomic

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

validity is unclear.

Confusingly, *B. utilis* ssp. *utilis* was described to have distributions in Gansu, Ningxia, Qinghai and Shaanxi according to Flora of China (Li & Skvortsov, 1999), where *B. utilis* ssp. *utilis* was not recorded in Ashburner and McAllister's monograph (Ashburner & McAllister, 2016) (Table 1). Recently, a 'diploid' *B. albosinensis* has been discovered from the Qinling Mountains in central China (Hu et al., 2019). A phylogenetic tree based on the internal transcribed spacer (ITS) region indicated a close relationship between the 'diploid' *B. albosinensis*, *B. ashburneri* and *B. costata* (Hu et al., 2019).

It remains unknown if the 'diploid' *B. albosinensis*, *B. ashburneri* and *B. costata* represent distinct genetic entities. Moreover, it remains unknown if the tetraploids *B. utilis* ssp. *albosinensis*, *B. ermanii* and *B. utilis* ssp. *utilis* described in Flora of China and in Ashburner and McAllister's monograph, respectively, represent distinct genetic entities (Li & Skvortsov, 1999; Ashburner & McAllister, 2016). For ease of reference, we abbreviate the 'diploid' *B. albosinensis* and *B. utilis* ssp. *utilis* described in Ashburner and McAllister's monograph and in Flora of China as *B. albosinenesis* [DA], *B. utilis* [AM] and *B. utilis* [FC], respectively.

To resolve taxonomic issues within section *Costatae*, we carried out morphological analysis, microsatellite genotyping and restriction-site associated DNA sequencing (RAD-seq). Our specific aims are to (1) identify the number of distinct genetic groups within section *Costatae*, with particular attention to (2) resolving the phylogenetic position of *B. utilis* [FC] and *B. albosinensis* [DA]; finally, (3) we integrate genetic

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

120 data with morphology and geographic distributions to present a revised treatment of
121 species boundaries within section *Costatae*. We consider the applicability of this
122 approach to other taxonomically complex genera.
123
124 **Materials and methods**
125 **Sampling**
126 Samples putatively identified (based on morphology) as *B. utilis* [AM], *B. utilis* ssp.
127 *albosinensis*, *B. ermanii*, *B. albosinensis* [DA], *B. utilis* [FC] and *B. costata* were
128 collected from between three and twelve populations each (Fig. 1). All species were
129 collected from naturally occurring woodland, meaning that they were not artificially
130 planted. Leaf samples were collected between May and September of 2018 and 2019,
131 with each sample separated by ~20m. A herbarium specimen was created for each
132 sample except for a subset of samples where branches were difficult to obtain. For
133 these samples, cambium tissue was collected. A GPS system (UniStrong) was used to
134 record the coordinate points of each population. Detailed species and population
135 sampling information is provided in Supplementary Data Table S1.
136 **Species identification**
137 *Betula utilis* [AM] is distributed through SE Tibet to Yunnan and Sichuan and *B. utilis*
138 ssp. *albosinensis* occurs in North Sichuan, Hubei, Shaanxi, Shanxi, Henan and Hebei
139 (Ashburner and McAllister, 2016). These two species co-occur in Sichuan province.
140 Due to a morphological continuum between *B. utilis* [AM] and *B. utilis* ssp.
141 *albosinensis*, we assigned our populations based on geographic origins, with

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

142 populations from northwestern Yunnan designated as *B. utilis* [AM] and populations
143 from south Shaanxi, Hubei, Shanxi, Henan and Hebei designated as *B. utilis* ssp.
144 *albosinensis*. *Betula utilis* [FC] occupies a higher altitude than *B. utilis* ssp.
145 *albosinensis* and can be distinguished from the latter by its leathery dark green leaves
146 (Ashburner & McAllister, 2016; Li & Skvortsov, 1999). *Betula costata* and *B. ermanii*,
147 having distributions in northeastern China, can be distinguished from leaf morphology
148 with the former having lanceolate leaves and the latter triangular-ovate leaves (Li &
149 Skvortsov, 1999; Ashburner & McAllister, 2016).

150 **Morphometric analyses**

151 For analyses of leaf shape among these species, we selected 6-27 individuals per taxa
152 and sampled 283 leaves. Leaves were scanned individually using a Hewlett-Packard
153 printer (LaserJet Pro MFP M128fn) with a resolution of 600 dpi. Thirteen landmarks
154 were selected from each scanned leaf according to the protocols of (Liu et al., 2018;
155 Hu et al., 2019). The 13 landmarks were converted to a configuration of 26 cartesian
156 coordinates using ImageJ (Abràmoff et al., 2004). A Generalized Procrustes Analysis
157 (GPA) was performed using the procGPA function in the R package “shapes” (Dryden,
158 2019). Eigenleaves were visualized using the “shapepca” function and principal
159 component scores, percentage variance and Procrustes-adjusted coordinates were
160 obtained from procGPA object values.

161 **DNA extraction and microsatellite genotyping**

162 High quality DNA was extracted from cambial tissues following a modified 2x CTAB
163 (cetyltrimethylammonium bromide) protocol (Wang et al., 2013). Extracted DNA was

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

assessed with 1.0% agarose gels. Fifteen microsatellite loci developed for *B. platyphylla* var. *japonica* (Wu et al., 2002), *B. pendula* (Kulju et al., 2004), *B. pubescens* ssp. *tortuosa* (Truong et al., 2005) and *B. maximowicziana* (Tsuda et al., 2009) were used to genotype our samples (Supplementary Data, Table S2), with the 5' terminus of the forward primers labeled with FAM, HEX or TAM fluorescent probes. These microsatellite loci have a good cross compatibility in multiple *Betula* species. Each microsatellite locus was amplified individually and was artificially combined into four multiplexes. The PCR protocol followed Hu et al. (2019). Microsatellite alleles were scored using GENEMARKER 2.4.0 (Softgenetics) and checked manually. Individuals with more than three missing loci were excluded for further analyses, resulting in 298 individuals in the final dataset.

RAD-seq

A subset of 34 DNA samples were selected for RAD-seq using an Illumina HiSeq 2500 and 150-bp pair-end sequencing with the restriction enzyme *PstI* (Personalbio company, Shanghai, China). These were combined with eight additional samples of section *Costatae* previously sequenced, using the same restriction enzyme in Wang et al. (2020). These samples represented six *B. costata*, six *B. utilis* [AM], six *B. ermanii*, twelve *B. utilis* ssp. *albosinensis*, seven *B. utilis* [FC] and one of each of *B. albosinensis* [DA], *B. ashburneri*, *B. ermanii* var. *lanata*, *B. albosinensis* var. *septentrionalis*, and *B. utilis* var. *prattii* (Supplementary Data, Table S3). The raw data were trimmed using Trimmomatic (Bolger et al., 2014) in paired-end mode. Reads with a quality of below 20 within the sliding-window of 5 bp and unpaired reads were

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

186 discarded. We performed LEADING and TRAILING to remove bases with a quality
187 below 20. Then we performed a SLIDINGWINDOW step to discard reads shorter
188 than 40 bp. Filtered reads of each sample were aligned to the whole genome sequence
189 of *B. pendula* (Salojärvi et al., 2017) using BWA-MEM v.0.7.17-r1188 algorithm in
190 BWA (v0.7.17) with default parameters (Li & Durbin, 2009). Non-specific mapped
191 reads were discarded. All subsequent analyses were performed using SAMtools v1.8
192 (Li et al., 2009) and GATK V4.1.4 (McKenna et al., 2010; DePristo et al., 2011).
193 These include conversion of alignments into indexed binary alignment map (BAM)
194 files, marking duplicates, calling genotypes and filtering SNPs (McKenna et al., 2010;
195 DePristo et al., 2011). SNPs within a 50 kb window with $r^2 > 0.5$ and a minimum
196 allele frequency (MAF) < 0.01 were removed to reduce linkage disequilibrium using
197 BCFtools v1.10.2 (Li, 2011). Prior to population structure analysis, we retained only
198 sites with no missing data, resulting in 82,137 SNPs for downstream analyses.

199 **Analyses of microsatellite data and SNPs**

200 A principal coordinate analysis (PCoA) was performed on microsatellite data of *B.*
201 *utilis* [AM], *B. utilis* [FC], *B. utilis* ssp. *albosinensis*, *B. albosinensis* [DA], *B. costata*
202 and *B. ermanii* using POLYSAT (Clark & Jasieniuk, 2011) implemented in R 4.0.2 (R
203 Core Team, 2020), based on Bruvo's genetic distances (Bruvo et al., 2004). For
204 nucleotide SNPs, a principal component analysis (PCA) was carried out using the
205 'adegenet' R package 2.1.1 (Jombart, 2008).

206 Microsatellite data were analyzed in STRUCTURE (Pritchard et al., 2000) to identify
207 the most likely number of genetic clusters (K) with a ploidy of four. Ten replicates

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

were performed with 1,000,000 iterations and a burn-in of 100,000 for each run at each value of K from 1 to 8. We used the admixture model, with an assumption of correlated allele frequencies among populations. Individuals were assigned to clusters based on the highest membership coefficient averaged over the ten independent runs. The number of genetic clusters was estimated using the "Evanno test" (Evanno et al., 2005) implemented in Structure Harvester (Earl & vonHoldt, 2012). Replicate runs were grouped based on a symmetrical similarity coefficient of >0.9 using the Greedy algorithm in CLUMPP (Jakobsson & Rosenberg, 2007) and visualized in DISTRUCT 1.1 (Rosenberg, 2004).

The filtered SNPs were analyzed in ADMIXTURE v1.3.0, a model-based approach to assessing population structure in a Maximum Likelihood framework (Alexander & Lange, 2011). We ran ADMIXTURE for K = 1-10 with 20 replicates for each K value and performed cross-validation error estimation in order to assess the most suitable value of K (Alexander & Lange, 2011). Replicate runs were aligned and visualised in pong v1.4.9 with the greedy algorithm (Behr et al., 2016).

ITS and SNP based phylogenetic analyses

To provide an additional line of evidence for the phylogenetic position of *B. utilis* [FC], *B. albosinensis* [DA], and *B. costata*, we generated ITS sequence and SNP based phylogenies.

First, we amplified the nuclear ribosomal internal transcribed spacer (nrITS) region (ITS1, 5.8S and ITS2) using primers ITS4 (White et al., 1990) and ITSLeu (Baum et al., 1998), with seven, ten, five and four individuals of *B. utilis* ssp. *albosinensis*

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

group II collected from the NSX, CKX, WLP and SNJ, respectively. The reaction mix and the PCR protocol followed that of Hu et al. (2019). PCR products were purified and sequenced at Tsingke Company (Qingdao, China). Sixty-four additional ITS sequences from Betulaceae (Wang et al., 2016) were included to infer the phylogenetic position of *B. utilis* ssp. *albosinensis* group II. In total, 90 sequences were aligned using BioEdit v7.0.9.0 (Hall, 1999) with default parameters.

Second, we collated RAD-seq data of 20 *Betula* taxa representing genus wide diploid species. The identity of the 20 sequenced *Betula* taxa was initially inferred via ITS sequences and genome size estimates (Wang et al., 2016). In addition, we included RAD-seq data of 17 samples generated in the present study. *Alnus inokumae* was selected as the outgroup (Supplementary Data, Table S3). SNPs of a total of 38 taxa were concatenated into a supermatrix for phylogenetic analysis. SNPs with a missing data > 50% were excluded, resulting in 2,285,620 SNPs.

For both the ITS alignment and the matrix of SNPs, we conducted a rapid bootstrap analysis under a GTR+GAMMA nucleotide substitution model, with 100 bootstraps and 10 searches using the maximum-likelihood method (ML) in RAxML v. 8.1.16 (Stamatakis, 2006). The phylogenetic trees were visualised in FigTree v.1.3.1.

Results

Morphometric analyses

Landmarks were first aligned using a GPA and then a principal component analysis (PCA) was conducted to visualise the major sources of shape variance of leaves from *B. albosinensis* [DA], *B. utilis* ssp. *albosinensis*, *B. utilis* [AM], *B. utilis* [FC], *B.*

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

costata and *B. ermanii*. PC1 and PC2 produce largely overlapping clusters among *B. albosinensis* [DA], *B. utilis* ssp. *albosinensis*, *B. utilis* [AM] and *B. utilis* [FC], but *B. costata* and *B. ermanii* overlapped to a much lesser extent (Fig. 2a). The shape variance, represented by PC1 and PC2, is mainly influenced by leaf width and marginally influenced by leaf length (Fig. 2b).

PCO and PCA analyses

PCO analysis based on microsatellite markers revealed five clusters, with the first three axes accounting for 43.8% of the total variation (Fig. 3a). *Betula utilis* ssp. *albosinensis* forms two groups: group I overlaps substantially with *B. utilis* [AM] and *B. ermanii* whereas group II separates from all the other species on coordinate 1 (Fig. 3a). *Betula utilis* [FC] and *B. albosinensis* [DA] overlap substantially whereas *B. costata* separates from the remaining species on coordinates 2 and 3 (Supplementary Data, Fig. S1a). *Betula ermanii* separates from *B. utilis* [AM] on coordinate 3 with *B. utilis* ssp. *albosinensis* group I intermediate (Supplementary Data, Fig. S1a).

For the sequenced individuals, between 12,234,848 and 28,155,092 reads were retained for each individual (mean 18,862,242) after trimming and filtering (Supplementary Data, Table S3). The number of variable sites of the sequenced individuals ranges from 5,520,333 to 9,735,507. A principal component analysis (PCA) based on genotype calls for 82,137 SNPs shows that both *B. utilis* ssp. *albosinensis* group II and *B. costata* separate from the remaining species and from each other (Fig. 3b). Two individuals of *B. utilis* ssp. *albosinensis* group I form a cluster and three individuals of *B. utilis* ssp. *albosinensis* group I form a cluster with

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the previously sequenced *B. utilis* var. *prattii* and *B. albosinensis* var. *septentrionalis* (Fig. 3b). *Betula utilis* [AM] forms a cluster with the previously sequenced *B. utilis* ssp. *albosinensis* whereas *B. ermanii* and *B. ermanii* var. *lanata* form a cluster from PC1 and PC2 (Supplementary Data, Fig. S1b). *Betula albosinensis* [DA] forms a cluster with one accession of *B. utilis* [FC] whereas the remaining accessions of *B. utilis* [FC] form another cluster. The two individuals of *B. utilis* ssp. *albosinensis* group I position between *B. ermanii* and three individuals of *B. utilis* ssp. *albosinensis* group I. *Betula ashburneri* forms a continuum with *B. utilis* [AM] and *B. utilis* [FC] on PC1 and PC3 (Supplementary Data, Fig. S1b).

STRUCTURE and ADMIXTURE analyses

STRUCTURE analyses based on microsatellite markers identified five clusters: (1) *B. utilis* ssp. *albosinensis* group I, (2) *B. albosinensis* [DA] and *B. utilis* [FC], (3) *B. costata*, (4) *B. utilis* [AM], (5) *B. ermanii* and (6) *B. utilis* ssp. *albosinensis* group II (Supplementary Data Figs. S2-3). *B. utilis* ssp. *albosinensis* group I is genetically similar to *B. ermanii* at all K values (Fig. 4a). *Betula utilis* ssp. *albosinensis* group II includes populations SNJ, WLP, NSX and CKX and separates with the remaining species (Supplementary Data, Fig. S3). Similarly, *B. albosinensis* [DA] and *B. utilis* [FC] are genetically similar at all supported K values (Supplementary Data, Fig. S3).

Admixture analysis based on the same set of SNPs showed that cross-validation error is smallest at K = 5 , but only with four out of twenty replicates having an average pairwise similarity of 0.98 (Supplementary Data, Fig. S4). At the value of K = 6, fourteen out of twenty replicates have an average pairwise similarity of 0.98. However,

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

296 the cross-validation error is slightly larger than that when K = 5 (Supplementary Data,
297 Fig. S4). At the value of K = 5, *B. utilis* ssp. *albosinensis* group I genetically
298 resembles *B. utilis* [AM] with exception of samples XLA01 and XLA32, which are
299 more genetically similar to *B. ermanii* (Fig. 4b). At the value of K = 6, *B. utilis* ssp.
300 *albosinensis* group I separates from *B. utilis* [AM] and XLA01 and XLA32 exhibit
301 genetic admixture from *B. ermanii* (Fig. 4b). *Betula utilis* ssp. *albosinensis* group II
302 separates from the remaining species at the value of K = 3 and onwards
303 (Supplementary Data, Fig. S5). Interestingly, this identified that *B. albosinensis* var.
304 *septentrionalis* and *B. utilis* var. *prattii* are genetically similar to *B. utilis* ssp.
305 *albosinensis* group I whereas the *B. utilis* ssp. *albosinensis* and *B. utilis* ssp. *utilis* are
306 genetically similar to *B. utilis* [AM] (Fig. 4b).
307
308 **Phylogenetic analyses**
309 **Identification of species novo *Betula buggsii***
310 Analyses of microsatellite data and SNPs indicate that *B. utilis* ssp. *albosinensis* group
311 II is genetically distinct from other species of section *Costatae* and therefore
312 represents a putative new species, namely *Betula buggsii*. Morphologically, despite
313 general similarity to *B. utilis* ssp. *albosinensis*, we found *B. buggsii* is characterised by
314 very elongated lenticels with bark peeling along lenticels into strips.
315 The phylogenetic tree based on ITS showed that *B. buggsii* samples formed a
316 monophyletic cluster, within a clade with species of section *Acuminatae*, *B. bomiensis*
317 and *B. nigra*. However, this clade received little support (Supplementary Data, Fig.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S6). The phylogenetic tree based on a matrix of 2,285,620 SNPs showed that the five individuals of *B. buggsii* from populations CKX, SNJ and WLP formed a monophyletic clade with 100% support, which was basal to a clade of *B. costata*, *B. ashburneri*, *B. albosinensis* [DA] and *B. utilis* [FC] (Fig. 5). The five individuals of *B. costata* formed a monophyletic clade whereas individuals of *B. utilis* [FC], *B. albosinensis* [DA] and *B. ashburneri* intermixed and together formed a monophyletic clade with 100% support (Fig. 5).

Taxonomic treatment

Betula buggsii N. Wang, sp. nov.

Diagnosis:—*Betula buggsii* is very similar with *B. utilis* ssp. *albosinensis* in leaf morphology but have very elongated lenticels on the bark of adult and elderly trees. Barks of *B. buggsii* peel along the elongated lenticels into strips. Adult or old *B. utilis* ssp. *albosinensis* trees exfoliate in large sheets (Fig. 6a). Seedlings of *B. buggsii* and *B. utilis* ssp. *albosinensis* (DBH < 5 cm) show no obvious difference in bark color and patterns of bark peeling.

Type:—CHINA. Chongqing: Chengkou County, elev. ca. 1600-2000 m, 108.7 E, 31.9 N, 6 October 2018 (holotype xx; isotypes xx).

Distribution and habitat:—*Betula buggsii* occurs in Chongqing, western Hubei and Shaanxi with five localities discovered. *Betula buggsii* grows in mixed forests with bamboos at an altitude of between 1500 and 2100 meters. At some localities, *B. buggsii* grows in parapatry with *B. luminifera* but at a higher altitude. We only founded a small number of *B. buggsii* individuals within each population. Given this

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

situation, we think *B. buggsii* needs conservation.

Etymology:—*Betula buggsii* is named after Prof. Richard J.A. Buggs, an evolutionary biologist from the Royal Botanical Gardens Kew and Queen Mary University of London, for his devotion to research on hybridisation, phylogenetics and conservation of the genus *Betula*. The Chinese name of *B. buggsii* is "年桦" (nián huá).

Discussion

Species delimitation within section *Costatae*

Here we have combined genetic, morphological and distribution data to revise species delimitation within section *Costatae* (genus *Betula*). Our results support six genetic units and thus prefer recognition of six taxa.

Cluster one — *B. albosinensis* [DA], *B. ashburneri* and *B. utilis* [FC].

Several lines of evidence jointly support the merging of *B. albosinensis* [DA], *B. ashburneri* and *B. utilis* [FC]. First, PCO and STRUCTURE analyses of microsatellite markers indicate an indistinguishable cluster of *B. albosinensis* [DA] and *B. utilis* [FC] (Figs. 3a, 4a). This was further corroborated by admixture analysis of SNPs, showing an indistinguishable cluster of *B. albosinensis* [DA] and *B. utilis* [FC] (Fig. 4b). However, admixture analysis of SNPs including *B. ashburneri* shows the same genetic cluster of *B. ashburneri* and *B. utilis* [AM] (Fig. 4b). By contrast, phylogenomic analysis based on a much larger number of SNPs shows a fully-supported monophyletic clade of *B. albosinensis* [DA], *B. ashburneri* and *B. utilis* [FC] (Fig. 5). The genetic similarity between *B. ashburneri* and *B. utilis* [AM]

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

362 based on admixture analyses of SNPs suggests *B. ashburneri* being a recent parent of
363 *B. utilis* [AM]. This has been confirmed based on a recent phylogenomic analysis
364 (Wang et al., 2020). In addition, gene flow between the two species may further result
365 in genetic similarity. *Betula ashburneri* is diploid based on chromosome number and
366 genome size analysis (Ashburner & McAllister, 2016; Wang et al., 2016), consistent
367 with the observation that *B. albosinensis* [DA] and *B. utilis* [FC] are also diploids
368 based on microsatellite markers. This is different from descriptions in Flora of China
369 that *B. utilis* [FC] was a tetraploid (Li & Skvortsov, 1999). *Betula ashburneri* was
370 described to occupy a higher altitude than *B. utilis* [AM] (McAllister & Rushforth,
371 2011), consistent with *B. utilis* [FC] or *B. albosinensis* [DA] occupying a higher
372 altitude than *B. utilis* ssp. *albosinensis* according to our field observations. In addition,
373 *B. ashburneri* was discovered from SE Tibet and reported to distribute in Sichuan and
374 Shaanxi provinces, overlapping with the distribution of *B. utilis* [FC] and *B.*
375 *albosinensis* [DA]. Based on these, we think *B. utilis* [FC], *B. albosinensis* [DA] and
376 *B. ashburneri* refer to the same species. *Betula ashburneri* was described to have a
377 multi-stemmed shrubby habit and grow up to four meters in height. However,
378 according to our field observations, it can reach 35 meters in height, consistent with
379 descriptions from Flora of China.

380 **Cluster two — *B. costata*.**

381 Both microsatellite and SNPs indicate that *B. costata* is genetically different from
382 other species of section *Costatae* (Figs. 3-4). Despite the fact that *B. costata* and *B.*
383 *ermanii* co-occur in some populations, the two are morphologically different in fruit,

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

leaf and bark color. In addition, *B. costata* is a diploid and occupies a lower altitude than *B. ermanii*, which is a tetraploid.

Cluster three — *B. utilis* [AM].

Despite occupying a morphological continuum with *B. utilis* ssp. *albosinensis* group I, molecular results support *B. utilis* [AM] as a genetically distinct unit. *Betula utilis* [AM] are described from the Himalayas, northwestern Yunnan and with an extension into western Sichuan where it coexists with *B. utilis* ssp. *albosinensis* group I.

Cluster four — *B. utilis* ssp. *albosinensis* group I.

Betula utilis ssp. *albosinensis* group I forms a morphological continuum with *B. utilis* [AM]. However, molecular analyses indicate that *B. utilis* ssp. *albosinensis* group I forms a distinct cluster with *B. utilis* [AM] (Fig. 4b). We also found that two individuals of *B. utilis* ssp. *albosinensis* group I (XLA01 and XLA32), collected from its northern distribution, show a genetic admixture between *B. utilis* ssp. *albosinensis* and *B. ermanii* (Fig. 4b), indicating their hybrid origin. The two individuals were close to the southern distribution of *B. ermanii*, making hybridisation potentially occur due to long-distance transportation of pollen by wind. In addition, the previously described *B. albosinensis* var. *septentrionalis* and *B. utilis* var. *prattii* are more genetically similar to *B. utilis* ssp. *albosinensis* group I; however, the previously described *B. utilis* ssp. *albosinensis* is genetically similar to *B. utilis* [AM] (Fig. 4b). This indicates some misidentification of these taxa. This was suggested by observations on the very limited number of provenances in cultivation in the UK which led Ashburner and McAllister to describe these taxa as subspecies. Interestingly,

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

406 the included *B. albosinensis* var. *septentrionalis*, *B. utilis* var. *prattii* and *B. utilis* ssp.
407 *albosinensis* are from Sichuan province where *B. utilis* [FC] and *B. utilis* ssp.
408 *albosinensis* were reported to co-occur. Great morphological variations exist within
409 some populations in Sichuan according to our field observations that even bark color
410 within population shows substantial variation. This made assigning individuals there
411 to either *B. utilis* ssp. *albosinesis* group I or *B. utilis* [AM] impossible based solely on
412 morphological characters.

413 **Cluster five —*B. ermanii*.**

414 *Betula ermanii* and *B. utilis* ssp. *albosinensis* group I are genetically similar based on
415 microsatellite markers but genetically distinct based on SNPs. This is possibly due to
416 very recent gene flow between *B. ermanii* and *B. utilis* ssp. *albosinensis* group I.
417 However, here we think *B. ermanii* should be recognised as a genetic unit on grounds
418 of morphological characters and distribution. Morphologically, *B. ermanii* shows
419 apparent differences in fruit, leaf shape, bark color and the pattern of bark peeling
420 with *B. utilis* ssp. *albosinensis* group I. Geographically, *B. ermanii* distributes around
421 the Changbai Mountains and its north in northeast China where *B. utilis* ssp.
422 *albosinensis* group I is absent there.

423 **Cluster six — *B. utilis* ssp. *albosinensis* group II (*B. buggsii* as discussed below)**

424 Our genetic analyses revealed a distinct cluster of *B. utilis* ssp. *albosinensis* (group II),
425 which was sufficiently differentiated when compared to other taxa in the genus to be
426 ranked as a new diploid species of section *Costatae*. Based on multiple lines of
427 evidence we describe this new species as *B buggsii*. Molecular analyses of

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

428 microsatellite markers and SNPs show that *B. buggsii* is genetically distinct from all
429 the other species of section *Costatae*. Phylogenetic analysis based on ITS sequences
430 shows that *B. buggsii* samples cluster together despite low support values
431 (Supplementary Data, Fig. S6). Furthermore, phylogenomic analysis including nearly
432 genus-wide diploid species shows a fully supported monophyletic clade of *B. buggsii*,
433 which was placed within section *Costatae* (Fig. 5). This allows us to confidently
434 establish *B. buggsii* as a new species of section *Costatae*. Interestingly, microsatellite
435 markers revealed two alleles at heterozygous sites for *B. buggsii* whereas three or four
436 alleles for *B. utilis* ssp. *albosinensis*, suggesting a difference in ploidy level. Apart
437 from these, *B. buggsii* shows morphological difference with *B. utilis* ssp. *albosinensis*
438 in bark color and the patterns of bark exfoliation (Fig. 6a). *Betula buggsii*'s bark color
439 is light brown and exfoliates along the elongated lenticels in stripes while *B. utilis* ssp.
440 *albosinensis*'s bark is red and exfoliates in large sheets or flakes (Fig. 6a). The overall
441 morphological similarity between *B. buggsii* and *B. utilis* ssp. *albosinensis* supports
442 the placement of *B. buggsii* within section *Costatae*. Unfortunately, we failed to
443 obtain fruiting catkins, however, we observed seedlings of *B. buggsii* in open habitats,
444 indicating its ability to regenerate and its regeneration depends on habitat disturbance
445 like *B. utilis* ssp. *albosinensis* (Guo et al., 2019).
446
447 **A framework for species delimitation within section *Costatae***
448 A combination of various sources of information (i.e. genetic data, morphological
449 characters, ploidy level and geographic origins) facilitates demarcating species within

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a morphological or genetic continuum (Fig. 6b). For example, ploidy level is useful in distinguishing a species complex of differing ploidy levels. Recognition of cytotypes would help for conservation purposes as different cytotypes may have different adaptive potentials and are often genetically differentiated. If species reveals a morphological continuum, genetic data and geographic origins would help for distinguishing. This is just the case for *B. utilis* [AM] and *B. utilis* ssp. *albosinensis* group I. Similarly, for species which occupy a genetic continuum, such as *B. utilis* ssp. *albosinensis* I and *B. ermanii*, both morphological data and geographic origin aid in identification. .

Finally, for the challenging tetraploids within section *Costatae*, we propose that the most practical taxonomy is to treat populations in north-west Yunnan and the eastern and central Himalaya as *B. utilis* ssp. *utilis*; those from the Qinling Mountains as *B. utilis* ssp. *albosinensis*; and those from northeastern China (e.g. Changbaishan) as *B. ermanii*. For the diploids, it is reasonable to recognise *B. ashburneri*, *B. buggsii* and *B. costata* based on genetic data, morphological characters and geographic origins. The tetraploids certainly hybridise in cultivation (obser.) and so are likely to hybridise where they co-occur in the wild, generating intermediates with various levels of genetic admixture. Hence, populations collected from region between northwestern Yunnan and the Qinling Mountains may be hybrids between *B. utilis* ssp. *utilis* and *B. utilis* ssp. *albosinensis* and these from between Hebei and northeast China, may be hybrids between *B. utilis* ssp. *albosinensis* and *B. ermanii*. Further research is needed

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

472 to characterise patterns of genetic admixture between these species within their
473 geographic distributions and to guide future management of genetic diversity.

474

475 **Acknowledgements**

476 This work was funded by the National Natural Science Foundation of China
477 (31770230 and 31600295) and Funds of Shandong 'Double Tops' Program
478 (SYL2017XTTD13).

479

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

Abràmoff MD, Magalhães PJ, Ram SJ. 2004. Image processing with ImageJ. *Biophotonics International* **11:** 36-42.

Alexander DH, Lange K. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12:** 246.

Anamthawat-Jónsson K, Thórsson AT. 2003. Natural hybridisation in birch: triploid hybrids between *Betula nana* and *B. pubescens*. *Plant Cell Tissue and Organ Culture* **75:** 99–107.

Anamthawat-Jónsson K, Tómasson T. 1999. High frequency of triploid birch hybrid by *Betula nana* seed parent. *Hereditas* **130:** 191–193.

Andújar C, Arribas P, Ruiz C, Serrano J, Gómez-Zurita J. 2014. Integration of conflict into integrative taxonomy: fitting hybridization in species delimitation of *Mesocarabus* (Coleoptera: Carabidae). *Molecular Ecology* **23:** 4344-4361.

Ashburner K, McAllister HA. 2016. *The genus Betula: a taxonomic revision of birches*. Kew Publishing: London.

Bacon CD, McKenna MJ, Simmons MP, Wagner WL. 2012. Evaluating multiple criteria for species delimitation: an empirical example using Hawaiian palms (Arecaceae: *Pritchardia*). *Bmc Evolutionary Biology* **12:** 23.

Baum DA, Small RL, Wendel JF. 1998. Biogeography and floral evolution of Baobabs (*Adansonia*, Bombacaceae) as inferred from multiple data sets. *Systematic Biology* **47:** 181–207.

Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. 2016. pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics* **32:** 2817-2823.

Bickford D, Lohman DJ, Sodhi NS, Ng PKL, Meier R, Winker K, Ingram KK, Das I. 2007. Cryptic species as a window on diversity and conservation. *Trends in Ecology & Evolution* **22:** 148-155.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30:** 2114–2120.

Bruvo R, Michiels NK, D'Souza TG, Schulenburg H. 2004. A simple method for the calculation of microsatellite genotype distances irrespective of ploidy level. *Molecular Ecology* **13:** 2101–2106.

Clark LV, Jasieniuk M. 2011. POLYSAT: an R package for polyploid microsatellite analysis. *Molecular Ecology Resources* **11:** 562–566.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43:** 491-498.

Dryden IL. 2019. shapes package. Vienna, Austria: R Foundation for Statistical Computing.

Earl DA, vonHoldt BM. 2012. STRUCTURE HARVESTER: a website and program

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4:** 359-361.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14:** 2611-2620.

Fišer C, Robinson CT, Malard F. 2018. Cryptic species as a window into the paradigm shift of the species concept. *Molecular Ecology* **27:** 613-635.

Guo Y, Zhao P, Yue M. 2019. Canopy disturbance and gap partitioning promote the persistence of a pioneer tree population in a near-climax temperate forest of the Qinling Mountains, China. *Ecology and Evolution* **9:** 7676-7687.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* **41:** 95–98.

Hu YN, Zhao L, Buggs RJA, Zhang XM, Li J, Wang N. 2019. Population structure of *Betula albosinensis* and *Betula platyphylla*: evidence for hybridization and a cryptic lineage. *Annals of Botany* **123:** 1179–1189.

Jakobsson M, Rosenberg NA. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23:** 1801-1806.

Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24:** 1403-1405.

Kulju KKM, Pekkinen M, Varvio S. 2004. Twenty-three microsatellite primer pairs for *Betula pendula* (Betulaceae). **4:** 471-473.

Leliaert F, Verbruggen H, Wysor B, Clerck OD. 2009. DNA taxonomy in morphologically plastic taxa: Algorithmic species delimitation in the *Boodlea* complex (Chlorophyta: Cladophorales). *Molecular Phylogenetics and Evolution* **53:** 122-133.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27:** 2987-2993.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25:** 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25:** 2078-2079.

Li PC, Skvortsov AK. 1999. Betulaceae. Flora of China. Beijing: Science Press **4:** 286-313.

Lissambou B-J, Couvreur Thomas LP, Atteke C, Stévart T, Piñeiro R, Dauby G, Monthe Franck K, Ikabanga Davy U, Sonké B, M'batchi B, Hardy Olivier J. 2019. Species delimitation in the genus *Greenwayodendron* based on morphological and genetic markers reveals new species. *Taxon* **68:** 442-454.

Liu Y, Li Y, Song J, Zhang R, Yan Y, Wang Y, Du FK. 2018. Geometric

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

567 morphometric analyses of leaf shapes in two sympatric Chinese oaks:
568 *Quercus dentata* Thunberg and *Quercus aliena* Blume (Fagaceae). *Annals of*
569 *Forest Science* **75:** 90.
570 **McAllister HA, Rushforth K. 2011.** *Betula ashburneri*. *Curtis's Botanical*
571 *Magazine* **28:** 111–118.
572 **McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A,**
573 **Garimella K, Altshuler D, Gabriel S, Daly M. 2010.** The Genome Analysis
574 Toolkit: a MapReduce framework for analyzing next-generation DNA
575 sequencing data. *Genome Research* **20:** 1297-1303.
576 **Nagamitsu T, Kawahara T, Kanazashi A. 2006.** Endemic dwarf birch *Betula*
577 *apoiensis* (Betulaceae) is a hybrid that originated from *Betula ermanii* and
578 *Betula ovalifolia*. *Plant Species Biology* **21:** 19–29.
579 **Pritchard JK, Stephens M, Donnelly P. 2000.** Inference of population structure
580 using multilocus genotype data. *Genetics* **155:** 945-959.
581 **R Core Team. 2020.** R: A Language and Environment for Statistical Computing.
582 Vienna, Austria: R Foundation for Statistical Computing.
583 **Rosenberg NA. 2004.** DISTRUCT: a program for the graphical display of
584 population structure. *Molecular Ecology Notes* **4:** 137-138.
585 **Salojärvi J, Smolander O-P, Nieminen K, Rajaraman S, Safronov O, Safdari P,**
586 **Lamminmäki A, Immanen J, Lan T, Tanskanen J, Rastas P, Amiryousefi**
587 **A, Jayaprakash B, Kammonen JI, Hagqvist R, et al. 2017.** Genome
588 sequencing and population genomic analyses provide insights into the
589 adaptive landscape of silver birch. *Nature Genetics* **49:** 904-912.
590 **Stamatakis A. 2006.** RAxML-VI-HPC: maximum likelihood-based phylogenetic
591 analyses with thousands of taxa and mixed models. *Bioinformatics* **22:**
592 2688–2690.
593 **Thomson AM, Dick CW, Pascoini AL, Dayanandan S. 2015.** Despite
594 introgressive hybridization, North American birches (*Betula* spp.) maintain
595 strong differentiation at nuclear microsatellite loci. *Tree Genetics &*
596 *Genomes* **11:** 101.
597 **Truong C, Palmé AE, Felber F, Naciri-Graven Y. 2005.** Isolation and
598 characterization of microsatellite markers in the tetraploid birch, *Betula*
599 *pubescens* ssp. *tortuosa*. *Molecular Ecology Notes* **5:** 96-98.
600 **Tsuda Y, Semerikov V, Sebastiani F, Vendramin GG, Lascoux M. 2017.**
601 Multispecies genetic structure and hybridization in the *Betula* genus across
602 Eurasia. *Molecular Ecology* **26:** 589–605.
603 **Tsuda Y, Ueno S, Ide Y, Tsumura Y. 2009.** Development of 14 EST-SSRs for *Betula*
604 *maximowicziana* and their applicability to related species. *Conservation*
605 *Genetics* **10:** 661-664.
606 **Wang N, Borrell JS, Bodles WJA, Kuttapitiya A, Nichols RA, Buggs RJA. 2014a.**
607 Molecular footprints of the Holocene retreat of dwarf birch in Britain.
608 *Molecular Ecology* **23:** 2771–2782.
609 **Wang N, Borrell JS, Buggs RJA. 2014b.** Is the Atkinson discriminant function a

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

610 reliable method for distinguishing between *Betula pendula* and *B.*
611 *pubescens* (Betulaceae)? *New Journal of Botany* **4:** 90-94.
612 **Wang N, Kelly LJ, McAllister HA, Zohren J, Buggs RJA. 2020.** Resolving
613 phylogeny and polyploid parentage using genus-wide genome-wide
614 sequence data from birch trees. *bioRxiv*:
615 https://doi.org/10.1101/2020.1107.1113.200444.
616 **Wang N, McAllister HA, Bartlett PR, Buggs RJA. 2016.** Molecular phylogeny and
617 genome size evolution of the genus *Betula* (Betulaceae). *Annals of Botany*
618 **117:** 1023–1035.
619 **Wang N, Thomson M, Bodles WJA, Crawford RMM, Hunt HV, Featherstone AW,**
620 **Pellicer J, Buggs RJA. 2013.** Genome sequence of dwarf birch (*Betula*
621 *nana*) and cross-species RAD markers. *Molecular Ecology* **22:** 3098–3111.
622 **White TJ, Bruns T, Lee S, Taylor T. 1990.** Amplification and direct sequencing of
623 fungal ribosomal RNA genes for phylogenetics *PCR protocols: a guide to*
624 *methods and applications*. New York: Academic press: 315–322.
625 **Whittall JB, Hellquist CB, Schneider EL, Hodges SA. 2004.** Cryptic species in an
626 endangered pondweed community (*Potamogeton*, Potamogetonaceae)
627 revealed by AFLP markers. *American Journal of Botany* **91:** 2022-2029.
628 **Wu B, Lian C, Hogetsu T. 2002.** Development of microsatellite markers in white
629 birch (*Betula platyphylla* var. *japonica*). *Molecular Ecology Notes* **2:** 413-415.
630 **Zohren J, Wang N, Kardailsky I, Borrell JS, Joecker A, Nichols RA, Buggs RJA.**
631 **2016.** Unidirectional diploid-tetraploid introgression among British birch
632 trees with shifting ranges shown by restriction site-associated markers.
633 *Molecular Ecology* **25:** 2413–2426.

634

635

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

636 **Figure legends**

637 **Figure 1** The distribution of samples used in the present study.

638 **Figure 2** (a) Principal component analysis (PCoA) of leaves of section *Costatae*
639 species. (b) 'Eigenleaves' showing leaf morphs represented by principal components
640 (PCs) at ± 3SD and shape variance explained by each PC. Each dot represents a leaf.

641 **Figure 3** Principal coordinate analysis (PCO) of section *Costatae* species at 15
642 microsatellite markers (a) and principal component analysis (PCoA) of section
643 *Costatae* species at 82,137 SNPs (b).

644 **Figure 4** STRUCTURE results of section *Costatae* at K values 5 and 6 based on 15
645 microsatellite markers (a) and admixture analysis of section *Costatae* at K values 5
646 and 6 at the 82,137 SNPs (b).

647 **Figure 5** Species tree from the maximum likelihood analysis of the diploid *Betula*
648 species using the supermatrix approach based on data from 2,285,620 SNPs.
649 Bootstrap support values of 100 were not shown. Numbers on the branches are
650 bootstrap support values between 60 and 100. The scale bar below indicates the mean
651 number of nucleotide substitutions per site. Species were classified according to Wang
652 et al. (2020).

653 **Figure 6** A schematic illustration of species delineation within section *Cosatate* (a)
654 and various sources of information used to distinguish species (b). Photos of each
655 species were placed below its names.

656 **Figure S1** Principal coordinate analysis (PCO) of section *Costatae* species at 15
657 microsatellite markers (a) and principal component analysis (PCoA) of section

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Costatae species at 82,137 SNPs (b).

Figure S2 The best number of clusters inferred using "Evanno test" method.

Figure S3 STRUCTURE results of section *Costatae* at K values from 2 to 6 based on 15 microsatellite markers.

Figure S4 The cross-validation error for each K value from 1 to 10.

Figure S5 Admixture results at K values from 2 to 10 based on 82,137 SNPs.

Figure S6 Phylogenetic tree from the maximum likelihood analysis of *B. buggsii* using ITS sequences. Species were classified according to Ashburner and McAllister (2016). Values above branches are bootstrap percentages of >50 %.

Table legends

Table 1 Detailed information on taxa of section *Costatae* used in the present study.

Table S1 Detailed information on populations used in the present study.

Table S2 Details of microsatellite primers used in the present study.

Table S3 Detailed information of samples used for ITS and RAD sequencing.

685 **Table 1** Detailed information on taxa of section *Costatae* used in the present study.

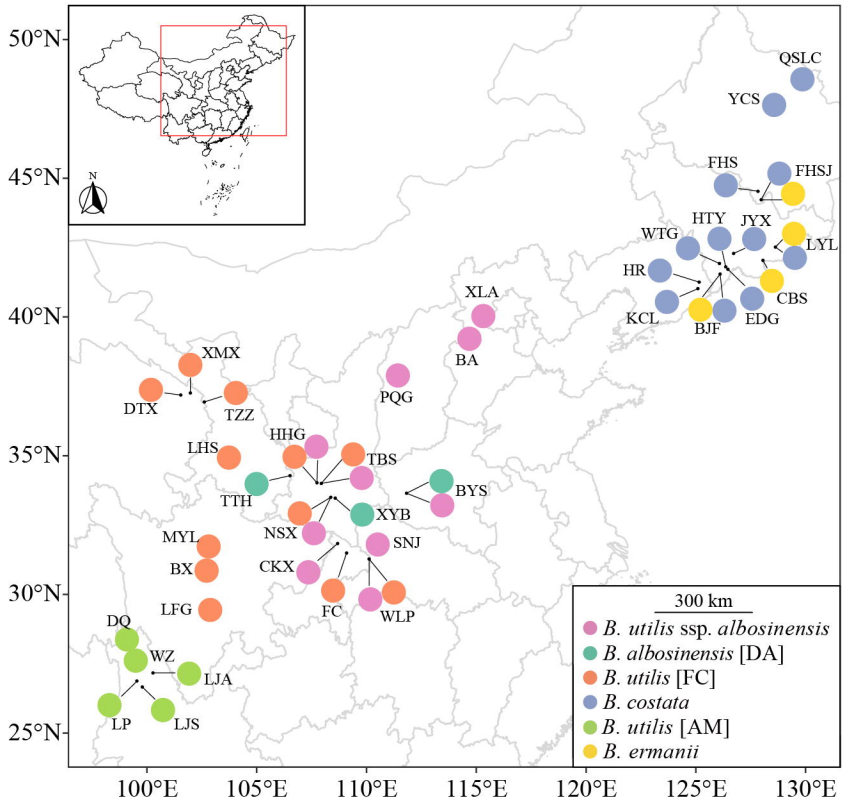
Species	Variety	Ploidy	Distribution	Reference
B. utilis	ssp. *utilis*	tetraploid	Nepal, eastwards through SE Tibet to Yunnan and west Sichuan where it merges with ssp. *albosinensis*	Ashburner and McAllister, 2016
	ssp. *utilis*	tetraploid	Gansu, Hebei, Ningxia, Qinghai, Shaanxi, west Sichuan, E and S Xizang, NW Yunnan	Li and Alexei, 1999, Flora of China
	var. *pratii*	tetraploid	Kangding, western Sichuan	Ashburner and McAllister, 2016
	ssp. *albosinensis*	tetraploid	North Sichuan, Hubei, south Gansu, south Ningxia, south Shaanxi, Shanxi, Henan and Hebei	Ashburner and McAllister, 2016
	ssp. *albosinensis*	diploid	south Shaanxi	Hu et al., 2019
B. albosinensis	ssp. *septentrionalis*	tetraploid	Western Sichuan	Ashburner and McAllister, 2016
B. ermanii	ssp. *ermanii*	tetraploid	Northeast China, Japan, Korea and the Russian Far East	Ashburner and McAllister, 2016
	var. *lanata*	tetraploid	Russia: from the eastern shores of Lake Baikal, eastward to the Pacific coast except Korea and north China	Ashburner and McAllister, 2016

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

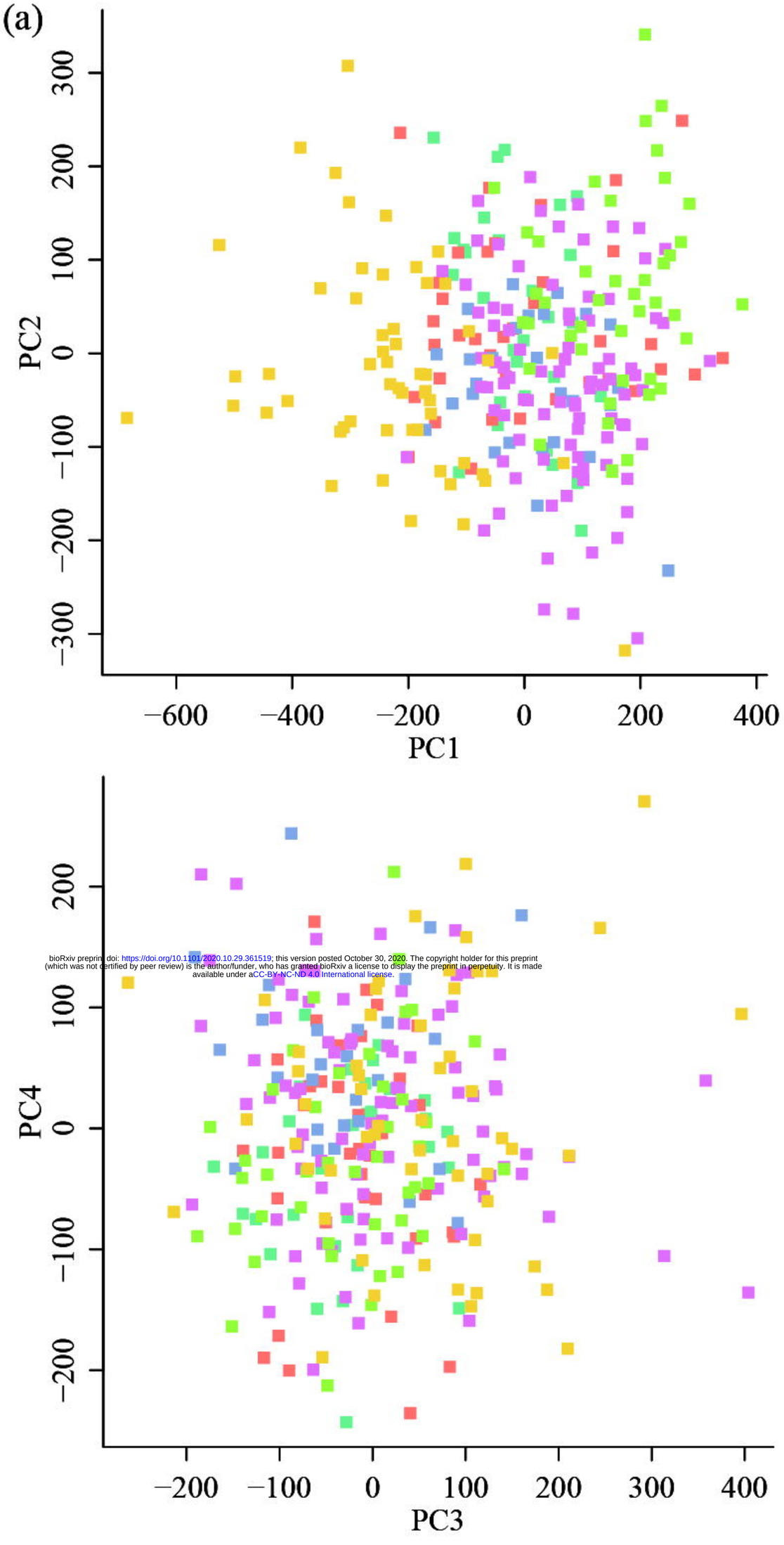
B. costata	NA	diploid	Northeast China	Ashburner and McAllister, 2016
B. ashburneri	NA	diploid	Southeast Tibet, Northwest Yunnan, Southwest Sichuan and possibly Shaanxi	McAllister, 2011; Ashburner and McAllister, 2016

686

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.


50°N
45°N
40°N
35°N
30°N
25°N
100°E
105°E
110°E
115°E
120°E
125°E
130°E
N
QSLC
YCS
FHS
FHSJ
HTY
JYX
WTG
LYL
HR
CBS
KCL
BJF
EDG
XLA
BA
XMX
DTX
TZZ
PQG
HHG
LHS
TBS
TTH
BYS
XYB
NSX
SNJ
MYL
BX
CKX
FC
WLP
LFG
DQ
WZ
LJA
LP
LJS
300 km
B. utilis ssp. albosinensis
B. albosinensis [DA]
B. utilis [FC]
B. costata
B. utilis [AM]
B. ermanii

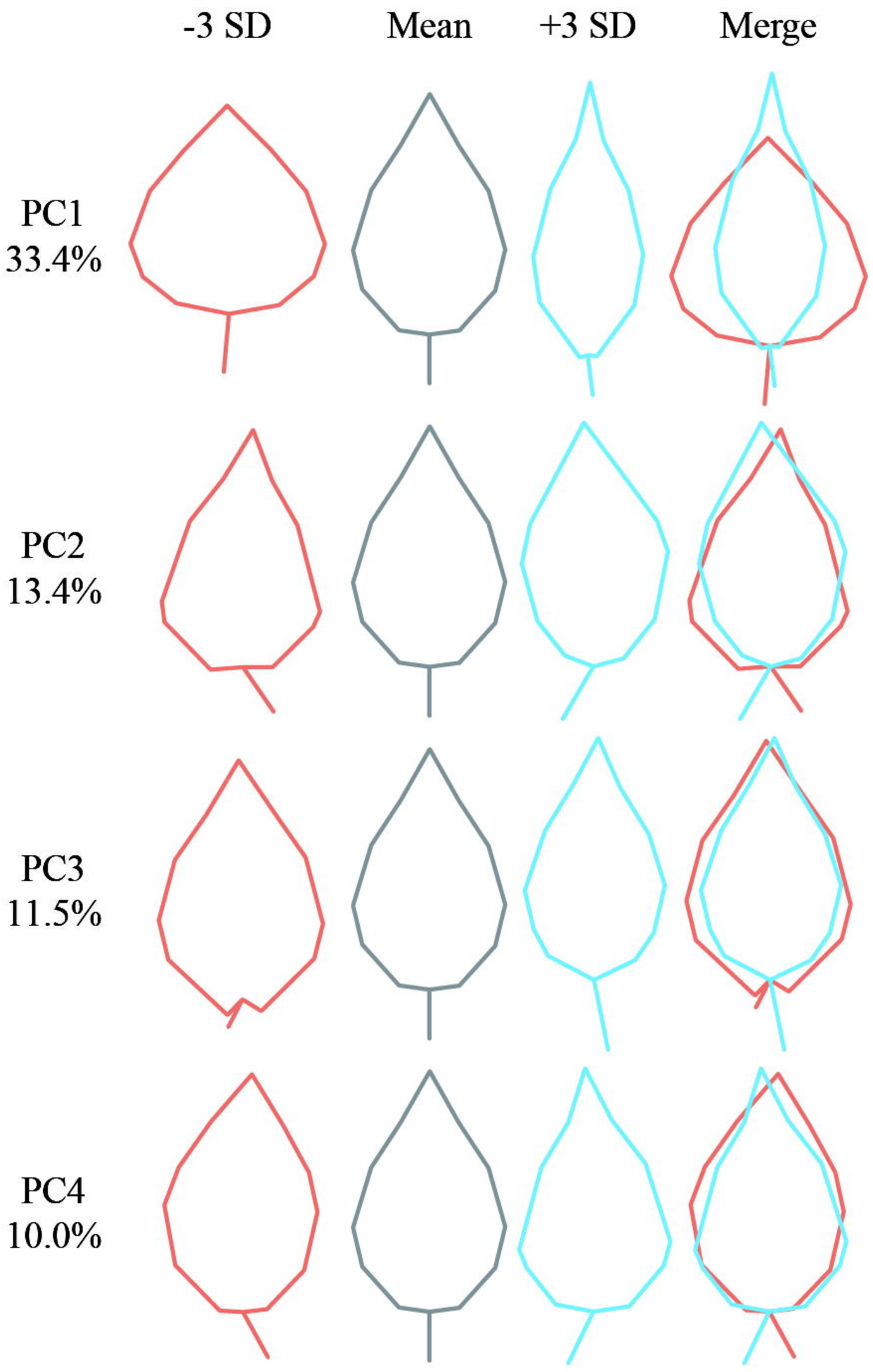
(a)



bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

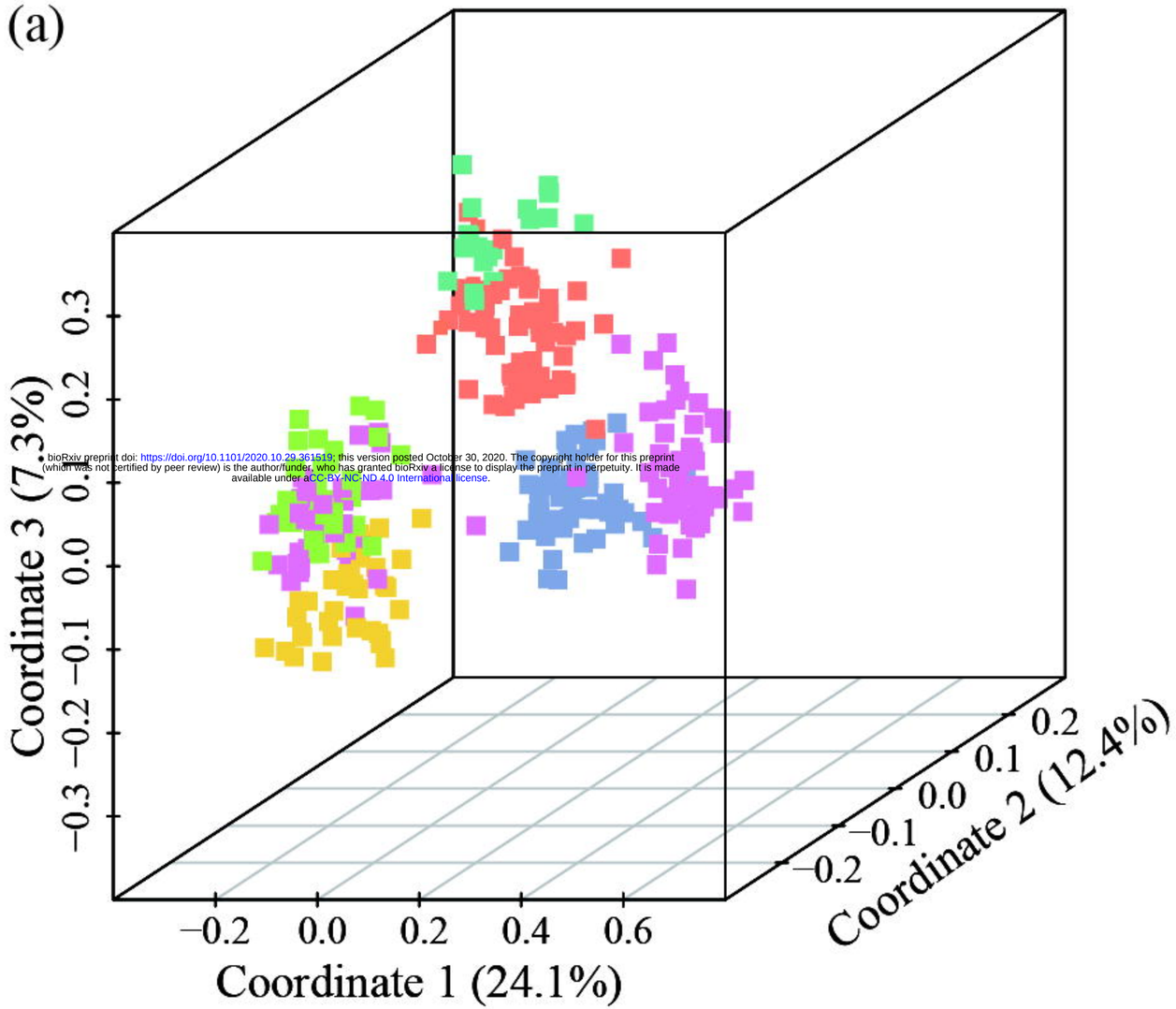
B. utilis ssp. *albosinensis*
B. albosinensis [DA]
B. utilis [FC]
B. costata
B. utilis [AM]
B. ermanii

(b)

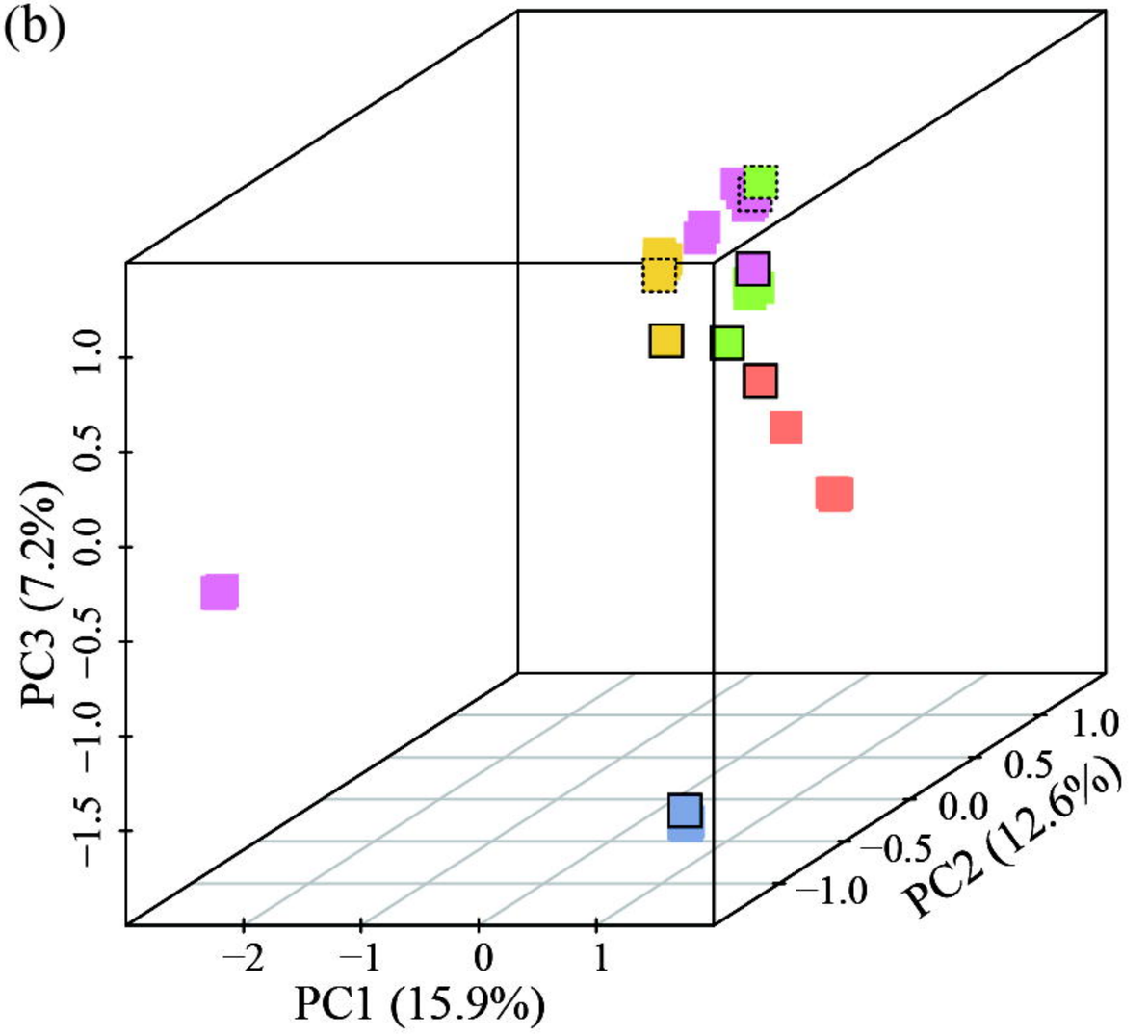


bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(a)



(b)



- *B. utilis* ssp. *albosinensis*
- *B. albosinensis* [DA]
- *B. utilis* [FC]
- *B. costata*
- *B. utilis* [AM]
- *B. ermanii*
- *B. ashburneri*
- *B. costata*
- *B. utilis* ssp. *albosinensis*
- *B. utilis* ssp. *utilis*
- *B. ermanii*
- *B. albosinensis* var. *septentrionalis*
- *B. utilis* var. *prattii*
- *B. ermanii* var. *lanata*

(a)

K = 5

K = 6

(b)

K = 5
4/20 runs

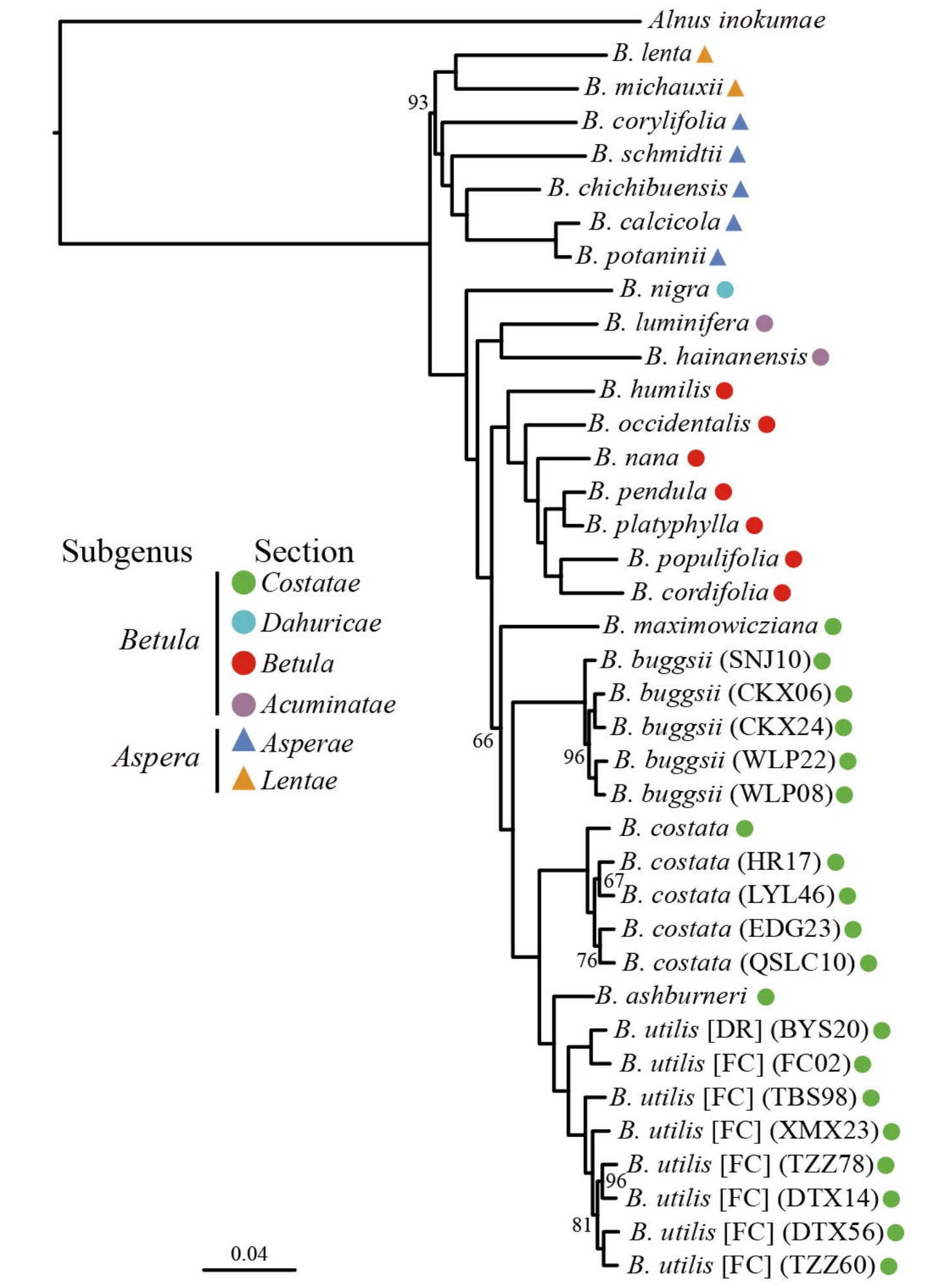
K = 6
14/20 runs

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361519; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

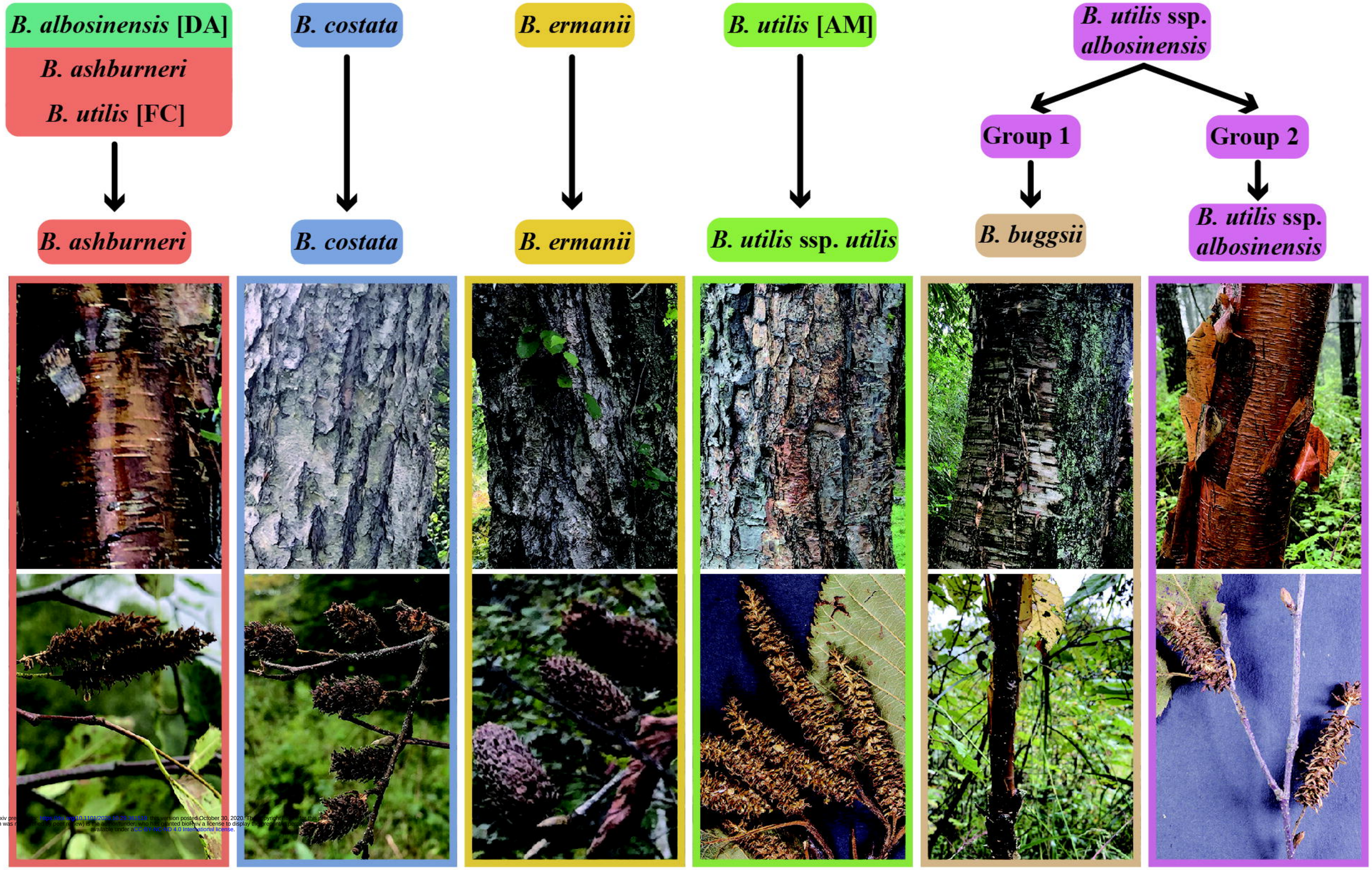
B. utilis ssp. *albosinensis*
B. albosinensis [DA]
B. utilis [FC]
B. costata
B. utilis [AM]
B. ermanii

B. ashburneri
B. costata
B. utilis ssp. *albosinensis*
B. utilis ssp. *utilis*
B. ermanii

B. albosinensis var. *septentrionalis*
B. utilis var. *prattii*
B. ermanii var. *lanata*


Alnus inokumae
B. lenta
B. michauxii
B. corylifolia
B. schmidtii
B. chichibuensis
B. calcicola
B. potaninii
B. nigra
B. luminifera
B. hainanensis
B. humilis
B. occidentalis
B. nana
B. pendula
B. platyphylla
B. populifolia
B. cordifolia
B. maximowicziana
B. buggsii (SNJ10)
B. buggsii (CKX06)
B. buggsii (CKX24)
B. buggsii (WLP22)
B. buggsii (WLP08)
B. costata
B. costata (HR17)
B. costata (LYL46)
B. costata (EDG23)
B. costata (QSLC10)
B. ashburneri
B. utilis [DR] (BYS20)
B. utilis [FC] (FC02)
B. utilis [FC] (TBS98)
B. utilis [FC] (XMX23)
B. utilis [FC] (TZZ78)
B. utilis [FC] (DTX14)
B. utilis [FC] (DTX56)
B. utilis [FC] (TZZ60)
93
66
96
67
76
96
81
Subgenus
Section
Betula
Costatae
Dahuricae
Betula
Acuminatae
Aspera
Asperae
Lentae
0.04

(a)



bioRxiv preprint doi: https://doi.org/10.1101/2020.10.29.361154; this version posted October 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(b)