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

Nian Wang¹,², Laura J. Kelly¹,³, Hugh A. McAllister⁴, Jasmin Zohren⁵, Richard J. A. Buggs¹,³*

¹School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, UK.
²School of Forestry, Shandong Agricultural University, Taian 271018, Shandong province, China.
³Royal Botanic Gardens Kew, Richmond, Surrey TW9 3AB, UK.
⁴Institute of Integrative Biology, Biosciences Building, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK.
⁵Sex Chromosome Biology Lab, the Francis Crick Institute, 1 Midland Road, London, NW1 1AT UK

*Corresponding author: Richard Buggs, Email: r.buggs@kew.org or r.buggs@qmul.ac.uk Tel: 02083325755
Abstract
Numerous plant genera have a history including frequent hybridisation and polyploidisation, which often means that their phylogenies are not yet fully resolved. The genus *Betula*, which contains many ecologically important allopolyploid tree species, is a case in point. We generated genome-wide sequence data for 27 diploid and 31 polyploid *Betula* species or subspecies using restriction site associated DNA (RAD) sequences assembled into contigs with a mean length of 675 bp. We reconstructed the evolutionary relationships among diploid *Betula* species using both supermatrix and species tree methods. We identified progenitors of the polyploids according to the relative rates at which their reads mapped to contigs from different diploid species. We sorted the polyploid reads into different putative sub-genomes and used the extracted contigs, along with the diploid sequences, to build new phylogenies that included the polyploid sub-genomes. This approach yielded a highly evidenced phylogenetic hypothesis for the genus *Betula*, including the complex reticulate origins of the majority of its polyploid taxa. The genus was split into two well supported clades, which differ in their seed-wing morphology. We propose a new taxonomy for *Betula*, splitting it into two subgenera. We have resolved the parentage of many widespread and economically important polyploid tree species, opening the way for their population genomic study.

Key words: polyploidy, whole genome duplication, hybridization, phylogenomics, *Betula*
1. Introduction

The evolution of plant diversity cannot be fully understood unless we can reconstruct evolutionary relationships for allopolyploids (hybrid species with duplicated genomes). While phylogenomic approaches that use thousands of loci can resolve the phylogenies of diploid taxa with a history of hybridization (Folk et al., 2017; Fontaine et al., 2015; Li et al., 2016), allopolyploids, which are common in plants (Barker et al., 2016), remain hard to place in phylogenetic trees (Oxelman et al., 2017). When molecular markers are sequenced from polyploids, it is difficult to phase them into their parental subgenomes (Eriksson et al., 2018), and it is easy to mistake homoeologs (genes duplicated in allopolyploidisation events) for paralogs (genes arising from duplication within a genome) and vice versa (Brysting et al., 2011; Linder and Rieseberg, 2004). Approaches to resolving the phylogenetic origins of tetraploids (but not higher ploidy levels) have determined parental genomes using heuristic methods (Jones et al., 2013; Lott et al., 2009) or long-read sequences (Rothfels et al., 2017). While phylogenomic approaches are sometimes used to detect the presence of hybrid polyploids (McKain et al., 2016; Morales-Briones et al., 2018), we are not aware of studies that have used phylogenomic data to resolve polyploid origins.

Resolving parental origins of polyploid subgenomes unlocks progress in their genomic characterisation. Knowledge of the origins of the allohexaploid genome of *Triticum aestivum* (bread wheat) allowed further characterisation to be assisted by sequencing of the close relatives of its parents: *Aegilops tauschii* and *Triticum turgidum* ssp. *dicoccoides* (Avni et al., 2017; Luo et al., 2017). Similarly, the allotetraploid genome of *Gossypium hirsutum* (Upland cotton) has been illuminated by sequencing of two close diploid relatives of its progenitor genomes (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 need accessible phylogenomic approaches to make this possible.
A relatively inexpensive method of generating genome-wide marker 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 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 assume that all loci have the same evolutionary history (Cariou et al., 2013; Cruaud et al., 2014; Eaton and Ree, 2013; Eaton et al., 2016; 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 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) (Eaton and Ree, 2013). DaCosta and Sorensen (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) (DaCosta and Sorenson, 2016; Hou et al., 2015).

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 can 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, but some birch species are endangered with
narrow distributions and 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
genesis (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
genesis 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 and has been extensively documented in Betula
(Anamthawat-Jónsson and Tómasson, 1990; Anamthawat-Jónsson and Tómasson,
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). Polyploidy is also common within Betula, with nearly 60% of species
being polyploids (Wang et al., 2016) and ploidy ranging from diploid to dodecaploid
(Ashburner and McAllister, 2016). Some species contain different cytotypes, such as
B.chinensis (6x and 8x) (Ashburner and McAllister, 2016). The origins of most
polyploids in the genus are unresolved. One of the best studied 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 reads assembled 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 origins 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. These approaches together yielded a well-resolved history for the genus 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 genome size of most of these taxa has been obtained (Wang et al., 2016) and 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 2 X CTAB (cetyltrimethylammonium bromide) protocol (Wang et al., 2013). The isolated DNA was assessed with a 1.0% agarose gel and measured with a Qubit
RAD libraries were prepared following the protocol of Etter et al. (2011) 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 hours 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. Briefly, 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 ddH₂O 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 100bp. 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 (Wang et al., 2013), 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 300bp was created for each sample. Betula glandulosa was excluded from further analysis because only 216 loci were mapped at a sufficient read depth. Multiple sequence alignments for 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. 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 species were represented by more than one sample (Table S2). RAD loci ≥ 300bp 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 bootstrapped gene trees (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. Phylogenetic Networks
We used the Species Networks applying Quartets (SNaQ) method (Solís-Lemus and Ané, 2016) implemented in the software PhyloNetworks 0.5.0 (Solís-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; the reference comprised 88,488 sequences in total, representing 5,045 loci. All the sequences were concatenated and separated with lengths of 500 Ns. Trimmed reads of polyploid taxa were mapped individually to this single reference containing loci from all diploid taxa using 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. Any region with a coverage of below three was removed and the consensus sequences for each polyploid with a minimum length of 300 bp were extracted. Variable sites were represented by ambiguity codes in the consensus. Given the fact that the number of loci available for each diploid species was variable (Table S2), we plotted 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 progenitor species of each polyploid species. We therefore sought to identify 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, 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 to 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 homoeologues for each RAD-locus 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 parents 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 the 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-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 7,002 present in only four of these (Fig. 1A), whereas for D2 99 loci are present in all 27 *Betula* diploid individuals and 6,078 present in only four (Fig. 1B). Contigs of ≥ 300bp, with an average length of 580.8bp - 755.8bp, varied in number between 13,597 in *A. inokumae* and 30,717 in *B. pendula* (Fig. 1C, D).

3.2. 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 (Fig. S4-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).

3.3. 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 hybridisation events involving four separate lineages within the largest of the major clades.

3.4. Read-mapping of polyploid species

For 28 of the polyploid species or varieties (80%) the mapping analysis identified two or more parental lineages (Fig. 4), represented by 17 diploid species (Table 1). Five polyploids, all with the ploidy level ≥ 8x, have putative progenitors from both of the two major diploid clades. *Betula nigra* seems not to be the putative progenitor of any polyploid species whereas *B. humilis* represents the putative parental lineage of up to nine polyploids (Table 1).

3.5. Phylogeny incorporating polyploid species

When we included phased homoeologues from polyploids for which we could identify putative parents in a phylogenetic analysis, for 22 of the 26 the polyploids their homoeologues form clades with each of the putative parental diploid 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 more unusually, species tree approaches to construct phylogenies based on RADseq data. We used longer contigs than is usual with RADseq (Tripp et al., 2017), with an average length of 675 bp; these contigs were generated from paired MiSeq reads, mapped to a reference genome (Salojarvi et al., 2017). The use of multiple gene trees also enabled us to detect evidence for hybridisation events among diploid species.

Two major clades were found in all analyses, which were not found 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, species of Clade 1 exclusively have no or very narrow seed wings and species of Clade 2 exclusively have obvious seed wings. The fact that some species of Clade 2 have very wide geographic distributions is likely due to their strong dispersal ability. Ashburner and McAllister’s taxonomic sections Asperae, Lentae and Nipponobetula were grouped exclusively into Clade 1 and sections Acuminatae, Dahuricae, Betula, Costatae, were grouped exclusively into Clade 2, but species of section Apterocaryon, which are dwarf in form, are split between both clades. Within Apterocaryon, B. michauxii, with a geographic distribution in North America, is nested into Clade 1 whereas B. humilis and B. nana, with a geographic distribution across Eurasia, are nested within Clade 2. Thus the shrubby dwarf forms are likely due to independent evolution. Independent evolution of dwarf forms has been occasionally observed in other genera, such as in Artemisia (Tkach et al., 2007) and Eucalyptus (Foster et al., 2007). Another trait that may have evolved independently in the genus is resistance to the bronze birch borer: species reported to be largely resistant to the bronze birch borer (Muilenburg and Herms, 2012) are split between the two clades.

On the basis of the above, 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. haninanensis form a clade, but B. maximowicziana is sister to section Costatae. The incongruence between morphology and molecular evidence for these three species is likely explained by hybridisation as indicated by our phylogenetic network analysis. We suggest that B. maximowicziana should be moved to section Costatae. These changes, taking into account the effects of hybridisation and convergent evolution, mean that the seven remaining sections of the genus Betula, and the three remaining subgenera proposed by Ashburner and McAllister, are all monophyletic in our diploid trees.

We also note that in our analyses, B. corylifolia, the single species of section Nipponobetula, was in a monophyletic group 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, reducing the genus to two sub-genera that correspond to the two major clades of our diploid phylogenies. These are subgenus Betula containing sections Acuminatae, Costatae, Dahuricae and Betula, 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). For example, tetraploids of section Costatae (excluding B. utilis ssp. occidentalis) have a high proportion of loci mapped to in 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). The parentage of B. pubescens, which is widely planted, 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 find B. pendula and its sister species B. platyphylla to be the most likely
parents of *B. pubescens*. We found a relatively low but still a considerable proportion of mapped loci from *B. pubescens* to *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 sequences from *B. nana* and *B. humilis* in *B. pubescens* genomes. Previous hypotheses for *B. lenta* c.f. as a parental species of *B. papyrifera* and *B. humilis* c.f. as a parental species of *B. ermanii* (Järvinen et al., 2004) were not supported by our results. Our results also show evidence of complex layering of hybridisation and polyploidisation events in the history of some taxa. For example, *B. luminifera* and *B. hainanensis* were identified as the extant representatives for the progenitors of two tetraploids, but they themselves have a possible hybrid origin (or have been subject to significant introgression) on the basis of the network analysis; the two tetraploids with *B. luminifera*/*B. hainanensis* identified as progenitors both have the next highest proportion of loci mapped to in *B. maximowicziana*, which might suggest the evidence for hybridisation between *B. luminifera*/*B. hainanensis* and *B. maximowicziana* identified from the network analysis occurred before the origin of the polyploids. Given such complex evolutionary histories, it is perhaps unsurprising that for eight of the 35 polyploids, we could not clearly identify all putative parents. This may also be because these eight 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*).

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 cheap and straightforward and could be applied to many other
plant groups where allopolyploidy has impeded evolutionary analyses.

Acknowledgements

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, 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.
References


sequencing. BMC Genomics 12, 304.


DaCosta, J.M. and Sorenson, M.D., 2016. ddRAD-seq phylogenetics based on
nucleotide, indel, and presence-absence polymorphisms: Analyses of two avian

RADseq Data: An Example from Flowering Plants (Pedicularis: Orobancheaceae).

missing data in RAD-seq phylogenetics with a deep-scale example from
flowering plants. Syst Biol 16, syw092.

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. and Holzapfel, C.M., 2010. Resolving postglacial phylogeography using
high-throughput sequencing. Proceedings of the National Academy of Sciences
of the United States of America 107, 16196–16200.

Eriksson, J.S., de Sousa, F., Bertrand, Y.J.K., Antonelli, A., Oxelman, B. and 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.

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.

Parallel Organellar Genome Capture Result in Extreme Phylogenomic Discord in

Fontaine, M.C., Pease, J.B., Steele, A., Waterhouse, R.M., Neafsey, D.E., Sharakhov,
I.V., Jiang, X.F., Hall, A.B., Catteruccia, F., Kakani, E., Mitchell, S.N., Wu, Y.C.,
M.W. and Besansky, N.J., 2015. Extensive introgression in a malaria vector
species complex revealed by phylogenomics. Science 347.


Katoh, K., Kuma, K., Toh, H. and Miyata, T., 2005. MAFFT version 5: improvement


Schenk, M.F., Thienpont, C.N., Koopman, W.J.M., Gilissen, L.J.W.J. and Smulders,


Wagner, C.E., Keller, I., Wittwer, S., Selz, O.M., Mwaiko, S., Greuter, L., Sivasundar,


**Figure legends**

**Figure 1** Detailed information of 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 300bp 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.

**Figure 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).

**Figure 3** Best network inferred from SNaQ analysis of the 20 *Betula* diploids of D1 with the number of hybridization events h=2. Blue lines indicate hybrid edges and values beside the blue line indicate estimated inheritance probabilities.

**Figure 4** Mapping patterns of polyploids to the diploid reference. Numbers on the x axis indicate number of mapped loci.

**Figure 5** Species tree incorporating polyploids from the maximum likelihood analysis using ASTRID. Species were classified according to Ashburner and McAllister (2013).

**Figure S1** Mapping patterns of the 20 diploid *Betula* species of D1. Numbers on the x axis indicate number of mapped loci.

**Figure S2** The schematic illustration of methods used for analyzing polyploids.

**Figure S3** Species tree from the maximum likelihood analysis of the 20 *Betula* diploids using the ASTRID approach based on 50,870 gene trees. Species were classified according to Ashburner and McAllister (2016).

**Figure S4** Species tree from the maximum likelihood analysis of the 27 *Betula*
diploids using the ASTRAL approach based on 58,442 gene trees. Asterisks on the branches indicate 1.00 local posterior probabilities. Species were classified according to Ashburner and McAllister (2016).

**Figure S5** Species tree from the maximum likelihood analysis of the 27 *Betula* diploids using the ASTRID approach based on 58,442 gene trees. Numbers above branches are support values. Marked with star indicates branches with low support values. Species were classified according to Ashburner and McAllister (2016).

**Figure S6** Species tree from the maximum likelihood analysis of the 27 *Betula* diploids using the supermatrix approach based on 58,442 gene trees. Numbers above or below branches are support values. The scale bar indicates the mean number of nucleotide substitutions per site. Species were classified according to Ashburner and McAllister (2016).

**Figure S7** The pseudolikelihood values for the number of hybridization events from 1 to 5.

**Table legend**

**Table 1** Putative diploid progenitors suggested for polyploids of *Betula* and included for phylogenetic analysis.

**Supplementary data**

**Table S1** Detailed information about *Betula* species used in this study and mapping results of RADseq.

**Table S2** Number of loci for each diploid species represented in the reference.

**Table S3** Detailed information about microsatellite markers mined from assembled contigs of species of *Betula*, *Alnus* and *Corylus*.
### Table 1. Putative diploid progenitors suggested for polyploids of *Betula* and included for phylogenetic analysis

<table>
<thead>
<tr>
<th>Species</th>
<th>Ploidy level</th>
<th>Putative diploid progenitors</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>B. pubescens</em> var. <em>pubescens</em></td>
<td>4</td>
<td><em>B. pendula</em>/<em>B. platyphylla</em></td>
</tr>
<tr>
<td><em>B. pubescens</em> var. <em>litwinowii</em></td>
<td>4</td>
<td><em>B. pendula</em>/<em>B. platyphylla</em></td>
</tr>
<tr>
<td><em>B. pubescens</em> var. <em>celtiberica</em></td>
<td>4</td>
<td><em>B. pendula</em>/<em>B. platyphylla</em></td>
</tr>
<tr>
<td><em>B. pubescens</em> var. <em>pumila</em></td>
<td>4</td>
<td><em>B. pendula</em>/<em>B. platyphylla</em></td>
</tr>
<tr>
<td><em>B. pubescens</em> var. <em>fragrans</em></td>
<td>4</td>
<td><em>B. pendula</em>/<em>B. platyphylla</em></td>
</tr>
<tr>
<td><em>B. papyrifera</em></td>
<td>6</td>
<td><em>B. cordifolia</em>/<em>B. populifolia</em>/<em>B. occidentalis</em></td>
</tr>
<tr>
<td><em>B. papyrifera</em> var. <em>commutata</em></td>
<td>6</td>
<td><em>B. cordifolia</em>/<em>B. populifolia</em>/<em>B. occidentalis</em></td>
</tr>
<tr>
<td><em>B. pumila</em></td>
<td>4</td>
<td><em>B. populifolia</em>/<em>B. occidentalis</em></td>
</tr>
<tr>
<td><em>B. albosinensis</em></td>
<td>4</td>
<td><em>B. ashburneri</em>/<em>B. costata</em></td>
</tr>
<tr>
<td><em>B. albosinensis</em> var. <em>septentrionalis</em></td>
<td>4</td>
<td><em>B. ashburneri</em>/<em>B. costata</em></td>
</tr>
<tr>
<td><em>B. utilis</em> ssp. <em>pratii</em></td>
<td>4</td>
<td><em>B. ashburneri</em>/<em>B. costata</em></td>
</tr>
<tr>
<td><em>B. utilis</em></td>
<td>4</td>
<td><em>B. ashburneri</em>/<em>B. costata</em></td>
</tr>
<tr>
<td><em>B. ermaninii</em></td>
<td>4</td>
<td><em>B. ashburneri</em>/<em>B. costata</em></td>
</tr>
<tr>
<td><em>B. ermaninii</em> var. <em>lanata</em></td>
<td>4</td>
<td><em>B. ashburneri</em>/<em>B. costata</em></td>
</tr>
<tr>
<td><em>B. cylindrostachya</em></td>
<td>4</td>
<td><em>B. luminifera</em>/<em>B. hainanensis</em></td>
</tr>
<tr>
<td><em>B. alnoides</em></td>
<td>4</td>
<td><em>B. luminifera</em>/<em>B. hainanensis</em></td>
</tr>
<tr>
<td><em>B. alleghaniensis</em></td>
<td>6</td>
<td><em>B. lenta</em></td>
</tr>
<tr>
<td><em>B. murrayana</em></td>
<td>8</td>
<td><em>B. lenta</em>/<em>B. occidentalis</em>/<em>B. populifolia</em></td>
</tr>
<tr>
<td><em>B. medwediewii</em></td>
<td>10</td>
<td><em>B. humilis</em>/<em>B. lenta</em></td>
</tr>
<tr>
<td>Species</td>
<td>Ploidy Level</td>
<td>Progenitors</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>B. megretica</td>
<td>12</td>
<td>B. humilis/B. lenta/B. maximowicziana/B. michauxii/B. luminifera</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. chinensis</td>
<td>6</td>
<td>B. calcicola/B. potaninii/B. chichibuensis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. chinensis</td>
<td>8</td>
<td>B. calcicola/B. potaninii/B. chichibuensis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. fargesii</td>
<td>10</td>
<td>B. calcicola/B. potaninii/B. chichibuensis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. delavayi</td>
<td>6</td>
<td>B. calcicola/B. potaninii</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. bomiensis</td>
<td>4</td>
<td>B. calcicola/B. potaninii</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. globispica</td>
<td>10</td>
<td>B. corylifolia/B. chichibuensis/B. lenta/B. michauxii/B. costata</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. insignis</td>
<td>10</td>
<td>B. corylifolia/B. chichibuensis/B. lenta/B. michauxii/B. luminifera/B.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>maximowicziana</td>
</tr>
</tbody>
</table>

1B. insignis (marked in blue) was provided with six putative progenitors due to the possible ploidy level either 10 or 12.
<table>
<thead>
<tr>
<th>Species</th>
<th>Inferior</th>
<th>Middle</th>
<th>Superior</th>
</tr>
</thead>
<tbody>
<tr>
<td>B. platycladus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. pendula</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. populifolia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. cordifolia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. occidentalis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. humilis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. aschersoniana</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. costata</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. globosa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. platanifolia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. alnoides</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. cylindrostachya</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. albo-sinensis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. ermanii</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. potaninii</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. calcicola</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. corylifolia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. michauxii</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. lentiscus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. nana</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. lenta</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. corylifolia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. luminifera</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. saundersii</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. papyrifera</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. pumila</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. calcitrapa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. potaninii</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. nigra</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. nana</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. lenta</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. corylifolia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. luminifera</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. saundersii</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. papyrifera</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. pumila</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. calcitrapa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. potaninii</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. nigra</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. nana</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. lenta</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. corylifolia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. luminifera</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. saundersii</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. papyrifera</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. pumila</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. calcitrapa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. potaninii</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. nigra</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. nana</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. lenta</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. corylifolia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. luminifera</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. saundersii</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. papyrifera</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. pumila</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. calcitrapa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. potaninii</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. nigra</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. nana</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. lenta</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>