MOLECULAR ECOLOGY

Molecular Ecology (2014) doi: 10.1111/mec.12837

Understanding the spectacular failure of DNA barcoding in willows (*Salix*): Does this result from a trans-specific selective sweep?

DIANA M. PERCY,*† GEORGE W. ARGUS,‡ QUENTIN C. CRONK,*† ARON J. FAZEKAS,§ PRASAD R. KESANAKURTI,§ KEVIN S. BURGESS,¶ BRIAN C. HUSBAND,§ STEVEN G. NEWMASTER,§ SPENCER C.H. BARRETT** and SEAN W. GRAHAM*†

**Department of Botany, University of British Columbia, Vancouver, BC, Canada V6T 1Z4, †Biodiversity Research Centre, University of British Columbia, Vancouver, BC, Canada V6T 1Z4, ‡Canadian Museum of Nature, PO Box 3443 Stn "D", Ottawa, ON, Canada K1P 6P4, §Department of Integrative Biology, University of Guelph, Guelph, ON, Canada N1G 2W1, ¶Department of Biology, Columbus State University, Columbus, GA 31907-5645, USA, **Department of Ecology & Evolutionary Biology, University of Toronto, 25 Willcocks Street, Toronto, ON, Canada M5S 3B2*

Abstract

Willows (*Salix*: Salicaceae) form a major ecological component of Holarctic floras and consequently are an obvious target for a DNA-based identification system. We surveyed two to seven plastid genome regions (~3.8 kb; ~3% of the genome) from 71 *Salix* species across all five subgenera, to assess their performance as DNA barcode markers. Although *Salix* has a relatively high level of interspecific hybridization, this may not sufficiently explain the near complete failure of barcoding that we observed: only one species had a unique barcode. We recovered 39 unique haplotypes, from more than 500 specimens, that could be partitioned into six major haplotype groups. A unique variant of group I (haplotype 1*) was shared by 53 species in three of five *Salix* subgenera. This unusual pattern of haplotype sharing across infrageneric taxa is suggestive of either a massive nonrandom coalescence failure (incomplete lineage sorting), or of repeated plastid capture events, possibly including a historical selective sweep of haplotype 1* across taxonomic sections. The former is unlikely as molecular dating indicates that haplotype 1* originated recently and is nested in the oldest major haplotype group in the genus. Further, we detected significant non-neutrality in the frequency spectrum of mutations in group I, but not outside group I, and demonstrated a striking absence of geographical (isolation by distance) effects in the haplotype distributions of this group. The most likely explanation for the patterns we observed involves recent repeated plastid capture events, aided by widespread hybridization and long-range seed dispersal, but primarily propelled by one or more trans-species selective sweeps.

Keywords: chloroplast capture, DNA barcoding, Malpighiales, molecular dating, phylogeography, selective sweep

Received 19 December 2012; revision received 29 May 2014; accepted 4 June 2014

Introduction

A partial failure of plastid data to track species boundaries is fairly common in phylogenetic plant studies (Percy *et al.* 2008; Starr *et al.* 2009; Hassel *et al.* 2013). This is often inferred by observations of strong incongruence between plastid and nuclear gene trees, or disagreements between gene trees and classically defined species (Xu *et al.* 2012; Yu *et al.* 2013). The processes underlying this incongruence likely contribute to the somewhat lower success of DNA barcoding markers as an identification tool in plants vs. animals (Fazekas *et al.* 2009), possibly reflecting more slowly evolving markers, or more problematic levels of hybridization,

Correspondence: Diana M. Percy, Department of Life Sciences, Natural History Museum, Cromwell Road, London SW7 5BD, UK, Fax: +44 (0)2079425229; E-mail: d.percy@nhm.ac.uk

© 2014 John Wiley & Sons Ltd

introgression or incomplete lineage sorting (Hollingsworth *et al.* 2011). Here, we report on patterns of variation in DNA barcoding markers in the genus *Salix*. The spectacular failure of DNA barcoding that we observe in willows may require an explanation involving trans-specific selection. This in turn may have important consequences for the interpretation of incongruence in other taxonomically complex plant groups.

Willows (*Salix*; Salicaceae), a genus of shrub and tree species, are a significant component of Holarctic ecosystems (Ager & Phillips 2008; Argus 2010; Myers-Smith *et al.* 2011). They are important indicators of riparian habitats, and many ecological studies include estimates of the diversity and abundance of willows because of their contribution to ecosystem function, community dynamics and assemblage (Myers-Smith *et al.* 2011). They are also extensively planted for habitat restoration, erosion control, and windbreaks, and the physiological adaptations and ecological resilience of willows make them valuable species for use in conservation and environmental projects (Kuzovkina & Quigley 2005; Kuzovkina & Volk 2009).

The genus contains ~450 species worldwide, including localized and widespread species with extensive circumpolar distributions. However, *Salix* species are notoriously difficult to identify based on morphology. Many species are vegetatively similar and exhibit substantial heteroblasty (the production of markedly different juvenile and adult leaves; Rechinger 1992; Zotz *et al.* 2011). Furthermore, *Salix* populations are dioecious, and taxonomic keys often require examination of both staminate and pistillate individuals. In many species, flowering (catkin production) occurs before leaf production, presenting an additional challenge when using keys that consider both vegetative and reproductive traits. Collectively, these factors can limit the characters available to identify individuals sampled at a single period of development. Finally, hybridization may blur boundaries between individual species, and hybrid offspring can exhibit highly variable morphologies (Mosseler 1990; Hardig *et al.* 2000).

The development of a molecular identification system for willows, such as DNA barcoding (CBOL Plant Working Group 2009), would be a useful tool for applied and basic research in the genus. The organellar genomes (plastid or mitochondrial) are currently the source of choice for plant and animal DNA barcodes. These genomes are haploid and typically uniparentally transmitted and have smaller effective population sizes than those in the diploid nuclear genome (and the dioecious sexual system of *Salix* would be expected to reduce effective population size further, Maynard Smith 1978). Loci in organellar genomes therefore undergo more rapid coalescence, which in turn should reduce confusion from incomplete lineage sorting and facilitate species fingerprinting. However, comparative studies have shown that while many plant groups may be amenable to reasonably precise species identification using DNA barcoding, others are less so (reviewed in Hollingsworth *et al.* 2011). Indeed, it is now acknowledged that plant species are generally harder to barcode than most animal species. This may reflect various phenomena, including large effective population sizes, periods of rapid speciation, larger disparity in dispersal rates of paternal and maternal haploid genomes and higher levels of hybrid speciation or introgression (Fazekas *et al.* 2009; Hollingsworth *et al.* 2011).

One or more of these phenomena may have an impact on the success of DNA barcoding in willows. Our overall goal in this study was to document the extent to which plastid DNA barcodes (CBOL Plant Working Group 2009) can be implemented in *Salix*. The more specific aims of this study were to: (i) establish the extent of species identification that is possible among western North American *Salix* using DNA barcode markers; (ii) to document haplotype patterns by sampling widely and deeply (within and among species); (iii) to establish a probable diversification timescale among plastid haplotypes (e.g. in the context of historic environmental and climatic shifts in North America); and (iv) to explore and evaluate hypotheses for the unusual patterns of molecular diversity that we observed.

Methods

Specimen sampling

Salix species are subdivided into five subgenera (four according to Chen *et al.* 2010) and are mostly found in arctic, boreal and temperate regions (Argus 2010). *Salix* is a large, diverse and widespread genus, and our sampling covers a wide range of the geographic (from North America, Asia, and Europe) and phenotypic diversity and includes a fair representation of the broad taxonomic diversity of the genus, with representatives from 27 of the ~40 sections (Table 1). Our species sampling focused on western North America, the most species-rich region of the North American willow flora. We sampled extensively within British Columbia, the province with the greatest willow diversity in Canada, with supplemental sampling in adjacent regions (e.g. Yukon, Alberta) and more widely across northern temperate regions, including species from central and eastern Canada, the United States, Europe, Mexico and Japan. We sampled a total of 546 individuals, representing 71 species, 10 identified hybrids and an additional 10 individuals that could not be identified to species due

© 2014 John Wiley & Sons Ltd

Table 1 *Salix* subgenera and sections (according to Flora of North America, Argus 2010) sampled for this study are shown (species sampled and number of individuals sampled per taxon relative to worldwide diversity are given in Appendix II) with assignment to the six major haploptype groups (G)

Subgenera	Section	G
Chamaetia	Chamaetia	I
	Diplodictyae	I
	Glaucae	I
	Herbella	I
	Lindleyanae	I
	Myrtilloides	I
	Myrtosalix	I
Longifoliae	Longifoliae	V, VI
Protitea	Humboldtianae	IV
Salix	Maccallianae	I
	Magnificae	I
	Salicaster	II, III, IV
	Salix	III
Vetrix	Arbuscella	I
	Canae	I
	Candidae	I, V
	Cinerella	I, V
	Cordatae	I, V
	Fulvae	I
	Geyerianae	I, V
	Hastatae	I, V
	Lanatae	I
	Mexicanae	I
	Phylicifoliae	I, V
	Sitchenses	I
	Villosae	I
	Viminella	I

to vegetative-only voucher material (referred to here as *Salix* spp. I-J). The sampling comprises 59% of the 113 species in the Flora of North America (Argus 2010). Our data set includes 29 species collected in British Columbia (BC), representing 57% of the willow flora of the province (BC has 51 native species: 53% of the native North American willow flora, and 65% of the native Canadian willow flora). The BC willow flora includes five introduced species (all are included in this analysis); none of these appear to be extensively invasive, although a few are naturalized in cultivated situations such as urban parks.

We typically sampled multiple individuals per species (a mean of eight individuals per species), with most (80%) represented by more than two individuals, and only 12 (eight of these non-native) represented by a single individual. Seven of the native, widespread, and highly variable species were represented by more than 20 individuals, with *Salix sitchensis*, one of the most widespread and phenotypically variable species in western North America, represented by 54 individuals. Our sampling included several voucher specimens collected between 1946 and 2000; individual herbarium samples of *Salix* with well preserved green leaves were selected, and they were all successfully amplified and sequenced (the oldest of these were collected in 1946 and 1947). We also sampled herbarium material from five hybrid willows from a study by Mosseler (1990), derived from experimental crosses made in 1983. We were able to sequence all five hybrids (see Appendix II) and the three parents for *rbc*L [*Salix interior* (maternal pistillate parent); *S. eriocephala* and *S. petiolaris* (paternal pollen donors)], and all of the hybrids and the two paternal parents for all of the remaining gene regions. Voucher specimens for new samples are housed at the University of British Columbia and the Canadian Museum of Nature, and specimen and collection details are publicly available on the BOLD database (project: SALIX; www.boldsystems.org; Ratnasingham & Hebert 2007).

The identity of each individual was carefully checked (by GWA and DMP), with assignments based on diagnostic morphological features of a given species, regardless of haplotype (although cases of highly variable hybrid morphologies are noted below). We also carefully checked for sequencing error by resequencing all individuals of species and hybrids that were placed in more than one haplotype group, or where hybrids with the same parents in the experimental crosses did not carry the same haplotype.

DNA extraction, amplification, sequencing and alignment

We stored field-collected leaf tissue in silica gel and extracted genomic DNA using a modified CTAB method (Doyle & Doyle 1987; Rai *et al*. 2003). For each individual, we surveyed up to seven plastid loci that have been considered as candidate DNA barcoding regions (Fazekas *et al*. 2008; CBOL Plant Working Group 2009). These comprised four coding regions (*mat*K, *rbc*L, *rpo*B, *rpo*C1) and three intergenic spacer regions (*atp*F-*atp*H, *psb*K-*psb*I, *trn*H-*psb*A). We amplified them with published primers (Appendix I) using the thermocycler and reaction conditions noted in Rai *et al*. (2003, 2008), but with the annealing temperature raised to 53 °C for the *trn*H-*psb*A region. Bidirectional sequencing of amplicons was performed at the Canadian Centre for DNA Barcoding (CCDB, University of Guelph; Ivanova *et al*. 2005). Our seven-region analyses used 145 individuals (39 species; Appendix II). We compiled a more extensive data set with an additional 401 samples for a total of 546 individuals (56 species) and sequenced these for the two core plant DNA barcoding regions, *mat*K and *rbc*L (CBOL Plant Working Group 2009) (532 of the 546 individuals were

© 2014 John Wiley & Sons Ltd

completely sequenced for both genes). We also sequenced cytochrome oxidase I (*COI*) for 30 samples (17 species and three hybrids) with the objective of testing a hypothesis that unusual aspects of the evolutionary history of the plastid genome (e.g. a selective sweep) may also be detectable for the mitochondrial genome (e.g. see Olson & McCauley 2000; Sun *et al.* 2014). Sequences were assembled in Sequencher 4.7 (Gene Codes Corp, Ann Arbor, MI), and Se-Al (Rambaut 1996) and are available from GenBank (see Data Accessibility section below for GenBank numbers). The plastid sequences are also archived on the BOLD database (www.boldsystems.org; Ratnasingham & Hebert 2007). Phylogenetic tree descriptions and the aligned matrices are also available from TreeBase (www.treebase.org). We had no detected sequence polymorphisms suggestive of multiple plastid types (e.g. heteroplasmy) or instances of DNA inversions. The length variations and indels were few, and we did not treat these as independent characters. Alignment was unproblematic, and we did not need to exclude regions.

Analyses of molecular data

We concatenated the plastid regions for analyses, as they belong to the same linkage group, and characterized major haplotype groups using heuristic maximum parsimony (MP), distance neighbour-joining (NJ), and maximum likelihood (ML) analyses in PAUP* (Swofford 2003) and a Bayesian analysis using BEAST (Drummond & Rambaut 2007), described further below. The major haplogroups as defined here represent nested sets of alleles whose phylogenetic relationships to each other are well supported, and whose relative times of divergence can be determined in a phylogenetic context. We restrict usage of the term 'haplotype' to refer to unique variants and use haplogroup or 'Group' when referring to sets of closely related haplotypes. The MP analysis employed 100 random addition replicates and NNI branch swapping, with MaxTrees set at 500; the NJ analysis used the Kimura 2-parameter (K2P) and the BioNJ option (Gascuel 1997), which is an improved NJ method for large DNA sequences data sets. The ML analysis considered a subset of taxa representing major haplotype groups (i.e. unique haplotypes only; see below) and used DNA substitution models and user-input model parameters indicated by the Akaike Information Criterion (AIC), as implemented in the program MODELTEST (Posada & Crandall 1998). The ML search used 10 random addition replicates and NNI branch swapping. Branch support for major haplotype groups (Table 2) was assessed using 200 bootstrap replicates (Felsenstein 1985) in RAXML (Stamatakis 2014). For MP, we used 10 random addition replicates for each of 200 bootstrap replicates and set MaxTrees to 100; for NJ, we used 1000 bootstrap replicates.

Maximum parsimony and NJ analyses of the plastid data were performed for: (i) each gene region independently (using the seven-region data set, Appendices II); (ii) the proposed core plant DNA barcoding combination (CBOL Plant Working Group 2009) *rbc*L + *mat*K (using the extended sampling, Appendix II), both with and without 13 individuals that were only sequenced for one of the two regions); (iii) a 145 sample data set in which all individuals are sequenced for all seven plastid regions (Appendix II). ML and Bayesian analyses were performed on a reduced version of the *rbc*L + *mat*K data set that excluded all identical haplotypes. The resulting 39 unique *Salix* haplotypes were aligned with sequences from 15 additional taxa in Salicaceae and Lacistemataceae obtained from GenBank, including representatives of the genera *Populus, Idesia, Poliothyrsis, Flacourtia, Xylosma, Casearia, Lunania, Scyphostegia* and *Lacistema* (see Appendix III for species and GenBank numbers), and henceforth, inclusion of all of these taxa (e.g. for the dating analyses described below) is referred to as a 'full' set of outgroup taxa (we also performed several dating analyses with a subset of outgroup taxa, see below). A haplotype network diagram of the seven-region data set for *Salix* was produced using Haploviewer (Salzburger *et al.* 2011) with the best ML tree topology ($-\ln L = -6295.796$) with uninformative/missing/ambiguous characters removed (see Fig. 1).

To aid in interpreting alternative processes (i.e. geographical isolation-by-distance effects vs. geographically widespread horizontal plastid capture) that may explain patterns observed in our data, we performed two types of statistical analyses. Although these methods, Mantel tests and the Tajima's *D* test statistic, are typically employed for within-species analyses, the extensive hybridization among willow species may support their applicability to the plastid data. Furthermore, we use these analyses to look specifically at differences between group I and the other major haplotype groups. We used Mantel tests (1000 iterations) performed with the 'Isolation-by-Distance Web Service' (IBDWS v 3.23; Jensen *et al.* 2005) to assess the strength of correlation between geographic distances (GPS coordinate point data transformed into pairwise distances using the Geographic Distance Matrix Generator; Ersts 2012) and plastid genetic distances (uncorrected and K2P distances obtained from PAUP*) among the North American individuals sampled on the seven plastid region data set, either with all taxa or with group I only. Non-North American samples were excluded to improve the likelihood of detecting any within-continent correlation. To take into consideration the much larger geographical distances vs. relatively small genetic distances, we also

© 2014 John Wiley & Sons Ltd

Table 2 Bootstrap support for six major haplotype groups (G) using MP, NJ, and ML analyses for (a) all seven plastid regions, (b) three plastid regions (*rbc*L + *mat*K + *trn*H-*psb*A) and (c) two plastid regions (*rbc*L + *mat*K)

Analysis/No. genes	G I	G II	I + II	G III	G IV	G V	G VI	IV + V + VI
MP 7	100	95	99	99	88	58	100	100
NJ 7	98	100	100	100	97	98	100	100
ML 7	98	99	98	100	98	80	100	100
MP 3	—	88	79	93	—	60	97	97
NJ 3	53	96	84	99	74	71	94	89
ML 3	—	89	89	97	—	87	100	98
MP 2	—	89	—	64	—	—	86	99
NJ 2	—	95	—	53	62	—	78	90
ML 2	—	75	—	75	74	—	88	97

performed the Mantel tests with and without log-transformation of the geographical distances. We used the Tajima's *D* test statistic to assess the frequency spectrum of selectively neutral mutations in the plastid data using DNASP v 5 (Librado & Rozas 2009) with all sites/mutations using the seven-region and two-gene data sets, with either a) all taxa, b) group II-VI, or c) group I only (Table 3). We selected a test of the overall frequency spectrum of polymorphisms rather than a genealogy based approach (e.g. HHT or HCT; Innan *et al.* 2005) due to the potentially confounding effects of lateral plastid transfer on estimating species boundaries.

To characterize the timescale of diversification of plastid markers in *Salix,* we estimated the ages of the major haplotype groups based on the *mat*K + *rbc*L data set (after reduction to nonidentical sequences). We first conducted a likelihood ratio test in PAUP*, comparing the model with and without the molecular clock enforced to assess whether there was significant rate heterogeneity in the *Salix* data, or in Saliceae (*Salix* + *Populus*) with two sets of additional outgroups (see below). We then performed molecular dating analyses using a Bayesian approach in BEAST and a maximum likelihood approach using r8s (version 1.71, Sanderson 2006). We confirmed that rate heterogeneity tests in PAUP* (described above) were consistent with the likelihood ratio test performed in r8s. Our BEAST analyses estimated mean rates of evolution, and trees and branch lengths from the data set with the full selection of outgroup taxa using the following parameters: substitution model GTR + Γ; clock model relaxed uncorrelated lognormal; tree prior Yule process with uniform distribution model; MCMC chain length of 20 million with 25% burn-in (multiple parallel analyses were run to check for stationarity, chain convergence and effective sample sizes). We visualized the results in the BEAST associated programs, TRACER, TREEANNOTATOR and FIGTREE. The ESS (effective sample size) values for all parameters estimated were >300. The r8s analyses used the ML tree recovered using PAUP* (described above). We ran the r8s analyses with two different outgroups, one with *Poliothyrsis* used to root a taxon set comprising Saliceae + *Idesia* + *Poliothyrsis*, and another with *Lacistema* used to root a taxon set that included the full set of outgroup taxa (in each case the outgroup taxon used to root the tree was pruned before divergence times were estimated; Sanderson 2006). We used the Langley–Fitch algorithm (LF) (for the taxon set comprising Saliceae + *Idesia* + *Poliothyrsis*) as these data were found to satisfy a molecular clock hypothesis. We used nonparametric rate smoothing (NPRS) and penalized likelihood (PL) methods (Sanderson 1997, 2002), with either Powell or TN (truncated Newton) algorithms for the full outgroup set, as these data violated a molecular clock. We established the optimum smoothing value for the PL analysis using the cross validation option in r8s. For all r8s analyses, we also used the CheckGradient option as a further confirmation of the correctness of the selected methods and algorithms.

We calibrated these analyses using several fossils. Fossil evidence and biogeographic studies suggest a possible warm temperate origin for *Salix* in North or Central America followed by early occupation of riparian habitats (Collinson 1992; Boucher *et al.* 2003; Abdollahzadeh *et al.* 2011). Subsequent range expansion into cooler northern hemisphere habitats was likely accompanied by repeated advances and retreats to refugia during glacial and interglacial periods (Ager *et al.* 2010). Because the leaf characters of salicoid (gland-tipped) teeth, camptodromous secondary venation, and elliptic, lanceolate or deltoid shape are not unique to *Salix* and *Populus* (Boucher *et al.* 2003; Cronk 2005), there are often problems associated with interpreting fossil material. However, the recent identification of the North American Eocene fossil *Pseudosalix* (bearing reproductive structures) as the immediate sister group to the tribe Saliceae (*Salix* + *Populus*; Boucher *et al.* 2003) has provided a useful additional calibration point for dating key events in the evolution of the family Salicaceae, the order Malpighiales, and the origins of

© 2014 John Wiley & Sons Ltd

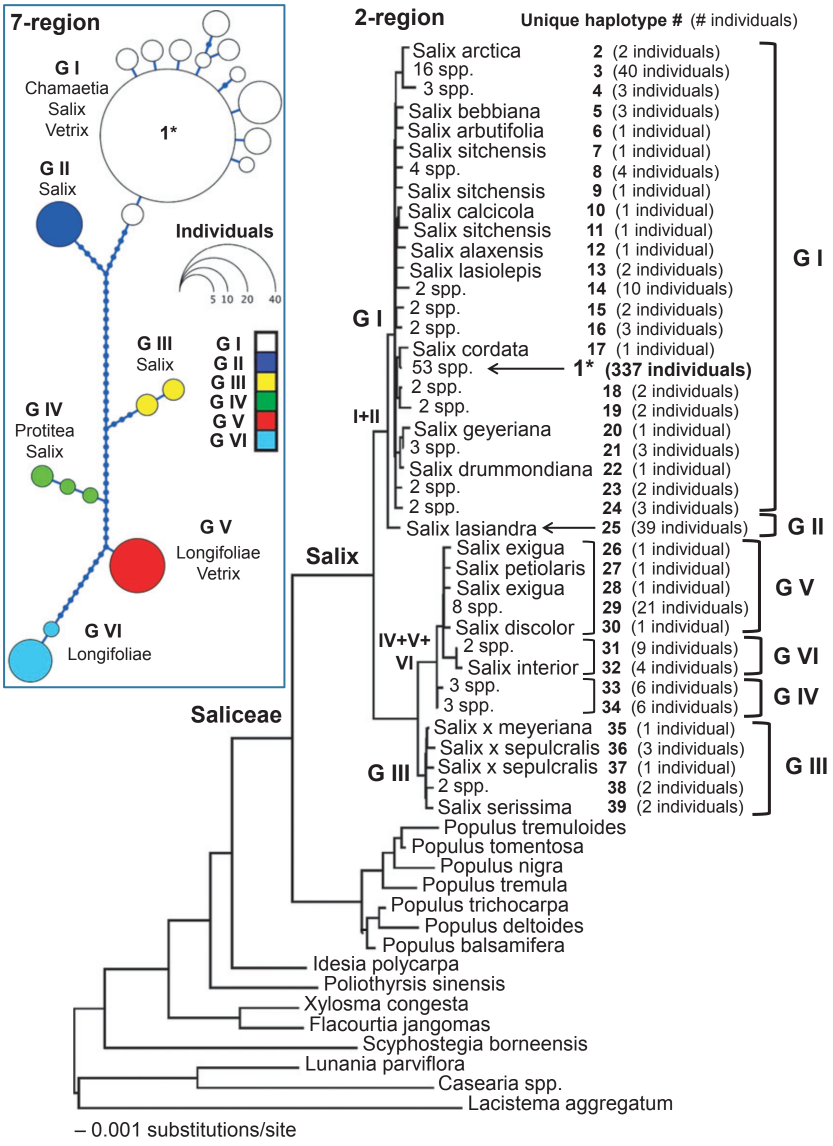

7-region
G I
Chamaetia
Salix
Vetrix
1*
G II
Salix
Individuals
5 10 20 40
G III
Salix
G I
G II
G III
G IV
G V
G VI
G IV
Protitea
Salix
G V
Longifoliae
Vetrix
G VI
Longifoliae
2-region
Unique haplotype # (# individuals)
Salix arctica 2 (2 individuals)
16 spp. 3 (40 individuals)
3 spp. 4 (3 individuals)
Salix bebbiana 5 (3 individuals)
Salix arbutifolia 6 (1 individual)
Salix sitchensis 7 (1 individual)
4 spp. 8 (4 individuals)
Salix sitchensis 9 (1 individual)
Salix calcicola 10 (1 individual)
Salix sitchensis 11 (1 individual)
Salix alaxensis 12 (1 individual)
Salix lasiolepis 13 (2 individuals)
2 spp. 14 (10 individuals)
2 spp. 15 (2 individuals)
2 spp. 16 (3 individuals)
Salix cordata 17 (1 individual)
53 spp. 1* (337 individuals)
2 spp. 18 (2 individuals)
2 spp. 19 (2 individuals)
Salix geyeriana 20 (1 individual)
3 spp. 21 (3 individuals)
Salix drummondiana 22 (1 individual)
2 spp. 23 (2 individuals)
2 spp. 24 (3 individuals)
Salix lasiandra 25 (39 individuals)
Salix exigua 26 (1 individual)
Salix petiolaris 27 (1 individual)
Salix exigua 28 (1 individual)
8 spp. 29 (21 individuals)
Salix discolor 30 (1 individual)
2 spp. 31 (9 individuals)
Salix interior 32 (4 individuals)
3 spp. 33 (6 individuals)
3 spp. 34 (6 individuals)
Salix x meyeriana 35 (1 individual)
Salix x sepulcralis 36 (3 individuals)
Salix x sepulcralis 37 (1 individual)
2 spp. 38 (2 individuals)
Salix serissima 39 (2 individuals)
G I
G II
G V
G VI
G IV
G III
G I
I+II
Salix
IV+V+
VI
G III
Saliceae
Populus tremuloides
Populus tomentosa
Populus nigra
Populus tremula
Populus trichocarpa
Populus deltoides
Populus balsamifera
Idesia polycarpa
Poliothyrsis sinensis
Xylosma congesta
Flacourtia jangomas
Scyphostegia borneensis
Lunania parviflora
Casearia spp.
Lacistema aggregatum
– 0.001 substitutions/site

Fig. 1 Maximum clade credibility tree from the BEAST analysis including 39 unique *mat*K + *rbc*L *Salix* haplotypes plus sequences obtained from GenBank for other genera (see Appendix III for details of outgroup taxon sampling). Major haplotype group positions (Groups I–VI) are shown and bootstrap support is given in Table 2. Inset: haplotype network diagram of the seven-region data set with major haplotype groups and *Salix* subgenera (see Table 1 and Appendix II for *Salix* classification).

modern tropical forests (Davis *et al.* 2005). The earliest unequivocal fossils of *Populus* and the fossil *Pseudosalix* are from the early Middle Eocene Green River Formation of Utah and Colorado, 46–50 Mya (Manchester *et al.* 1986; Boucher *et al.* 2003). These fossils consist of leafy shoots with unisexual inflorescences, and capsular infructescences. Earlier leaf records from the Cretaceous and Palaeocene are thought to combine features of *Populus*, *Idesia* and *Poliothyrsis* (Iljinskaya & Chelebaeva 2002; Boucher *et al.* 2003), and there are plausible records for *Populus* from the late Palaeocene (Collinson 1992). There are, however, numerous Eocene records for Salicaceae s.l., and by this period, *Populus* seems to have become well established in North America and Asia (Sun *et al.* 2004). It still remains unclear whether the extant genus *Salix* was also established in the Eocene period, despite some reports of *Salix* fossils from the Green River Formation (Brown 1934; MacGinitie 1969; Boucher *et al.* 2003). We therefore used two fossil age constraints: the Middle Eocene (Green River Formation) *Pseudosalix handleyi* fossil age of 48 Mya (Boucher *et al.* 2003) as a mean node age (in BEAST, standard deviation 0.5) and minimum age (in r8s) constraint for the crown clade of Saliceae [Saliceae (*Populus* + *Salix*) vs. *Idesia* split], and the *Casearia*-type Late-Middle Eocene (Panama) pollen age of 37 Mya (Graham 1985)

© 2014 John Wiley & Sons Ltd

Table 3 Results of the Tajima's *D* test statistic (Tajima 1989) which tests the hypothesis that the frequency spectrum of mutations is selectively neutral

Data set	S	Pi	Theta	Tajima's *D*
All taxa, 7 genes	103	0.494	0.625	−0.67 NS
All taxa, 2 genes	31	0.128	0.406	−1.8*
G I, 7 genes	38	0.043	0.229	−2.51***
G I, 2 genes	15	0.016	0.192	−2.16**
G II-VI, 7 genes	79	0.72	0.577	0.9 NS
G II-VI, 2 genes	20	0.383	0.296	0.85 NS

S = number of polymorphic sites; Pi = nucleotide diversity per 100 sites; Theta = estimate of mutation rate ($2N_e\mu$) per 100 sites; Tajima's *D* = significance of rejection of neutrality: NS, $P > 0.10$; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

as a mean node age (in BEAST, standard deviation 0.5) and a fixed-age (in r8s) constraint for the *Casearia–Lunania* split. Uncertainty in the age of the calibrated node can be partly accommodated in BEAST by applying a prior distribution, considering the fossil age as a minimum bound. An underlying assumption in applying fossil constraints at all is that splits in the gene tree correspond to splits in the species tree. This assumption may only hold approximately within *Salix* and relatives. Using relatively small standard deviations on the distribution priors in BEAST, and minimum/fixed ages in r8s may be sufficient for our purposes, as we are primarily interested in understanding the scale and order of gene tree splits, not the precise dates of these splits.

Results

Characteristics of individual and combined plastid regions

The overall sequencing success among the seven plastid regions was lowest for *mat*K at 90%, similar to other comparative studies (Fazekas *et al.* 2008; Kress *et al.* 2009). The sequencing success for all other regions was >95%, with the highest success for *rbc*L at 98%, and *trn*H-*psb*A at 99%. Sequence characteristics for each of the seven plastid gene regions (*mat*K, *rbc*L, *rpo*C1, *rpo*B, *trn*H-*psb*A, *atp*F-*atp*H, *psb*K-*psb*I) and the single mitochondrial gene region (*COI*), including the number of individuals sequenced, aligned sequence length, mean sequence length and parsimony informative characters (PIC) for each region are provided in Appendices IV and V. The *mat*K gene provided the largest number of phylogenetically informative characters, followed by the noncoding regions *psb*K-*psb*I, *atp*F-*atp*H and *trn*H-*psb*A (the latter region is relatively short in *Salix*; mean 238 bp) and then *rbc*L. The remaining two coding regions (*rpo*B and *rpo*C1) were the most conserved. We found only a single nucleotide polymorphism (SNP) in the *COI* region, a synapomorphy for two individuals that are in the same plastid haplotype group (I); otherwise, the mitochondrial sequences are identical. We observed six major haplotype groups in various individual and combined plastid gene analyses that were in general moderately to strongly supported by bootstrap analysis (groups I-VI in Table 2, Fig. 1, Appendix II). As expected, the seven regions belonging to the same plastid linkage group yield congruent gene trees, albeit with greater or lesser degrees of branch support (Table 2). To illustrate the information present in each data set, Appendix V shows a single maximum parsimony tree for each of the seven plastid regions analysed separately, the number of parsimony informative characters (PIC) for each gene region (see also Appendix IV), and pairwise matrices that show the predicted number of nucleotide substitutions along internal branches separating the six major haplotype groups.

Not surprisingly, the combined seven-region analysis (145 individuals; 3798 plastid nucleotides; 109 informative sites) provides the best support for relationships among the six haplotype groups, although four to five of these groups have at least 50% bootstrap support in analysis of two- and three-region multilocus barcoding combinations (i.e. *mat*K + *rbc*L, and *mat*K + *rbc*L + *trn*H-*psb*A) (Table 2). For all of these combinations, it is clear that the plastid phylogenies do not delimit taxonomic species (here based on Argus 2010). Four species in subgenus *Vetrix* (*S. candida*, *S. cordata*, *S. eriocephala* and *S. planifolia*) belonging to different sections, have different individuals that come out in more than one major haplotype group (in all four cases from groups I and V; Appendix II), and even within major haplotype groups, the individuals of a particular species may not have unique or identical haplotypes (e.g. group I and haplotype 1*, Appendix II). In the 145-sample, seven-region data set (which comprises 40 *Salix* species), there are 43 unique haplotypes, and no less than 52 individuals in 20 species carry the same variant of haplotype (type 1* in Appendix II; identical across 3798 bp and seven plastid regions) for some or all of the individuals sampled for those species.

Expanded sampling of the matK + rbcL data set

The two-gene analysis used a much greater number of samples (an additional 401 individuals), and although less well resolved (Table 2), this broader sampling confirms the overall patterns found in the seven-region analysis and suggests that the haplotype groupings based on only two genes are highly representative of plastid distributions in willows (Appendix II). With this expanded sampling, the number of

© 2014 John Wiley & Sons Ltd

individuals (337) and species (57 spp.) found with an identical haplotype (type 1*) surpasses those found elsewhere in group I (although the lower number of informative characters in the two-region vs. seven-region analysis likely also contributes to a higher number of individuals with identical sequence). Henceforth, when we refer to haplotype 1*, we mean in regard to the expanded sampling used for the two-region analysis, unless otherwise stated, which incorporates the taxa in the seven-region analysis (Appendix II). With less sequence data per taxon in the two-gene analysis, haplotype groups IV, V and VI are less clearly distinguishable (Table 2, Fig. 1).

The expanded sampling of individuals in the *mat*K + *rbc*L analysis provides additional evidence that divergent haplotypes are present in the four species already identified as carrying haplotype groups I and V in the less densely sampled seven-region analysis (i.e. *S. candida*, *S. cordata*, *S. eriocephala*, and *S. planifolia*) (Table 1, Appendix II). A subtle difference occurred for two individuals that had only *rbc*L sequence data, when incorporating those individuals in a combined gene analysis; *rbc*L by itself lacks sufficient information to differentiate haplotype group I from II, and group V from VI (Appendix V). Two other single gene analyses failed to differentiate all six haplotype groups: *rpoC1* with only four informative characters failed to differentiate group II from III, and group IV from V; and *mat*K, despite having 28 informative characters, failed to differentiate groups IV from V (Appendix V). The Asian species, *Salix arbutifolia*, has haplotype group I in our analysis based on the placement of GenBank sequences for *mat*K (EU790701) and *rbc*L (AB012776). This supports other studies that place this species, previously placed in the genus *Chosenia*, within *Salix* (Ohashi 2001; Chen *et al.* 2010; Hardig *et al.* 2010).

Remarkably, of 71 willow species, only one species, *S. lasiandra*, could be barcoded consistently and with confidence using either the seven- or two-region barcode (it has a distinct and unique haplotype, designated group II, that has strong bootstrap support, ≥ 95% with seven-regions; Fig. 1, Table 2, Appendix II). Another species, *S. interior*, is the only species in haplotype group VI, and it groups with most of the hybrid crosses using this species as the maternal parent; there is some intra-specific sequence variability within this haplotype group and strong support for group VI (bootstrap support 100% with seven-regions; Fig 1, Table 2, Appendix II).

Evidence from statistical tests

The results from the Mantel tests indicate that there is a significant correlation between geographic and genetic distance (the results using either uncorrected or K2P distances were similar, and we report only the latter here) using the seven plastid regions when all taxa are included ($P < 0.001$, $R^2 = 0.343$), but within group I alone, there is no such correlation ($P = 0.49$, $R^2 = 0.0002$). When the geographical distance axis was log transformed, the R^2 differed (all taxa: $R^2 = 0.133$; group I: $R^2 = 0.002$), but the significance (or lack) of the correlation did not change (all taxa: $P < 0.001$; group I: $P = 0.2$). These results add support to the hypothesis that there is taxonomically indiscriminate and widespread lateral gene transfer and spread of haplotypes, especially haplotype 1* in group I. Here, we use the term lateral (or horizontal) transfer to include processes involving hybridization and introgression.

The Tajima's *D* test statistic (Tajima 1989) tests the hypothesis that the frequency spectrum of mutations is selectively neutral. A significantly negative Tajima's *D* is expected when the data depart from neutral expectations. Our results using this test support the presence of strong positive selection located in group I as indicated by the highly significant negative Tajima's *D* results for this haplotype group ($P < 0.001$ and $P < 0.01$ in the seven-region and two-gene analyses, respectively) and the contrastingly positive Tajima's *D* results when group I is excluded (Table 3). These results demonstrate an excess of low-frequency polymorphisms in group I, which is indicative of a non-random process, such as a selective sweep of the plastid genome. An alternative explanation of an ancestral population expansion as the source of these polymorphisms (Muir & Filatov 2007) seems improbable, as group I consists of individuals from many different species distributed across the Palaearctic.

Molecular dating analyses

Our comparative dating analyses based on the *mat*K + *rbc*L reduced data set (i.e. eliminating identical sequences), contained 54 unique haplotypes (39 *Salix* and 15 samples from other genera obtained from GenBank) and comprised 1550 aligned nucleotides. The age estimates and order of chronological events for the divergence of the major haplotype groups are mostly consistent between the two dating methods that we employed (Appendix VI shows the age estimates from the BEAST and r8s analyses with the expanded outgroup sampling, and the r8s analyses that were run with a reduced outgroup configuration, see Methods). Although r8s-based dates are younger than those obtained from BEAST (the latter shown on Fig. 2), the relative order of chronological events of clade origins is the same between methods, and the 95% HPD (highest posterior density interval) range obtained from the BEAST

© 2014 John Wiley & Sons Ltd

analysis typically includes the r8s age estimates (Appendix VI). We were able to obtain older dates in r8s, similar to those obtained in BEAST, by lowering the smoothing factor in the PL analyses, which allows increased relaxation of the clock assumptions (relaxation of the clock by lowering the gradient factor below 20 caused the GradientCheck to fail). BEAST nearly always gave older age estimates for nodes, a pattern also found in a comprehensive review of comparative dating methods (Goodall-Copestake *et al.* 2009). We report on the BEAST ages here (Fig. 2; see Appendix VI for comparison to the r8s-based analyses). The mean rates of evolution (substitutions per site per year) as estimated separately in BEAST for *mat*K (3.5×10^{-10}) and *rbc*L (2.3×10^{-10}) accord with mean plastid rates for other angiosperms (e.g. legumes, Lavin *et al.* 2005). Our dating analyses places the diversification of tribe Saliceae (*Populus* and *Salix*) at around 35 Mya (26–42 Mya crown age for the clade comprising these two genera), which is consistent with the date established for the *Populus*-*Salix* split at the end of the Eocene (ca. 34 Mya) by Davis *et al.* (2005) using a much broader range of taxa and fossil constraints across the Malpighiales. We estimate diversification within extant *Salix* at ca. 20 Mya (13–28 Mya crown age). The younger than expected age estimate for *Populus* (ca. 17 Mya crown age; 10–26 Mya) may be due to the limited sampling for this genus in our data set. Of the major haplotype groups within *Salix*, group I was the oldest (crown age: 9.6 Mya), followed by group III (crown age: 5.6 Mya) and group V (crown age: 4.6 Mya), with haplotype group IV (crown age: 3.7 Mya) and group VI (crown age: 1.2 Mya) the youngest. Group II is represented by only a single haplotype, but the group I-group II split is estimated by BEAST as 12.9 Mya (Fig. 2).

Investigation of artificial hybrids and the influence of hybridization on the data

A complication to interpreting the expected distribution of haplotypes using plastid-based identification markers is the prospect of rare paternal transmission of plastids, which occurs in some gymnosperms and angiosperms (Muschner *et al.* 2006; Bouillé *et al.* 2011) and may occur in willows. Our data, derived from herbarium voucher samples of the artificial hybrid crosses of Mosseler (1990) (see Appendix II), suggest that 'leaky' paternal plastid inheritance may take place in willows, but any

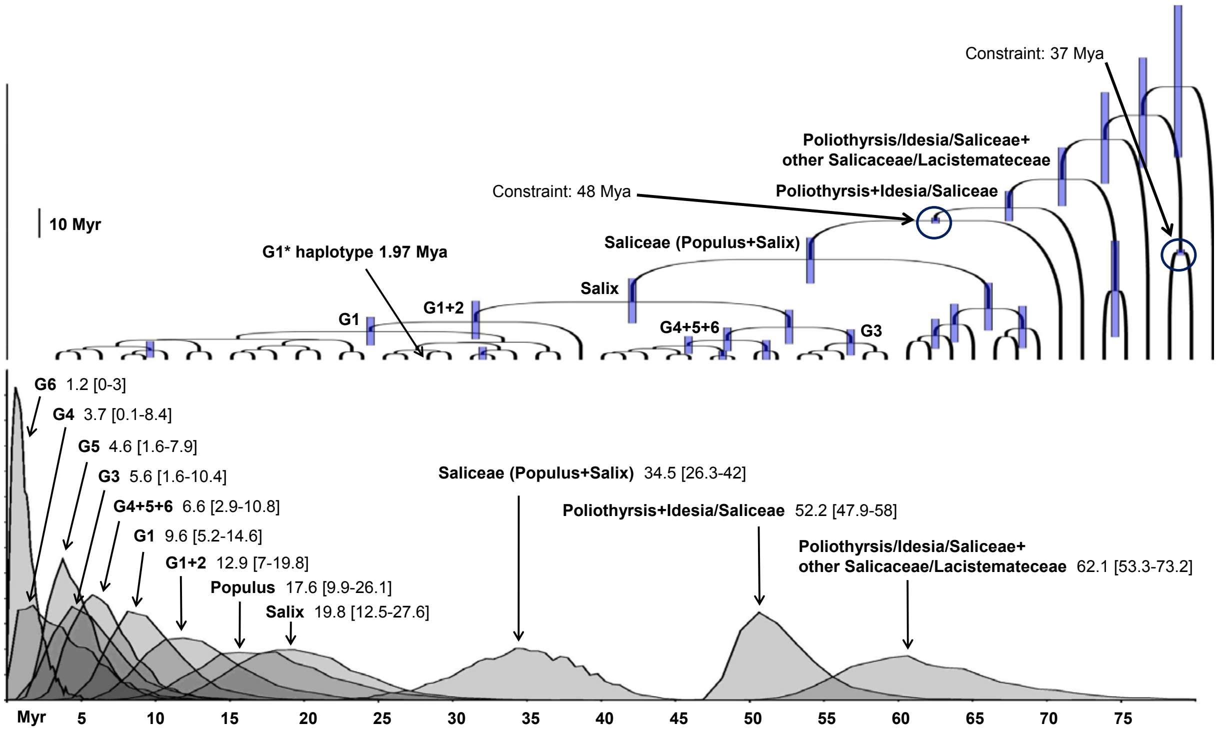


Fig. 2 Bayesian molecular dating analysis using the *rbc*L + *mat*K data set with 39 unique haplotypes sampled for *Salix*, plus sequences obtained from GenBank for *Populus* (seven species), *Idesia*, *Poliothyrsis*, *Flacourtia*, *Xylosma*, *Scyphostegia*, *Casearia*, *Lunania* and *Lacistema* (see Appendix III for details of outgroup taxon sampling). Haplotype groups are shown with the mean node age and the 95% highest posterior density (HPD) interval.

© 2014 John Wiley & Sons Ltd

inferences about the degree and mode of inheritance are limited here because we could only sequence the individual maternal parent (*S. interior*) for *rbc*L, and this was distinguishable only as an undifferentiated group V + VI haplotype (providing insufficient information to place it conclusively in either of these groups). However, the five hybrid progeny derived from the same maternal individual and two different paternal individuals are clearly assigned to two different haplotype groups: either group V or VI; two are in group V [one *interior* × *eriocephala* and one *interior* × *petiolaris*], and three are in group VI [two *interior* × *eriocephala* and one *interior* × *petiolaris*] see Appendix II). The paternal *S. eriocephala* individual is one of the samples of this species that carries haplotype group I, and the paternal *S. petiolaris* individual is in group V. So, although we cannot place the maternal parent definitively in haplotype group V vs. VI, we can at least say that hybrid progeny derived from the same parental crosses are not always found in the same haplotype group, supporting at least occasional biparental inheritance of the plastid genome. Repeating this experiment on a larger scale would be needed to clearly demonstrate leaky paternal inheritance of plastids in willows. We did not encounter any evidence of heteroplasmy when examining the DNA sequence chromatograms.

There are 18 taxonomic sections for which multiple species were sampled (Appendix II), and although the majority, 11, were found to consistently carry the same haplotype group among the species sampled in that section, various sections have at least one species in a highly divergent major haplotype group (Table 1). Section *Salicaster* within subgenus *Salix* was the most heterogeneous section with individuals from three different haplotype groups, including *S. lasiandra* (group II), *S.* x *meyeriana* (*S. pentandra* x *euxina*) and *S. serissima* (group III), and *S. lucida* (group IV). *Salix lasiandra* and *S. lucida*, once considered to be the same species and still thought to be closely related based on morphology, are found in different, highly divergent haplotype groups (also found in Hardig *et al.* 2010). In addition to section *Salicaster*, six other sections, five within subgenus *Vetrix*, were found to have species carrying multiple major haplotype groups, and in all cases they carried haplotype groups I and V. The prevalence for mixed sections and species to carry these two particular haplotype groups (groups I and V) appears non-random, and group I is always more commonly carried by more species within a taxonomic section than group V.

Discussion

Several recent studies have attempted to understand why DNA barcode identification systems may be unable to assign species in some groups of organisms (Fazekas *et al.* 2009; von Cräutlein *et al.* 2011; Arca *et al.* 2012). Willows have wind and insect-dispersed pollen (Argus 1974; Vroege & Stelleman 1990), and their seeds may travel large distances compared to species with more locally restricted dispersal (Steyn *et al.* 2004). In principle, therefore, the presence of wind pollination in tribe Saliceae (*Populus* and *Salix*; Boucher *et al.* 2003), combined with their efficiently wind and water-dispersed seeds should make them good candidates for barcoding identification methods. This is because high dispersability of genes among populations within species should help purge introgressed neutral genes from other species (Currat *et al.* 2008; Petit & Excoffier 2009; Hollingsworth *et al.* 2011). However, our study shows very low species-specific identity for plastid markers, and so these features of willow biology do not appear to be sufficient to purge foreign genes and ensure the success of DNA barcoding in the genus. Here, we compare expectations and evidence for phenomena that may contribute to the particularly poor success of DNA barcoding markers in willows.

Retention of ancestral polymorphisms

Massive coalescence failure (incomplete lineage sorting) and subsequent haplotype extinction events may explain widespread sharing of particular haplotypes in willows, especially haplotype 1* (Fig. 1), via expansions and contractions during a succession of interglacial periods up to and including after the last glacial maximum. Furthermore, coalescence failure could result from rapid radiation during the early diversification of the genus, and/or extremely large effective plastid population sizes in individual species (however, it is worth noting that both the haploid nature of the plastid organellar genome and the dioecious sexual system of *Salix* decrease the effective population size and increase the speed and rate of expected coalescence times compared to nuclear genes or hermaphroditic species). For this explanation to be tenable, it would require a large number of extinction events within major taxonomic groups to leave the current dispersed but uneven distribution of group I and 1* haplotypes across the genus. Coalescence failure during the early diversification of *Salix* would be expected to exhibit random sharing or extinction of genotypes among and across taxon groups. It is difficult to see how subsequent lineage sorting could have led to the observed pattern of variation among taxa, as this would require repeated sorting of one variant into terminal taxa. It also seems highly unlikely that variant 1* would not have diversified and accumulated variation since the origin of the genus, considering the deep diversification evident in the dated gene tree for *Salix*. The observed pattern of some haplotypes being limited to within taxonomic groups (Table 1), and

© 2014 John Wiley & Sons Ltd

of other haplotypes being found in many different taxonomic groups, is not consistent with retention of ancestral polymorphisms across numerous successive speciation events from the early diversification of *Salix*, unless sorting and extinction of ancestral polymorphisms occurred consistently in a non-random fashion. Such non-randomness seems highly unlikely. In addition, the argument for coalescence failure or incomplete lineage sorting is unconvincing, given the dated phylogeny, because the origin of haplotype 1* occurred recently. We estimated diversification of the crown clade of *Salix* at ca. 20 Mya (13–28 Mya). The age estimates for the six major haplotype groups in *Salix* range from ca. 9 Mya (group I) to 1 Mya (group VI). Haplotype 1* is the most widespread haplotype and appears to have originated around 2 Mya, but may be considerably younger given that the sequences are identical across many divergent species. It is not possible, given our data, to determine the relative ages of the haplotype groups vs. the ages of the species within those groups, but we can make some assumptions that, where a morphologically determined taxonomic section is uniquely represented by a major haplotype group (e.g. *Humboldtianae* represented by group IV; Table 1), that this haplotype group may be at least as old as the species within it. In contrast, haplotype 1*, within group I, may be considerably younger than any of the species sharing this haplotype, which would not support the coalescence-failure hypothesis.

Horizontal gene transfer and a selective sweep hypothesis

If coalescence failure can be ruled out as an explanation for the large-scale sharing of haplotypes in *Salix*, the most obvious remaining explanation for this pattern is that it reflects repeated capture and spread of plastids across species and continental barriers. In principle, this is consistent with extensive literature reports of hybridization in willows, which are primarily within subgenera, but can also be between subgenera (Argus 2010). However, if this spread across species boundaries involved neutral genes, we might expect a more random distribution of shared haplotypes, although this depends on whether hybridization patterns are truly random. Instead we observed an overwhelming dominance of certain haplotypes, especially variant 1* (Fig. 1). This hypothesis is difficult to reconcile with the breadth of both taxonomic and geographic samples with identical or near-identical haplotypes. For example, willows with haplotype 1* are mixed phenotypically, taxonomically (from many different subgenera and sections) and geographically (from North America and Europe; Table 1 and Appendix II). The most widely shared haplotype (type 1*) appears to have a very recent origin (<2 Myr), and it is nested within one of the oldest haplotype groups (group I; ca. 9 Myr). Given the large number of different morphospecies (53 species) sharing this identical haplotype, and the apparent rapidity and recency of its spread, it seems doubtful that introgression (or hybrid speciation) alone can explain the observed pattern of haplotype diversity here.

Several recent studies have suggested that organellar genomes can undergo selective sweeps in plants and animals (Muir & Filatov 2007; Lack *et al.* 2011). We propose that haplotype 1* may be the result of a trans-specific selective sweep (see Muir & Filatov 2007 for an example in *Silene*, Caryophyllaceae). The spread of an adaptive plastid may be facilitated by initial plastid capture (Tsitrone *et al.* 2003; Kapralov & Filatov 2007); but how well can selection propel one or more plastid types across multiple species boundaries? A number of studies have looked at the spread of advantageous alleles across populations within species (Gross *et al.* 2007; Presgraves *et al.* 2009; Blackman *et al.* 2010), and within-species selective sweeps may contribute to the cohesiveness of species (Morjan & Rieseberg 2004). But few studies have found evidence for trans-specific selective sweeps (e.g. Muir & Filatov 2007; Lack *et al.* 2011; Brand *et al.* 2013; Twyford *et al.* 2013). Nonetheless, this particular process should carry a detectable and testable signature (Muir & Filatov 2007). The Tajima's *D* test results (Table 3) support this hypothesis for *Salix*. These imply that, for group I at least, introgression of haplotype 1* is unlikely to have occurred via neutral processes. Horizontal gene transfer, whether by hybridization or other means, may result in a faster accumulation of genetic novelties than through mutation alone, and if selected on also contribute to evolutionary processes (Muir & Filatov 2007; Lucek *et al.* 2010; Hudson *et al.* 2011; Richards *et al.* 2011). However, whether hybridization is actually adaptive remains controversial (Schemske & Morgan 1990; Mallet 2005; Twyford & Ennos 2012).

Geographically widespread and promiscuous willows, such as present-day *S. candida, S. eriocephala, S. glauca, S. pedicellaris, S. planifolia*, may have aided the transmission of a plastid type over a wide geographic (and taxonomic) range. An example of a modern-day willow whose ancestor may have facilitated the initial stages of plastid transmission is *S. pedicellaris* (group I). This species currently has a wide native geographic range within North America, forms natural hybrids with at least six other species in group I, and is known to hybridize with two of the species that have multiple haplotype groups, *S. candida* and *S. eriocephala* (see Argus 2010; both carry groups I + V). In addition, *S. eriocephala* (groups I + V) forms natural hybrids with at least seven species from groups I, II, IV, VI (see Argus 2010) and *S. glauca* has a Holarctic distribution and forms natural hybrids with at least eight species in group I. Finally, *S. candida* and *S. planifolia*, both widespread in North America, and both

© 2014 John Wiley & Sons Ltd

carrying groups I + V, hybridize naturally with eight or more species (Argus 2010). The ancestors of these or other modern-day species may have been similarly promiscuous, and the long-distance seed dispersal typical of *Salix* species may further promote plastid spread on a large geographic scale (Palmé *et al.* 2003). In tandem with a selective sweep, this may explain the lack of geographical (isolation by distance) effects observed here in haplotype group I using the Mantel test.

In theory, a putative selective sweep of the plastid genome should also affect the mitochondrial genome when both are maternally inherited (Olson & McCauley 2000; McCauley *et al.* 2007), but the very different modes and tempos of evolution in plant organelles make detection of the same selective forces across both organellar genomes difficult (Palmer 1990; Soria-Hernanz *et al.* 2008). Our mitochondrial data were too invariable to provide evidence for or against a plastid selective sweep. Although selection on the plastid is feasible due to the number of functionally important genes encoded by this organelle (e.g. Kapralov & Filatov 2007), the observed pattern could also be caused by selection on linkage groups co-inherited with the plastid. In *Salix*, these not only comprise the other cytoplasmically inherited genome, the mitochondrion, but also recorded cases of cyto-nuclear linkage disequilibrium (Latta *et al.* 2001; Fields *et al.* 2014). Such patterns of disequilibrium may result from numerous processes, including migration, hybridization, drift, but also selection (Burke *et al.* 1998; Edmands & Burton 1999). Therefore, selection on the mitochondrion or a part of the nucleus that is maternally inherited could be responsible for the pattern we observe and may have swept the plastid by coinheritance.

To date, there are no existing, well sampled and robustly resolved nuclear phylogenies for the genus *Salix*. Two recent studies that include plastid and nuclear data for a limited number of taxa and gene regions (Hardig *et al.* 2010; Abdollahzadeh *et al.* 2011) have both found nonmonophyly of taxa, and in the case of Hardig *et al.*, incongruence between plastid and nuclear data and better agreement between the taxonomic classification of *Salix* and nuclear data. Our current understanding of *Salix* taxonomy comes from extensive morphological studies (e.g. Argus 2010) and current molecular data remains too limited to assist in improving species classifications (an exception being the synonomization of the genus *Chosenia* with *Salix*). Therefore, addressing the genetic/phylogenetic extent of species boundaries of *Salix* taxa should be a high priority of future research as it affects our interpretation of when gene transfers are truly lateral. This is an important caveat of any study, like ours, that relies on classical species definitions to understand trans-specific sharing of alleles. To further test the hypothesis of a selective sweep in *Salix* will require extensive sampling of taxa and nuclear regions. Additional data from the nuclear genome is needed to fully test our hypothesis that the pattern observed here is a result of a trans-specific selective sweep affecting (at least) the plastid genome.

Significance of the mode of inheritance

Our analysis of archived material used in experimental crosses suggests that there is primarily maternal inheritance in willows, but likely accompanied by limited paternal transmission. The material we had access to was not suitable for definitively answering the mode of plastid inheritance in *Salix*, but resolving this issue would be useful for tracking which parent contributed captured organellar genomes, and for better understanding the dynamics of gene flow during hybridization. The latter may differ if the 'invading' genome came maternally (through seeds) or paternally (through pollen). However, while a selective sweep could be affected by the fine details of the inheritance mode (e.g. whether seeds or pollen travel further), the overall pattern that we observed could reasonably be expected to occur with either mode of inheritance. We do not know if introgression is expected to be more frequent given a predominant mode of organellar inheritance via pollen or seed.

Conclusions

This study highlights the serious challenges to the use of plastid data for the barcoding of willows. The only willow species in our sample that was consistently distinguishable using plastid barcoding regions was *S. lasiandra*. The lack of interspecific variation, the occurrence of multiple species sharing identical haplotypes, combined, in some cases, with considerable intraspecific variation and species with multiple divergent haplotypes, all serve to confound the use of plastid data to identify willow species. The willows provide an extreme example of how DNA barcoding can fail. At the same time, this study illustrates how the drive to barcode the world's organisms can lead to insights into, not only the extent of failure or success expected when barcoding a particular group, but also the possible evolutionary mechanisms of this taxon-specific variability (or lack thereof). It is clear that a reliable barcode for every willow species will never be achieved using the plastid as the sole source for DNA markers. Species assignment in the willows using plastid data is simply not possible for nearly all species, and so plant DNA barcoding with plastid loci will not be useful for applications like ecological surveys, identification of riparian indicator and/or rare species. At most, these data provide haplotype group assignments that are nonrandom and have some consistency at the subgeneric

© 2014 John Wiley & Sons Ltd

and sectional levels. The unusual patterns of haplotype group assignment in willows can likely be explained by a combination of factors that include interglacial demographic history of populations, patterns and frequency of hybridization, and theories of haplotype spread such as selective sweeps. We have described here the potential patterns and signature of a trans-specific selective sweep, in particular the presence of a recently evolved trans-species haplotype. Past studies that have attributed the failure of plastid data to track species boundaries to processes such as lack of variation, hybridization, introgression, and incomplete lineage sorting, might also usefully look at the relative ages of the haplotypes involved to assess how widespread this phenomenon may be in taxonomically complex plant groups.

Acknowledgements

We are grateful to Genome Canada for financial support through the Ontario Genomics Institute to the Canadian Barcode of Life Network. We thank those who assisted with sequencing: Maria Kuzmina (University of Guelph) and Dorothy Cheung (University of British Columbia). We thank four anonymous reviewers for help improving the manuscript, and for useful discussions on population genetics and molecular dating, we are grateful to Armando Geraldes and William Illes. For assistance in the field we thank Cecilia Alstrom-Rapaport, Bruce Bennett, Ying Chang, Dominique Collett, Robert Foottit, Carl-Eric Granfelt, Jason Grant, Eric Maw, Nancy Moore, Jeffrey Saarela, John Wardell, and David White.

References

Abdollahzadeh A, Kazempour Osaloo S, Maassoumi AA (2011) Molecular phylogeny of the genus *Salix* (Salicaceae) with an emphasize to its species in Iran. *Iranian Journal of Botany*, **17**, 244–253.

Ager TA, Phillips RL (2008) Pollen evidence for Late Pleistocene Bering land and bridge environments from Norton Sound, northeastern Bering Sea, Alaska. *Arctic, Antarctic, and Alpine Research*, **40**, 451–461.

Ager TA, Carrara PE, McGeehin JP (2010) Ecosystem development in the Girdwood area, south-central Alaska, following late Wisconsin glaciation. *Canadian Journal of Earth Sciences*, **47**, 971–985.

Arca M, Hinsinger DD, Cruaud C *et al.* (2012) Deciduous trees and the application of universal DNA barcodes: a case study on the circumpolar *Fraxinus*. *PLoS ONE*, **7**, e34089.

Argus GW (1974) An experimental study of hybridization and pollination in *Salix* (willows). *Canadian Journal of Botany*, **52**, 1613–1619.

Argus GW (2010) *Salix*. In: Flora of North America Editorial Committee, eds. 1993 + . *Flora of North America North of Mexico*. 16 + vols. Flora of North America Association, New York and Oxford. Vol. 7, 23–162.

Blackman BK, Strasburg JL, Michaels SD *et al.* (2010) The role of recently derived FT paralogs in sunflower domestication. *Current Biology*, **20**, 629–635.

Boucher LD, Manchester SR, Judd WS (2003) An extinct genus of Salicaceae based on twigs with attached flowers, fruits and foliage from the Eocene Green River Formation of Utah and Colorado. *American Journal of Botany*, **90**, 1389–1399.

Bouillé M, Senneville S, Bousquet J (2011) Discordant mtDNA and cpDNA phylogenies indicate geographic speciation and reticulation as driving factors for the diversification of the genus *Picea*. *Tree Genetics and Genomes*, **7**, 469–484.

Brand CL, Kingan SB, Wu L *et al.* (2013) A selective sweep across species boundaries in *Drosophila*. *Molecular Biology and Evolution*, **30**, 2177–2186.

Brown RW (1934) The recognizable species of the Green River flora. *United States Geological Survey Professional Paper*, **185-C**, 45–77.

Burke JM, Voss TJ, Arnold ML (1998) Genetic interactions and natural selection in Louisiana *Iris* hybrids. *Evolution*, **52**, 1304–1310.

CBOL Plant Working Group (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences, USA*, **106**, 12794–12797.

Chen J-H, Sun H, Wen J *et al.* (2010) Molecular phylogeny of *Salix* L. (Salicaceae) inferred from three chloroplast datasets and its systematic implications. *Taxon*, **59**, 29–37.

Collinson ME (1992) The early fossil history of Salicaceae: a brief review. *Proceedings of the Royal Society of Edinburgh*, **98B**, 155–167.

von Cräutlein M, Korpelainen H, Pietiläinen M *et al.* (2011) DNA barcoding: a tool for improved taxon identification and detection of species diversity. *Biodiversity and Conservation*, **20**, 373–389.

Cronk QCB (2005) Plant eco-devo: the potential of poplar as a model organism. *New Phytologist*, **166**, 39–48.

Currat M, Ruedi M, Petit RJ *et al.* (2008) The hidden side of invasions: massive introgression by local genes. *Evolution*, **62**, 1908–1920.

Davis CC, Webb CO, Wurdack KJ *et al.* (2005) Explosive radiation of Malpighiales supports a mid-cretaceous origin of modern tropical rain forests. *The American Naturalist*, **165**, E36–E65.

Doyle JJ, Doyle JL (1987) A rapid isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*, **19**, 11–15.

Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, **7**, 214.

Edmands S, Burton RS (1999) Cytochrome C oxydase activity in interpopulation hybrids of a marine copepod: a test of nuclear-nuclear or nuclear-cytoplasmic coadaptation. *Evolution*, **53**, 1972–1978.

Ersts PJ (2012) Geographic Distance Matrix Generator (version 1.2.3). American Museum of Natural History, Center for Biodiversity and Conservation. Available from http://biodiversityinformatics.amnh.org/open_source/gdmg. Accessed on 29 July 2012.

Fazekas AJ, Burgess KS, Kesanakurti PR *et al.* (2008) Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS ONE*, **3**, e2802.

Fazekas AJ, Kesanakurti PR, Burgess KS *et al.* (2009) Are plant species harder to discriminate than animal species using DNA barcoding markers? *Molecular Ecological Resources*, **9**, 130–139.

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

© 2014 John Wiley & Sons Ltd

Fields PD, McCauley DE, McAssey EV *et al.* (2014) Patterns of cyto-nuclear linkage disequilibrium in *Silene latifolia*: genomic heterogeneity and temporal stability. *Heredity*, **112**, 99–104.

Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, **14**, 685–695.

Goodall-Copestake WP, Harris DJ, Hollingsworth PM (2009) The origin of a mega-diverse genus *Begonia* (Begoniaceae) using alternative datasets, calibrations and relaxed clock methods. *Botanical Journal of the Linnean Society*, **159**, 363–380.

Graham A (1985) Studies in neotropical paleobotany. IV. The Eocene communities of Panama. *Annals of the Missouri Botanical Garden*, **72**, 504–534.

Gross BL, Turner KG, Rieseberg LH (2007) Selective sweeps in the homoploid hybrid species *Helianthus deserticola*: evolution in concert across populations and across origins. *Molecular Ecology*, **16**, 5246–5258.

Hardig TM, Brunsfeld SJ, Fritz RS *et al.* (2000) Morphological and molecular evidence for hybridization and introgression in a willow (*Salix*) hybrid zone. *Molecular Ecology*, **9**, 9–24.

Hardig TM, Anttila CK, Brunsfeld SJ (2010) A phylogenetic analysis of *Salix* (Salicaceae) based on *mat*K and ribosomal DNA sequence data. *Journal of Botany*, ID 197696, 12.

Hassel K, Segreto R, Ekrem T (2013) Restricted variation in plant barcoding markers limits identification in closely related bryophyte species. *Molecular Ecology Resources*, **13**, 1047–1057.

Hollingsworth PM, Graham SW, Little DP (2011) Choosing and using a plant DNA barcode. *PLoS ONE*, **6**, e19254.

Hudson AG, Vonlanthen P, Seehausen O (2011) Rapid parallel adaptive radiations from a single hybridogenic ancestral population. *Proceedings of the Royal Society B-Biological Sciences*, **278**, 58–66.

Iljinskaya IA, Chelebaeva AI (2002) A new fossil genus *Utkholokia* combining leaf characters of *Populus* (Salicaceae) with those of *Idesia* and *Poliothyrsis* (Flacourtiaceae). *Botanicheskii Zhurnal*, **87**, 101–110.

Innan H, Zhang K, Marjoram P *et al.* (2005) Statistical tests of the coalescent model based on the haplotype frequency distribution and the number of segregating sites. *Genetics*, **169**, 1763–1777.

Ivanova NV, DeWaard JR, Hajibabaei M *et al.* (2005) Protocols for high volume DNA barcoding. http://www.dnabarcoding.ca/ (accessed 20 March 2014).

Jensen JL, Bohonak AJ, Kelley ST (2005) Isolation by distance, web service. *BMC Genetics* **6**, 13. http://ibdws.sdsu.edu/ (accessed 20 March 2014).

Kapralov MV, Filatov DA (2007) Widespread positive selection in the photosynthetic Rubisco enzyme. *BMC Evolutionary Biology*, **7**, 73.

Kress WJ, Erickson DL, Jones FA *et al.* (2009) Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. *Proceedings of the National Academy of Sciences, USA*, **106**, 18621–18626.

Kuzovkina YA, Quigley MF (2005) Willows beyond wetlands: uses of *Salix* L. species for environmental projects. *Water, Air and Soil Pollution*, **162**, 183–204.

Kuzovkina YA, Volk TA (2009) The characterization of willow (*Salix* L.) varieties for use in ecological engineering applications: co-ordination of structure, function and autecology. *Ecological Engineering*, **35**, 1178–1189.

Lack JB, Nichols RD, Wilson GM *et al.* (2011) Genetic signature of reproductive manipulation in the phylogeography of the bat fly, *Trichobius major*. *Journal of Heredity*, **102**, 705–718.

Latta RG, Linhart YB, Mitton JB (2001) Cytonuclear disequilibrium and genetic drift in a natural population of ponderosa pine. *Genetics*, **158**, 843–850.

Lavin M, Herendeen P, Wojciechowski MF (2005) Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the Tertiary. *Systematic Biology*, **54**, 530–549.

Librado P, Rozas J (2009) DNASP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**, 1451–1452.

Lucek K, Roy D, Bezault E *et al.* (2010) Hybridization between distant lineages increases adaptive variation during a biological invasion: stickleback in Switzerland. *Molecular Ecology*, **19**, 3995–4011.

MacGinitie HD (1969) The Eocene Green River flora of northwestern Colorado and northeastern Utah. *University of California Publications in Geological Sciences*, **83**, 1–203.

Mallet J (2005) Hybridization as an invasion of the genome. *Trends in Ecology and Evolution*, **20**, 229–237.

Manchester SR, Dilcher DL, Tidwell WD (1986) Interconnected reproductive and vegetative remains of *Populus* (Salicaceae) from the Middle Eocene Green River Formation, northeastern Utah. *American Journal of Botany*, **73**, 156–160.

Maynard Smith J (1978) *The Evolution of Sex*, p. 236. Cambridge University Press, Cambridge, UK.

McCauley DE, Sundby AK, Bailey MF *et al.* (2007) Inheritance of chloroplast DNA is not strictly maternal in *Silene vulgaris* (Caryophyllaceae): evidence from experimental crosses and natural populations. *American Journal of Botany*, **94**, 1333–1337.

Morjan CL, Rieseberg LH (2004) How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. *Molecular Ecology*, **13**, 1341–1356.

Mosseler A (1990) Hybrid performance and species crossability relationships in willows (*Salix*). *Canadian Journal of Botany*, **68**, 2329–2338.

Muir G, Filatov D (2007) A selective sweep in the chloroplast DNA of dioecious *Silene* (Section *Elisanthe*). *Genetics*, **177**, 1239–1247.

Muschner VC, Lorenz-Lemke AP, Vecchia M *et al.* (2006) Differential organellar inheritance in *Passiflora*'s (Passifloraceae) subgenera. *Genetica*, **128**, 449–453.

Myers-Smith IH, Forbes BC, Wilmking M *et al.* (2011) Shrub expansion in tundra ecosystems: dynamics, impacts and research priorities. *Environmental Research Letters*, **6**, 045509.

Ohashi H (2001) Salicaceae of Japan. *Scientific Report, Tohoku University, 4th Series Biology*, **40**, 269–396.

Olson MS, McCauley DE (2000) Linkage disequilibrium and phylogenetic congruence between chloroplast and mitochondrial haplotypes in *Silene vulgaris*. *Proceedings of the Royal Society, Biological Sciences*, **267**, 1801–1808.

Palmé AE, Semerikov V, Lascoux M (2003) Absence of geographical structure of chloroplast DNA variation in sallow, *Salix caprea* L. *Heredity*, **91**, 465–474.

Palmer JD (1990) Contrasting modes and tempos of genome evolution in land plant organelles. *Trends in Genetics*, **6**, 115–120.

Percy DM, Garver AM, Wagner WL *et al.* (2008) Progressive island colonization and ancient origin of Hawaiian *Metrosideros* (Myrtaceae). *Proceedings of the Royal Society, B-Biological Sciences*, **275**, 1479–1490.

© 2014 John Wiley & Sons Ltd

Petit RJ, Excoffier L (2009) Gene flow and species delimitation. *Trends in Ecology & Evolution*, **24**, 386–393.

Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.

Presgraves DC, Gérard PR, Cherukuri A, Lyttle TW (2009) Large-scale selective sweep among segregation distorter chromosomes in African populations of *Drosophila melanogaster*. *PLoS Genetics*, **5**, e1000463.

Rai HS, O'Brien HE, Reeves PA *et al.* (2003) Inference of higher-order relationships in the cycads from a large chloroplast data set. *Molecular Phylogenetics and Evolution*, **29**, 350–359.

Rai HS, Reeves PA, Peakall R *et al.* (2008) Inference of higher-order conifer relationships from a multilocus plastid data set. *Botany*, **86**, 658–669.

Rambaut A (1996) *Se-Al: Sequence Alignment Editor, Version 2*. Oxford University, Oxford.

Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System. *Molecular Ecology Notes*, **7**, 355–364.

Rechinger KH (1992) *Salix* taxonomy in Europe – problems, interpretations, observations. *Proceedings of the Royal Society of Edinburgh*, **98B**, 1–12.

Richards TA, Soanes DM, Jones MDM *et al.* (2011) Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. *Proceedings of the National Academy of Sciences, USA*, **108**, 15258–15263.

Salzburger W, Ewing G, Haeseler A (2011) The performance of phylogenetic algorithms in estimating haplotype genealogies with migration. *Molecular Ecology*, **20**, 1952–1963.

Sanderson MJ (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution*, **14**, 1218–1231.

Sanderson MJ (2002) Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Molecular Biology and Evolution*, **19**, 101–109.

Sanderson MJ (2006) r8s version 1.71. Analysis of rates ('r8s') of evolution. Section of Evolution and Ecology. Univeristy of California, Davis, http://loco.biosci.arizona.edu/r8s/ (accessed 15 January 2012).

Sang T, Crawford DJ, Stuessy TF (1997) Chloroplast DNA phylogeny, reticulate evolution and biogeography of *Paeonia* (Paeoniaceae). *American Journal of Botany*, **84**, 1120–1136.

Schemske DW, Morgan MT (1990) The evolutionary significance of hybridization in *Eucalyptus*. *Evolution*, **44**, 2150–2151.

Soria-Hernanz DF, Braverman JM, Hamilton MB (2008) Parallel rate heterogeneity in chloroplast and mitochondrial genomes of Brazil nut trees (Lecythidaceae) is consistent with lineage effects. *Molecular Biology and Evolution*, **25**, 1282–1296.

Stamatakis A (2014) *RAXML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies*. In Bioinformatics, 2014, open access.

Starr JR, Naczi RF, Chouinard BN (2009) Plant DNA barcodes and species resolution in sedges (*Carex*, Cyperaceae). *Molecular Ecology Resources*, Suppl. s1, 151–163.

Steyn EMA, Smith GF, Van Wyk AE (2004) Functional and taxonomic significance of seed structure in *Salix mucronata* (Salicaceae). *Bothalia*, **34**, 53–59.

Sun B, Yan D, Xie S, Cong P, Xin C, Yun F (2004) Palaeogene fossil *Populus* leaves from Lanzhou Basin and their palaeoclimatic significance. *Chinese Science Bulletin*, **49**, 1494–1501.

Sun Y, Abbott RJ, Li L *et al.* (2014) Evolutionary history of purple cone spruce *Picea purpurea* in the Qinghai-Tibet Plateau: homoploid hybrid origin and Pleistocene expansion. *Molecular Ecology*, **23**, 343–359.

Swofford DL (2003) *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4*. Sinauer, Sunderland, Massachusetts.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Tsitrone A, Kirkpatrick M, Levin DA (2003) A model for chloroplast capture. *Evolution*, **57**, 1776–1782.

Twyford AD, Ennos RA (2012) Next-generation hybridization and introgression. *Heredity*, **108**, 179–189.

Twyford AD, Kidner CA, Harrison N *et al.* (2013) Population history and seed dispersal in widespread Central American *Begonia* species (Begoniaceae) inferred from plastome-derived microsatellite markers. *Botanical Journal of the Linnean Society*, **171**, 260–276.

Vroege PW, Stelleman P (1990) Insect and wind pollination in *Salix repens* L. and *Salix caprea* L. *Israel Journal of Botany*, **39**, 125–132.

Xu B, Wu N, Gao XF *et al.* (2012) Analysis of DNA sequences of six chloroplast and nuclear genes suggests incongruence, introgression, and incomplete lineage sorting in the evolution of *Lespedeza* (Fabaceae). *Molecular Phylogenetics and Evolution*, **62**, 346–358.

Yu WB, Huang PH, Li DZ *et al.* (2013) Incongruence between nuclear and chloroplast DNA phylogenies in *Pedicularis* section *Cyathophora* (Orobanchaceae). *PLoS ONE*, **8**, e74828.

Zinovjev AG (2011) *Salix* x *meyeriana* (= *Salix pentandra* x *S. euxina*) – a forgotten willow in Eastern North America. *Phytotaxa*, **22**, 57–60.

Zotz G, Wilhelm K, Becker A (2011) Heteroblasty – a review. *The Botanical Review*, **77**, 109–151.

D.P., Q.C., S.G. and S.B. planned and wrote the first draft of the manuscript. All authors contributed to subsequent revisions and editing of the manuscript. The following authors contributed to specimen sampling and identification (D.P., G.A., Q.C., A.F., P.K., K.B., B.H., S.N.); lab work and sequencing (D.P., A.F., P.K., K.B.); combining molecular genetics and ecological theory (Q.C., B.H.).

Data accessibility

DNA sequences: Genbank accessions KM001945-KM002483 (*mat*K), KM002817-KM003354 (*rbc*L), KM002484-KM002650 (*rpo*B), KM002651-KM002816 (*rpo*C1), KM003552-KM003706 (*atp*F-*atp*H), KM003707-KM003869 (*psb*K-*psb*I), KM003355-KM003551 (*trn*H-*psb*A), KM001915-KM001944 (*COI*).

Final DNA sequence assembly: aligned sequences available from BOLD database and GenBank.

Phylogenetic data: TreeBASE Study accession no. TB2:S15979.

Specimen information, sampling locations, and DNA sequences and alignments uploaded to Dryad (doi:10.5061/dryad.qf6t5).

© 2014 John Wiley & Sons Ltd

Appendix I

Primers used in the amplification and sequencing of seven plastid regions and one mitochondrial region. The *trn*H-*psb*A primers are published in Sang *et al.* (1997), all other plastid and *COI* primers in Fazekas *et al.* (2008), except *mat*K reverse primer 'EquiR' which is published online at http://www.kew.org/barcoding/update.html as *mat*K primer: R (Equisetum)

Gene region name	Mean sequence length bp	Primer name and sequence 5′ – 3′
*trn*H-*psb*A	238	trnH – GGCGCATGGTGGATTCACAAATC psbA – GTTATGCATGAACGTAATGCTC
*rpo*B	358	2F – ATGCAACGTCAAGCAGTTCC 3R – CCGTATGTGAAAAGAAGTATA
*rpo*C1	471	1F – GTGGATACACTTCTTGATAATGG 4R – CCATAAGCATATCTTGAGTTGG
*mat*K	888	Xf – TAATTTACGATCAATTCATTC EquiR – GTACTTTTATGTTTACGAGC
*rbc*L	628	80F – ATGTCACCACAAACAGAAACTAAAGCAAGT ajf634R1 – GAAACGGTCTCTCCAACGCAT
*atp*F-*atp*H	567	atpF – ACTCGCACACACTCCCTTTCC atpH – GCTTTTATGGAAGCTTTAACAAT
*psb*K-*psb*I	470	psbK – TTAGCCTTTGTTTGGCAAG psbI – AGAGTTTGAGAGTAAGCAT
COI	649	cox42F – GGATCTTCTCCACTAACCACAA cox1ajf699R – CCGAAAGAGATGCTGGTATA

Appendix II

Salix species and hybrids sampled for this study with subgeneric and section assignment (approximate number of species worldwide given in {}) according to Flora of North America (Argus 2010). A superscript ^{E} indicates the experimental hybrids of Mosseler (1990). The number in brackets beside species names is the number of individuals sampled for this study. The assignment of individuals to the six major haplotype groups (G) is given (seven-region data set/two-region data set), and species or hybrids carrying more than one haplotype group are in bold. Native or introduced [N/I (native range given)] status follows Argus (2010), unless provenance of material is Europe (Eu), or of garden origin [GO (native range given)]. *Salix* x *meyeriana* (= *Salix pentandra* × *S. euxina*) (Zinovjev 2011)

Taxa	N/I	GI 1*	GI others	GII	GIII	GIV	GV	GVI
subg. Chamaetia {142}								
sec. Chamaetia {4}								
S. nivalis (5)	N	3/5	1/–					
S. reticulata (38)	N	–/30	–/8					
S. vestita (2)	N		–/2					
sec. Diplodictyae {5}								
S. arctica (19)	N	3/12	8/7					
sec. Diplodictyae x Glaucae								
S. arctica x glauca (1)	N		–/1					
sec. Diplodictyae x Herbella								
S. arctica x polaris (2)	N	–/1	–/1					
S. petrophila (1)	N		–/1					
sec. Glaucae {8}								
S. brachycarpa (2)	N	–/2						
S. glauca (23)	N	–/16	5/6					
S. glauca x *brachycarpa* (2)	N	–/1	–/1					
S. nakamurana (1)	GO (Japan)	–/1	1/–					
S. niphoclada (3)	N	–/2	–/1					
S. x *glauca* (1)	N		–/1					
sec. Herbella {7}								
S. polaris (1)	N	–/1						
S. rotundifolia (2)	N	–/1	–/1					

Appendix II *Continued*

Taxa	N/I	GI 1*	GI others	GII	GIII	GIV	GV	GVI
sec. Lindleyanae {21}								
S. lindleyana (1)	GO (Asia)	–/1						
sec. Myrtilloides {5}								
S. athabascensis (2)	N	–/1	–/1					
S. pedicellaris (2)	N		–/2					
sec. Myrtosalix {20}								
S. uva-ursi (1)	N	1/1						
subg. Longifoliae {7}								
sec. Longifoliae {7}								
S. exigua (6)	N						–/6	
S. interior (11)	N							7/11
subg. Longifoliae x Vetrix								
sec. Longifoliae x Cordatae								
S. interior x *eriocephala*^{E} (3)	N							2/3
subg. Longifoliae x Vetrix								
sec. Longifoliae x Geyerianae								
***S. interior* x *petiolaris*^{E} (2)**	N						1/1	–/1
subg. Protitea {32}								
sec. Humboldtianae {15}								
S. amygdaloides (2)	N					1/2		
S. amygdaloides x *gooddingii* (2)	N					–/2		
S. bonplandiana (2)	N					–/2		
S. gooddingii (2)	N					–/2		
S. laevigata (2)	N					–/2		
S. nigra (1)	N					1/1		
subg. Salix {84}								
sec. Maccallianae {1}								
S. maccalliana (6)	N	–/6						
sec. Magnificae {8}								
S. magnifica (1)	GO (Japan)	–/1	1/–					
sec. Salicaster {9}								
S. lasiandra (37)	N			9/37				
S. lucida (2)	N					2/2		
S. x *meyeriana* (1)	I (Eu)				–/1			
S. serissima (2)	N				2/2			
sec. Salix {8}								
S. alba (1)	I (Eu/Asia)				1/1			
S. x *fragilis* (3)	I (Eu)				–/3			
S. x *sepulcralis* (5)	I (Eu/Asia)				2/5			
subg. Vetrix {211}								
sec. Arbuscella {13}								
S. arbusculoides (4)	N	1/4	1/–					
sec. Canae {1}								
S. elaeagnos (1)	I (Eu)	1/1						
sec. Candidae {2}								
***S. candida* (11)**	N	3/9	1/1				1/1	
sec. Cinerella {36}								
S. discolor (6)	N						3/6	
S. hookeriana (13)	N	2/12	1/1					
S. humilis (2)	N						2/2	
S. pedicellata (1)	Eu	–/1						
S. scouleriana (35)	N	3/26	9/9					
sec. Cordatae {1}								
***S. eriocephala* (4)**	N	1/2	1/–				2/2	
S. famelica (2)	N	1/2	1/–					
S. ligulifolia (4)	N	–/4						

© 2014 John Wiley & Sons Ltd

Appendix II *Continued*

Taxa	N/I	GI 1*	GI others	GII	GIII	GIV	GV	GVI
S. lutea (2)	N	–/2						
S. prolixa (13)	N	1/11	–/2					
sec. Fulvae {7}								
S. bebbiana (34)	N	7/24	4/8					
sec. Geyerianae {4}								
S. geyeriana (2)	N	–/1	–/1					
S. lemmonii (2)	N	–/2						
S. petiolaris (4)	N						4/4	
sec. Hastatae {25}								
S. arizonica (3)	N	–/2	–/1					
S. barclayi (32)	N	7/24	4/5					
S. boothii (8)	N	–/5	1/3					
S. commutata (3)	N	1/3	1/–					
***S. cordata* (5)**	N	–/2	1/1				–/2	
S. eastwoodiae (2)	N	–/1	–/1					
S. farriae (2)	N	–/2						
S. monticola (3)	N	–/2	–/1					
S. myricoides (2)	N	–/1	–/1					
S. myrtillifolia (4)	N	2/2	–/2					
S. pseudomonticola (7)	N	–/7	2/–					
S. pseudomyrsinites (6)	N	2/6	1/–					
S. pyrifolia (2)	N	1/2						
sec. Lanatae {5}								
S. calcicola (2)	N		–/2					
S. richardsonii (5)	N	–/3	–/2					
sec. Mexicanae {5}								
S. irrorata (2)	N	–/2						
S. lasiolepis (4)	N	–/2	–/2					
sec. Phylicifoliae {11}								
S. drummondiana (18)	N	–/14	1/3					
***S. planifolia* (8)**	N	–/6	–/1				–/1	
S. pulchra (2)	N	–/2						
sec. Sitchenses {4}								
S. jepsonii (3)	N	–/3						
S. sitchensis (54)	N	7/45	8/9					
sec. Villosae {6}								
S. alaxensis (8)	N	1/7	–/1					
S. barrattiana (7)	N	1/5	3/8					
sec. Villosae x Phylicifoliae								
S. alaxensis x *drummondiana* (1)	N	–/1						
sec. Viminella {13}								
S. viminalis (1)	Eu	–/1	1/–					

© 2014 John Wiley & Sons Ltd

Appendix III

Taxon sampling from GenBank (GB) for *rbc*L and *mat*K from Salicaceae (Sal) and Lacistemataceae (Lac)

Taxon (Family)	GB *rbc*L	GB *mat*K
Salix arbutifolia (Sal)	AB012776	EU790701
Populus balsamifera (Sal)	EU676955	EU749348
Populus deltoids (Sal)	AJ418829	EU790702
Populus nigra (Sal)	AJ418828	AB038186
Populus tomentosa (Sal)	AF527489	AY177666
Populus tremula (Sal)	AJ418827	AJ506082
Populus tremuloides (Sal)	AF206812	JF429913
Populus trichocarpa (Sal)	NC009143	NC009143
Idesia polycarpa (Sal)	AB021924	AB233831
Poliothyrsis sinensis (Sal)	AJ402991	EF135586
Flacourtia jangomas (Sal)	AF206768	EF135541
Xylosma congesta (Sal)	AB233938	AB233834
Scyphostegia borneensis (Sal)	AJ403000	EF135594
Casearia (Sal)	AF206746 *C. sylvestris*	EF135516 *C. nitida*
Lunania parviflora (Sal)	AB233936	EF135561
Lacistema aggregatum (Lac)	AY935746	FJ670025

Appendix IV

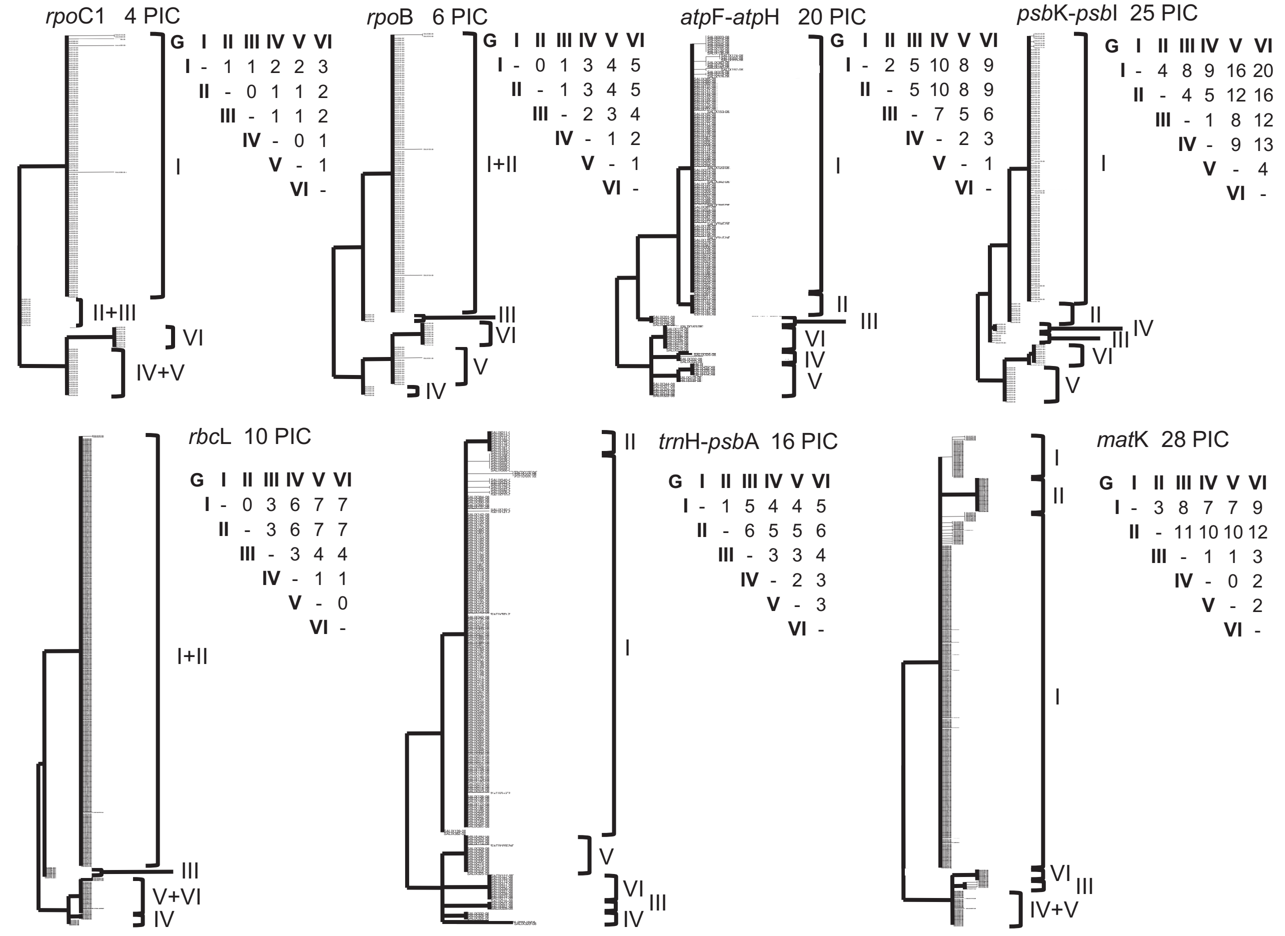
Sequence characteristics of each of the seven plastid regions (*mat*K, *rbc*L, *rpo*C1, *rpo*B, *trn*H-*psb*A, *atp*F-*atp*H, *psb*K-*psb*I), and the single mitochondrial gene (*COI*), including parsimony informative characters (PIC) for each region

Plastid region	No. individuals sequenced	aligned sequence length	mean sequence length	PIC
*mat*K	539	898	888	28
*rbc*L	538	634	628	10
*rpo*C1	166	474	471	4
*rpo*B	167	358	358	6
*trn*H-*psb*A	197	291	238	16
*atp*F-*atp*H	155	609	567	20
*psb*K-*psb*I	163	534	470	25
COI	30	656	649	1

© 2014 John Wiley & Sons Ltd

Appendix V

Maximum parsimony analysis for individual gene regions are illustrated with a single parsimony tree for each of the seven plastid regions, the number of parsimony informative characters (PIC) for each region or gene are given, and the placement of the six major haplotype groups are indicated. The matrices show the number of nucleotide changes along branches separating each of the six major groups.



*rpo*C1 4 PIC

G	I	II	III	IV	V	VI
I	-	1	1	2	2	3
II		-	0	1	1	2
III			-	1	1	2
IV				-	0	1
V					-	1
VI						-

*rpo*B 6 PIC

G	I	II	III	IV	V	VI
I	-	0	1	3	4	5
II		-	1	3	4	5
III			-	2	3	4
IV				-	1	2
V					-	1
VI						-

*atp*F-*atp*H 20 PIC

G	I	II	III	IV	V	VI
I	-	2	5	10	8	9
II		-	5	10	8	9
III			-	7	5	6
IV				-	2	3
V					-	1
VI						-

*psb*K-*psb*I 25 PIC

G	I	II	III	IV	V	VI
I	-	4	8	9	16	20
II		-	4	5	12	16
III			-	1	8	12
IV				-	9	13
V					-	4
VI						-

*rbc*L 10 PIC

G	I	II	III	IV	V	VI
I	-	0	3	6	7	7
II		-	3	6	7	7
III			-	3	4	4
IV				-	1	1
V					-	0
VI						-

*trn*H-*psb*A 16 PIC

G	I	II	III	IV	V	VI
I	-	1	5	4	4	5
II		-	6	5	5	6
III			-	3	3	4
IV				-	2	3
V					-	3
VI						-

*mat*K 28 PIC

G	I	II	III	IV	V	VI
I	-	3	8	7	7	9
II		-	11	10	10	12
III			-	1	1	3
IV				-	0	2
V					-	2
VI						-

Appendix VI

Comparative estimates of node ages in BEAST and r8s for the *mat*K + *rbc*L data set, using additional taxa sampled from GenBank (Appendix III). The BEAST mean node ages [median and the 95% highest posterior density (HPD) interval] are given. The BEAST analysis and the last two r8s analyses given were run with the expanded outgroup taxon set (i.e. *Salix* plus all taxa in Appendix III), the first three r8s analyses given were run with *Salix* plus two *Populus* taxa, *Idesia* and *Poliothyrsis*. The smoothing factor used in the penalized likelihood (PL) analyses is given in parentheses

	BEAST – Bayesian	r8s – LF (Powell)	r8s – PL (smooth 320)	r8s – PL (smooth 32)	r8s – PL (smooth 20)	r8s – NPRS
G6	1.2 [1; 0–3]	0.49	0.44	0.39	0.31	1.38
G3	5.6 [5.3; 1.6–10.4]	1.86	1.73	1.83	1.45	8.66
G4 + 5 + 6	6.6 [6.3; 2.9–10.8]	2.40	2.21	2.11	1.66	7.9
G1	9.6 [9.2; 5.2–14.6]	2.54	2.48	8.16	3.32	10.88
G1 + G2	12.9 [12.5; 7–19.8]	4.07	3.98	11.65	5.28	18.02
G3 + 4 + 5 + 6	11 [10.6; 5.5–17.1]	5.33	5.03	5.47	4.22	13.91
Salix	19.8 [19.5; 12.5–27.6]	14.75	14.36	17.88	12.23	22.76
Saliceae	34.5 [34.6; 26.3–42]	fixed 34	fixed 34	fixed 34	32.02	36.16

© 2014 John Wiley & Sons Ltd