Molecular Phylogenetics and Evolution 160 (2021) 107126

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Resolving phylogeny and polyploid parentage using genus-wide genome-wide sequence data from birch trees

Nian Wang [a,b], Laura J. Kelly [a,c], Hugh A. McAllister [d], Jasmin Zohren [e], Richard J.A. Buggs [a,c,*]

[a] School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, UK
[b] 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
[c] Royal Botanic Gardens Kew, Richmond, Surrey TW9 3AB, UK
[d] Institute of Integrative Biology, Biosciences Building, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK
[e] Sex Chromosome Biology Lab, the Francis Crick Institute, 1 Midland Road, London, NW1 1AT, UK

ARTICLE INFO

Keywords:
Polyploidy
Whole genome duplication
Hybridisation
Phylogenomics
Betula

ABSTRACT

Numerous plant genera have a history including frequent hybridisation and polyploidisation (allopolyploidisation), which means that their phylogeny is a network of reticulate evolution that cannot be accurately depicted as a bifurcating tree with a single tip per species. The genus *Betula*, which contains many ecologically important tree species, is a case in point. We generated genome-wide sequence reads for 27 diploid and 36 polyploid *Betula* species or subspecies using restriction site associated DNA (RAD) sequences. These reads were assembled into contigs with a mean length of 675 bp. We reconstructed the evolutionary relationships among diploid *Betula* species using both supermatrix (concatenation) and species tree methods. We identified the closest diploid relatives of the polyploids according to the relative rates at which reads from polyploids mapped to contigs from different diploid species within a concatenated reference sequence. By mapping reads from allopolyploids to their different putative diploid relatives we assembled contigs from the putative sub-genomes of allopolyploid taxa. We used these to build new phylogenies that included allopolyploid sub-genomes as separate tips. This approach yielded a highly evidenced phylogenetic hypothesis for the genus *Betula*, including the complex reticulate origins of the majority of its polyploid taxa. Our phylogeny divides the genus into two well supported clades, which, interestingly, differ in their seed-wing morphology. We therefore propose to split *Betula* into two subgenera.

1. Introduction

The evolution of plant diversity cannot be fully understood unless we can reconstruct phylogenetic relationships for allopolyploids, which are hybrid species with duplicated genomes. Each parental sub-genome present within an allopolyploid can be represented as its own tip within a phylogenetic tree. To do this is challenging, because it is difficult to phase molecular markers sequenced from allopolyploids and allocate them into their different parental subgenomes (Eriksson et al., 2018; Oxelman et al., 2017). It is easy to mistake homoeologs (duplicated genes derived from the different parents of an allopolyploid) for paralogs (genes arising from duplication within a genome) and vice versa (Brysting et al., 2011; Linder and Rieseberg, 2004). In addition, over time allopolyploids tend to be diploidised (Buggs et al., 2012; Mandáková and Lysak, 2018), losing parts of each sub-genome, meaning that studies of single genes may not yield information about all parental progenitors.

Resolving parental origins of polyploid sub-genomes allows for a deeper understanding of the evolutionary history of an allopolyploid. For instance, knowledge of the identity of the closest diploid relatives of the allohexaploid *Triticum aestivum* (bread wheat) allowed accurate assembly of its genome by comparison with assembled genomes of those relatives (Avni et al., 2017; Luo et al., 2017). Similarly, the structure of the allotetraploid origin of *Gossypium hirsutum* (Upland cotton) has been better resolved by the discovery and sequencing of two close diploid relatives of its progenitors (Li et al., 2015). However, for many polyploids of economic and ecological importance, we do not know the identity of the closest living relatives of their progenitor genomes. We

* Corresponding author at: School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, UK.
E-mail addresses: r.buggs@kew.org, r.buggs@qmul.ac.uk (R.J.A. Buggs).

https://doi.org/10.1016/j.ympev.2021.107126
Received 8 January 2020; Received in revised form 15 February 2021; Accepted 22 February 2021
Available online 27 February 2021
1055-7903/

need accessible approaches to make this possible.

The genomic composition of some polyploids has been determined using small numbers of genes (Dauphin et al., 2018; Jones et al., 2013; Lott et al., 2009; Marcussen et al., 2015; Rothfels et al., 2017). Genome-wide approaches have been used to detect the presence of hybrid polyploids and suggest their closest relatives (Fitz-Gibbon et al., 2017; Kamneva et al., 2017; McKain et al., 2016; Morales-Briones et al., 2018), but we are not aware of studies that have thus far used genome-wide data to place the different parental sub-genomes of allopolyploids as their own tips within phylogenies.

A relatively inexpensive method of generating genome-wide data for phylogenomics from large numbers of individuals is through sequencing of restriction-site associated DNA (RAD) libraries with short reads of 50–150 bp. This is widely used as a method of genome-wide single nucleotide polymorphism (SNP) genotyping in non-model organisms for population genomic analyses (Andrews et al., 2016; Barchi et al., 2011; Emerson et al., 2010; Etter et al., 2011; Hohenlohe et al., 2010; Zohren et al., 2016). SNP data from RAD-seq has also been used in phylogenetic reconstruction using supermatrix approaches which concatenate all loci (Cariou et al., 2013; Cruaud et al., 2014; Eaton and Ree, 2013; Eaton et al., 2016; Fitz-Gibbon et al., 2017; Gonen et al., 2015; Hipp et al., 2014; Massatti et al., 2016; Pante et al., 2015; Rubin et al., 2012; Wagner et al., 2013). A few RAD studies have used species tree approaches, which take into account the possibility of different evolutionary histories for separate loci, for analysis of short read RAD-seq data. For example, Eaton and Ree (2013) used RAD loci inferred from single end reads to build a species tree in the genus *Pedicularis* (lousewort), DaCosta and Sorenson (2016) used single end reads to construct species trees in two avian genera and Hou et al. (2015) used paired-end reads to build a species tree for the genus *Diapensia* (pincushion plant). Using short reads it is hard to assign homology reliably, and the loci assembled are likely to be short, yielding little phylogenetic information. We reasoned that extra power for phylogenetic analysis may be gained by sequencing RAD libraries with 300 bp paired-end reads and assembling these reads against a reference genome to generate longer contigs spanning restriction enzyme and variable sites. These contigs could be aligned to each other and individual phylogenies reconstructed for each locus, for input into species tree methods, or the alignments combined, for a supermatrix approach. So far, we are not aware of any studies which have sequenced longer RAD loci in an attempt to gain greater power for species tree methods.

The genus *Betula* (birches) includes about 65 species and subspecies with ranges across the Northern Hemisphere (Ashburner and McAllister, 2016). Some act as keystone species of forests across Eurasia and North America (Ashburner and McAllister, 2016). Various birch species are planted for timber, paper, carbon sequestration and ecological restoration. Some birch species are endangered with narrow distributions. In addition, there is concern about the increasing threat posed by the bronze birch borer (Muilenburg and Herms, 2012; Shaw et al., 2014). Previous phylogenetic analyses of *Betula* using nuclear genes (ITS, NIA and ADH), chloroplast genes (matK and rbcL) and AFLPs provided limited resolution of relationships among species and partly contradicted each other (Järvinen et al., 2004; Li et al., 2005; Li et al., 2007; Schenk et al., 2008). In addition, molecular phylogenies based on nuclear genes contradict some species groupings proposed in a recent monograph based on morphology, such as the placement of the ecologically and economically important *B. maximowicziana* (monarch birch) (Ashburner and McAllister, 2016; Wang et al., 2016).

Hybridisation is frequent within *Betula* (Anamthawat-Jónsson and Tómasson, 1990, 1999; Anamthawat-Jónsson and Thórsson, 2003; Anamthawat-Jónsson et al., 2010; Barnes et al., 1974; Eidesen et al., 2015; Johnsson, 1945; Tsuda et al., 2017; Wang et al., 2014) and nearly 60% of *Betula* species are polyploids (Wang et al., 2016) with ploidy ranging from diploid to dodecaploid (Ashburner and McAllister, 2016). Some species contain different cytotypes, such as *B. chinensis* (6× and 8×) (Ashburner and McAllister, 2016). The parental origins of most polyploids in the genus remain unresolved. One of the most studied polyploids is the tetraploid *B. pubescens* (downy birch), with different lines of evidence suggesting as candidate parents: *B. pendula* based on RAPD markers (Howland et al., 1995), *B. humilis* or *B. nana* based on ADH (Järvinen et al., 2004), *B. humilis* based on morphology (Walters, 1968) or *B. lenta* based on SNPs (Salojarvi et al., 2017). This uncertainty hinders genomic research on *B. pubescens*, the most widespread birch tree in Europe and western Asia.

Here, in order to better resolve the phylogeny of *Betula* and elucidate the parental origins of its polyploid species we use a RAD-seq approach with 300 bp paired-end reads mapped against the *B. pendula* reference genome (Salojarvi et al., 2017). We construct the phylogeny of diploid species using supermatrix and species tree methods. As a heuristic method for analyzing the parental origin of the polyploid species, we create a reference using contigs from all diploid species and compare the genomic similarity between each polyploid species and all diploids by mapping reads of each polyploid species to the reference. Polyploid taxa should have a higher level of genetic similarity to diploids closely related to their ancestors and hence a higher number of mapped reads. By mapping reads from allopolyploids to their different putative diploid relatives we assembled contigs from the putative sub-genomes of allopolyploid taxa, allowing us to place them as their own tips within the phylogeny of the genus. These approaches together yielded a well-resolved phylogenetic history for *Betula*, including polyploid taxa.

2. Materials and methods

2.1. Sample collection

Samples were obtained from living collections in Stone Lane Gardens (SL hereafter), Ness Gardens (N hereafter), the Royal Botanic Garden Edinburgh (RBGE) or collected from the wild by the research group (Table S1). The samples in the present study were largely from the same individuals as those used previously for genome size estimation (Wang et al., 2016). Morphological characters were used to confirm the identity of each taxon sampled. *Alnus inokumae* was chosen as the outgroup as *Alnus* has been shown to be sister to *Betula* (Li et al., 2007). In addition, *A. orientalis* and *Corylus avellana* were included for marker development. Herbarium specimens of most of these samples have been deposited at the Natural History Museum London and RBGE with accession numbers provided in Table S1.

2.2. DNA extraction, RAD library preparation and Illumina sequencing

Genomic DNA was isolated from silica-dried cambial tissue or leaves following a modified 2X CTAB (cetyltrimethylammonium bromide) protocol (Wang et al., 2013). The isolated DNA was assessed with a 1.0% agarose gel and measured with a Qubit 2.0 Fluorometer (Invitrogen, Life technologies) using Broad-range assay reagents. RAD libraries were prepared following a previous protocol with slight modifications (Etter et al., 2011). Briefly, 0.5–1.0 μg of genomic DNA for each sample was heated at 65 °C for 2–3 h prior to digestion with PstI (New England Biolabs, UK). This enzyme has a 6 bp recognition site and leaves a 4 bp overhang. Digestion was followed by ligation of barcoded P1 adapters. Ligated DNA was sheared using a Bioruptor (KBiosciences, UK) instrument in 1.5 mL tubes (high intensity, duration 30 s followed by a 30 s pause which was repeated eight times). Sheared fragments were evenly distributed between 100 bp and 1500 bp and fragments of ~600 bp were selected using Agencourt AMPure XP Beads (New England Biolabs) following a protocol of double-size selection. A ratio of bead:DNA solution of 0.55 was used to remove large fragments and then a second round of size selection was conducted, using 5 μl of bead solution concentrated from a starting volume of 20 μl. After end-repair and A-tailing, the size-selected DNA was ligated to P2 adapters (400 nm) and PCR amplified. PCR amplification was carried out in 25 μl reactions consisting of 0.46 vol ddH_2O and template DNA (4–5 ng), 0.5 vol 2 ×

Phusion Master Mix (New England Biolabs), and 0.04 vol P1 and P2 amplification primers (10 nm), using the following cycling parameters: 98 °C for 30 s followed by 12 cycles of 98 °C for 10 s and 72 °C for 60 s. Three or four independent PCR replicates were conducted for each sample to achieve a sufficient amount of the library. The final library was quantified using a Bioanalyzer and a Qubit 2.0 Fluorometer (Invitrogen, Life Technologies) using high-sensitivity assay reagents and was normalized prior to sequencing. The quantified library was sequenced on an Illumina MiSeq machine using MiSeq Reagent Kit v3 (Illumina) at the Genome Centre of Queen Mary University of London.

2.3. RAD data trimming and demultiplexing

The raw data were trimmed using Trimmomatic (Bolger et al., 2014) in paired-end mode with the following steps. First, LEADING and TRAILING steps were used to remove bases from the ends of a read if the quality is below 20. Then a SLIDINGWINDOW step was performed with a window size of 1 and a required quality of 20. Finally, a MINLENGTH step was used to discard reads shorter than 100 bp. FastQC was used to check various parameters of sequence quality in both raw and trimmed datasets (Andrews, 2014). The trimmed data were demultiplexed, using the process_radtags module of Stacks (Catchen et al., 2013).

2.4. Reads mapping, sequence alignment and trimming

The whole genome assembly of *B. pendula* (Salojarvi et al., 2017) was used as a reference for mapping our RAD data, to separate orthologous loci (i.e. mapped segments of DNA) from paralogous loci, and to anchor reads with adjacent restriction cutting sites. Mapping of trimmed reads for each sample was conducted using the 'Map Reads to Reference' tool in the CLC Genomics Workbench v. 8. A similarity value of 0.8 and the fraction value of 0.8 were applied as the threshold. Reads with non-specific matches were discarded and any regions with coverage of below three were removed. A consensus sequence with a minimum contig length of 300 bp was created for each sample. This resulted in 13,597 (*Alnus inokumae*) to 30,717 (*B. pendula*) contigs assembled per sample. *Betula glandulosa* was excluded from further analysis because only 216 loci were mapped at a sufficient read depth. Multiple sequence alignments for the individual loci were generated using MAFFT v.6.903 (Katoh et al., 2005) with default parameters. Aligned sequences were trimmed using trimAl v1.2rev59 (Capella-Gutierrez et al., 2009); gaps present in 40% of taxa or above were removed (-gt 0.6).

2.5. Diploid species tree inference

Two datasets were used for phylogenetic analysis: dataset 1 (D1 hereafter) includes 20 diploid *Betula* samples and dataset 2 (D2 hereafter) 27 diploid samples. In D2, some diploid species were represented by more than one sample (Table S2). RAD loci $\geq$ 300 bp in length that occurred in a minimum of four *Betula* samples were used for gene tree inference. The gene tree for each locus was estimated using the maximum-likelihood method (ML) in RAxML v. 8.1.16 (Stamatakis, 2006). A rapid bootstrap analysis with 100 bootstraps and 10 searches was performed under a GTR + GAMMA nucleotide substitution model. The species tree was estimated from the gene trees with ASTRAL-II v5.5.7 (Mirarab and Warnow, 2015) and ASTRID (Vachaspati and Warnow, 2015). Branch support in the ASTRAL and ASTRID trees was assessed via calculation of local posterior probabilities based on the gene tree quartet frequencies (Sayyari and Mirarab, 2016) and multi-locus bootstrapping (Vachaspati and Warnow, 2015), respectively. All loci used for building gene trees and inferring the species trees were concatenated into a supermatrix, using custom shell scripts, which was analysed in RAxML v. 8.1.16 using the same settings as above. The consensus tree generated above was visualised in FigTree v.1.3.1 (http://tree.bio.ed.ac.uk/software/figtree).

2.6. Diploid phylogenetic networks

We used the Species Networks applying Quartets (SNaQ) method (Solís-Lemus and Ané, 2016) implemented in the software PhyloNetworks 0.5.0 (Solis-Lemus et al., 2017) to investigate whether the species tree without hybridisation events or a phylogenetic network with one or more hybridisation events better describes the diploid species relationships within *Betula*. Phylogenetic trees generated in RAxML were used to estimate quartet concordance factors (CFs), which represent the proportion of genes supporting each possible relationship between each set of four species. These CFs were then used to reconstruct phylogenetic networks under incomplete lineage sorting (ILS) and differing numbers of hybridisation events, and to calculate their respective pseudolikelihoods. To determine whether a tree accounting for ILS or a network better fits the observed data, we estimated the best phylogenetic network with hybridisation events (h) ranging from 0 to 5 using the phylogeny obtained with ASTRID as a starting tree. Using a value of h = 0 will yield a tree without reticulation, h = 1 will yield a network with a maximum of one reticulation and so forth. The fit of trees and networks to the data was evaluated based on pseudo-deviance values, and estimated inheritance probabilities (i.e. the proportion of genes contributed by each parental population to a hybrid taxon) were visualised. This test compares the score of each network based on the negative log-pL, where the network with the lowest value has the best fit.

2.7. Identification of putative diploid progenitors of polyploid species

We sought to infer the putative origins of polyploid species of *Betula* by read-mapping. First, we used RAD loci present in at least 15 out of 20 diploid *Betula* taxa to create a reference. A consensus assembly for each of these RAD loci was produced for each diploid taxon that contained them, with heterozygous sites within diploids represented by ambiguity codes. These locus assemblies were then concatenated together in a single large reference sequence for each diploid, containing thousands of loci. Each locus was separated from the next by 500 "N" bases. The reference sequences for all diploid taxa were placed in a single multi-sequence FASTA file, which was used as a reference sequence for mapping of polyploid reads.

All trimmed RAD-seq reads from each polyploid taxon were mapped to this reference containing loci from all diploid taxa. We used strict parameters in CLC Genomic Workbench: a fraction value of 0.9 and a similarity value of at least 0.995. Reads with non-specific matches were discarded: thus we retained only those reads from polyploid taxa that mapped uniquely to one locus in one diploid taxon. We built a consensus assembly of the mapped reads from each polyploid taxon, excluding regions of the reference with a mapped read coverage of below three. We also excluded contigs with a length of less than 300 bp. Variable sites were represented by ambiguity codes in the consensus assembly for each polyploid.

For each polyploid taxon, we counted the number of loci in each diploid taxon that yielded a contig made from mapped reads from that polyploid taxon. We reasoned that the diploid taxa most closely related to the progenitors of each polyploid would yield the highest numbers of contigs from the mapped polyploid reads. As the number of loci available from each diploid species in the reference was variable, for each diploid species we calculated the proportion of loci which yielded contigs, and plotted these in charts for comparison (Table S2). We plotted histograms of the proportion of loci in the reference for each diploid species to which reads from each polyploid species mapped. We expected a higher proportion of consensus loci to be mapped in those diploids that were closest relatives of the progenitor species of each polyploid species. We therefore sought to identify representatives of the diploid progenitors for each polyploid based on the number of mapped consensus loci assuming that the number of progenitors could not be more than half the ploidy level of each polyploid. In addition, as a control, we mapped reads from diploid species to the reference containing loci from all diploid taxa using the same parameters as described above. We found a small number of loci of each diploid were mapped by reads from other diploid species, with the exception of *B. calcicola* and *B. potaninii*, for which a relatively high number of loci (>1000) were mapped in each species by reads from the other (Fig. S1).

2.8. Phylogeny incorporating polyploid species

For those polyploids for which we could identify putative diploid parental species, we separated their RAD-loci into different sub-genomes using another concatenated single reference, similar to the one described in the paragraph above, but this time containing all 50,870 loci present in a minimum of four *Betula* taxa of D1. For each of these polyploids, we extracted a RAD consensus sequence from each mapped diploid locus, with a minimum length of 200 bp; we excluded any sequence where the polyploid had not mapped to their putative parental representatives for that locus. We then excluded loci where reads from one diploid had mapped to another diploid species (see above - Identification of putative diploid progenitors of polyploid species). We constructed phylogenies including these phased polyploid RAD loci with the diploid RAD loci (Fig. S2). Individual gene trees were constructed in RAxML v. 8.1.16 using the same parameters as described above (see - Species tree inference) and the species tree was inferred using ASTRID. The putative diploid progenitors which we included for phylogenetic analysis are provided in Table S2.

2.9. Simple sequence repeat analysis

To develop markers for future use in all *Betula* species for population genetic analyses, mapped consensus sequences with length equal to or greater than 300 bp were mined for simple sequence repeats (SSR) using the QDD pipeline version 3.1.2 (Meglécz et al., 2014). Consensus sequences with a repeat motif of 2–5 bp, and repeated a minimum of five times, were screened using the Downstream QDD pipeline version 3.1.2. Primer pairs were designed within 200 bp flanking regions using PRIMER3 software (Untergasser et al., 2012). The primer table output by the QDD version 3.1.2 pipeline allows selection of the best primer pair design for each SSR locus. We filtered primer pairs according to parameters provided by QDD version 3.1.2. The selected SSR loci had: a minimum number of 7 motif repeats within the SSR sequence; a maximum primer alignment score of 5; a minimum of 20 bp forward and reverse flanking region between SSR and primer sequences; and a high-

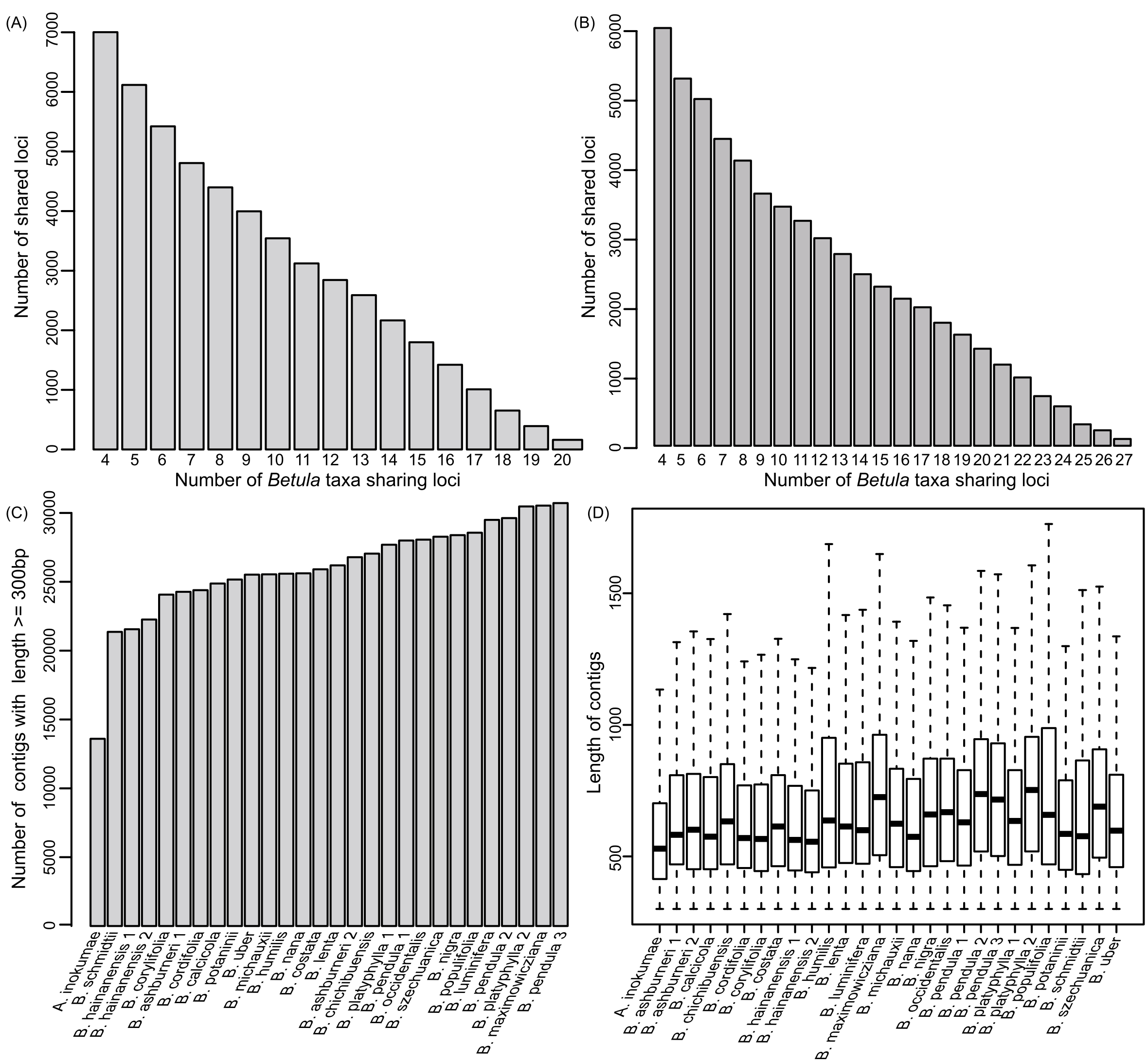

Fig. 1. Detailed information on the number of shared loci and number of contigs: (A) number of shared loci only in between four and 20 of the diploid *Betula* species of D1; (B) number of shared loci only in between four and 27 of the diploid *Betula* species of D2; (C) number of contigs with length above 300 bp for *A. inokumae* and each of the 27 diploid *Betula* species of D2; (D) length of contigs for *A. inokumae* and each of the 27 diploid *Betula* species of D2. The whiskers of the boxplot from the bottom to the top indicate the minimum, the first quartile, the median, the third quartile and the maximum value of contig length excluding outliers.

quality primer design without homopolymer, nanosatellite and microsatellite sequence in the primer or flanking sequences. For polyploid species of *Betula*, *A. inokumae*, *A. orientalis* and *C. avellana*, we selected SSR loci with a minimum number of 5 motif repeats as a majority of loci had 5 or 6 motif repeats within the SSR sequence.

3. Results

3.1. RAD data description

The number of trimmed reads per diploid taxon ranges from 1,065,196 to 2,560,486 (average of 1,508,904) with between 881,333 and 2,252,171 (80.60% – 90.75%) mapped to the *B. pendula* genome for each of the 27 diploid *Betula* and 707,914 (51.64%) for the outgroup *A. inokumae* (Table S1). In D1, 162 loci are present in all 20 *Betula* diploid taxa and 7002 present in only four of these (Fig. 1A), whereas for D2 99 loci are present in all 27 *Betula* diploid individuals and 6078 present in only four (Fig. 1B). Contigs of ≥300 bp, with an average length of 580.8 bp – 755.8 bp, varied in number between 13,597 in *A. inokumae* and 30,717 in *B. pendula* (Fig. 1C, D). The BioProjectID for the raw reads on the NCBI Sequence Read Archive repository is PRJNA679451.

3.2. Diploid phylogenetic inference

The concatenated D1 (50,870 loci) and D2 (58,442 loci) datasets include 31,815,738 and 35,859,769 nucleotides with 60.25% and 63.12% missing data (gaps and undetermined characters), respectively. The three approaches used for phylogenetic analysis of D1 (ASTRAL, ASTRID and supermatrix) all produced well resolved trees that split the genus into two major clades. The ASTRAL species tree (Fig. 2A) and concatenation tree (Fig. 2B) have identical topologies, whereas the ASTRID tree for this dataset differs in the placement of *B. cordifolia* (Fig. S3). Phylogenetic trees for D2 inferred with the species tree methods also separate the genus into two major clades, similar to the D1 trees, but with some differences in the placement of a small number of taxa within the largest clade (Figs. S4 and S5). The concatenation tree of D2 does not recover the same major clades as the other analyses, although these differences are not well supported (Fig. S6).

We compared the two major clades found in most of our analyses with the morphology-based taxonomy of *Betula* in Ashburner and McAllister (2016). The sections of Ashburner and McAllister are shown in Fig. 2 for each diploid species. While our major clades did not fit neatly with the Ashburner and McAllister taxonomy, there was a correspondence between seed wing morphology as recorded by Ashburner and McAllister and our major clades. Species in Clade 1 had no or very narrow seed wings, but species in Clade 2 had large seed wings.

3.3. Diploid phylogenetic networks

The pseudolikelihood values of hybrid nodes decreased sharply from h = 0 to h = 2, with only marginal improvements when further increasing the number of hybridisation events (Fig. S7), suggesting the best-fitting phylogenetic model is one involving two main hybridisation events. The D1 phylogenetic network when h = 2 is similar to the phylogenetic trees for this dataset (Fig. 3), but with evidence for introgressive hybridisation between *B. nana* and *B. pendula* and between *B. maximowicziana* and the lineage leading to *B. hainanensis* and *B. luminifera*.

3.4. Read-mapping of polyploid species

Of the RAD loci we assembled, 5045 were present in at least 15 out of 20 diploid *Betula* taxa. Each of these were present in a mean of 17.54 diploid taxa. Thus, the multi-sequence FASTA file with the concatenated diploid reference sequence contained 88,488 assembled loci with a combined length of 69,805,419 bp. Each diploid species was represented by a mean of 4424 loci (minimum 3731 loci, maximum 5045 loci).

For each polyploid taxon, between 33,057 and 222,703 reads mapped to the diploid concatenated reference, giving 2674–17,104 assembled loci for each polyploid taxon. For 28 of the polyploid species or varieties the distribution of the number of loci assembled from each diploid taxon allowed identification of two or more parental lineages (Fig. 4, Table 1). Five polyploids, all with the ploidy level ≥8×, have putative progenitors from both of the two major diploid clades.

3.5. Phylogeny incorporating polyploid species' sub-genomes

When we included phased homoeologues from polyploids for which we could identify putative parental relatives in a phylogenetic analysis,

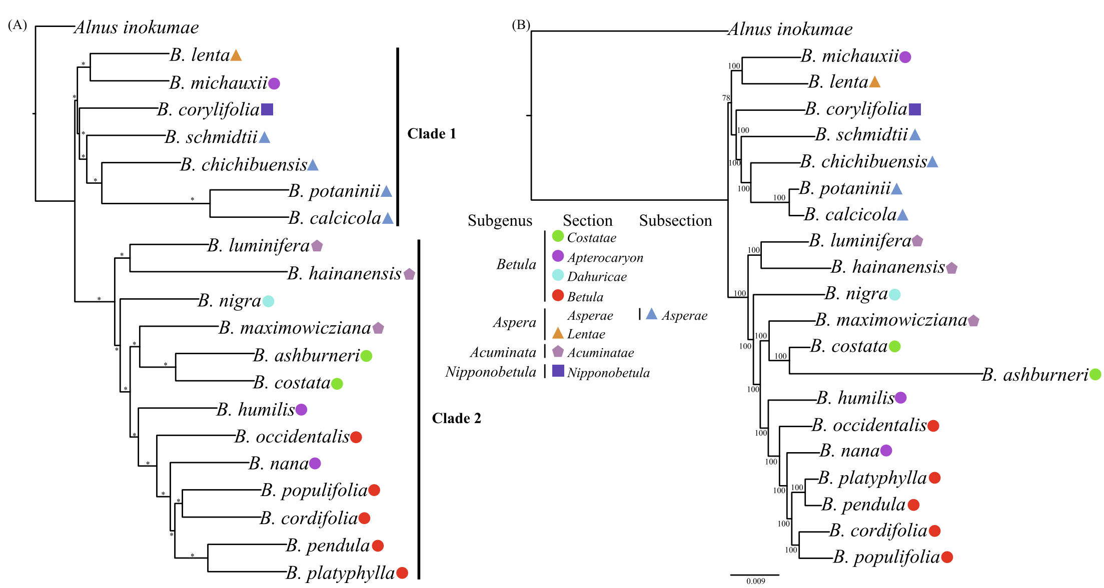

Fig. 2. Species tree from the maximum likelihood analysis of the 20 *Betula* diploids of D1 using ASTRAL (A) and the supermatrix (B) approach based on data from 50,870 loci. Asterisks on the branches of (A) indicate local posterior probabilities of 1 and numbers on the branches of (B) are bootstrap support values. The scale bar below (B) indicates the mean number of nucleotide substitutions per site. Species were classified according to Ashburner and McAllister (2016).

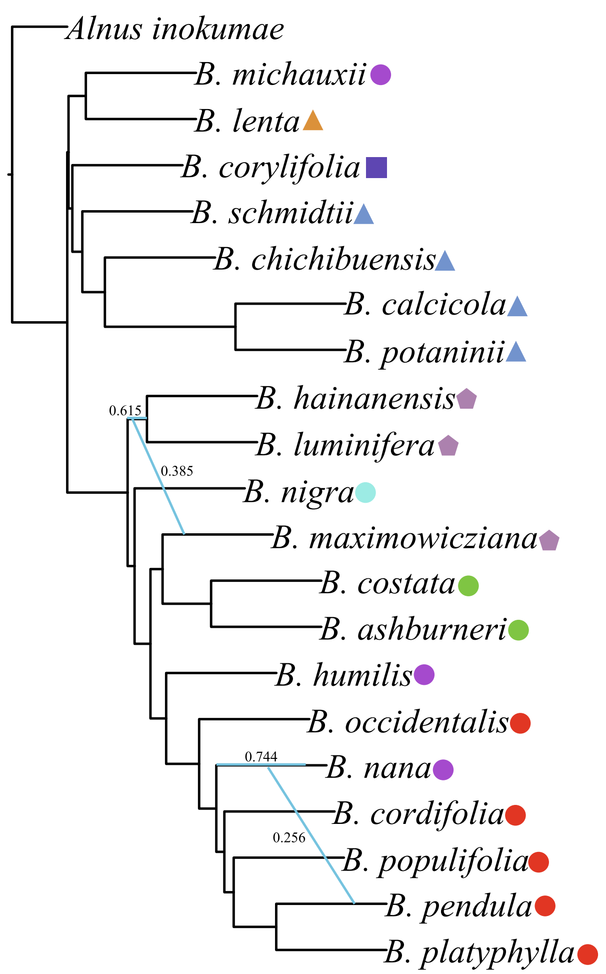

Fig. 3. Best network inferred from SNaQ analysis of the 20 *Betula* diploids of D1 with the number of hybridisation events h = 2. Blue lines indicate hybrid edges and values beside the blue line indicate estimated inheritance probabilities. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

for 23 of the 27 polyploids their homoeologues formed clades with each of the putative parental diploid relative species (Table 1; Fig. 5). For example, subgenomes of *B. pubescens* and its varieties formed monophyletic clades which were sister to *B. pendula* and *B. platyphylla*, respectively.

3.6. Simple sequence repeat analysis

We developed between 58 and 565 microsatellite primer pairs for the diploid *Betula* taxa and between 40 and 633 for polyploid *Betula* taxa. In addition, 100, 84 and 41 microsatellite primers pairs were developed for *A. inokumae*, *A. orientalis* and *C. avellana*, respectively (Table S3).

4. Discussion

4.1. A well resolved diploid phylogeny for Betula

We used both supermatrix and species tree approaches to construct diploid phylogenies based on RADseq data. The use of multiple gene trees not only enabled us to detect evidence for hybridisation events between diploid species but also resolved the phylogeny more robustly compared to previous studies.

Two major clades were revealed in all our analyses, which were not indicated by previous molecular or morphological analyses (Ashburner and McAllister, 2016; Bina et al., 2016; Järvinen et al., 2004; Li et al., 2005; Li et al., 2007; Nagamitsu et al., 2006; Schenk et al., 2008; Wang et al., 2016). Interestingly, we noted that species of Clade 1 exclusively have no or very narrow seed wings and species of Clade 2 exclusively have obvious seed wings. We also noted that species of Clade 1 have geographically narrow distributions and species of Clade 2 have wide geographic distributions, indicating a possible link between seed wings and geographic distributions. For example, *B. pendula*, *B. platyphylla*, *B. humilis* and *B. nana* of Clade 2 have a distribution across Eurasia; *B. costata* and *B. ashburneri* of Clade 2 have a distribution across eastern Asia and *B. cordifolia*, *B. occidentalis* and *B. populifolia* of Clade 2 have a distribution across North America (Ashburner and McAllister, 2016). In contrast, all the diploid species of Clade 1 have a restricted distribution and *B. calcicola*, *B. chichibuensis* and *B. corylifolia* were reported to be endangered (Ashburner and McAllister, 2016). Further research could test the hypothesis suggested by our results that seed wing morphology has had a long term impact on the range size of *Betula* species.

We found evidence for convergent evolution of dwarfism in *Betula*. Ashburner and McAllister's (2016) taxonomic section *Apterocaryon*, which comprises three dwarf species, is split between our major clades with *B. michauxii* nested into Clade 1 and *B. humilis* and *B. nana* nested within Clade 2. Independent evolution of dwarf forms has been observed in other genera, such as in *Artemisia* (Tkach et al., 2007) and *Eucalyptus* (Foster et al., 2007).

Based on our phylogenetic results, we suggest that section *Apterocaryon* should be dissolved, and *B. michauxii* placed in section *Lentae*, and *B. humilis* and *B. nana* in section *Betula*. In the case of section *Acuminatae*, our analysis shows *B. luminifera* and *B. hainanensis* form a clade, but *B. maximowicziana* is sister to section *Costatae* (Fig. 2). The incongruence between morphology and molecular evidence for these three species is likely explained by hybridisation as indicated by our phylogenetic network analysis (Fig. 3). We suggest that *B. maximowicziana* should be moved to section *Costatae*. After making these taxonomic changes, taking into account the effects of hybridisation and convergent evolution, five sections (*Acuminatae*, *Betula*, *Costatae*, *Asperae* and *Lentae*) represented by multiple individuals are monophyletic in our diploid trees (Figs. 2 and 3). Although section *Dahuricae* only has one diploid individual in this study, a previous study represented by multiple individuals showed that section *Dahuricae* is monophyletic (Wang et al., 2016).

We also note that in our analyses, *B. corylifolia*, the single species of section *Nipponobetula*, was in a monophyletic clade with species of section *Asperae*. Such a relationship has previously been indicated based on ITS (Nagamitsu et al., 2006; Wang et al., 2016). We therefore suggest that this species is placed in section *Asperae*. We can therefore reduce the genus *Betula* to two sub-genera that correspond to the two major clades of our diploid phylogenies: these are subgenus *Betula* containing sections *Acuminatae*, *Betula*, *Costatae* and *Dahuricae*, and subgenus *Aspera* containing sections *Asperae* and *Lentae*.

4.2. Inferring polyploid parentage of Betula species

Our results provide novel insights into parental species for a majority of *Betula* polyploids (Table 1). Below we discuss putative parental species of *Betula* polyploids following the classification proposed in this study.

4.2.1. Section Asperae (subgenus Aspera)

We found the closest diploid relatives of *B. delavayi* to be *B. calcicola* and *B. potaninii*, consistent with the results suggested by ITS sequences (Wang et al., 2016). Interestingly, we found that *B. chinensis* (6× and 8×) and *B. fargesii* (10×), close relatives of *B. delavayi*, have *B. chichibuensis* as a third diploid relative in addition to *B. calcicola* and *B. potaninii*. *Betula chinensis*/*B. fargesii* occur in central, north and northeast China whereas *B. delavayi* occurs in southwest China. The genetic contribution from *B. chichibuensis*, an endemic species to Japan, may have helped *B. chinensis* colonise north and northeast China as

Fig. 4. Mapping patterns of reads from polyploids to the diploid reference. Numbers on the x axis indicate proportion of mapped loci. Species names are coloured to show their taxonomic sections, using the colour scheme used for the section symbols in Figure 2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Putative diploid progenitors suggested by rates of read mapping for polyploids of *Betula* and included in subsequent phylogenetic analysis including polyploid sub-genomes.

Species[1]	Ploidy level	Putative diploid progenitors
B. pubescens var. *pubescens*	4	*B. pendula/B. platyphylla*
B. pubescens var. *litwinowii*	4	*B. pendula/B. platyphylla*
B. pubescens var. *celtiberica*	4	*B. pendula/B. platyphylla*
B. pubescens var. *pumila*	4	*B. pendula/B. platyphylla*
B. pubescens var. *fragrans*	4	*B. pendula/B. platyphylla*
B. papyrifera	6	*B. cordifolia/B. populifolia/B. occidentalis*
B. papyrifera var. *commutata*	6	*B. cordifolia/B. populifolia/B. occidentalis*
B. pumila	4	*B. populifolia/B. occidentalis*
B. albosinensis	4	*B. ashburneri/B. costata*
B. albosinensis var. *septentrionalis*	4	*B. ashburneri/B. costata*
B. utilis subsp. *prattii*	4	*B. ashburneri/B. costata*
B. utilis	4	*B. ashburneri/B. costata*
B. ermanii	4	*B. ashburneri/B. costata*
B. ermanii var. *lanata*	4	*B. ashburneri/B. costata*
B. cylindrostachya	4	*B. luminifera/B. hainanensis*
B. alnoides	4	*B. luminifera/B. hainanensis*
B. alleghaniensis	6	*B. lenta*
B. murrayana	8	*B. lenta/B. occidentalis/B. populifolia*
B. medwediewii	10	*B. humilis/B. lenta/B. maximowicziana/ B. michauxii/B. luminifera*
B. megrelica	12	*B. humilis/B. lenta/B. maximowicziana/ B. michauxii/B. luminifera*
B. chinensis	6	*B. calcicola/B. potaninii/B. chichibuensis*
B. chinensis	8	*B. calcicola/B. potaninii/B. chichibuensis*
B. fargesii	10	*B. calcicola/B. potaninii/B. chichibuensis*
B. delavayi	6	*B. calcicola/B. potaninii*
B. bomiensis	4	*B. calcicola/B. potaninii*
B. globispica	10	*B. corylifolia/B. chichibuensis/B. lenta/ B. michauxii/B. costata*
B. insignis	10 or 12	*B. corylifolia/B. chichibuensis/B. lenta/ B. michauxii/B. luminifera/ B. maximowicziana*

[1] Six putative progenitors are indicated for *B. insignis* due to its ploidy level possibly being either 10 or 12.

B. chichibuensis is drought tolerant with limestone outcrops in the mountains as its habitat (McAllister, 2019). Similarly, *B. chinensis* usually occupies mountain slopes according to our field observations. The decaploid *B. globispica* has a high proportion of loci mapped to from species such as *B. corylifolia* and *B. chichibuensis*, which are close relatives of *B. calcicola* and *B. potaninii*. This may explain the close relationship between *B. globispica* and *B. chinensis* based on ITS sequences (Wang et al., 2016).

4.2.2. Section Lentae (subgenus Aspera)

Both *B. medwediewii* (10×) and *B. megrelica* (12×) have *B. humilis*, and likely also *B. lenta*, *B. maximowicziana*, *B. michauxii* and *B. luminifera* as close relatives of their diploid progenitors (note that *B. lenta* and *B. michauxii* are sister species (Fig. 2)). Another polyploid in section *Lentae*, *B. insignis* (10×), an endemic species to central and south China, has *B. lenta* as a close relative of one of its diploid parents, a species from North America, suggesting possible past range overlap between the two species. The octoploid species *B. murrayana* has previously been suggested as a recent allopolyploid derivative from *B. x purpusii*, an inter-subgenus hybrid between *B. alleghaniensis* (6×) and *B. pumila* (4×) (Barnes and Dancik, 1985), consistent with our read mapping that both *B. murrayana* and *B. alleghaniensis* have the highest proportion of loci mapped to from *B. lenta*. In addition, *B. murrayana* has a higher proportion of loci mapped to from *B. occidentalis* and *B. populifolia*, similar to the read mapping pattern of *B. pumila* (Fig. 4). The hexaploid species *B. grossa* has been placed in section *Lentae* due to morphological similarities consistent with *B. lenta* being one of its parents (Ashburner and McAllister, 2016). However, *B. grossa* is clustered with species of subgenus *Betula* by ITS sequences (Nagamitsu et al., 2006; Wang et al., 2016), indicating that gene flow has occurred from a species of subgenus *Betula*. Our read mapping shows that *B. grossa* has a relatively higher proportion of loci mapped to from species of subgenus *Betula*, such as *B. maximowicziana*, indicating potential gene flow from it (Fig. 4). Interestingly, *B. alleghaniensis*, a very closely-related hexaploid species to *B. grossa*, also has *B. lenta* as its parent relative, but shows a much lower proportion of loci mapped to from any species of subgenus *Betula*.

4.2.3. Section Acuminatae (subgenus Betula)

The two tetraploids *B. alnoides* and *B. cylindrostachya* have *B. luminifera* and *B. hainanensis* identified as close relatives of their diploid progenitors. Interestingly, the two species have a high proportion of loci mapped to from *B. maximowicziana*, indicating gene flow from *B. maximowicziana*, as evidenced by the phylonetwork analysis (Figs. 3 and 4). *Betula maximowicziana* is restricted in Japan whereas *B. luminifera* is widespread in western China, from Yunnan as far north as Gansu, and east to Jiangsu (Ashburner and McAllister, 2016). We hypothesize that *B. maximowicziana* was historically present in China and hybridised with *B. luminifera*, resulting in genetic footprints in *B. luminifera*.

4.2.4. Section Betula (subgenus Betula)

The parentage of *B. pubescens* has been controversial for decades and has been suggested as *B. pendula* (Howland et al., 1995; Walters, 1968), *B. humilis*, *B. nana* (Howland et al., 1995; Järvinen et al., 2004; Walters, 1968) and *B. lenta* c.f. (Salojarvi et al., 2017). Here we found *B. pendula* and its sister species *B. platyphylla* to be the most likely relatives of the parents of *B. pubescens*. For *B. pubsecens* we also found a low but still a considerable proportion of mapped loci from *B. nana* and *B. humilis*, but several studies have found evidence for introgressive hybridisation among these species since the formation of *B. pubescens* (Bona et al., 2018; Jadwiszczak et al., 2012; Thórsson et al., 2001) which could account for this pattern. Previous hypotheses for *B. lenta* c.f. as a parental species of *B. papyrifera* and *B. humilis* cf. as a parental species of *B. ermanii* (Järvinen et al., 2004) were not supported by our results. The tetraploid species *B. tianshanica*, *B. microphylla*, *B. halophila* and *B. ovalifolia* have a high proportion of loci mapped to from *B. humilis* (Fig. 4), indicating *B. humilis* is a close relative of their diploid progenitors. The four species intergrade to each other and have no clear morphological boundaries with *B. humilis* (Ashburner and McAllister, 2016). Despite the fact that *B. tianshanica*, *B. microphylla* and *B. halophila* occur in northwestern China and *B. ovalifolia* occurs in northeastern China, the wide distribution of *B. humilis* spanning Eurasia consistent with it being a close diploid relative of their progenitors. Furthermore, *B. humilis*, like these four species, grows in open wetlands according to our field observations.

4.2.5. Section Costatae (subgenus Betula)

Tetraploids of section *Costatae* (excluding *B. utilis* subsp. *occidentalis*) have a high proportion of loci mapped to from *B. ashburneri* and *B. costata*, indicating their likely parentage. This is consistent with morphological characters, based on which Ashburner and McAllister placed these species within section *Costatae* (Ashburner and McAllister, 2016). Previous hypotheses for *B. humilis* cf. as a parental species of *B. ermanii* (Järvinen et al., 2004) were not supported by our results. Interestingly, *B. ermanii* has a higher proportion of loci mapped to from *B. costata* whereas *B. utilis* and *B. albosinensis* and their subspecies have a higher proportion of loci mapped to from *B. ashburneri*. This indicates that *B. costata* and *B. ashburneri* are major parents to *B. ermanii* and *B. utilis*, respectively. This is supported by geographic distributions, with *B. costata* and *B. ermanii* distributed in northeast China and its adjacent regions and with *B. ashburneri* and *B. utilis* occurring in southwest China.

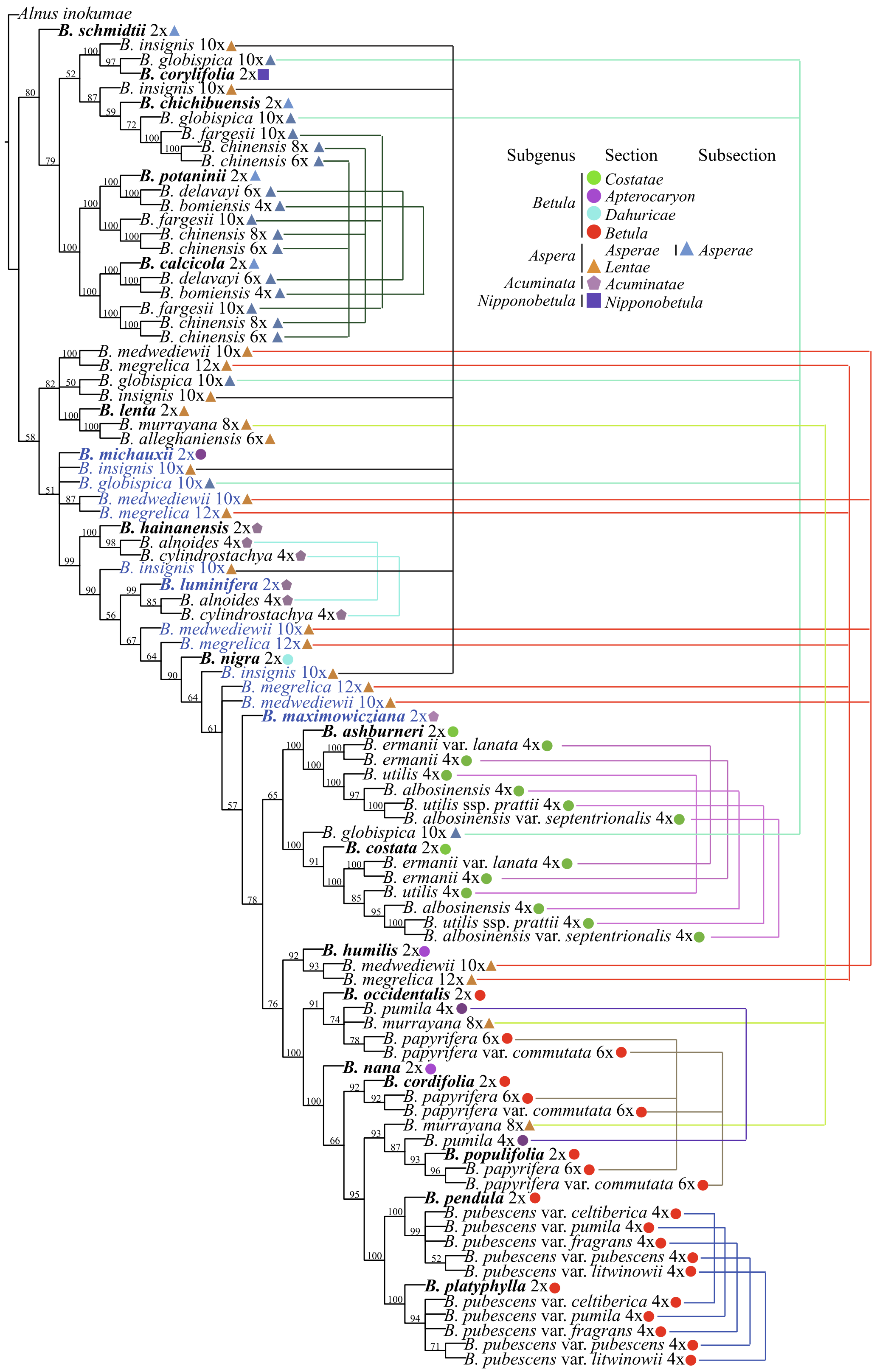

Fig. 5. Species tree incorporating polyploid sub-genomes as separate tips, connected by super-imposed coloured lines. The tree is from the maximum likelihood analysis using ASTRID. Diploid species are shown in bold. Coloured shapes by each species show the taxonomic section each taxon is classified into according to Ashburner and McAllister (2016). Numbers on the branches show bootstrap support values. Species names in blue are sub-genomes of polyploids which did not form clades with the putative parental diploid representative. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Another plausible explanation is hybridisation and subsequent introgression between *B. costata* and *B. ermanii* and between *B. ashburneri* and *B. utilis*, resulting in different number of mapped loci. According to our field observations, *B. costata* and *B. ermanii* often occur sympatrically and *B. ashburneri* and *B. utilis* often occur sympatrically. However, we have not noted any populations where *B. costata* co-occurs with *B. utilis* or *B. ashburneri* co-occurs with *B. ermanii* (Wang N., unpublished data).

4.2.6. *Section Dahuricae (subgenus Betula)*

Section *Dahuricae* includes *B. nigra* (2×), *B. dahurica* (6× and 8×)

and *B. raddeana* (6×). However, since we started work on the *B. dahurica* (6×) individual it has been re-identified as a possible hybrid between *B. dahurica* (8×) and *B. ermanii* (4×) (Hugh McAllister, unpublished data). We found *B. dahurica* (8×) and *B. raddeana* (6×) to have the highest proportion of loci mapped to from *B. humilis* and a very low proportion of loci mapped to from *B. nigra*. The highly similar bark (including bark color and patterns of bark peeling) of *B. nigra* and *B. dahurica* led Ashburner and McAllister (2016) to place them into the same section and to regard *B. nigra* as a diploid progenitor of *B. dahurica*. However, our results indicated that *B. nigra* is unlikely to be a parent to *B. dahurica* and the morphological similarity between *B. nigra* and *B. dahurica* likely reflects phenotypic convergence. We found that the putative hybrid mislabeled *B. dahurica* (6×) has a high proportion of loci mapped to from *B. populifolia* and *B. humilis*, suggesting that *B. populifolia* rather than *B. ermanii* (4×) could have been involved in its formation.

Given such complex evolutionary histories, it is perhaps unsurprising that for nine of the 36 polyploids, we could not clearly identify all putative parents. This may also be because these nine polyploids are older than the polyploids for which we have identified putative parents, and thus more divergent from the diploid species. Another possibility is that they are derived from diploid species which are now extinct, have not yet been discovered or were not included in our phylogenetic analyses (i.e. *B. glandulosa*).

4.3. High frequency of polyploidy within Betula and its implications

Polyploids within *Betula* account for nearly 60% of the described taxa, with the ploidy level reaching dodecaploid (Wang et al., 2016). The high frequency of polyploidy, and especially allopolyploidy, within *Betula* may be for several reasons. The haploid genome size of *Betula* is relatively small (Wang et al., 2016) meaning that polyploidisation may be less costly than in species with larger haploid genomes (Guignard et al., 2017; Leitch and Leitch, 2012). *Betula* species are wind-pollinated and frequent hybridisation has been recorded between many species. For example, natural hybrids have been found between *B. pubescens* (4×) × *B. nana* (2×), *B. papyrifera* (6×) × *B. alleghaniensis* (6×), *B. alleghaniensis* (6×) × *B. lenta* (2×) and *B. platyphylla* (2×) × *B. albosinensis* (4×) (Anamthawat-Jónsson and Thórsson, 2003; Hu et al., 2019; Thomson et al., 2015). Hybridisation is an intrinsic part of allopolyploid formation, and may itself promote unreduced gamete formation (Ramsey and Schemske, 1998). Many *Betula* species are pioneer species of high latitude and high altitude environments, which also may promote unreduced gamete formation (Mason et al., 2011; Ramsey, 2007).

5. Conclusion

Here, by generating a new phylogenetic hypothesis for *Betula*, and providing new evidence for the progenitors of many of its polyploid taxa, we have provided a framework within which the evolution and systematics of the genus can be understood. Knowledge of the parentage of the allopolyploids, some of which are widespread and economically important, opens the way for their genomic analysis. The approach we have used is relatively cost effective and straightforward and could be applied to many other plant groups where allopolyploidy has impeded evolutionary analyses.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Dr. Yu Feng from Zhejiang University for helpful discussions on phylogenetic network analysis. This work was funded by Natural Environment Research Council Fellowship NE/G01504X/1 to R.J.A.B. and was funded by the National Natural Science Foundation of China (31770230 and 31600295) to N.W.

Author contributions

NW and RB conceived the project. NW, RB and HM collected samples. HM identified the samples based on morphology. NW carried out lab work. NW, JZ and LK analysed data. NW, RB and LK wrote the manuscript. All the authors contributed to editing the manuscript.

Data accessibility

All sequences are deposited in the NCBI-Sequence Read Archive (SRA) repository under the BioProjectID PRJNA679451. Phylogenetic trees used for species tree inference and the code are in the Dryad Digital Repository https://doi.org/10.5061/dryad.fj6q573tb.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2021.107126.

References

Anamthawat-Jónsson, K., Tómasson, T., 1990. Cytogenetics of hybrid introgression in Icelandic birch. Hereditas 112, 65–70.
Anamthawat-Jónsson, K., Tómasson, T., 1999. High frequency of triploid birch hybrid by Betula nana seed parent. Hereditas 130, 191–193.
Anamthawat-Jónsson, K., Thórsson, A.T., 2003. Natural hybridisation in birch: triploid hybrids between *Betula nana* and *B. pubescens*. Plant Cell, Tissue Organ Cult. 75, 99–107.
Anamthawat-Jónsson, K., Thórsson, A.T., Temsch, E.M., Greilhuber, J., 2010. Icelandic birch polyploids-the case of perfect fit in genome size. J. Bot. 347254.
Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G., Hohenlohe, P.A., 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat. Rev. Genet. 17, 81–92.
Andrews, S., 2014. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
Ashburner, K., McAllister, H.A., 2016. The genus *Betula*: A Taxonomic Revision of Birches. Kew Publishing, London.
Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, S.O., Gundlach, H., Hale, I., Mascher, M., Spannagl, M., Wiebe, K., Jordan, K.W., Golan, G., Deek, J., Ben-Zvi, B., Ben-Zvi, G., Himmelbach, A., MacLachlan, R.P., Sharpe, A.G., Fritz, A., Ben-David, R., Budak, H., Fahima, T., Korol, A., Faris, J.D., Hernandez, A., Mikel, M.A., Levy, A.A., Steffenson, B., Maccaferri, M., Tuberosa, R., Cattivelli, L., Faccioli, P., Ceriotti, A., Kashkush, K., Pourkheirandish, M., Komatsuda, T., Eilam, T., Sela, H., Sharon, A., Ohad, N., Chamovitz, D.A., Mayer, K.F.X., Stein, N., Ronen, G., Peleg, Z., Pozniak, C.J., Akhunov, E.D., Distelfeld, A., 2017. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science 357, 93–96.
Barchi, L., Lanteri, S., Portis, E., Acquadro, A., Valè, G., Toppino, L., Rotino, G.L., 2011. Identification of SNP and SSR markers in eggplant using RAD tag sequencing. BMC Genom. 12, 304.
Barnes, B.V., Bruce, P.D., Sharik, T.L., 1974. Natural hybridization of yellow birch and white birch. For. Sci. 20, 215–221.
Barnes, B.V., Dancik, B.P., 1985. Characteristics and origin of a new birch species, *Betula murrayana*, from southeastern Michigan. Can. J. Bot. 63, 223–226.
Bina, H., Yousefzadeh, H., Ali, S.S., Esmailpour, M., 2016. Phylogenetic relationships, molecular taxonomy, biogeography of *Betula*, with emphasis on phylogenetic position of Iranian populations. Tree Genet. Genomes 12, 84.
Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.
Bona, A., Petrova, G., Jadwiszczak, K.A., 2018. Unfavourable habitat conditions can facilitate hybridisation between the endangered *Betula humilis* and its widespread relatives *B. pendula* and *B. pubescens*. Plant Ecol. Divers. 11, 295–306.
Brysting, A.K., Mathiesen, C., Marcussen, T., 2011. Challenges in polyploid phylogenetic reconstruction: a case story from the arctic-alpine *Cerastium alpinum* complex. Taxon 60, 333–347.
Buggs, R.J.A., Chamala, S., Wu, W., Tate, J.A., Schnable, P.S., Soltis, D.E., Soltis, P.S., Barbazuk, W.B., 2012. Rapid, repeated, and clustered Loss of duplicate genes in allopolyploid plant populations of independent origin. Curr. Biol. 22, 248–252.

Capella-Gutierrez, S., Silla-Martinez, J.M., Gabaldon, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973.

Cariou, M., Duret, L., Charlat, S., 2013. Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. Ecol. Evol. 3, 846–852.

Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A., Cresko, W.A., 2013. Stacks: an analysis tool set for population genomics. Mol. Ecol. 22, 3124–3140.

Cruaud, A., Gautier, M., Galan, M., Foucaud, J., Saune, L., Genson, G., Dubois, E., Nidelet, S., Deuve, T., Rasplus, J.Y., 2014. Empirical assessment of RAD sequencing for interspecific phylogeny. Mol. Biol. Evol. 31, 1272–1274.

DaCosta, J.M., Sorenson, M.D., 2016. ddRAD-seq phylogenetics based on nucleotide, indel, and presence–absence polymorphisms: analyses of two avian genera with contrasting histories. Mol. Phylogenet. Evol. 122–135.

Dauphin, B., Grant, J.R., Farrar, D.R., Rothfels, C.J., 2018. Rapid allopolyploid radiation of moonwort ferns (Botrychium; Ophioglossaceae) revealed by PacBio sequencing of homologous and homeologous nuclear regions. Mol. Phylogenet. Evol. 120, 342–353.

Eaton, D.A.R., Ree, R.H., 2013. Inferring phylogeny and introgression using RADseq Data: an example from flowering plants (*Pedicularis*: Orobanchaceae). Syst. Biol. 62, 689–706.

Eaton, D.A.R., Spriggs, E.L., Park, B., Donoghue, M.J., 2016. Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. Syst. Biol. 16, syw092.

Eidesen, P.B., Alsos, I.G., Brochmann, C., 2015. Comparative analyses of plastid and AFLP data suggest different colonization history and asymmetric hybridization between *Betula pubescens* and *B. nana*. Mol. Ecol. 24, 3993–4009.

Emerson, K.J., Merz, C.R., Catchen, J.M., Hohenlohe, P.A., Cresko, W.A., Bradshaw, W. E., Holzapfel, C.M., 2010. Resolving postglacial phylogeography using high-throughput sequencing. Proc. Natl. Acad. Sci. U.S.A. 107, 16196–16200.

Eriksson, J.S., de Sousa, F., Bertrand, Y.J.K., Antonelli, A., Oxelman, B., Pfeil, B.E., 2018. Allele phasing is critical to revealing a shared allopolyploid origin of *Medicago arborea* and *M. strasseri* (Fabaceae). BMC Evol. Biol. 18, 9.

Etter, P.D., Bassham, S., Hohenlohe, P.A., Johnson, E.A., Cresko, W.A., 2011. SNP discovery and genotyping for evolutionary genetics using RAD sequencing. In: Orgogonzo, V., Rockman, M.V. (Eds.), Molecular Methods for Evolutionary Genetics. Humana Press, NY.

Fitz-Gibbon, S., Hipp, A.L., Pham, K.K., Manos, P.S., Sork, V.L., 2017. Phylogenomic inferences from reference-mapped and de novo assembled short-read sequence data using RADseq sequencing of California white oaks (*Quercus* section *Quercus*). Genome 60, 743–755.

Foster, S.A., McKinnon, G.E., Steane, D.A., Potts, B.M., Vaillancourt, R.E., 2007. Parallel evolution of dwarf ecotypes in the forest tree *Eucalyptus globulus*. New Phytol. 175, 370–380.

Gonen, S., Bishop, S.C., Houston, R.D., 2015. Exploring the utility of cross-laboratory RAD-sequencing datasets for phylogenetic analysis. BMC Res. Notes 8, 299.

Guignard, M.S., Leitch, A.R., Acquisti, C., Eizaguirre, C., Elser, J.J., Hessen, D.O., Jeyasingh, P.D., Neiman, M., Richardson, A.E., Soltis, P.S., Soltis, D.E., Stevens, C.J., Trimmer, M., Weider, L.J., Guy Woodward, G., Leitch, I.J., 2017. Impacts of nitrogen and phosphorus: from genomes to natural ecosystems and agriculture. Front. Ecol. Evol. 5, 70.

Hipp, A.L., Eaton, D.A.R., Cavender-Bares, J., Fitzek, E., Nipper, R., Manos, P.S., 2014. A framework phylogeny of the American Oak clade based on sequenced RAD data. PLoS One 9, e93975.

Hohenlohe, P.A., Bassham, S., Etter, P.D., Stiffler, N., Johnson, E.A., Cresko, W.A., 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet. 6, e1000862.

Hou, Y., Nowak, M.D., Mirré, V., Bjorå, C.S., Brochmann, C., Popp, M., 2015. Thousands of RAD-seq loci fully resolve the phylogeny of the highly disjunct Arctic-Alpine genus *Diapensia* (Diapensiaceae). PLoS One 10, e0140175.

Howland, D.E., Oliver, R.R., Davy, A.J., 1995. Morphological and molecular variation in natural populations of *Betula*. New Phytol. 130, 117–124.

Hu, Y.N., Zhao, L., Buggs, R.J.A., Zhang, X.M., Li, J., Wang, N., 2019. Population structure of *Betula albosinensis* and *Betula platyphylla*: evidence for hybridization and a cryptic lineage. Ann. Bot. 123, 1179–1189.

Järvinen, P., Palmé, A., Morales, L.O., Lännenpää, M., Keinänen, M., Sopanen, T., Lascoux, M., 2004. Phylogenetic relationships of *Betula* species (Betulaceae) based on nuclear ADH and chloroplast matK sequences. Am. J. Bot. 91, 1834–1845.

Jadwiszczak, K.A., Banaszek, A., Jabłońska, E., Sozinov, O.V., 2012. Chloroplast DNA variation of *Betula humilis* Schrk. in Poland and Belarus. Tree Genet. Genomes 8, 1017–1030.

Johnsson, H., 1945. Interspecific hybridization within the genus *Betula*. Hereditas 31, 163–176.

Jones, G., Sagitov, S., Oxelman, B., 2013. Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. Syst. Biol. 62, 467–478.

Kamneva, O.K., Syring, J., Liston, A., Rosenberg, N.A., 2017. Evaluating allopolyploid origins in strawberries (*Fragaria*) using haplotypes generated from target capture sequencing. BMC Evol. Biol. 17.

Katoh, K., Kuma, K., Toh, H., Miyata, T., 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33, 511–518.

Leitch, A.R., Leitch, I.J., 2012. Ecological and genetic factors linked to contrasting genome dynamics in seed plants. New Phytol. 194, 629–646.

Li, F.G., Fan, G.Y., Lu, C.R., Xiao, G.H., Zou, C.S., Kohel, R.J., Ma, Z.Y., Shang, H.H., Ma, X.F., Wu, J.Y., Liang, X.M., Huang, G., Percy, R.G., Liu, K., Yang, W.H., Chen, W. B., Du, X.M., Shi, C.C., Yuan, Y.L., Ye, W.W., Liu, X., Zhang, X.Y., Liu, W.Q., Wei, H. L., Wei, S.J., Huang, G.D., Zhang, X.L., Zhu, S.J., Zhang, H., Sun, F.M., Wang, X.F., Liang, J., Wang, J.H., He, Q., Huang, L.H., Wang, J., Cui, J.J., Song, G.L., Wang, K.B., Xu, X., Yu, J.Z., Zhu, Y.X., Yu, S.X., 2015. Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. Nat. Biotechnol. 33, 524–530.

Li, J.H., Shoup, S., Chen, Z.D., 2005. Phylogenetics of *Betula* (Betulaceae) inferred from sequences of nuclear ribosomal DNA. Rhodora 107, 69–86.

Li, J.H., Shoup, S., Chen, Z.D., 2007. Phylogenetic relationships of diploid species of *Betula* (Betulaceae) inferred from DNA sequences of nuclear nitrate reductase. Syst. Bot. 32, 357–365.

Linder, C.R., Rieseberg, L.H., 2004. Reconstructing patterns of reticulate evolution in plants. Am. J. Bot. 91, 1700–1708.

Lott, M., Spillner, A., Huber, K.T., Petri, A., Oxelman, B., Moulton, V., 2009. Inferring polyploid phylogenies from multiply-labeled gene trees. BMC Evol. Biol. 9, 216.

Luo, M.C., Gu, Y.Q., Puiu, D., Wang, H., Twardziok, S.O., Deal, K.R., Huo, N.X., Zhu, T.T., Wang, L., Wang, Y., McGuire, P.E., Liu, S.Y., Long, H., Ramasamy, R.K., Rodriguez, J. C., Van, S.L., Yuan, L.X., Wang, Z.Z., Xia, Z.Q., Xiao, L.C., Anderson, O.D., Ouyang, S. H., Liang, Y., Zimin, A.V., Pertea, G., Qi, P., Ennetzen, J.L.B., Dai, X.T., Dawson, M. W., Muller, H.G., Kugler, K., Rivarola-Duarte, L., Spannagl, M., Mayer, K.F.X., Lu, F. H., Bevan, M.W., Leroy, P., Li, P.C., You, F.M., Sun, Q.X., Liu, Z.Y., Lyons, E., Wicker, T., Salzberg, S.L., Devos, K.M., Dvorak, J., 2017. Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. Nature 551, 498–502.

Mandáková, T., Lysak, M.A., 2018. Post-polyploid diploidization and diversification through dysploid changes. Curr. Opin. Plant Biol. 42, 55–65.

Marcussen, T., Heier, L., Brysting, A.K., Oxelman, B., Jakobsen, K.S., 2015. From gene trees to a dated allopolyploid network: insights from the angiosperm genus *Viola* (Violaceae). Syst. Biol. 64, 84–101.

Mason, A.S., Nelson, M.N., Yan, G., Cowling, W.A., 2011. Production of viable male unreduced gametes in *Brassica* interspecific hybrids is genotype specific and stimulated by cold temperatures. BMC Plant Biol. 11, 103.

Massatti, R., Reznicek, A.A., Knowles, L.L., 2016. Utilizing RADseq data for phylogenetic analysis of challenging taxonomic groups: a case study in *Carex* sect. *Racemosae*. Am. J. Bot. 103, 337–347.

McAllister, H.A., 2019. *Betula chichibuensis*. Curtis's Bot. Mag. 36, 365–374.

McKain, M.R., Tang, H., McNeal, J.R., Ayyampalayam, S., Davis, J.I., Depamphilis, C.W., Givnish, T.J., Pires, J.C., Stevenson, D.W., Leebens-Mack, J.H., 2016. A phylogenomic assessment of ancient polyploidy and genome evolution across the poales. Genome Biol. Evol. 8, 1150–1164.

Meglécz, E., Pech, N., Gilles, A., Dubut, V., Hingamp, P., Trilles, A., Grenier, R.e.a., 2014. QDD version 3.1: a user-friendly computer program for microsatellite selection and primer design revisited: experimental validation of variables determining genotyping success rate. Mol. Ecol. Resour. 14, 1302–1313.

Mirarab, S., Warnow, T., 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31, i44–i52.

Morales-Briones, D.F., Liston, A., Tank, D.C., 2018. Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae). New Phytol. 218, 1668–1684.

Muilenburg, V.L., Herms, D.A., 2012. A review of bronze birch borer (Coleoptera: Buprestidae) life history, ecology, and management. Environ. Entomol. 41, 1372–1385.

Nagamitsu, T., Kawahara, T., Kanazashi, A., 2006. Endemic dwarf birch *Betula apoiensis* (Betulaceae) is a hybrid that originated from *Betula ermanii* and *Betula ovalifolia*. Plant Spec. Biol. 21, 19–29.

Oxelman, B., Brysting, A.K., Jones, G.R., Marcussen, T., Oberprieler, C., Pfeil, B.E., 2017. Phylogenetics of allopolyploids. Annu. Rev. Ecol. Evol. Syst. 48, 543–557.

Pante, E., Abdelkrim, J., Viricel, A., Gey, D., France, S.C., Boisselier, M.C., Samadi, S., 2015. Use of RAD sequencing for delimiting species. Heredity 114, 450–459.

Ramsey, J., 2007. Unreduced gametes and neopolyploids in natural populations of *Achillea borealis* (Asteraceae). Heredity 98, 143–150.

Ramsey, J., Schemske, D.W., 1998. Pathways, mechanisms, and rates of polyploid formation in flowering plants. Annu. Rev. Ecol. Syst. 29, 467–501.

Rothfels, C.J., Pryer, K.M., Li, F.W., 2017. Next-generation polyploid phylogenetics: rapid resolution of hybrid polyploid complexes using PacBio single-molecule sequencing. New Phytol. 213, 413–429.

Rubin, B.E.R., Ree, R.H., Moreau, C.S., 2012. Inferring phylogenies from RAD sequence data. PLoS One 7, e33394.

Salojarvi, J., Smolander, O.P., Nieminen, K., Rajaraman, S., Safronov, O., Safdari, P., Lamminmaki, A., Immanen, J., Lan, T.Y., Tanskanen, J., Rastas, P., Amiryousefi, A., Jayaprakash, B., Kammonen, J.I., Hagqvist, R., Eswaran, G., Ahonen, V.H., Serra, J. A., Asiegbu, F.O., Barajas-Lopez, J.D., Blande, D., Blokhina, O., Blomster, T., Broholm, S., Brosche, M., Cui, F.Q., Dardick, C., Ehonen, S.E., Elomaa, P., Escamez, S., Fagerstedt, K.V., Fujii, H., Gauthier, A., Gollan, P.J., Halimaa, P., Heino, P.I., Himanen, K., Hollender, C., Kangasjarvi, S., Kauppinen, L., Kelleher, C. T., Kontunen-Soppela, S., Koskinen, J.P., Kovalchuk, A., Karenlampi, S.O., Karkonen, A.K., Lim, K.J., Leppala, J., Macpherson, L., Mikola, J., Mouhu, K., Mahonen, A.P., Niinemets, U., Oksanen, E., Overmyer, K., Palva, E.T., Pazouki, L., Pennanen, V., Puhakainen, T., Poczai, P., Possen, B.J.H.M., Punkkinen, M., Rahikainen, M.M., Rousi, M., Ruonala, R., van der Schoot, C., Shapiguzov, A., Sierla, M., Sipila, T.P., Sutela, S., Teeri, T.H., Tervahauta, A.I., Vaattovaara, A., Vahala, J., Vetchinnikova, L., Welling, A., Wrzaczek, M., Xu, E.J., Paulin, L.G., Schulman, A.H., Lascoux, M., Albert, V.A., Auvinen, P., Helariutta, Y., Kangasjarvi, J., 2017. Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. Nat. Genet. 49, 904–912.

Sayyari, E., Mirarab, S., 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol. Biol. Evol. 33, 1654–1668.

Schenk, M.F., Thienpont, C.N., Koopman, W.J.M., Gilissen, L.J.W.J., Smulders, M.J.M., 2008. Phylogenetic relationships in *Betula* (Betulaceae) based on AFLP markers. Tree Genet. Genomes 4, 911–924.

Shaw, K., Stritch, L., Rivers, M., Roy, S., Wilson, B., Govaerts, R., 2014. The Red List of Betulaceae. BGCI, Richmond, UK.
Solís-Lemus, C., Ané, C., 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLoS Genet. 12, e1005896.
Solis-Lemus, C., Bastide, P., Ane, C., 2017. PhyloNetworks: a package for phylogenetic networks. Mol. Biol. Evol. 34, 3292–3298.
Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.
Thórsson, T.H., Salmela, E., Anamthawat-Jónsson, K., 2001. Morphological, cytogenetic, and molecular evidence for introgressive hybridization in birch. J. Hered. 92, 404–408.
Thomson, A.M., Dick, C.W., Pascoini, A.L., Dayanandan, S., 2015. Despite introgressive hybridization, North American birches (*Betula* spp.) maintain strong differentiation at nuclear microsatellite loci. Tree Genet. Genomes 11, 1–12.
Tkach, N.V., Hoffmann, M.H., Röser, M., Korobkov, A.A., von Hagen, K.B., 2007. Parallel evolutionary patterns in multiple lineages of arctic *Artemisia* L. (Asteraceae). Evolution 62, 184–198.
Tsuda, Y., Semerikov, V., Sebastiani, F., Vendramin, G.G.M.L., 2017. Multispecies genetic structure and hybridization in the *Betula* genus across Eurasia. Mol. Ecol. 26, 589–605.
Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., Rozen, S. G., 2012. Primer3-new capabilities and interfaces. Nucleic Acids Res. 40, e115.
Vachaspati, P., Warnow, T., 2015. ASTRID: accurate species trees from internode distances. BMC Genom. 16, S3.
Wagner, C.E., Keller, I., Wittwer, S., Selz, O.M., Mwaiko, S., Greuter, L., Sivasundar, A., Seehausen, O., 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22, 787–798.
Walters, S.M., 1968. *Betula* L. in Britain. Proc. Bot. Soc. Brit. Isles 7, 179–180.
Wang, N., Borrell, J.S., Bodles, W.J.A., Kuttapitiya, A., Nichols, R.A., Buggs, R.J.A., 2014. Molecular footprints of the Holocene retreat of dwarf birch in Britain. Mol. Ecol. 23, 2771–2782.
Wang, N., McAllister, H.A., Bartlett, P.R., Buggs, R.J.A., 2016. Molecular phylogeny and genome size evolution of the genus *Betula* (Betulaceae). Ann. Bot. 117, 1023–1035.
Wang, N., Thomson, M., Bodles, W.J.A., Crawford, R.M.M., Hunt, H.V., Featherstone, A. W., Pellicer, J., Buggs, R.J.A., 2013. Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers. Mol. Ecol. 22, 3098–3111.
Zohren, J., Wang, N., Kardailsky, I., Borrell, J.S., Joecker, A., Nichols, R.A., Buggs, R.J.A., 2016. Unidirectional diploid-tetraploid introgression among British birch trees with shifting ranges shown by restriction site-associated markers. Mol. Ecol. 25, 2413–2426.