Genomic Tools for Identification of Medicinal Plants

A thesis submitted by

Marco Kreuzer

for the degree of Doctor of Philosophy

School of Biological Sciences

University of Reading

August, 2017

To my parents Liseli and Karl

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Marco Kreuzer

Reading, August, 2017

Abstract

DNA barcoding of herbal medicines has raised awareness of species substitution and adulteration, highlighting issues surrounding their safety and quality. Regulation of herbal medicines is a pressing issue for regulatory agencies and, in response, DNA barcodes have recently been incorporated into the British Pharmacopoeia. Previous studies have found that DNA barcoding to species-level may be impaired by evolutionary mechanisms. This thesis investigates evolutionary relationships of genus *Berberis* and their impacts on DNA barcoding. Phylogenetic relationships within genus *Berberis* in the Himalayas and the Hengduan Mountains are studied using whole plastid genomes and hundreds of nuclear loci. The phylogenies reveal pronounced biogeographic structures in the Sino-Himalayan region and suggest that the relatively recent orogeny of the Hengduan Mountains has a strong impact on *in situ* diversification of *Berberis* species. Low phylogenetic resolution at species-level may be explained by incomplete lineage sorting. The phylogenies suggest that evolutionary mechanisms hinder DNA barcoding to species-level and, therefore, a method is devised for identifying evolutionary lineages. A strategy for generating DNA barcodes based on diagnostic nucleotides using whole plastid genomes is presented. These barcodes are tested on commercial samples, and their utility for regulatory purposes outlined. Furthermore, species substitution and adulteration in global trade are evaluated with two different specimen identification methods. The first uses the phylogenetic placements of commercial samples of *Berberis* for specimen identification. The second approach applies DNA metabarcoding to commercial samples of *Phyllanthus amarus*. The results of these analyses show that congeneric species are in trade and further reveal a high congruence between species in global and local markets, emphasizing the dependency of global medicinal plant trade on local trade systems. Finally, sequencing data from genus *Arabidopsis* is analysed to identify the effect of assembling nuclear loci that belong to paralogous clusters on phylogenomic inference. Read mapping from cognate paralogues in *Arabidopsis* has little to no effect on outcomes from phylogenomic inference.

Acknowledgments

This thesis would not have been possible without the support from my supervisors Julie Hawkins, Caroline Howard and Colin Pendry. I would like to thank Julie Hawkins for her support and inspiration during this PhD. Her scientific expertise, positive attitude and patience were invaluable. I would like to thank Caroline Howard for giving me the opportunity to work at the National Institute of Biological Standards and Control (NIBSC), South Mimms. Her expertise in medicinal plant barcoding and her numerous methodological inputs contributed greatly to this work. I would also like to thank Colin Pendry from the Royal Botanic Garden Edinburgh (RBGE), who hosted me many times at the RBGE. His expertise about the flora of Nepal was immensely important for conceptualizing this PhD and his guidance for planning and conducting field work in Nepal was invaluable.

This project was part of the MedPlant ITN and received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 606895. I am grateful for this support and would like to thank the people who were involved in acquiring the funding. I would like to thank all people from the MedPlant network for the interesting and valuable scientific inputs and the friendship.

I would like to thank Bhaskar Adhikari from the RGBE for his continuous help during the completion of this project and for providing many of the samples that were used. I would also like to thank him for the great time we spent together on a field expedition in Nepal in 2014, for offering help in taxonomic questions on genus *Berberis*, and for the scientific discussions during my PhD. I would like to thank Leo Gibson, Christina Gkouva, and Mina Kalantarzadeh from the herbals group at NIBSC. Special thanks go to Claire Lockie-Williams with whom I shared many hours in the lab and who was always helpful when I needed her. I would further like to thank the members of the NGS core facility, Nadine Holmes and Martin Fritzsche for their support in the lab and Mark Preston for valuable discussions on project design and bioinformatics. I would like to thank Edward Mee for his continuous support in the lab and for his inputs in the project. Julian Harber provided *Berberis* samples from the Hengduan Mountains and I thank him for his contribution to this work and his inputs on *Berberis* taxonomy and diversity. I would also like to thank Logan Kistler, University of Warwick for the many

discussions we had about the laboratory aspects of this work. His inputs were immensely important for the success of the lab work. Furthermore, I thank Umer Zeeshan Ijaz for letting me attend his excellent course on metagenomics at the University of Glasgow, where I learned valuable bioinformatics skills.
I would like to thank Hugo de Boer and Anneleen Kool for hosting me at the Natural History Museum, Oslo. I would also like to thank Audun Schrøder-Nielsen with whom I spent time in the lab. Furthermore, I would like to thank Vincent Manzanilla for the great collaboration we had and that is ongoing.
Big thanks go to my colleagues and friends here at the University of Reading. I would like to thanks Andrew Meade for providing me help in phylogenetics and for helpful discussions about this project. Big thanks to Irene Teixidor Toneu, Paul Wennekes, Estevão Fernandes de Souza and Mazhani Binti Muhammad for the many discussions about biology and medicinal plants and, most importantly, for their friendship. Thanks to Ilias Tzimourakas for help in Python programming.
I would like to thank my family who were always there for me and supported me wherever they could, despite being geographically apart. My mother and my father had such a great way of sending me their energy. My brother Frank with his wife Sandra and my sister Andrea with her husband Jan were always supportive and their interest motivated me immensely. My nieces Luisa, Hannah and Ella deserve a special note because they made me smile on so many occasions.
My biggest thanks go to Andrea who I met at the beginning of my PhD and who was always at my side during these years. It’s difficult to express how grateful I am to have her. Her smiles and her words always motivated me to continue.

Table of contents

Abstract iii
Acknowledgments iv
Table of contents vi
List of tables ix
List of figures 1
Chapter 1 Introduction 3
1.1 Preface 3
1.2 Methodological considerations 4
1.2.1 From phylogenetics to phylogenomics 4
1.2.2 DNA barcoding in the era of next-generation sequencing 5
1.3 Study organisms 7
1.3.1 *Berberis* L. 8
1.3.2 *Phyllanthus* L. 9
1.3.3 Genomic resources for *Arabidopsis* (DC.) Heynh. 9
1.4 Thesis organisation 10
Chapter 2 A phylogenetic hypothesis for *Berberis* (Berberidaceae) in the Himalayas and the Hengduan Mountains 12
2.1 Introduction 12
2.2 Material and methods 16
2.2.1 Sampling 16
2.2.2 Laboratory work 16
2.2.3 Bioinformatics 19
2.3 Results 28
2.3.1 Nuclear DNA marker assembly 28
2.3.2 Filtering of nuclear DNA markers 28
2.3.3 Plastid assembly and alignment 30
2.3.4 Phylogenetic hypotheses 31
2.3.5 Ancestral range estimation 32
2.4 Discussion 38
2.4.1 Species tree inference from nuclear data 38
2.4.2 Phylogenetic relationships 40
2.4.3 Conflict between nuclear and plastid hypotheses 42

2.4.4 Biogeography and evolution 43
2.4.5 Conclusion 45
Chapter 3 New approaches for DNA barcoding herbal medicines: a case study of genus *Berberis* 46
3.1 Introduction 46
3.2 Material and methods 50
3.2.1 Sampling 50
3.2.2 Sequencing 50
3.2.3 Plastid genome reconstructions 50
3.2.4 Annotation of plastid sequence 51
3.2.5 Universal barcode reconstruction 52
3.2.6 Barcoding analysis and phylogenies 52
3.2.7 Test dataset 54
3.3 Results 55
3.3.1 Whole plastid phylogeny 55
3.3.2 Identifying informative barcodes 57
4.3.2 Testing barcodes 63
3.4 Discussion 64
Chapter 4 Perspectives on global trade and on the regulation of medicinal plants revealed by DNA barcoding 68
4.1 Introduction 68
4.2 Material and methods 71
4.2.1 Sampling 71
4.2.2 *Berberis* phylogeny 72
4.2.3 *Phyllanthus*: DNA metabarcoding 73
4.3 Results 73
4.4 Discussion 76
4.4.1 Species composition of globally traded products 76
4.4.2 Global trade mirrors local markets 77
4.4.3 Generic complexes and concepts of substitution 78
Chapter 5 Impact of targeting paralogues on phylogenomic inference 81
5.1 Introduction 81
5.2. Materials and methods 84
5.2.1 Sampling 84
5.2.3 Marker assembly and allele reconstruction 84

5.2.5 Tree reconstruction 86
5.3 Results 87
5.4 Discussion 91
Chapter 6 General discussion 95
6.1 Summary of findings 95
6.2 The future of phylogenomics 96
6.3 The future of medicinal plant barcoding 99
6.4 Research questions emerging from this study 101
Bibliography 103
Appendices 120
Appendix figures 120
Appendix tables 123

List of tables

Table 2-1 Summary of Berberis samples used for shotgun sequencing. Voucher specimens are deposited at the Royal Botanic Garden Edinburgh (RBGE). 17

Table 2-2 Output from BioGeoBears analysis. The parameters are d=dispersal, e=extinction and j=founder-event speciation. The model was chosen according to the Akaike information criterion (AIC). 37

Table 3-1 In silico mixtures of B. aristata and B. asiatica samples. 55

Table 3-2 Barcode selection resulting from investigating variability patterns across whole plastid alignment. ITS2, matK and rbcL were not identified as highly variable but included in the study. Var = Variable sites; PIS = parsimony informative sites; "aristata recovered" and "asiatica recovered" indicates whether the clades were recovered in the respective phylogeny. 57

Table 3-3 Results from the model test for molecular evolution. The GTR model was only favoured for the barcode ndhI-ndhG. GTM = General Time Reversible Model; TVM = Transversion Model; TIM = Transition Model; TPM = 3-parameter Model. ... 60

Table 3-4 Top: Matrix of informative barcode positions. The positions are relative to the consensus of the multiple sequence alignments of each barcode. "SA clade" stands for South American clade. Bottom: Results of the test samples. Market1, Market2 and Market11 are commercial samples and Mixture1 and Mixture2 are in silico mixtures. Numbers below multiple base calls represent the ratio of nucleotides in the mapping. . 63

Table 4-1 Berberis trade samples. 71

Table 4-2 Phyllanthus trade samples 72

Table 5-1 Arabidopsis samples used in this study. The number of reads comprises forward and reverse reads. 85

Appendix Table AT-1 Table with specimen information. 123

Appendix Table AT-2 Sequencing information. The seqeuencing strategy describes whether the sample was target enriched (TE), shotgun sequenced (SG) or borth (TE + SG). Numbers in the row "Capture" indicates which samples were pooled in the hybridization capture. Furthermore, the average coverage and standard deviation (Stdev) are displayed. 126

List of figures

Figure 2-1 Average coverage per sample across all loci. 28

Figure 2-2 The heatmap shows the fraction recovered of each loci (n=607) for samples that were included in the phylogenetic analysis. Grey bars indicate loci where no sequence could be retrieved. 29

Figure 2-3 Top left: The plot shows the average sequence similarity and standard deviation (grey bars) per gene. The red line is the set arbitrary threshold. Top right: Plot of sequence similarity and standard deviation per locus (black dot). Loci outside the red rectangle were discarded for further analysis (mean <= 94, sd <= 6). Bottom left: Average pairwise phylogenetic distance per loci with standard deviation (grey bars). Genes that exceeded the threshold of 0.625 were discarded. Bottom right: Plot of the mean pairwise phylogenetic distance and standard deviation of each loci (black dots). Loci outside the red rectangle were discarded (average > 0.625, sd > 0.4). 30

Figure 2-4 Quality filtering and coverage plots. Note that mapping quality is usually not applied to indels. 31

Figure 2-5 Top: Maximum likelihood tree of concatenated gene alignments. Only bootstrap values below 100 are shown above branches. The tree scale describes the mean substitutions per site. Numbers in circles indicate the major clades. Bottom: Map with specimen localities. Colours correspond to clades in the phylogeny. 33

Figure 2-6 Left: Bayesian phylogeny inferred from concatenated marker alignments. All nodes have a posterior probability of 1. Right: Phylogeny based on the MSC inferred with ASTRAL-II. Numbers above branches are quartet scores, no displayed number stands for full support. 34

Figure 2-7 *Top:* The traces of the likelihoods of the15 independent runs. The burnin of 10 million generations is not shown. The runs were run for a minimum of 45 million generations. *Bottom:* The marginal probabilities are displayed as a density plot. Note that the likelihood curve of one run is slightly shifted, indicating that the run has not converged. The results from this run were excluded from further analysis. 35

Figure 3-1 ML phylogeny based on whole plastid sequences. Note that B. aristata, in the aristata clade, is a polyphyletic species, but the B. asiatica in the asiatica groups are monophyletic. Numbers above branches are bootstrap values between 51 and 99. Branches with support < 50 were collapsed to polytomies, bootstrap values of 100 are not shown. 56

Figure 3-2 SNP density along the plastid genome (red histograms). The outer circle describes the boundaries of the large single copy (LSC), the inverted repeats (IRa and IRb) and the small single copy (SSC). Regions that are coloured green in the inner circle are coding regions, blue are RNA genes (rRNA and tRNA genes) and white is noncoding sequence. Red colour below the outer circle shows regions that have been masked and are thus coded as “N”. 58

Figure 3-3 Subselection of barcode regions with the SSC_noncoding2 region. The newly determined barcode is marked in red. 58

Figure 3-4 Maximum likelihood phylogenies and haplotype networks of individual barcodes. The Roman numerals indicate different haplotypes and the size of the circles corresponds to the number of samples sharing this haplotype. A: SSC_noncoding2, B: matK, C: ndhI-trnG. 61

Figure 3-5 Maximum likelihood tree from the concatenated barcodes matK, SSC_noncoding2 and ndhI-ndhG. Nodes with bootstrap support <50 were collapsed to polytomies. Bootstrap values beween 50 – 99 are shown above branches. No number indicates a bootstrap value of 100. Numbered circles indicate groups that were recovered in the whole plastid phylogeny (see. Fig 3-1). 62

Figure 4-1 Phylogeny of 43 Berberis species and 16 market samples. The labels of market samples are coloured in red. Only bootstrap values < 100 are shown (numbers above branches). Coloured dots represent where the samples have been bought. In the case of UK samples, the provenance of the raw material is unknown and were purchased in the UK. 74

Figure 4-2 Heatmap of species identified. The relative abundance of reads per product mapping to species are represented with the intensity of colour. 75

Figure 5-1 The boxplots describe the average coverage of orthologous (left) and paralogous markers (right) per sample. For reasons of scaling the individual plots, the highest 66 data points per sample were removed prior to producing the boxplots (final data set: n = 1,100 markers per sample). 88

Figure 5-2 Density plot of sums of branch length. 88

Figure 5-3 A: Mean and standard deviation of sequence similarity of each marker, calculated from the pairwise distances between pairs of alleles. B: Mean and standard deviation of phylogenetic distance of pairwise alleles of each marker. Density plots on top and right of the scatter plots describe the distribution of points along the respective axis. 89

Figure 5-4 Phylogenetic trees of concatenated orthologous, paralogous and the combined marker dataset. Top row: Maximum likelihood tree with 100 bootstrap replicates. Numbers below branches are shown when the bootstrap support is lower than 100. Bottom row: Majority rule extended consensus tree. Branch lengths are ignored. Numbers above branches are gene support frequencies (GSF) in percent. 90

Appendix Figure AF-1 Consensus network from a sample of 984 trees from the Bayesian analysis. The network is a 3D representation drawn in 2D. No conflicting splits were detected. 120

Appendix Figure AF-2 Gene map of the plastid genome of Berberis aristata. Genes on the outside of the circle are transcribed clockwise and genes on the inside anti-clockwise. The dark grey histograms in the inner circle show the GC content. 121

Appendix Figure AF-3 Phylogenies of the selected barcodes ndhI-ndhG, matK and SSC_noncoding2 under different models of evolution. The aristata and asiatica clades were both recovered, leading to the same conclusion as under the GTRCAT model. .122

Chapter 1 Introduction

1.1 Preface

Since Darwin's *On the Origin of Species*, biologists have recognized the crucial importance of heredity, variation and natural selection as forces in creating biological diversity. Through the development of novel DNA sequencing techniques, researchers nowadays have unprecedented opportunities for analysing this diversity imprinted in the genetic material of organisms. Genomics has played a subordinate role in medicinal plant research, but is receiving growing attention. Phylogenetic tools have been proposed to be useful in predicting lineages for bioprospecting (Ernst et al., 2016; Saslis-Lagoudakis et al., 2012), whole genome sequences of medicinal plants are published with the aim of exploring biosynthetic pathways of active compounds (e.g. Zhao et al., 2017), and DNA barcoding is emerging as a routine tool for quality control of marketed herbal medicines (de Boer et al., 2015; Sgamma et al., 2017).

Medicinal plant use demonstrates the utility of biological diversity. According to the World Health Organisation (WHO), between 70 – 95% of populations in developing countries depend on traditional medicines including herbal medicines (Robinson and Zhang, 2011). The WHO defines herbal medicines as "herbs, herbal materials, herbal preparations and finished herbal products, that contain as active ingredients parts of plants, or other plant materials, or combinations" (WHO, 2002). The use of these medicines is divided into 'Traditional Medicine' (TM) and 'Complementary Medicine' (CM; WHO, 2014). TM is defined as "the sum total of the knowledge, skill, and practices based on the theories, beliefs, and experiences indigenous to different cultures […]". CM is defined as "a broad set of healthcare practices that are not part of that country's own tradition […] and are not fully

integrated into the dominant healthcare system […]". The use of CM as a category reflects the integration and commercialization of traditional and herbal medicine on a global scale. The market for herbal medicines is growing, with an estimated annual global market value of US$ 83 billion in 2008 (Robinson and Zhang, 2011). The raw materials contributing to this market are mainly collected in the wild by local harvesters, and are traded regionally before entering global trade (Mander, 1998; Olsen, 1998).

The growing use of medicinal plant products raises concerns about their safety and efficacy. In response, the European Union (EU) has published several directives addressing these issues (Directive 2001/83/EC, 2001; Directive 2004/83/EC 2004). Standards of herbal medicines are represented in pharmacopoeias (e.g. British Pharmacopoeia, 2016) and are mainly based on anatomical, physical and chemical properties. An integral aspect of quality control is species authentication (European Medicines Agency, 2006), which is now complemented with DNA barcoding techniques (British Pharmacopoeia Commission, 2017).

1.2 Methodological considerations

The rise of the field of genomics offers opportunities for fundamental and applied research. This thesis explores ways of using genomics for the study of evolution, as well as providing tools for using genomics in DNA barcoding of herbal medicines, and identifies areas of further development within the field.

1.2.1 From phylogenetics to phylogenomics

Phylogenetics is the reconstruction of evolutionary relationships between organisms and is fundamental practice for virtually all evolutionary studies (Delsuc et al., 2005). Providing that homologous characters are used, any type of data (e.g.

morphological or molecular) can be used for phylogeny estimation. The field of molecular phylogenetics uses mainly DNA sequences. In recent years, DNA sequencing technology has made dramatic steps forward. The most commonly-used Sanger method, where relatively short, targeted sequences are produced, is being replaced by so-called next-generation sequencing (NGS) technologies, where high-throughput parallel sequencing enables researchers to sequence whole genomes within short periods at relatively low cost. The impact of NGS in non-model organismic biology is immense and revolutionizes fields such as molecular ecology (Ekblom and Galindo, 2011; Tautz et al., 2010) or crop genetics (Varshney et al., 2009). The vast amount of data that can be generated with new sequencing methods is transforming the discipline of phylogenetics into phylogenomics, where genome-scale data is used for the reconstruction of the tree of life (Delsuc et al., 2005). The assembly of whole genome sequences are labour- and resource-intensive and researchers studying non-model organisms usually use a range of techniques to target specific parts of the genome (Cronn et al., 2012). One such technique uses in-solution hybridization capture of specific regions with biotinylated oligonucleotides, where hundreds of nuclear genes can be targeted (Lemmon et al., 2012). Prior to sequencing, the biotinylated oligonucleotides are hybridized to the sequencing library and the hybridized fraction is then sequenced. Phylogenomic inference depends on targeting orthologous, single-copy genes, since comparing paralogous sequences may produce misleading signals (Philippe et al., 2011; Struck, 2014).

1.2.2 DNA barcoding in the era of next-generation sequencing

DNA barcoding refers to the identification of taxa based on short, unique and standardized DNA sequences. The concept of genetic identification of species using

sequence data was first applied to microorganisms, where morphological differentiation of species can be challenging (Nanney, 1982). The practice has subsequently been applied to many different organisms (Eggert et al. 2002; Floyd et al. 2002). The term 'barcoding' was introduced in a paper about the identification of strains/lineages of parasites (Arnot et al., 1993) but did not receive much attention as a new concept from the scientific community. The work of Hebert et al. (2003) later led to a wider appreciation for the potential of barcoding practices. Within their work, the authors promoted the use of the cytochrome oxidase1 (*CO1*) as a taxonomic tool for species identification across the animal kingdom. It is important to note that their work is focused on identification and is not, as suggested by Tautz et al. (2003), a proposition for DNA taxonomy. Hebert et al. (2003) demonstrated that a single region in the mitochondrial genome could serve as a universal sequence to distinguish between animal taxa in a standardized procedure, and began to build up a shared *CO1* gene database. Moreover, they argued that DNA barcoding can aid the delineation of species by applying genetic distance thresholds. The paper provoked mixed responses. One criticism was of the single barcode approach due to low resolution among closely-related species made inclusion of several markers necessary (Mallet and Willmott, 2003). Others supported incorporating DNA barcodes in taxonomic identification (Blaxter, 2003; Janzen, 2004), and several groups confirmed the usefulness of the *CO1* gene as an animal barcode (e.g. Smith et al. 2008; Smith et al. 2006; Ward et al. 2005; Clare et al. 2007). In contrast to the situation for animals, there was no single region for barcoding found in plants and no easy consensus about the set of regions that might be selected for plant identification (Taylor and Harris, 2012).

NGS approaches are not yet commonly used by the DNA barcoding community and concerns about the "continued resistance to improvement" of the DNA barcoding

enterprise has been expressed (Taylor and Harris, 2012). With Sanger sequencing approaches, plant researchers typically use small regions of the plastid genome or the nuclear ITS region for barcoding a species. In contrast, NGS techniques allow sequencing of whole plastid genomes in a single sequencing run. For plant DNA barcoding, whole plastid DNA (cpDNA) and complete ITS sequencing has been proposed to be a valuable source for identification at species- and even population-level (Coissac et al., 2016; Kane et al., 2012). Several plastid genomes have been sequenced using long-range Polymerase Chain Reaction (PCR) with subsequent multiplex sequencing (Cronn et al., 2008; Parks et al., 2009; Whittall et al., 2010). Other studies show the potential of target enrichment strategies, where the whole genomic DNA is reduced to a genomic fraction of interest (for a review, see Cronn et al. 2012). Furthermore, the method of genome skimming (Straub et al., 2012) – sometimes referred to as 'Ultra-Barcoding' (Kane et al., 2012) – is particularly appealing for DNA barcoding, because of the simplicity of the laboratory workflow. Genome skimming is a shallow sequencing approach and takes advantage of the high abundance of plastid DNA in total genomic DNA and the repetitive nature of the ITS region, which ensures enough sequencing depth for the regions of interest. Numerous whole plastid genomes have been sequenced with this approach (e.g. *Theobroma* sp., Kane et al. 2012; *Asclepias* sp., Straub et al. 2011).

1.3 Study organisms

This thesis is primarily focused on the evolution and DNA barcoding of genus *Berberis* from the family Berberidaceae (Chapters 2, 3 and 4). Data from genus *Phyllanthus* (Phyllanthaceae) are used for investigating commercial, internationally

traded samples (Chapter 4). Data from genus *Arabidopsis* (Brassicaceae) are used for investigating the effect of paralogy on target enrichment studies (Chapter 5).

1.3.1 *Berberis* L.

Genus *Berberis* from the family of Berberidaceae contains more than 600 species (incl. *Mahonia* Nutt.; Mabberley, 2008). There have been conflicting views on the delineation of *Berberis* and *Mahonia* and they are commonly now treated as one genus (Mabberley, 2008; Marroquin and Laferriere, 1997). Here, *Berberis sensu lato (s.l.)* is referred to genus *Berberis* including *Mahonia*, and *Berberis sensu stricto* (*s.s.*) is used for simple–leaved *Berberis* (in the sense of Ahrendt, 1961). *Berberis s.s.* is divided into two groups: Septentrionales, with ca. 300 species, is distributed in Eurasia, and group Australes, with ca. 169 species, in South America (Ahrendt, 1961). The taxonomy of *Berberis* is still changing, with several instances where taxa recognized by Ahrendt (1961) were combined to single species (Adhikari et al., 2012; Landrum, 1999) and where new species are described (Adhikari et al., 2012; Harber, 2017a, 2017b). Most of the species of the genus are diploid (Rounsaville and Ranney, 2010). The antitropical disjunction of *Berberis s.s.* has drawn considerable attention from biogeographers and the debate of how this pattern arose is ongoing (Adhikari et al., 2015; Li et al., 2010).

Several species of *Berberis* are used in traditional medicine (e.g. Manandhar 2002), among which *B. aristata* DC. seems most important. *B. aristata* is a diploid species and is widely distributed in the Himalayas at elevations between 1,300 to 3,400 m. The species is included in the British Pharmacopoeia (2016) and the the Ayurvedic Pharmacopoeia of India (2001). *Berberis* species produce the benzylisoquinoline alkaloid Berberine, which, in modern medicine, has drawn

considerable attention for its cholesterol-lowering properties (Kong et al., 2004) and its potential efficacy as a hypoglycemic agent for patients with type 2 diabetes mellitus (Yin et al., 2008).

1.3.2 *Phyllanthus* L.

Genus *Phyllanthus* L. (Phyllanthaceae) has a pantropical distribution (Mabberley, 2008). The main species in focus here is *P. amarus* Schumach. & Thonn, which is likely to be native to the tropical Americas but exhibits a pantropic distribution (Mabberley, 2008). The plant is traditionally used in many tropical and subtropical regions of the world (Patel et al., 2011). It is also used in Ayurvedic practice (Ayurvedic Pharmacopoeia of India, 2001), where it is considered a cure for problems relating to the stomach, genitourinary system, liver, kidney and spleen (Patel et al., 2011). The plant raised interest within modern biomedicine because of its potential to treat Hepatitis B patients (Blumberg et al., 1989) and was biochemically thoroughly investigated (Patel et al., 2011).

1.3.3 Genomic resources for *Arabidopsis* (DC.) Heynh.

Arabidopsis thaliana (L.) Heynh. (Brassicaceae) is a well-studied model organism in plant biology and was the first plant from which the complete genome was sequenced (Kaul et al., 2000). This landmark publication was followed with a series of large-scale projects intending to understand gene functions (Bevan and Walsh, 2005) and the evolution of this species (Long et al., 2013). Evolutionary studies were soon extended to genus *Arabidopsis* (Novikova et al., 2016) and vast amounts of raw sequencing data is available in public databases such as the Short Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra; last accessed 14/08/2017). These data, in conjunction

with such detailed knowledge about gene functions or gene clusters, provide excellent opportunities for testing analysis methods developed for non-model organisms.

1.4 Thesis organisation

This thesis is organized into four data chapters, all demonstrating the use of genome-scale sequence data. The main findings are summarized in Chapter 6, where emerging fields for phylogenomics and DNA barcoding are also discussed.

In Chapter 1, I describe the evolution of genus *Berberis* in the Himalayas and the Hengduan Mountains. The study is based on phylogenies inferred from target-enrichment of hundreds of nuclear genes and whole plastid genomes, and gives an unprecedented view on the evolution of the genus in these mountain systems.

In Chapter 2, I investigate new barcoding approaches for the Himalayan species *B. aristata* and closely related taxa using whole plastid genomes. The study focuses on providing suitable DNA barcodes for regulatory purposes.

In Chapter 3, genomic identification techniques are applied to commercial *Berberis* and *Phyllanthus* samples in global trade. This study aims to give insights into the diversity of traded species and further explores the structure of global herbal medicines trade. Results from *Phyllanthus* samples were published in the article “DNA Barcoding for Industrial Quality Assurance” (Sgamma et al., 2017), where I authored the next-generation sequencing section. The results in this chapter are slightly modified from the publication and discussed in a different context.

Chapter 4 describes an *in silico* target enrichment experiment on genus *Arabidopsis.* I address the potential impact of capturing reads from paralogous copies on phylogenomic inference. This study indicates how our understanding of comparative genome evolution intersects with pipelines for handling next-generation sequence data,

and highlights how some of the methodologies used in Chapter 1 and 3 are at an early stage of development. This chapter was produced in collaboration with Vincent Manzanilla (University of Oslo), who performed the raw read processing (quality filtering) and the read mapping.

Chapter 2 A phylogenetic hypothesis for *Berberis* (Berberidaceae) in the Himalayas and the Hengduan Mountains

2.1 Introduction

The phylogeny and biogeography of Berberidaceae have drawn considerable attention from botanists and several studies have investigated phylogenetic patterns within the family (Adhikari et al., 2015; Kim et al., 2004; Wang et al., 2007). The family Berberidaceae comprises 14 genera , mainly distributed in the Northern Hemisphere, with only the genus *Berberis* extending to the Southern Hemisphere in South America (Mabberley, 2008). *Berberis s.l.* contains more than 600 species (Mabberley 2008), including the simple-leaved *Berberis s.s.* and the compound-leaved species formerly included in *Mahonia* Nutt. For many years, authors disagreed about whether to consider *Berberis* and *Mahonia* as one genus or two, but most now support the transfer of *Mahonia* species to *Berberis* (Mabberley, 2008; Marroquin and Laferriere, 1997; *Berberis* including *Mahonia* is henceforth referred to as *Berberis s.l.*). Two groups of compound-leaved species are recognized, Occidentales that grow in North and Central America and Orientales from China and the Himalayas. *Berberis s.s.* has two major centers of diversity: the ca. 169 species placed in Australes are distributed in South America and the ca. 300 species placed in Septentrionales are distributed in Eurasia (Ahrendt, 1961). The actual number of species is likely to change, since recent revisions have synonymized several described taxa (Adhikari et al., 2012; Landrum, 1999) and new species are described (Adhikari et al., 2012; Harber, 2017a, 2017b). Simple-leaved *Berberis* have an antitropical distribution, and the debate over how this pattern emerged is ongoing (Adhikari et al., 2015; Li et al., 2010).

This study focuses on the evolution of *Berberis* in a mountain system, specifically the Himalayan and the Hengduan Mountains which form the southern and eastern border regions of the Qinghai-Tibetan Plateau (QTP). Both harbour a spectacular biodiversity and are listed among the biodiversity hotspots in the Northern Hemisphere (Myers et al., 2000). The uplifts of the QTP and the Himalayas resulted from the collision of the Indian and Eurasian continental plates. The elevation history of the Himalayas is still uncertain (Miehe and Weidinger, 2015; Mulch and Chamberlain, 2006), but the main uplift of the Himalayas is thought to have occurred 21-13 Myr ago (Searle, 2011). Available data on the elevation history of the QTP suggests that 40 Myr ago, the plateau was already at an elevation of 4,000 m (Royden et al., 2008; Wang et al., 2008). The Hengduan mountains are considerably younger than the Himalayas and the QTP, with major uplifts in the late Miocene and late Pliocene (Favre et al., 2015; Wang et al., 2012). *Berberis* species are found in the Himalayas and the younger Hengduan Mountains in montane habitats at elevations between 1,000 m (*B. asiatica*) to as high as 4,700 m (*B. tsarica*, Adhikari et al., 2012). The distribution within two mountain systems of different age raises the question of how this distribution pattern was formed. The distribution could either arise by frequent dispersal events between the two mountain systems or by infrequent colonization in conjunction with *in situ* diversification.

Until recently, the poor resolution at shallow phylogenetic levels has precluded asking precise questions about the evolution of genus *Berberis* in the Himalayan and Hengduan Mountains. However, the development of methods for generating large amounts of DNA sequences via high-throughput sequencing technologies is revolutionizing molecular phylogenetics. The inference of evolutionary relationships among organisms from genome-scale data has given rise to the field of phylogenomics

(Delsuc et al., 2005; Eisen and Fraser, 2003). Phylogenomics applies the well-established principles of phylogenetics, using homologous characters to reconstruct evolutionary relationships among organisms, but using genome-scale data. For model organisms, whole genome sequences are generally available (e.g. *Arabidopsis thaliana*; Arabidopsis Genome Initiative, 2000). However, phylogeneticists studying non-model organisms focus on subsets of genomic regions for phylogenetic inference by enriching specific regions of the genome (Cronn et al., 2012). Two main strategies for target enrichment have emerged in recent years (for a comparison, see Harvey et al., 2016). The first strategy encompasses enrichment of anonymous sequences in the genome where no prior knowledge of the DNA sequence is necessary. These methods usually use enzyme-based genomic DNA restriction for selecting appropriate DNA fragments, such as the restriction site associated DNA (RAD) tags (Baird et al., 2008). The second category uses the polymerase chain reactions (PCR) or hybridization capture to enrich known regions (Prum et al., 2015). The latter uses hybridization probes to separate the genomic sequences of interest. One hybridization enrichment approach is to use probes designed to target hundreds of genetic loci, which are then sequenced (e.g. McCormack et al., 2013; Weitemier et al., 2014). Genome skimming, the shallow sequencing of a shotgun library, effectively selects part of the genome, delivering sufficient read coverage for sequence reconstruction of multi-copy genes and multi-copy genomes such as the plastid genome (Straub et al., 2012). Plant phylogenomic studies usually include whole plastid sequences (Parks et al., 2012), a set of nuclear markers (De Sousa et al., 2014) or a combination of both (Folk et al., 2016; Weitemier et al., 2014).

The premise of including hundreds of low-copy genomic sequences and fully sequenced organellar genomes, rather than a few gene sequences, is to increase the number of informative characters for phylogenetic inference. Methodologically, two

types of analysis have emerged to handle massive multi-gene datasets. Concatenation of regions into large, total-evidence alignments prior to analysis of the combined data usually leads to a single, well-supported phylogeny (e.g. Rokas et al., 2003). However, several studies have shown that gene trees may differ substantially from so-called total-evidence trees, and so the second type of analysis summarizes evidence from multiple gene-trees (Kubatko and Degnan, 2007; Salichos and Rokas, 2013). As the number of phylogenomic studies of plants in the literature has increased, so has awareness of the impact of complex evolutionary processes such as incomplete lineage sorting (ILS) or chloroplast capture on phylogenomic datasets (Folk et al., 2016; Liu et al., 2015; Salichos and Rokas, 2013).

This chapter describes the generation of a phylogenetic hypothesis for *Berberis* in the Himalayas and Hengduan Mountains using several phylogenetic inference techniques for hybridization-captured nuclear loci and plastid genomes. The methods of data collection for phylogenomic analyses are emphasized, as are the analysis pipeline and the investigation of phylogenetic discord between genomes and between nuclear partitions.

Both the datasets and phylogenetic hypotheses find application in the regulation and authentication of medicinal plants (Chapters 4 and 5). However, the phylogenetic hypotheses generated for this study also have great potential to address questions about the origins of the montane flora found in the mountain ranges adjoining the QTP. Although finalising robust time-calibrated phylogenetic analyses and identifying shifts in diversification rate is beyond the scope of this thesis work, preliminary ancestral distribution analyses are performed, and the phylogenetic hypotheses for Himalayan/Hengduan Mountains *Berberis* are discussed in the context of their possible contribution to the emerging view of montane diversification in this area.

2.2 Material and methods

2.2.1 Sampling

Five silica-dried leaf samples from *B. aristata* (n=4) and *B. asiatica* (n=1) were used for hybridization probe design (Table 2-1). For the phylogeny, silica-dried leaf material from 85 samples, representing 53 species were included in this study (Appendix Table AT-1). One sample was extracted twice and was used as a technical replicate (*B. petiolaris1* and *B. petiolaris2*). This study focuses on *Berberis* species from the Himalayas and the Hengduan Mountains, which belong to a previously identified clade within the group Septentrionales (Adhikari et al., 2015). A total of 73 samples representing 44 species were included from this clade. This corresponds to about 14 percent of the known species from the group Septentrionales. In addition, nine samples representing eight species from the group Australes and three compound-leaved *Berberis* samples were included as outgroups. We used up-to-date taxonomic treatments for identification of Himalayan *Berberis* species (Adhikari et al., 2012). However, the identification of specimens is often difficult when only vegetative characters are available and some of the specimens could not be identified to species level. A monograph of *Berberis* species from China is in the process of completion (Harber, pers. communication) and, therefore, the new species (*B. new_sspA*, *B. new_sspB*) in the phylogeny are not formally described and published yet.

2.2.2 Laboratory work

2.2.2.1 DNA extraction

DNA was extracted using either the Qiagen DNeasy Plant Kit following the manufacturer's protocol or the CTAB method (Doyle and Doyle, 1987). The quality of the extractions was checked for the degree of degradation on 1% or 1.5% agarose gels.

Furthermore, we performed PCR amplifications of the *rbcL* gene in different dilutions (1:1, 1:10 and 1:100) and finally we measured the DNA concentration on a Qubit® Fluorometer (Life Technologies, Carlsbad, CA, USA), using the dsDNA High Sensitivity kit. The concentrations after extraction ranged from 1.5 ng/µl to 34.8 ng/µl.

2.2.2.2 Library preparation and Sequencing

For DNA marker development, shotgun sequencing libraries were prepared for six samples (Table 2-1). We used the Nextera XT kit according to the manufacturer's guidelines. The average fragment length of the libraries was between 500 – 700 bp. The samples were sequenced on an Illumina MiSeq® with a MiSeq v2 reagent kit with the paired-end option and 500 cycles (resulting in 250 bp paired-end sequences). The six samples comprised 95% of the final pooled library. These samples were used for marker development (see below).

Table 2-1 Summary of *Berberis* samples used for shotgun sequencing. Voucher specimens are deposited at the Royal Botanic Garden Edinburgh (RBGE).

Sample	**Species**	**Voucher (RBGE)**
B. marker1	*B. aristata*	EA243
B. marker2	*B. aristata*	EA249
B. marker3	*B. aristata*	WP21.1
B. marker4	*B. aristata*	WP21.5
B. marker5	*B. aristata*	EA109

The library preparation for the target-enrichment and shotgun sequencing was performed according to Meyer and Kircher (2010). The libraries were sequenced in two runs on a MiSeq® (run 1) and a NextSeq® (run2). Depending on their integrity, the DNA samples were shared mechanically to a fragment size of approximately 400 bp using a Covaris © sonicator with peak incident power of 75; duty factor of 10%, and

200 cycles per burst. The duration of treatment was chosen according to the observed fragment size on agarose gels and ranged between 30s (medium degradation) and 40s (genomic DNA).

We followed the protocol for blunt-end repair, adapter ligation and adapter fill-in. After each of these steps, the DNA was cleaned-up with AMPure® XP beads (Agencourt®). Before the indexing PCR, the DNA quantity was measured on a Qubit ©. Depending on the concentration of adapter-ligated libraries, we aimed to use between 50 – 100 ng of DNA as input for the indexing PCR where possible. Higher concentrations may impair the PCR reaction. In order to avoid high duplication levels in target-enriched libraries, a minimal number of PCR cycles were applied. Libraries with concentrations lower than 40 ng were amplified with 16 PCR cycles. If more than 40 ng of library was used for the PCR, 12 cycles were applied. We used the index sequences ("barcodes") as suggested by the protocol. The final libraries were washed using AMPure® XP beads (Agencourt®). We then measured for concentration with Qubit © and assessed the fragment size using Bioanalyzer® (Agilent). Libraries with similar concentration levels were then pooled for target enrichment in equimolar concentrations to a total of eight pools. The number of samples per pool varied from 8 to 22, depending on the library concentration of samples after indexing PCR (Appendix Table AT-2). Generally, samples with higher concentrations were pooled with more other samples. Several samples that were captured are not described in this chapter. In-solution hybrid capture was conducted following the MYbaits v. 3.02 protocol, where 7 µl of pooled libraries is the starting point. The total amount of DNA per pool used for the capture varied between 147 ng to 400 ng. The incubation time was 30 hours. After the cleanup of the captured library, we applied 14 cycles of reamplification using the reamplification primers suggested by Meyer and Kircher (2010). A large part of the

libraries was used to achieve the necessary concentration levels for the target enrichment. However, libraries of 64 samples contained enough DNA for shotgun sequencing / genome skimming. These libraries were diluted to 10 mM and pooled together..

The libraries were sequenced in two runs on a MiSeq® (run 1) and a NextSeq® (run2). Target-enriched libraries of six samples were sequenced on an Illumina MiSeq in run 1. Shotgun libraries of 5 of these samples were sequenced in run2 on the Illumina NextSeq run. The target enrichment libraries of the remaining 79 samples and 63 shotgun libraries were sequenced in run 2. In total, 85 target enrichment libraries were sequenced of which 63 shotgun libraries were sequenced in parallel (Appendix Table AT-2).

2.2.3 Bioinformatics

2.2.3.1 Baits design

This section describes how the reference markers and the corresponding hybridization probes ("baits") for in-solution target enrichment were designed. *De novo* assemblies of five samples (Table 2-1) and the transcriptome of *Nandina domestica* (scaffold-YHFG-2011734-Nandina_domestica, Wong 2013, www.onekp.com) were used for developing DNA markers. Raw reads from the shotgun sequencing were trimmed using Trimmomatic v.0.33 (Bolger et al., 2014) with the options LEADING:3, TRAILING:3, SLIDINGWINDOW:4:20. This step ensures that only high quality reads are used in down-stream analyses. Reads shorter than 50 bp were discarded. The read quality was checked with FastQC (Andrews, 2010). Reads that map to organellar genomes were removed. Initially, all reads were mapped to an *Arabidopsis thaliana* (L.) Heynh. mitochondrium reference (GenBank accession: NC_001284.2) with Burrows-Wheeler Alignment tool (BWA; Li and Durbin, 2009). The reference sequence was

indexed with *'bwa index'* and mapped with the command *'bwa mem'* with default options. The sequence alignment in SAM format was transformed to its binary version BAM with SAMtools (Li et al., 2009). This format stores information on read mapping, such as which reads mapped to the reference. The unmapped reads were then extracted with the '*bam2fastq'* tool from BEDtools (Quinlan and Hall, 2010), resulting in a file where the mitochondrial reads are discarded. The mapping and filtering process was repeated for filtering plastid reads against the reference plastid genome of *Berberis bealei* Fortune (GenBank accession: NC_022457.1) and ribosomal reads against the ITS sequence from an *Arabidopsis thaliana* accession (GenBank accession LC089989.1). The final set of reads only contained nuclear sequences that were used for a *de novo* assembly using SOAPdenovo2 (Luo et al., 2012). Before running the assembly, the optimal k-mer size was estimated with kmgergenie (Chikhi and Medvedev, 2014). The *de novo* assembly was run with the 123mer version of SOAPdenovo with the options 'pair_num_cutoff=30', where an overlap of at least 30 bp is needed for making connections between two contigs or pre-scaffolds; 'avg_ins=600', which sets the estimated average fragment length of libraries to 600 bp; and 'asm_flags=3', which sets to run a contig and a scaffold assembly. The quality of the assemblies was checked with QUAST (Gurevich et al., 2013).

The selection of markers by comparing transcriptome data and *de novo* assembled contigs followed a script written by Vincent Manzanilla (University of Oslo, unpublished). In summary, contigs from the *de novo* assembly that were shorter than 400 bp were removed with the python script python_cleaner.py (http://biopython.org/wiki/Sequence_Cleaner; last accessed 16/08/2017). The contigs from the *de novo* assembly were then clustered using the program cd-hit (Li and Godzik, 2006) and contigs that shared a sequence similarity >80% were removed. This

prevents that the baits target genetically similar regions in the genome. The same two steps were applied to the transcriptome sequences, where sequences shorter that 119 bp were discarded. The transcriptome sequences were then mapped against the *de novo* contigs using BLAT (Kent, 2002) and single hits were extracted. The extracted transcriptome sequences were mapped against the *de novo* contigs with BWA (Li and Durbin, 2009) with default options and *de novo* contigs that exhibited a coverage > 1 were removed. This step prohibits contigs with duplicated copies in the genome from being used for marker design. Only *de novo* contigs that were longer than 400 bp were used as markers and, if applicable, were trimmed to 980 bp. The resulting *de novo* contigs were used as reference markers comprising of 607 sequences with lengths between 400 to 980 bp. The selected DNA markers were used to produce MYbaits® bait probes (MYcroarray©; Ann Arbor, Michigan, USA), which are RNA sequences with a length of 120 bp each. The baits were designed to cover each marker four times (4x tiling), which resulted in a total of 13,248 unique baits.

2.2.3.2 Raw read processing and quality control

Samples were sequenced on an Illumina MiSeq or NextSeq sequencer. Adapters were removed either with the built-in Illumina software on sequencers or using cutadapt v. 1.10 (Martin, 2011). Raw reads were trimmed using Trimmomatic v.0.33 (Bolger et al., 2014) with the options LEADING:3, TRAILING:3, SLIDINGWINDOW:4:20. Reads from Illumina NextSeq were discarded when shorter than 30 bp and from MiSeq when shorter than 50 bp. The read quality was checked with FastQC (Andrews, 2010).

2.2.3.3 Nuclear DNA marker assembly

The reference DNA marker file was indexed with the command ‘bwa index’ in BWA and paired-end reads from each sample were mapped to the reference with ‘bwa mem’ with default options. The average read coverage was calculated with SAMtools

('samtools depth'). The resulting BAM files were sorted and indexed with SAMtools ('samtools sort', 'samtools index'). In order to extract two alleles per sample for each marker, the command 'samtools phase' was applied to the sorted and indexed BAM files. The algorithm extracts two alleles per sequence alignment (He et al., 2010), resulting in two BAM files (allele0.bam, allele1.bam). Single-nucleotide polymorphisms on these alleles were called using 'samtools mpileup' and 'bcftools call'. The final sequence of alleles was called with the command 'vcfutils.pl vcf2fq' from VCFtools (Danecek et al., 2011). The sequences in fastq format were transformed to fasta with seqtk (https://github.com/lh3/seqtk). The final sequence was generated by calling the consensus of the allele sequences. The fraction of recovered sequence compared to the length of the reference sequence was calculated as

$$f = \frac{L-N}{L},$$

where L is the length of the consensus sequence per locus and sample and N is the number of missing data in the consensus sequence.

2.2.3.4 Filtering of nuclear DNA markers

The reference DNA markers were designed using whole genome draft assemblies of *B. aristata* and *B. asiatica*. However, a draft genome is partial, and it is possible that at least some targeted loci may have paralogues in the reference genome, or that gene duplication events in species other than *B. aristata* may have occurred. During capture, the targeted loci may therefore be contaminated with reads from paralogous copies. We addressed this issue by developing a pipeline for filtering loci that are potentially contaminated with reads that derive from paralogous copies (Chapter 2). Through phasing read alignments with SAMtools, two putative allelic sequences per locus were extracted. We used two approaches for identifying outlier loci by analyzing the putative allelic copies of a locus. The first approach depended on

calculating the sequence similarity of each pair of alleles. The reasoning behind this step was that reads from different paralogous copies will be represented in each of the phased alleles. The assumption is that higher divergence between a pair of alleles is indicative of contaminant reads from different paralogous copies. Average sequence divergence and standard deviation for allele pairs across all loci was calculated. In the second approach, gene trees for each locus were built from alignments containing all allelic copies. Maximum likelihood (ML) gene trees were inferred using RAxML v. 8.2.9 with 100 rapid bootstrap replicates, resulting in 607 gene trees each containing 170 alleles from 85 samples. For each gene tree, we calculated the pairwise distance between pairs of alleles with the cophenetic.phylo function in the R package ape (Paradis et al., 2004), which uses branch lengths to calculate pairwise distances. The assumption is that distance on a phylogenetic tree between true allelic copies is smaller than the distance between alleles that represent paralogous copies. All distances between pairs of alleles from each gene tree were averaged and the standard deviation calculated. With these methods, we retrieved for each marker the average and the standard deviation for allelic sequence divergence and allelic phylogenetic distance. A threshold was applied for both measures and loci that did not meet the criteria were discarded.

2.2.3.5 Nuclear marker phylogeny

The first approach was to infer species phylogenies based on concatenation of gene alignments. Aligned DNA markers were concatenated using phyutility v.2.2.6 (Smith and Dunn, 2008), resulting in a data matrix of 303,754 bp length. The data matrix was analysed with RAxML v. 8.2.10 (Stamatakis, 2014) with 1,000 fast bootstrap replicates (option '-f a'). The concatenated alignment was partitioned where each of the 396 individual loci represents an independent partition. The best fitting

model of substitution was inferred with jModeltest2 and was GTR+G for 70% of loci (versus 26% GTR+I+G and 4% GTR). The model of substitution in RAxML was therefore set to GTRGAMMA for all partitions. Members of the compound-leaved *Berberis* were set as outgroup (*B. nervosa*, *B. polyodonta* and *B. nevinii*). Clades with bootstrap support lower than 50 were collapsed to polytomies.

For Bayesian phylogenetic inference, the data matrix was partitioned by locus and the best-fitting model of substitution assessed with PartitionFinder 2.1.1 (Lanfear et al., 2016). The data matrix was analyzed in Bayes Phylogenies (Pagel and Meade, 2006) with 15 independent chains where each was run for a minimum of 45 million generations. The burn-in was set to 10 million. Convergence of the chains was checked in Tracer v. 1.6 (Rambaut et al., 2014,). In order to avoid autocorrelation, only a fraction of the sampled trees were used for further analysis, resulting in 984 trees. The MCMC samples from the posterior distribution were summarized to a consensus tree with minimal clade frequency of 95% using SumTrees (Sukumaran and Holder, 2015). A Bayesian consensus network was calculated with the R package phangorn (Schliep, 2011).

Recent studies have shown that concatenation of genes may produce misleading results and researchers therefore use alternative approaches using gene trees. The incorporation of numerous genes for estimating phylogenies has found considerable discordance across gene trees which is often accounted to incomplete lineage sorting (ILS, e.g. Degnan and Rosenberg, 2009; Kubatko and Degnan, 2007). In order to account for topological variation in gene trees and to compare the results to the concatenation approach, we applied the multi-species coalescence (MSC) method in ASTRAL-II (Mirarab and Warnow, 2015). The algorithm implemented in ASTRAL provies a statistically consistent estimate of the species tree, calculated from unrooted

gene trees under the multi-species coalescent model. The algorithm finds the species tree that agrees with the largest number of quartet partitions in the unrooted gene trees (Mirarab et al., 2014; Mirarab and Warnow, 2015). Unrooted gene trees were inferred with RAxML, using 100 rapid bootstrap replicates (option '-f a'). The best trees of each gene and the corresponding bootstrap trees were used as input in ASTRAL and species trees were estimated with 100 bootstrap replicates.

2.2.3.6 Plastid assembly and alignment

The quality filtered paired-end reads were mapped to a reference genome of *B. aristata* (Kreuzer et al., unpublished) with Burrows-Wheeler Alignment tool (BWA, ver. 0.7.12, Li and Durbin, 2009). The reference genome was indexed using option 'bwa index'. Read pairs that survived the quality check were mapped with default options of the command 'bwa mem'. The resulting SAM file was converted to BAM format with 'samtools view' and sorted with 'samtools sort' in SAMtools v. 1.2. The average coverage was calculated with 'samtools stats'. Optical read duplicates were removed with Picard tools (http://broadinstitute.github.io/picard; last accessed 30/06/17). We used the SNP calling workflow in GATK (McKenna et al., 2010; Van der Auwera et al., 2013). Regions that contain insertions and deletions are often badly aligned. Therefore, a local realignment process was applied with the command '–T IndelRealigner' in GATK. Variant calling was performed on the realigned BAM files with the '–T HaploTypeCaller' module with haploid settings ('-ploidy 1'). The output is a "genomic VCF" file (GVCF) that contains base call information for all sites of the markers. The variant calls were then exported with '–T GenotypeGVCFs' to the standard variant call format (VCF). SNP and indel variants were then filtered separately. The first SNP filter applied is quality by depth (QD), which can be considered as the quality of the variant call standardized by the depth of coverage. QD

avoids inflation of the Phred quality score for the variant call caused by deep coverage. Variants that had a QD < 2 were filtered out as recommended by Van der Auwera et al. (2013). The FisherStrand (FS) quality filter is a Phred-scaled probability that strand bias exists at a specific site. Specifically, the score is a measure for whether an alternate allele was seen more or less often on either forward or reverse reads. The mapping quality (MQ) in GATK is calculated as the root mean square quality over all reads at a given site. Variants with an MQ score <M 40 were removed from the dataset. The final sequence was reconstructed with the command '–T FastaAlternateReferenceMaker' in GATK. We checked our pipeline by visual comparison of the final plastid sequence with the BAM file for selected samples.

The reconstructed plastid genomes were then aligned using MAFFT v7.215 with default options. The inverted repeats were removed from the alignment. SNP calling on inverted repeat regions is not straight-forward since reads with polymorphisms in only one region will map to the other repeat as well. Random mapping to inverted repeat regions often results in apparently heterozygous read alignments, precluding unique assignments of SNPs to a specific inverted repeat. The alignment was checked manually and badly-aligned regions were removed.

2.2.3.7 Plastid phylogeny

The best model of substitution was calculated under the Aikaike Information Criterion in jModeltest2. The ML phylogeny was estimated with 1,000 bootstrap replicates under the GTRGAMMA + I substitution model in RAxML using the online CIPRES portal. The whole alignment was considered as a single partition. Members of the compound-leaved *Berberis* were set as outgroup (*B. nervosa*, *B. polyodonta* and *B. nevinii*). In order to calculate how many nuclear gene trees are in agreement with the plastid phylogeny, we used ASTRAL-II to produce branch support values (Mirarab and

Warnow, 2015). The support value shows how many of the quartet trees in the gene trees support the quartet tree in the species tree.

2.2.3.8 Ancestral range estimation

We used a concatenated ML tree, pruned to include one exemplar per species, to infer ancestral areas with the R package BioGeoBEARS. The package implements the models Dispersal-Extintion-Cladogenesis (DEC; Ree et al., 2005); DIVALIKE, a modified version of DIVA (Ronquist, 1997); and BAYAREALIKE from BayArea (Landis et al., 2013). The program allows for estimating ancestral areas with an extra free parameter *j*, which considers founder-event speciation (Matzke, 2014). The areas were coded to SA = South America, NA = North America, HE = Hengudan Mountains and HI = Himalayas. The data were run under all three models considering only dispersal and extinction (*d* and *e* parameters) and in a second calculation, the parameter *j* was estimated. The likelihood scores were compared using the Aikaike Information criterion (AIC) and data interpreted under the model with the highest AIC value.

2.3 Results

2.3.1 Nuclear DNA marker assembly

The average depth of coverage of the 86 samples is shown in Figure 2-1. The minimum number of loci per sample, where at least part of the sequence could be reconstructed, is 602 (*B. microphylla2*). For 66 out of 85 samples, all 607 genes could be at least partly reconstructed. The breadth of coverage of each locus for the 85 samples is displayed in the heatmap as the fraction of loci recovered (Figure 2-2).

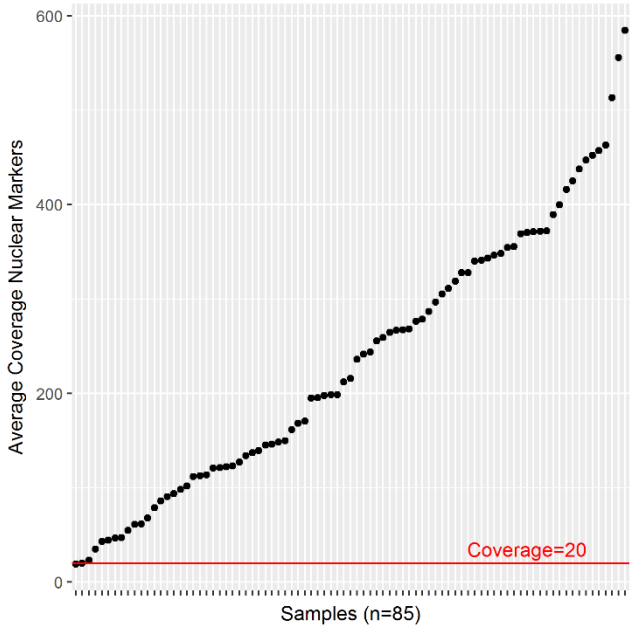


Figure 2-1 Average coverage per sample across all loci.

2.3.2 Filtering of nuclear DNA markers

The allelic divergences and pairwise distances are shown in Figure 2-3. After inspection of the plots, a threshold for sequence similarity (mean <= 94, standard deviation <= 6) and phylogenetic distance (mean <= 0.0625, standard deviation <=

0.04) was determined. From the initial 607 DNA markers, 210 were discarded (34.6%), resulting in 396 DNA markers for further analysis.

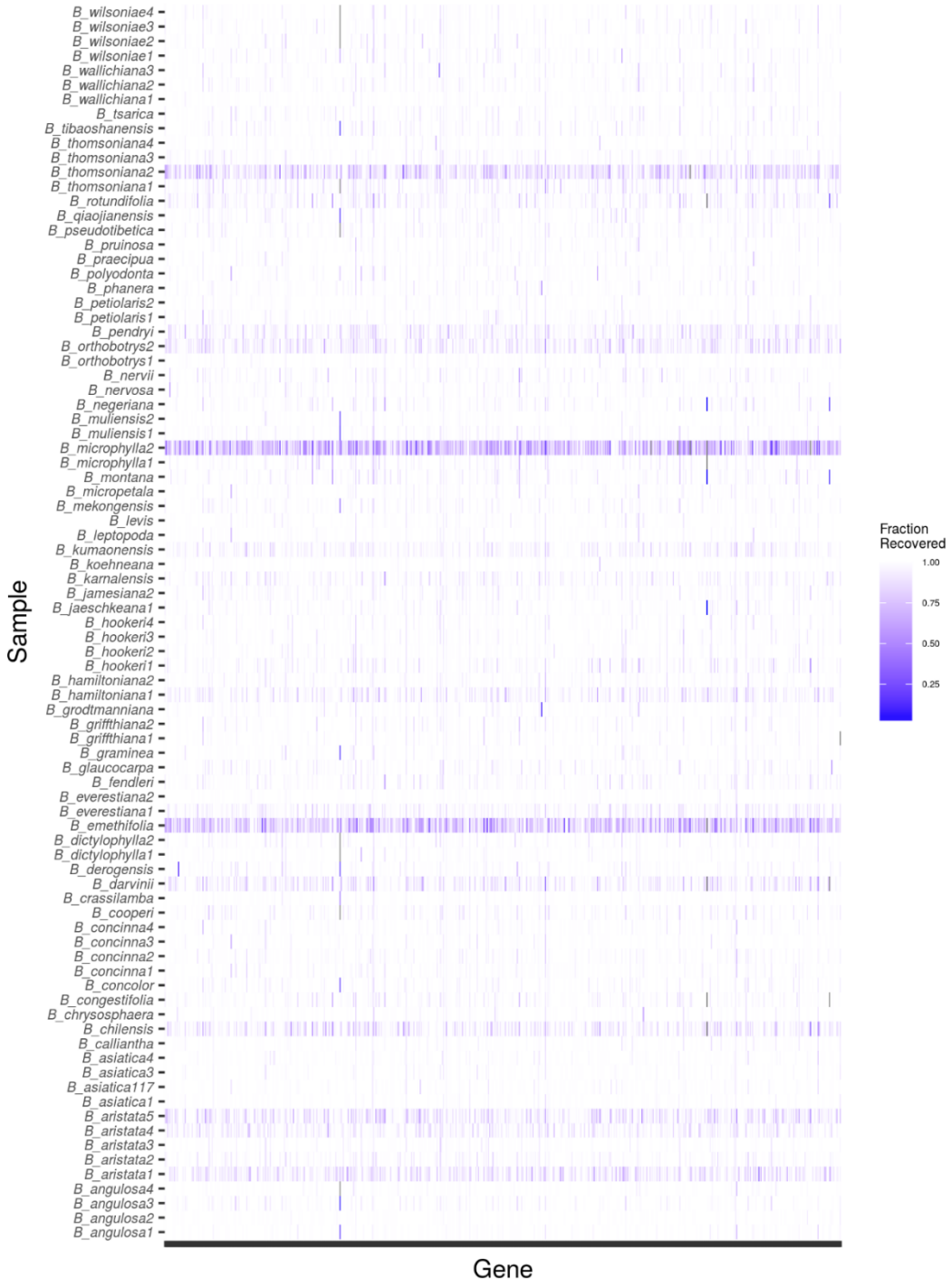


Figure 2-2 The heatmap shows the fraction recovered of each loci (n=607) for samples that were included in the phylogenetic analysis. Grey bars indicate loci where no sequence could be retrieved.

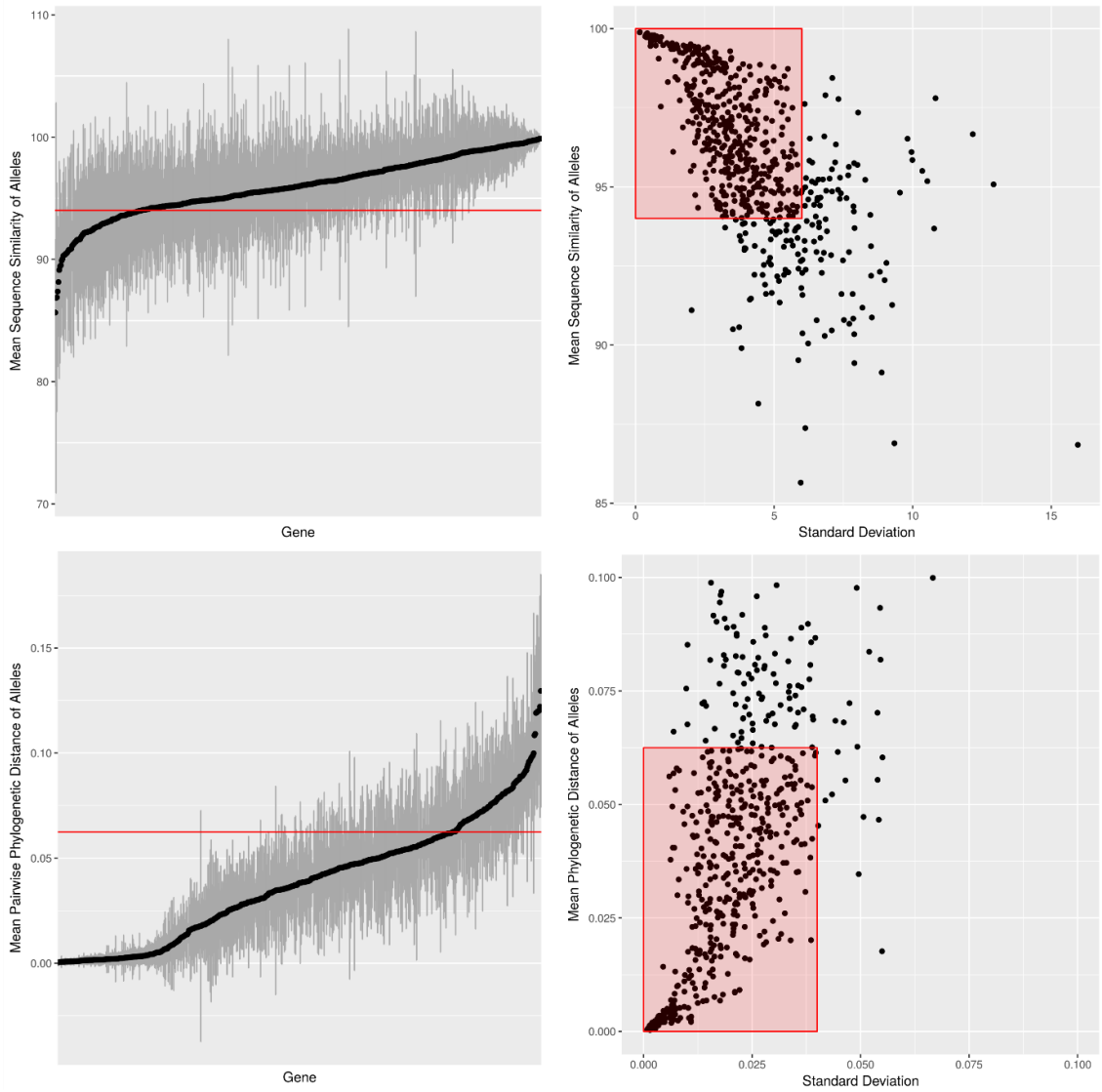


Figure 2-3 *Top left*: The plot shows the average sequence similarity and standard deviation (grey bars) per gene. The red line is the set arbitrary threshold. *Top right*: Plot of sequence similarity and standard deviation per locus (black dot). Loci outside the red rectangle were discarded for further analysis (mean <= 94, sd <= 6). *Bottom left*: Average pairwise phylogenetic distance per loci with standard deviation (grey bars). Genes that exceeded the threshold of 0.625 were discarded. *Bottom right*: Plot of the mean pairwise phylogenetic distance and standard deviation of each loci (black dots). Loci outside the red rectangle were discarded (average > 0.625, sd > 0.4).

2.3.3 Plastid assembly and alignment

The average coverage of the mapping is shown in Figure 2-4 and ranged from 16 to 664. The SNP filtering step removed 60 polymorphisms (54 SNPs and 6 Indels),

leaving 29,785 polymorphisms (22,123 SNPs and 7602 indels) across 64 samples (Figure 2-4). After removing the inverted repeats and badly-aligned regions, the alignment of the plastid genomes resulted in a data matrix of 93,697 bp length and contained a total of 1,229 parsimony informative sites.

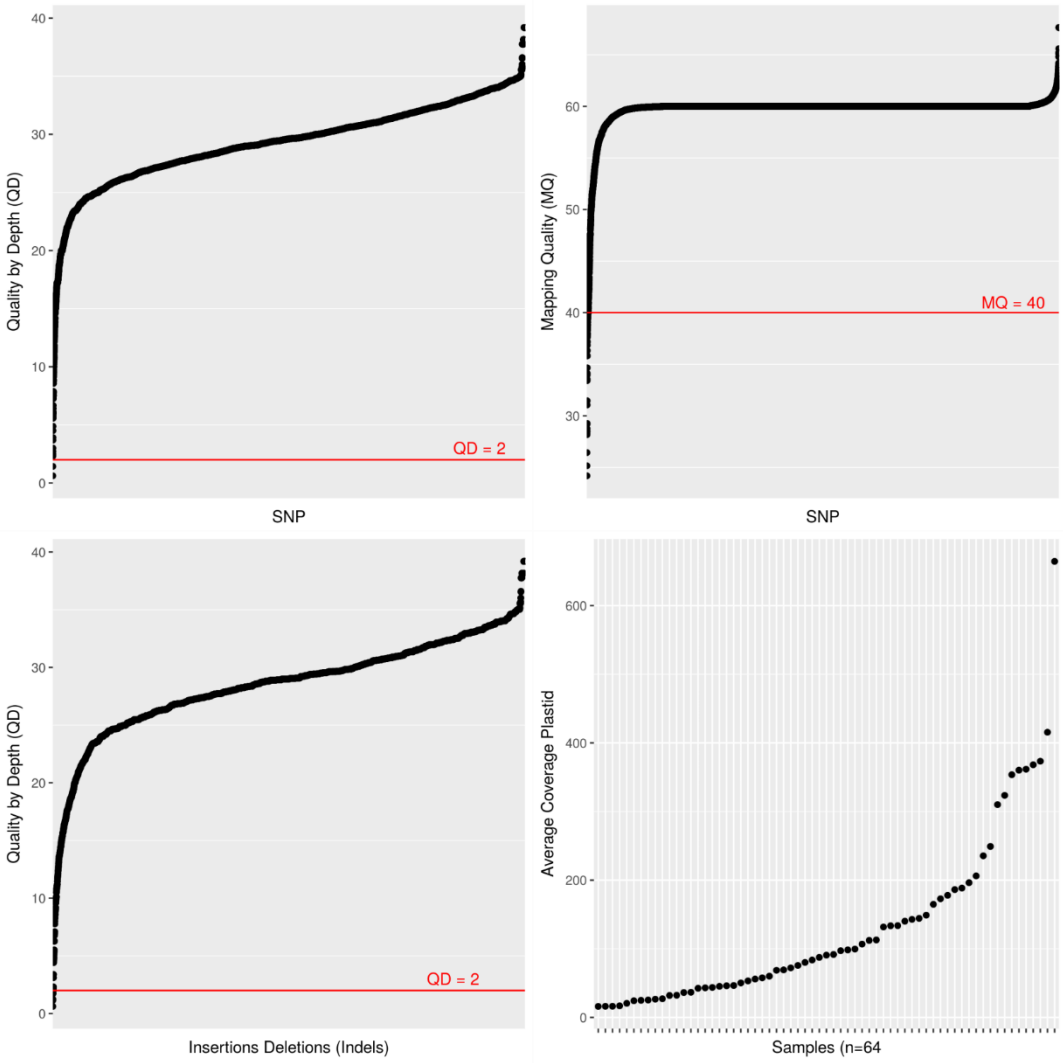


Figure 2-4 Quality filtering and coverage plots. Note that mapping quality is usually not applied to indels.

2.3.4 Phylogenetic hypotheses

Phylogenetic hypotheses generated here are as follows: firstly, the concatenated ML phylogeny, the best-scoring tree is presented, with bootstrap values, alongside a map

with the locality information of the samples from the Himalayas and the Hengduan Mountains. (Figure 2-5); secondly, the phylogeny inferred under the MSC is shown alongside the majority rule consensus tree from Bayesian analysis (Figure 2-6)thirdly, the plastid ML phylogeny mirrored to a reduced ML phylogeny with matching sampling (Figure 2-8). The MCMC sampling for the Bayesian inference was performed for at least 45 million generations. The trace file and the marginal probability plot for the 15 independent runs are shown in Figure 2-7. One of the 15 runs was discarded due to a lower likelihood score, suggesting misconvergence. The consensus network is shown in Appendix Figure AF-1.

2.3.5 Ancestral range estimation

The data used for ancestral range estimation in BioGeoBEARS is best explained by the model DIVALIKE+J (lowest AIC, Table 2-2). The ancestral ranges estimated under this model favour the Himalayan Mountain range with high probability as the ancestral range of Himalayan and Hengduan Mountain *Berberis* species (Node 4, Figure 2-9), deciduous (Node 10) and evergreen species (Node 7). The same result holds for deciduous Hengduan species, where *B. tsarica* is sister and the only member of the clade which is distributed in the Himalayas (Node 11). However, the ancestral range of the ingroup of this clade is with high probability the Hengduan Mountains. Within the evergreen clade, Node 9 is highly ambiguous with almost equal probabilities for the ancestral range being either of the two regions. Thus, no assumption of founder-effect speciation can be made. The ancestral area for the South American clade is, unsurprisingly, well-supported (Node 3). The sampling of species is equilibrated for Himalayan/Hengduan mountain species, but not for all other clades, which may influence ancestral area reconstruction at deeper phylogenetic levels.

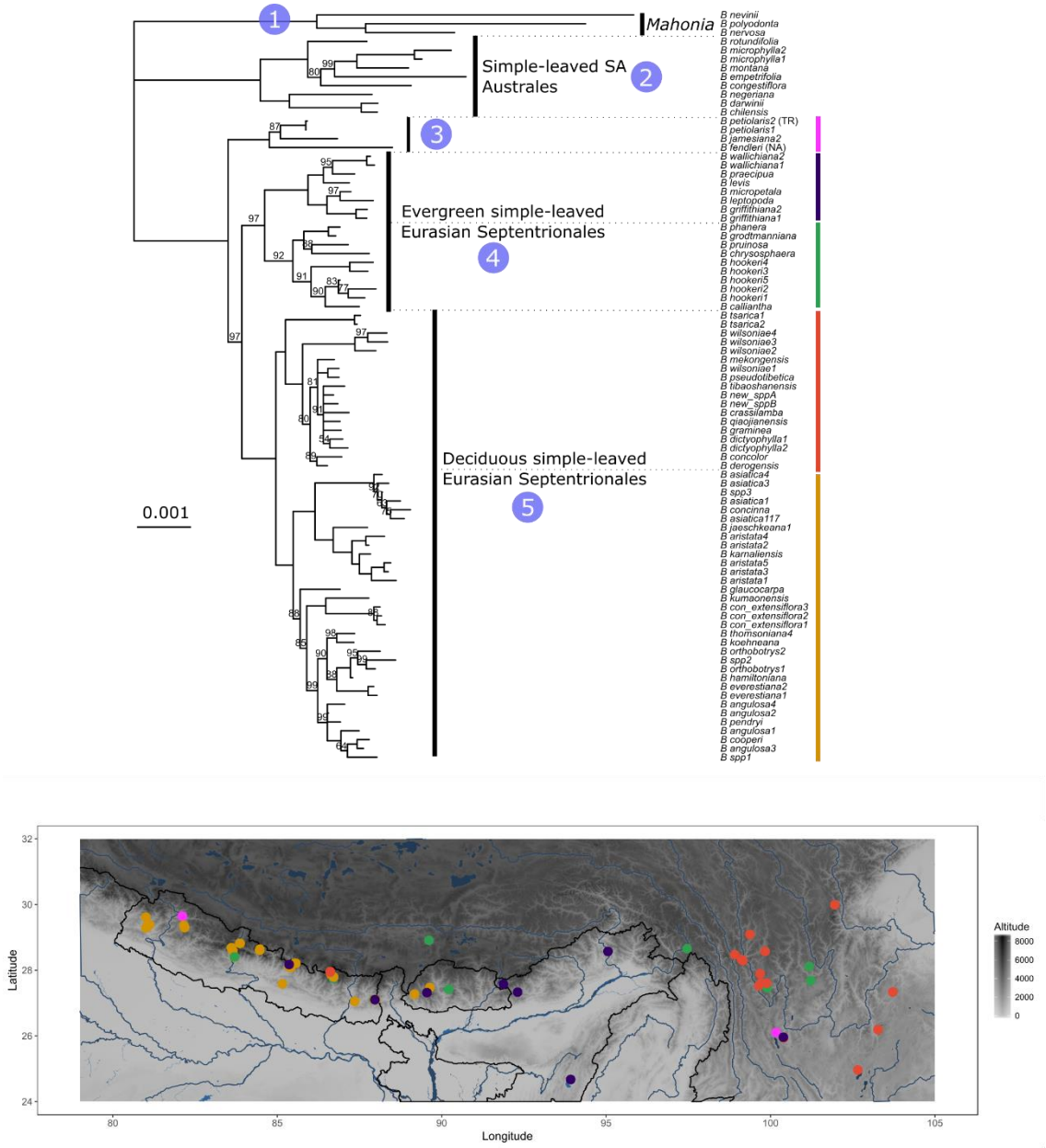


Figure 2-5 *Top*: Maximum likelihood tree of concatenated gene alignments. Only bootstrap values below 100 are shown above branches. The tree scale describes the mean substitutions per site. Numbers in circles indicate the major clades. *Bottom*: Map with specimen localities. Colours correspond to clades in the phylogeny.

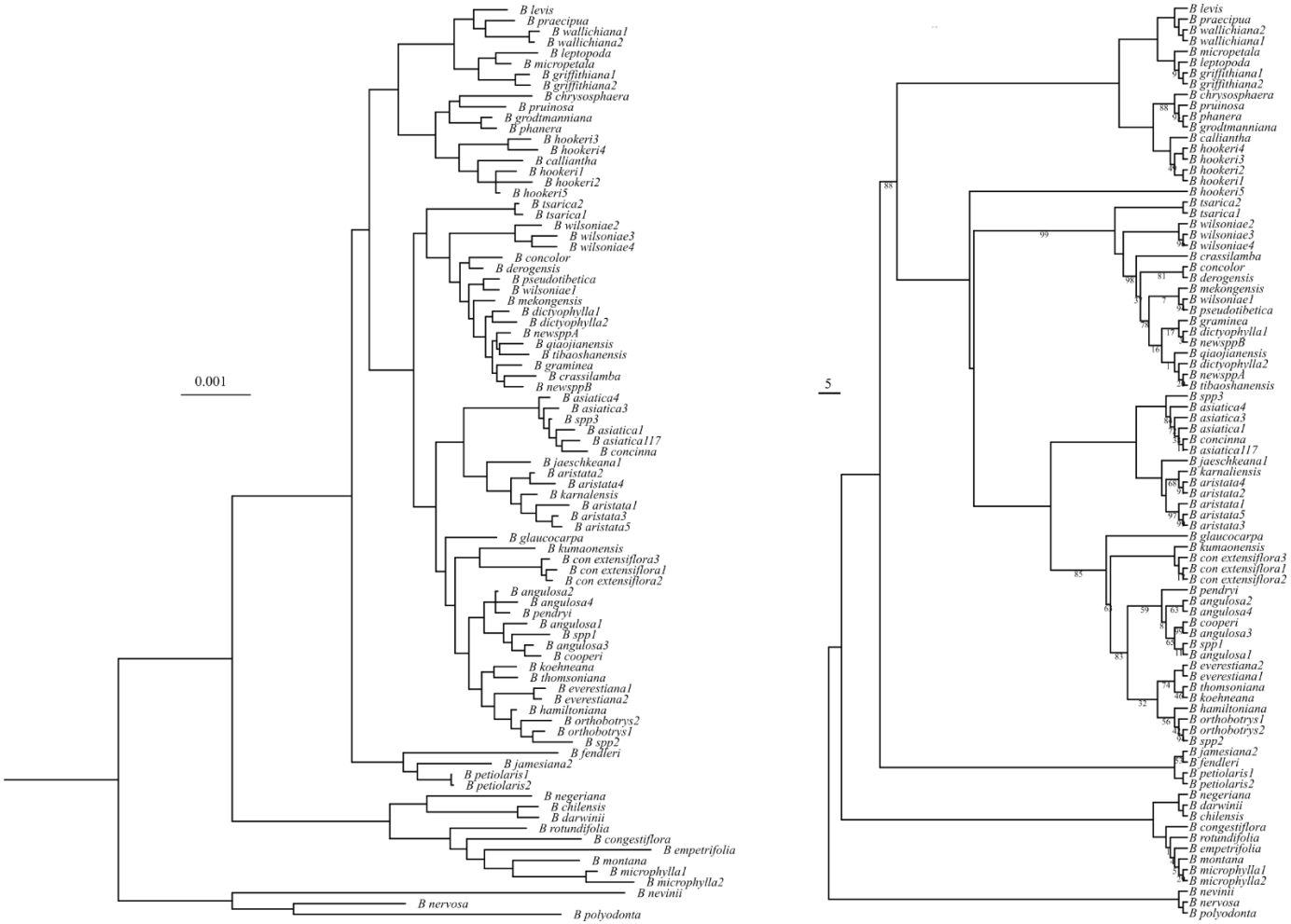


Figure 2-6 *Left*: Bayesian phylogeny inferred from concatenated marker alignments. All nodes have a posterior probability of 1. *Right*: Phylogeny based on the MSC inferred with ASTRAL-II. Numbers above branches are quartet scores, no displayed number stands for full support.

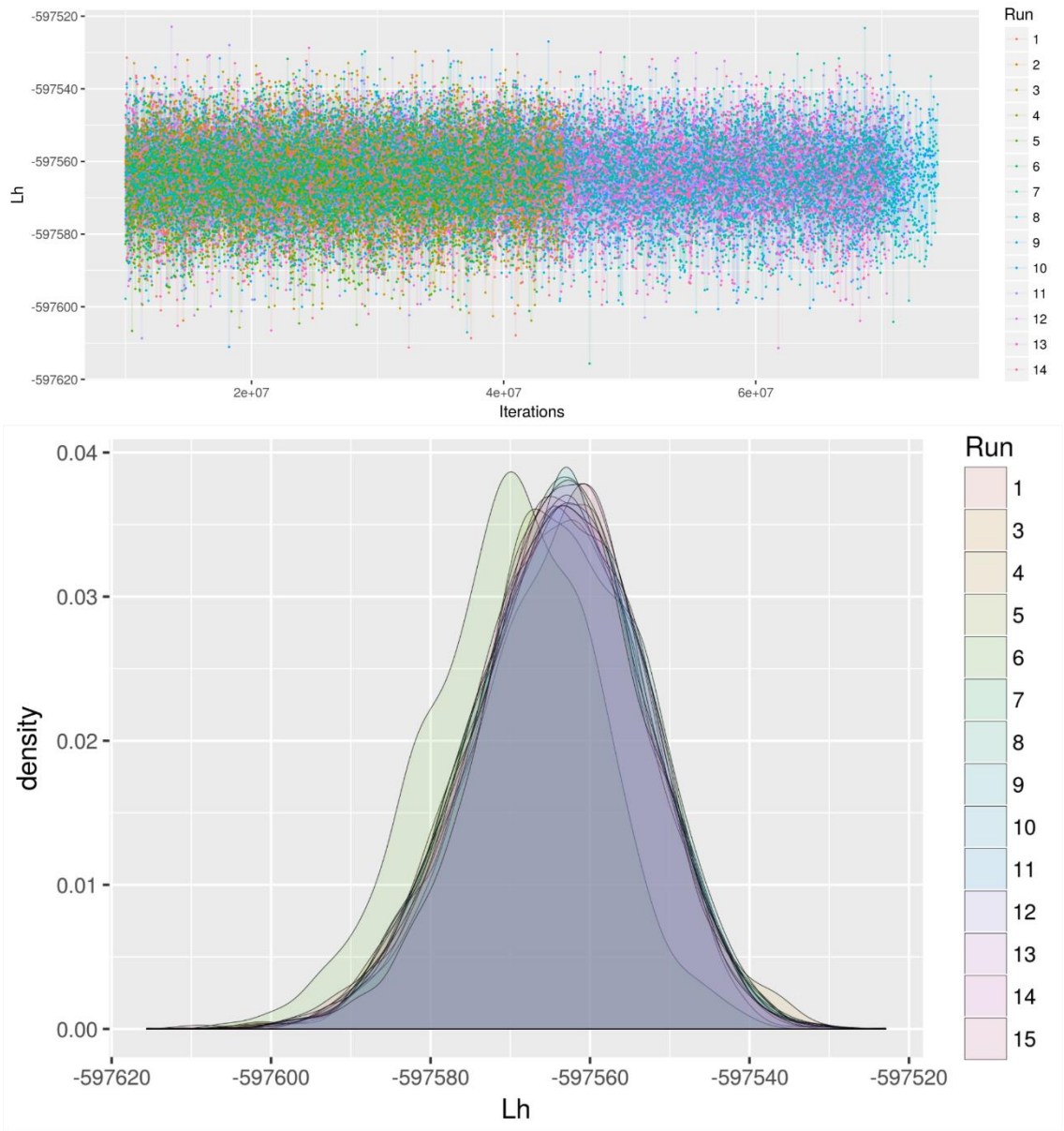


Figure 2-7 *Top:* The traces of the likelihoods of the15 independent runs. The burnin of 10 million generations is not shown. The runs were run for a minimum of 45 million generations. *Bottom:* The marginal probabilities are displayed as a density plot. Note that the likelihood curve of one run is slightly shifted, indicating that the run has not converged. The results from this run were excluded from further analysis.

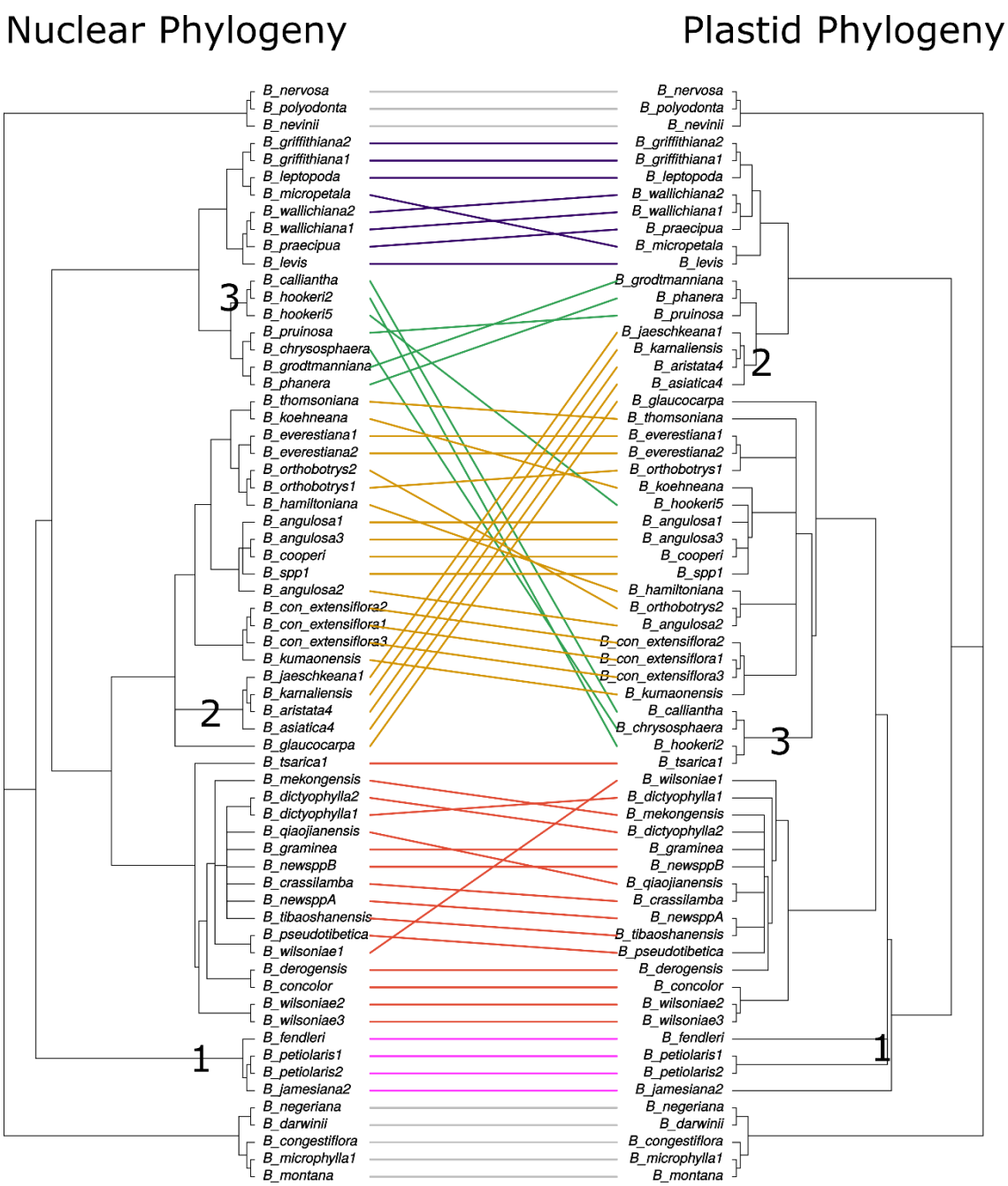


Figure 2-8 Mirrored plastid and nuclear phylogeny showing the connections between samples (coloured lines). Numbers on clades show whole clade shifts, as discussed in the text.

Table 2-2 Output from BioGeoBears analysis. The parameters are d=dispersal, e=extinction and j=founder-event speciation. The model was chosen according to the Akaike information criterion (AIC).

	LnL	Num. Parameters	d	e	j	AIC
DEC	-44.54	2	5	1.00E-12	0	93.08
DEC+J	-35.19	3	3.01	8.37E-02	0.025	76.39
DIVALIKE	-49.22	2	5	7.03E-07	0	102.44
DIVALIKE+J	**-34.64**	**3**	**4.92**	**8.88E-04**	**0.026**	**75.29**
BAYAREALIKE	-97.62	2	5	5.00E+00	0	199.25
BAYAREALIKE+J	-38.85	3	4.47	5.00E+00	0.034	83.72

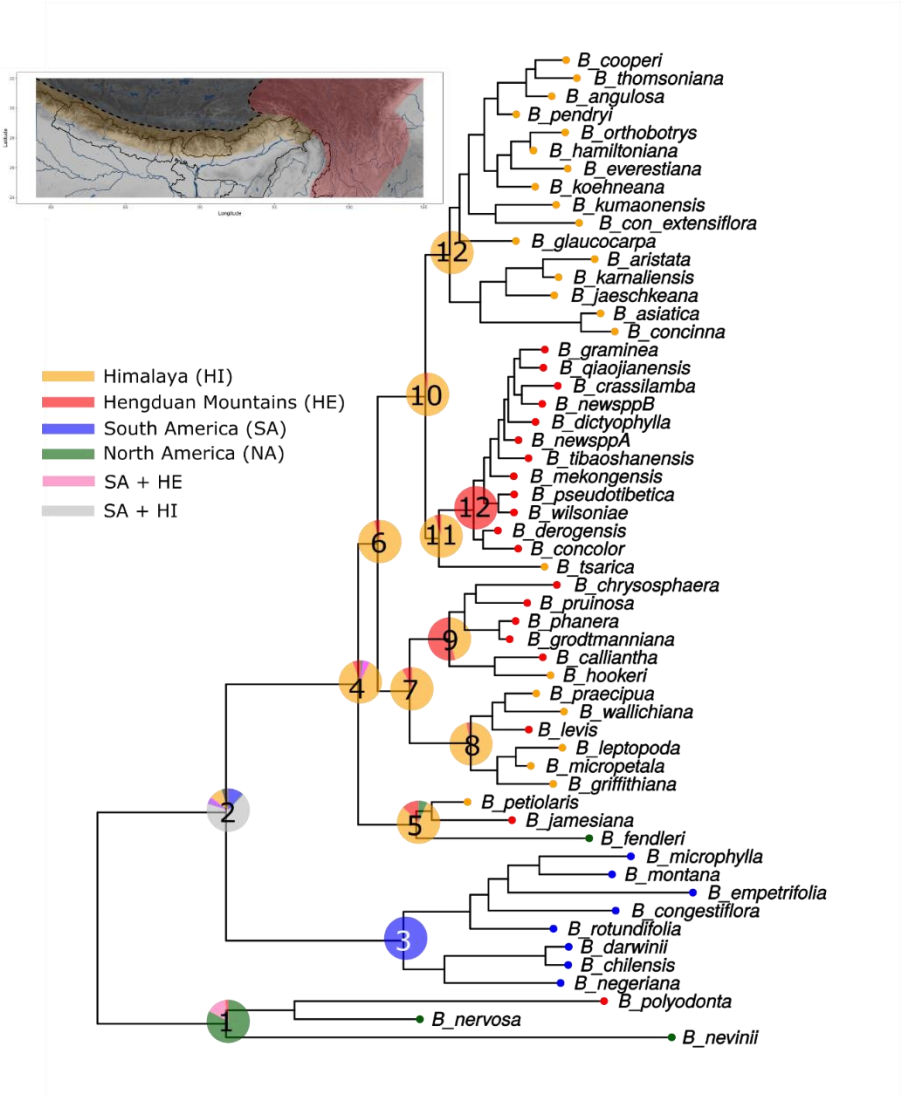


Figure 2-9 Ancestral range estimation of *Berberis* using a pruned ML tree. Note that Node 11 favours with high probability the Himalayan Mountains as the ancestral area. Furthermore, the ancestral area for Node 9 is highly ambiguous, favouring slightly an ancestral range in the Hengduan Mountains.

2.4 Discussion

2.4.1 Species tree inference from nuclear data

Phylogenetic hypotheses are generated here using multiple approaches (Figures 2-5 – 2-8). Sources of discordance between gene trees and the species tree include evolutionary processes such as incomplete lineage sorting (ILS) or gene flow through hybridization, as well as paralogy (Chapter 5), and appropriate methods at different stages of a phylogenetic study are needed to account for them. Considering phylogenetic reconstruction using concatenation or coalescent-based approaches, Folk et al. (2016) argued that using alternative approaches is a reasonable strategy for analysis of multi-locus data, since different approaches make different assumptions, and for empirical systems there are few grounds for making these *a priori* (Folk et al., 2016). McVay and Carstens (Mcvay and Carstens, 2013) noted that coalescent-based approaches are generally preferred in phylogeographic study since in this case incomplete sorting can be very marked; conversely concatenation is often used by those working at deeper taxonomic levels. However, McVay and Carstens (2013) challenged the implicit reasoning for using concatenation at higher levels – that processes of incomplete lineage sorting are less relevant at deeper levels - noting that populations-level processes occurred throughout the history of life. On these grounds, Edwards (2009) argued that coalescent-based approaches are preferable on philosophical grounds. This view is upheld by several authors who have highlighted deficiencies when using phylogenetic inference from concatenated multi-locus data (Degnan and Rosenberg, 2009; Edwards et al., 2016; Kubatko and Degnan, 2007). It can be misleading due to discordance between gene trees and the species tree (Kubatko and Degnan, 2007; Salichos and Rokas, 2013), and commonly-used node support metrics

such as bootstrap values or posterior probabilities are often overestimated, giving the impression of fully resolved species trees without conflicting signals (Rokas et al., 2003). In this study, concurring with Edwards (2009) that coalescent-based approaches are more suitable, the phylogeny of *Berberis* was estimated under the multispecies coalescent (MSC) implemented in ASTRAL-II, which models ILS. Simulating gene tree distributions directly from species trees can indicate whether discordances are likely to occur under the coalescent alone (Garcia et al., 2017 de Portugal et al., 2017). If not, and once paralogues are excluded, hybridization rather than ILS is inferred to explain discordance. Analyses of this type are not carried out here, so we consider two sets of evidence that point towards conflict resulting from ILS or hybridization: whether MSC and concatenated conflicts are deep or shallow, and whether plastid and nuclear hypotheses conflict. Although the overall topologies between MSC and concatenated topologies were largely congruent, quartet scores were lower than support calculated in concatenated analyses for some terminal clades, and where the concatenated analyses failed to resolve relationships, suggestive of ILS at shallow phylogenetic levels. This was often true when several members of a single species contributed to low support. The alternative placements of the specimen of *B. pendryi*, as sister to or nested in a *B. angulosa* clade, may reflect ILS since it is at a shallow phylogenetic level. A deeper conflict between concatenated and MSC topologies is observed for one of the specimens of *B. hookeri*. This species is one that shows alternative placements in the plastid and nuclear topologies, apparently the result of chloroplast capture in this individual. This conflict supports the interpretation of deep conflict between MSC and concatenated topologies as the result of hybridization.

2.4.2 Phylogenetic relationships

The phylogeny presented here gives an unprecedented view of the phylogenetic structure of *Berberis* species from the Himalayan and Hengduan Mountains. The nuclear and plastid phylogeny resulted in very different topologies (see 2.4.3) and, given the problematic nature of plastid phylogenies for reconstructing species relatedness and evolution (e.g. Rieseberg and Soltis, 1991), we consider the nuclear phylogenies more likely to best reflect species relationship. Examination of nuclear phylogenies reveals five clades (see clade numbers in Figure 2-5): compound-leaved *Berberis* (Clade 1); simple-leaved South American Australes (Clade 2); a clade with species belonging to the Septentrionales group (Clade 3); within the remainder of the Septentrionales, an evergreen clade (Clade 4) and a deciduous clade (Clade 5). The simple-leaved South American *Berberis* species form a strongly supported clade here and in the studies of Adhikari et al. (2015) and Kim et al. (2004). The Septentrionales clade was also recovered by Adhikari et al., (2015). However, the evergreen and deciduous clades are recovered here for the first time. We find that the clade comprising *B. petiolaris*, *B. jamesiana* and the North American species *B. fendleri*, is sister to the remainder of the Septentrionales. Thus, our results confirm the close relatedness of *B. fendleri* to Eurasian species, and support the long-distance dispersal hypothesis for this species.

Previous phylogenies based on a few genetic markers (*ndhF* and ITS) exhibit low resolution at shallow phylogenetic levels and poor support overall (Adhikari et al., 2015; Kim et al., 2004). The data set in this study comprises 396 nuclear loci and, in addition to recovering monophyletic groups representing the evergreen and deciduous traits for the first time, reveals considerable phylogenetic structure at species level. Both the evergreen and deciduous clades comprise species from the mountain ranges

surrounding the Qinghai–Tibetan Plateau (QTP), specifically the Hengduan Mountains and Himalayan ranges. Figure 2-5 shows that both the deciduous and the evergreen clades comprise two subclades. The two subclades of the deciduous clade show a strong geographical signal for either the Himalayan mountain range (orange colour in Figure 2-5) or the Hengduan Mountain (red colour in Figure 2-5). The geographical signal is less pronounced in the "evergreen" clade. One "evergreen" subclade consists of *B. hookeri* from the Himalayan mountain range, *B. calliantha* from the QTP and four species from the Hengduan mountains. Its sister subclade is mainly distributed in the Himalayan mountain range. The biogeographic hypotheses suggested by these distributions are further explored in section 3.4.4.

The taxonomy of *Berberis* is complicated, with revisionary work challenging the species numbers and delineations proposed by Ahrendt (1961). For example, Adhikari et al., (2012) revised the Nepalese species and Landrum (1999) the Chilean species, resulting in a reduction of the number of species. Revisonary studies for Chinese *Berberis* species are ongoing, and a monograph is in preparation (Harber, pers. communication). All our specimens were identified by taxonomic experts using the most recent published or draft taxonomic treatments. However, in complex groups there can be synergy between taxonomic and phylogenetic studies, and for at least one taxon our results challenge current taxonomy. Our phylogeny suggests *B. concinna* var. *extensiflora* should be raised to species rank, since it is only distantly related to *B. concinna* var. *concinna*. Three paraphyletic species are recovered, *B. asiatica*, *B. aristata*, *B. angulosa* and *B. wilsoniae*. Many species are paraphyletic (Rieseberg and Brouillet, 1994) and whether a phylogenetic species concept recognizing only monophyletic species should be applied is controversial (Agapow et al., 2004).

Certainly, the polyphyletic nature of the economically-important species *B. asiatica* and *B. aristata* may have repercussions for their recognition in trade (Chapter 4).

2.4.3 Conflict between nuclear and plastid hypotheses

Incongruence between nuclear and plastid phylogenies is a well-documented phenomenon (e.g. Rieseberg and Soltis, 1991), beginning to be reported using phylogenomic data (Folk et al., 2016). Nuclear and plastid phylogenies differ dramatically for *Berberis* species from the Himalayan and Hengduan Mountains. In the plastid phylogeny, evergreen and deciduous species do not form distinct clades as suggested by the nuclear phylogeny. We identify two distinct patterns of incongruence, whole-clade shifts and single-species shifts. Whole clade shifts have strong effects on the backbone of the phylogeny. For example, Clade 3 in the plastid phylogeny is not sister to the rest of the Himalayan and Hengduan Mountain species (see Figure 2-8). Another dramatic clade shift encompasses the clade with *B. aristata* which groups with evergreen species from the Hengduan Mountains in the plastid phylogeny (Clade 2; Figure 2-8). Similarly, the nuclear clade with *B. hookeri* species (Clade 2; Figure 2-8) groups with deciduous species in the plastid phyologeny. Whole clade shifts indicate ancient hybridization and introgression of the plastid genome between ancestral species that occurred in sympatry. Our data suggests that a common ancestor of species in the *"aristata"* clade (Clade 2; Figure 2-8) occurred in sympatry with a common ancestor of *B. grodtmanniana, B. phanera* and *B. pruinosa.* Similar patterns of incongruence have been shown in genus *Heuchera,* where an organellar capture event between an ancestral species of section *Heuchera* with a member of section *Holochloa* has been reported (Folk et al., 2016). The shifts of single species that we identify here point towards more recent events of chloroplast capture. For example, the evergreen *B. hookeri5* groups

with deciduous specimens *B. angulosa1+3, B_cooperi* and *B. koehneana*, rather than with conspecifics, as in the nuclear phylogenies. *B. angulosa1+3, B_cooperi* and *B. koehneana* are distantly related to *B. hookeri* in the nuclear phylogenies, but all specimens in this plastid clade are in geographic proximity (Eastern Nepal or Bhutan). The phenomenon of geographical grouping in plastid phylogenies has been documented, for example in genus *Nothofagus* Blume (Acosta and Premoli, 2010). Hybridization has been reported in *Berberis* (e.g. Adhikari et al., 2012) and it is likely that chloroplast capture happened via hybridization and introgression of organellar DNA. Models for chloroplast capture suggest that this process can occur beyond species barriers, and is in fact promoted by nuclear genome incompatibilities under certain conditions (Tsitrone et al., 2003). A further mechanism that may act as driver of chloroplast capture on species that occur in sympatry is grafting (Stegemann et al., 2012). The example of incongruence between nuclear and plastid phylogenies in *Berberis* confirms that plastid phylogenies are not a reliable source for reconstructing the species evolution, but in conjunction with nuclear phylogenies can be used to track ancient and recent chloroplast capture events. In terms of the development of markers for species identification (Chapter 4), our results strongly caution against using plastid markers.

2.4.4 Biogeography and evolution

The phylogeny and historical biogeography analyses of *Berberis* species from the Hengduan Mountains and the Himalayas shed light on how species are recruited to these two mountain systems, and whether frequent dispersal or *in situ* diversification explain this pattern. The geographical split of the two clades within deciduous species points towards *in situ* diversification within both of these clades, rejecting the

hypothesis of frequent dispersal events between the two mountain systems in deciduous *Berberis*. A recent study has compared *in situ* diversification rates versus colonization rates in both mountain systems across several plant genera, finding that the Hengduan Mountains flora was mainly assembled by *in situ* diversification (Xing and Ree, 2017). Although diversification rates were not calculated, genetic differentiation between species in the Hengduan Mountains is low, suggesting a young clade where lineage sorting has not yet completed. Sister to the Hengduan Mountains clade is *B. tsarica*, a species that is reportedly the highest growing *Berberis* distributed in the Himalayan chain (Adhikari et al., 2012) and in the southern margins of the QTP (www.efloras.org; last accessed 17/07/2017; Harber, pers. communication). The current distribution of this taxon and its phylogenetic placement suggests that the Hengduan Mountains were colonized by species that were at least partly distributed in the QTP and further diversified within this mountain system. In the case of *Berberis,* the QTP acted as a bridge between the already established Himalayan Mountains and the uplifting Hengduan mountains. The orogenesis of the Hengduan Mountains and rapid radiations have been shown for several taxa, for example for lineages *Ciliatae* and *Porphyrion* of the family Saxifragaceae, where the QTP acted as source for species colonization (Ebersbach et al., 2017). The emergence of mountainous habitats is known to trigger speciation by the emergence of island-like systems, so-called ‘sky islands’ (Hughes and Atchison, 2015). Although the uplift of the Hengduan Mountains may have played an important role in *Berberis* diversification, *Berberis* is not a typical alpine genus where radiations happen in island-like alpine habitats. The complex physiographic structure of these mountains may facilitate allopatric speciation below the tree line.

2.4.5 Conclusion

Phylogenomic data sets are powerful for resolving phylogenies to species level and highlight the importance of inferring phylogenies with multiple approaches. The nuclear phylogenies of *Berberis* show that evergreen and deciduous species form two distinct lineages and that deciduous species show a strong geographical structure with two lineages either in the Himalayas or the Hengduan Mountains. Furthermore, dramatic differences between plastid and nuclear phylogenies in genus *Berberis* were revealed, suggesting ancient and more recent chloroplast capture events.

Chapter 3 New approaches for DNA barcoding herbal medicines: a case study of genus *Berberis*

3.1 Introduction

DNA barcoding has two major objectives: specimen identification, where an unknown sequence is matched to a sequence of a known species, and species discovery, which is equivalent to species delimitation and species description (DeSalle, 2006). DNA barcoding of herbal medicines is mainly concerned with authentication, the identification of specimens for quality assurance (Sgamma et al., 2017). In the last decade, DNA barcoding of herbal medicines has raised awareness of species substitution and adulteration, highlighting issues surrounding the quality of herbal medicines in the global market (Newmaster et al., 2013; Srirama et al., 2017). Regulation of herbal medicines is a pressing issue for regulatory agencies (Directive 2001/83/EC, 2001; Directive 2004/83/EC, 2004; Vlietinck et al., 2009). Published pharmacopoeial standards for authentication predominantly rely on chemical and anatomical methods (e.g British Pharmacopoeia, 2016). DNA barcoding offers new tools for regulatory purposes (de Boer et al., 2015) and DNA barcodes have recently been incorporated into the British Pharmacopoeia for the first time (British Pharmacopoeia Commission, 2017). Here we investigate opportunities and limits of DNA barcoding using next-generation sequence data of an evolutionarily complex genus as a case study. The aim is to provide methodological approaches for producing DNA barcodes for regulatory purposes, pharmacovigilance and quality assurance.

The initial proposition of DNA barcoding using the small, single DNA sequence of the cytochrome c oxidase subunit 1 (*CO1*) for species identification (Hebert et al., 2003) complies with its core principles of standardization, minimalism and scalability

(Hollingsworth et al., 2011). Whilst DNA barcoding with the *CO1* region has proven to be effective for many groups in the animal kingdom (e.g. Hebert et al., 2004; Smith et al., 2008), research shows that mtDNA of plants evolves at a much slower rate than the plastid and nuclear DNA (Wolfe et al., 1987), and that the divergence of the *CO1* gene among plants is very low (Cho et al., 2004, 1998). A single DNA barcode for land plants has not been identified (Hollingsworth et al., 2011), although several propositions have been made (e.g. CBOL Plant Working Group et al., 2009; Chase et al., 2007; Kress et al., 2005). Following Hollingsworth et al. (2011), most studies use a combination of the plastid regions *matK, rbcL,* the intergenic spacer *trnH-psbA* and the nuclear ITS2. Advances in sequencing technology have encouraged the barcoding community to augment the standard barcoding approach (Coissac et al., 2016; Kane et al., 2012; Vaughn et al., 2014). In the era of next-generation sequencing, some researchers have argued for the use of whole plastid genomes as barcodes (Coissac et al., 2016; Kane et al., 2012; Vaughn et al., 2014).

Methodological approaches for specimen identification using DNA barcodes commonly rely on either distance-based measures or phylogenetic methods (Austerlitz et al., 2009). The former are based on the assumption that intra- and interspecific variation does not overlap (e.g. Hebert et al., 2004), also referred to as the barcoding gap (Meyer and Paulay, 2005). Accurate specimen identification using distance-based approaches such as BLAST are highly dependent on a well-curated database where ideally all members of a group are represented by several individuals (Meyer and Paulay, 2005). Drawbacks of using distance-based approaches are that there is no objective distance threshold criterion and that the nearest neighbour is not always the closest relative (Moritz and Cicero, 2004). Specimen identification using phylogenetic methods is based on membership of a query sequence to a specific clade (Casiraghi et

al., 2010). One difficulty associated with using tree-based barcoding methods is that phylogenies inferred from the barcode sequence might not be resolved sufficiently for an individual to be allocated to a clade and that clades may exhibit poor support, questioning the robustness of any phylogenetic hypothesis (Moritz and Cicero, 2004). The use of concatenated DNA sequences for species tree inference has been shown to produce more robust phylogenetic hypotheses (Rokas et al., 2003). However, phylogenetic methods to DNA barcoding are not suitable when the underlying system is not based on strictly hierarchical ancestor-descendant relations structures, such as in nested structures (Goldstein and DeSalle, 2005). A general criticism to both, distance-based and phylogenetic methods, is that these methods are not compatible with taxonomic decision circles where several lines of diagnostic evidence is needed for describing a taxon, which is circumvented by using diagnostic molecules in DNA barcodes (DeSalle et al., 2005).

Whether specimens of different species can be discriminated between depends on the choice of the DNA barcode and the relatedness of species under study. Although relatively high success for the identification of genera has been reported when using common barcodes in plants, limited sequence variation is often the cause of the failure to distinguish between closely related species (Braukmann et al., 2017; Parmentier et al., 2013; Seberg and Petersen, 2009). One incentive for employing genomic approaches to barcoding is that broader genome coverage increases the variation in the barcoding data set (Coissac et al., 2016). However, closely related species may not exhibit a DNA barcoding gap even when the most variable regions are employed. In the case of incipient speciation where lineage sorting is incomplete, species are likely to be paraphyletic (Fazekas et al., 2009; Rieseberg and Brouillet, 1994). Furthermore, cytoplasmic genomes can have different evolutionary histories compared to nuclear

genomes through processes such as chloroplast capture (Rieseberg and Soltis, 1991), and specimens may group geographically rather than taxonomically (Acosta and Premoli, 2010). The biology and evolutionary history of several plant groups may therefore limit the success of DNA barcoding (Percy et al., 2014).

Genus *Berberis* is a case where DNA barcoding with few regions has limited success (Roy et al., 2010). Similarly, a phylogeny of *Berberis* based on *ndhF* and ITS loci failed to resolve species boundaries (Adhikari et al., 2015). However, *Berberis aristata* is a medicinal plant that has been in traditional use in India for centuries and is nowadays traded throughout the world (Srirama et al., 2017). Local market studies suggest that several species are traded under the same vernacular name (see Chapter 4, Srivastava and Rawat, 2013), including *B. aristata* and *B. asiatica. B. aristata* is described in several pharmacopoeias (Ayurvedic Pharmacopoeia of India, 2001; British Pharmacopoeia, 2016) and although chemical and anatomical tests are published, there is incentive for producing a DNA barcoding method for identification.

The aim of this study is to investigate whole plastid sequences of genus *Berberis* as a resource for barcode design, and to examine the evolutionary relationships of the species that might contribute to understanding the difficulties of using barcoding as a means for specimen identification. We present a method for producing short, informative plastid barcode regions based on diagnostic nucleotides. These barcodes, which are informative of clade membership in a phylogenetic context, are tested on commercial samples, and their utility for regulatory purposes outlined.

3.2 Material and methods

3.2.1 Sampling

This study includes 85 specimens from 57 species (Appendix Table AT-1). The dataset includes sequences from two putatively new species (named in this study as B_newsppA & B_newsppB) and one unidentified species (B_spp). The samples are partly the same as used in Chapter 2. The numbering of samples is congruent between the chapters.

3.2.2 Sequencing

The methods for library preparation and sequencing are described in Chapter 2.2.2. The specific sequencing platforms used are documented in the Appendix Table AT-2.

3.2.3 Plastid genome reconstructions

The reference genome for *B. aristata7* was reconstructed using a hybrid strategy of read mapping and *de novo* assembly. All reads were mapped to the reference plastid genome of the closely related *Berberis bealei* (Ma et al., 2013 GenBank reference KF176554), using the Geneious medium-low sensitivity 'Map to Reference' function with five iterations. The resulting contig was then checked manually for low coverage and low pairwise identity regions. One read from each of these regions was extracted and all reads were then mapped against these individual reads using the same settings as above. The iterations lead to an extension of the read to a contig (typically up to 2,500 bp). The consensus sequences were then mapped to the reference obtained from the first read mapping. This method allowed large indels in the *B. aristata* reference that were not detected by the read mapping algorithm to be identified. The built-in *de novo*

algorithm in Geneious 7.1.7 was used for the *de novo* assembly of the plastid genome. We performed the assembly only with reads that matched to the reference sequence of *B. bealei.* The ten largest contigs, ranging in length from 1,132 bp to 29,132 bp, were then mapped to the *B. aristata* reference and checked for ambiguities. All reads were then mapped again to the new consensus sequence.

The reconstruction of the plastid sequences using the newly generated *B. aristata7* reference is described in section 2.2.3.6. The plastids were aligned using the MAFFT aligner (Katoh and Standley, 2013) with default options. The alignment of repetitive regions such as poly A sequences was not straight-forward, therefore two alignment files were created: the first alignment was used for phylogenetic inference, and blocks where no unambiguous alignment could be constructed were removed. Furthermore, the inverted repeats were removed, since SNP calling on these repeats was difficult to address (see section 2.2.3.6 for further explanations). The second alignment was used for the barcoding analysis. Regions were masked (coded as "N") where no unambiguous alignment was possible

3.2.4 Annotation of plastid sequence

The online platforms DOGMA (Wyman et al., 2004) and CpGAVAS (Liu et al., 2012) were used for the annotation of the genome of *B. aristata7.* The full genome sequences were imported into Apollo (Lee et al., 2009). The annotation of *B. aristata* was compared with the previously published annotation of *Berberis bealei* (Ma et al., 2013). Start and stop codons were checked manually. The annotation was visualized using OGdraw.

3.2.5 Universal barcode reconstruction

3.2.5.1 Extraction of plastid barcode sequences

The sequences of *matK*, *rbcL* and *trnH-psbA* of *B. aristata* were extracted from the annotated reference *B. aristata0299*. The sequences were then aligned to the plastid genomes using BLAT (Kent, 2002). The output was parsed to produce a BED file, which denotes the start and end position of an alignment. The respective sequence was then extracted with the 'getfasta' option in BEDTools (Quinlan and Hall, 2010).

3.2.5.2 Reconstruction of ITS2

A two-step pipeline was devised to reconstruct the ITS2 from shotgun sequencing data. Firstly, reads that map to the reference were filtered and then a *de novo* assembly was performed using these reads. Filtering prior to *de novo* assembly reduces computation time substantially. The reference sequence of ITS2 (*Berberis repens*, BOLD accession: HIMS1138-12) was indexed with BWA (Li and Durbin, 2009) using the command 'bwa index'. Trimmed and filtered reads were mapped to the reference with 'bwa mem'. Mapped reads were then separated from unmapped reads with SAMtools (Li et al., 2009) 'samtools view –b –F 4', resulting in a BAM file with only mapped reads. The mapped reads were then extracted to fastq format using Picard tools (http://broadinstitute.github.io/picard, last accessed 30/06/17) with the command 'SamToFastq'. The reads were then used for *de novo* assembly using SPAdes v3.7.0 (Bankevich et al., 2012) and the longest contig extracted.

3.2.6 Barcoding analysis and phylogenies

The phylogeny for the whole plastid genome was inferred using the same parameters described in 2.2.3.7. Phylogeny reconstruction was performed on the online portal CIPRES (Miller et al., 2010).

The barcoding analysis aimed to find set of informative nucleotides that are unique to clades of interest. A barcoding method based on diagnostic characters was preferred over distance or purely phylogenetic approaches, because of its ease of application for regulatory purposes and to provide an alternative approach in an evolutionary complex group. Potential novel *Berberis*-specific barcodes were explored by extracting SNP positions of the multiple sequence alignment of whole plastid genomes with the program SNP-sites (Page et al., 2016). The SNPs were summarized in 500 bp windows and their distribution plotted with Circos (Krzywinski et al., 2009). Potential barcodes were selected spanning regions where a 500 bp window had a sequence variability of > 5%, and a maximum amount of missing/masked data < 3%. The 500 bp regions were then compared to the annotated plastid genome and the barcodes were constructed to correspond with genomic regions, such as intergenic spacers that are flanked by conservative regions suitable for primer design.

The commonly-used barcodes ITS2, *rbcL*, *matK* and *trnH-psbA* and the *Berberis* specific barcodes derived from the whole plastid alignment were evaluated. The individual barcode regions were aligned using MAFFT v7.215 (Katoh and Standley, 2013) with default options and were then manually trimmed. A first step was to infer a maximum likelihood tree of the barcode with RAxML v.8.2.9 (Stamatakis, 2014) with 1,000 rapid bootstrap replicates ('-f a') under the GTRCAT model. Haplotype networks were constructed with the function 'haploNet' in the R package pegas (Paradis, 2010). The potential barcodes were sorted according to the percent variable sites, percent parsimony informative sites, recovery of *B. aristata* and *B. asiatica* groups and the recovery of groups present in the whole plastid phylogeny. The selected barcodes were concatenated and a maximum likelihood phylogeny was built with the same parameters as described above. The best fitting substitution model was

inferred with jModelTest 2 (Darriba et al., 2012) by calculating the likelihood scores and by evaluating the models with the Akaike information criterion. Phylogenies of the selected barcodes were inferred under the GTRCAT model, but were also inferred under the Jukes-Cantor and the Hasegawa–Kishino–Yano models in RAxML v. 8.2.9 (Stamatakis, 2014). Furthermore, the barcodes were examined by inferring a phylogeny under the maximum parsimony criterion using the R package phangorn (Schliep, 2011). The implemented algorithm finds the tree with the lowest parsimony score using nearest-neighbor interchanges and subtree pruning and regrafting. The topology of the whole plastid genome phylogeny was used to determine meaningful groups of species according to evolutionary relationships of their plastid genomes. Barcodes were then constructed for identifying these evolutionary entities, rather than individual species. The alignment of each selected barcode was then reduced to SNP sites only and synapomorphic polymorphisms were identified for each group in order to delimit a minimal barcode.

3.2.7 Test dataset

The first data set for testing the barcode consisted of three commercial samples, putatively derived from *B. aristata* (Table 3-1). The sequences for the market samples were produced in the target enrichment experiment (Chapter 4). Although the samples were enriched for nuclear loci before sequencing, a certain amount of off-target reads were sequenced as well (Weitemier et al., 2014). The second data set consisted of *in silico* mixtures of samples from species *B. aristata* and *B. asiatica*. The mixtures were produced using shotgun sequencing data from unambiguously identified *Berberis* leaf samples (see 2.2.2.2). The quality filtered reads were mapped to the reference plastid genome (*B. aristata7*) and the mapped reads were extracted with SAMtools and

bedtools vcf2fastq. The resulting fastq files were then subsampled randomly and combined (Table 3-1).

The reads of these five test samples were mapped to the consensus sequence of the barcodes with BWA (Li and Durbin, 2009) and sorted with 'samtools sort' in SAMtools (Li et al., 2009). The base call(s) and base frequencies of the respective barcode positions were extracted from the BAM files using the program bam-readcount (https://github.com/genome/bam-readcount; last accessed 02/06/17).

Table 3-1 *In silico* mixtures of *B. aristata* and *B. asiatica* samples.

	Number of Reads (forward and reverse)			***aristata:asiatica*** (Ratio)
Sample	***B. aristata7***	***B. asiatica5***	***B. aristata8***	
Mixture1	90K	60K	30K	2:1
Mixture2	470K	72K	-	5.5:1

3.3 Results

3.3.1 Whole plastid phylogeny

The annotated plastid sequence of *B_aristata7* is shown in Appendix Figure AF-2. As described in Chapter 2.2.3, the plastid and nuclear phylogenies differ, and neither the evergreen nor deciduous groups are monophyletic in the plastid phylogeny. The whole plastid phylogeny is shown in Figure 3-1. The groups 1 + 2 (Figure 3-1, circled numbers) consist mainly of deciduous species. Sister to these groups are *B. petiolaris, B. jamesiana* and the North American species *B. fendlerii.* The deciduous species in the *aristata* clade (4) and the *asiatica* clade are nested within evergreen species (groups 5 + 6). The plastid phylogeny reveals that within the *aristata* clade, *B. aristata* is not monophyletic since *B. jaeschkeana, B. karnaliensis* and *B. mucrifolia* are nested within

the polyphyletic species. The South American species (8) and *Mahonia* species (9) form monophyletic groups with high support.

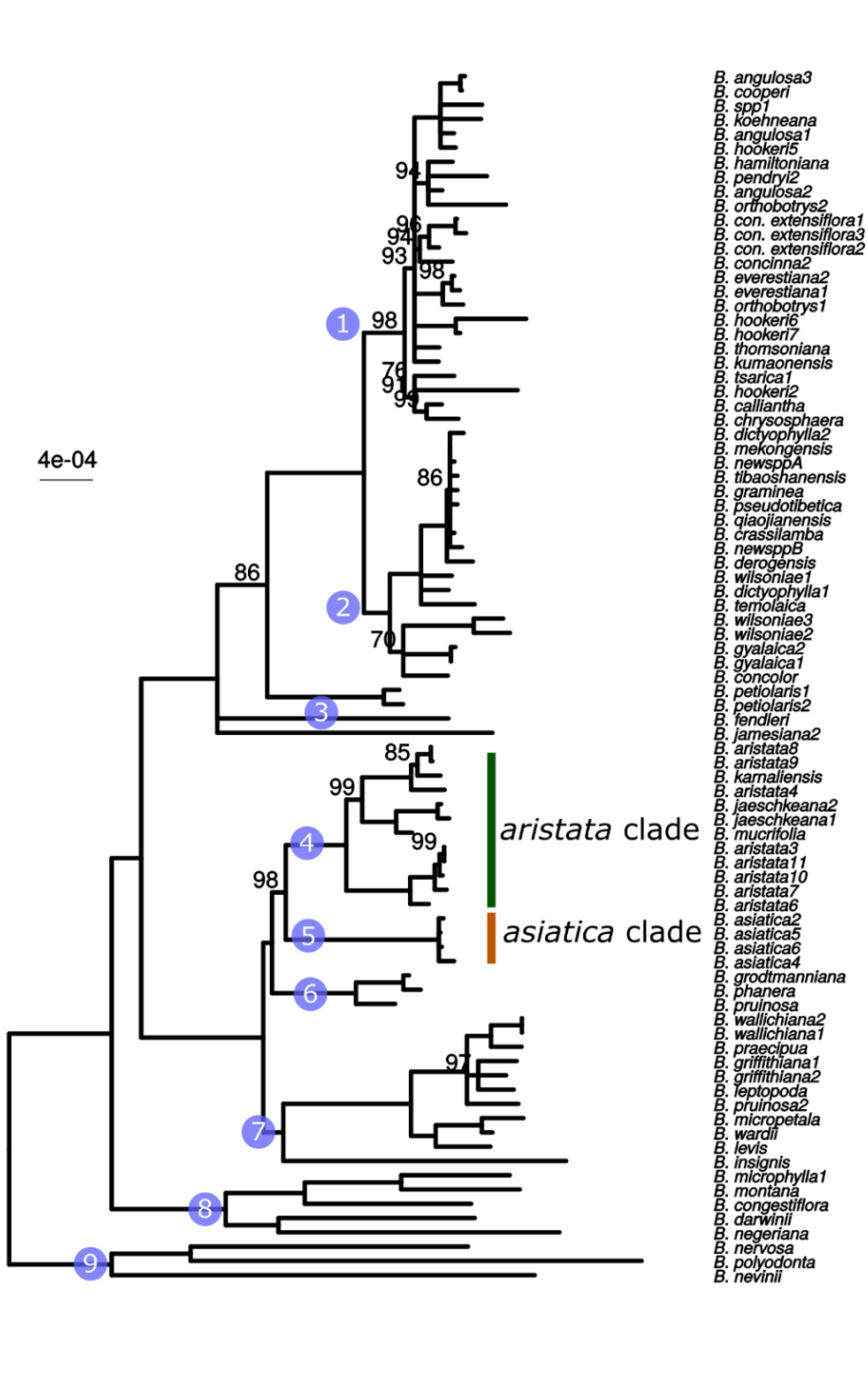


Figure 3-1 ML phylogeny based on whole plastid sequences. Note that *B. aristata*, in the *aristata* clade, is a polyphyletic species, but the *B. asiatica* in the *asiatica* groups are monophyletic. Numbers above branches are bootstrap values between 51 and 99. Branches with support < 50 were collapsed to polytomies, bootstrap values of 100 are not shown.

3.3.2 Identifying informative barcodes

The density of SNPs in 500 bp windows along the whole plastid alignment is shown in Figure 3-2. The bins contained between 0 and 124 variable sites per 500 bp. The inspection of bins with > 25 SNPs (5%) resulted in 21 potential barcode regions. Several of the highly variable bins fell into regions where the alignment was partly masked, leaving 13 bins for further inspection. Two neighbouring bins were combined into a single potential barcode of 1,000 bp, and a set of four bins combined into a 2000 base pair barcode. The barcode of 2,000 bp (SSC_noncoding2) was further examined by partitioning the alignment into 50 bp windows and reducing the barcode size (SSC_noncoding2, Figure 3-3). The *trnH-psbA* intergenic spacer was identified among one of the seven highly variable regions, and together with the matK, *rbcL* and ITS2 barcodes, eleven barcode candidates were investigated (Table 3-2).

Table 3-2 Barcode selection resulting from investigating variability patterns across whole plastid alignment. ITS2, *matK* and *rbcL* were not identified as highly variable but included in the study. Var = Variable sites; PIS = parsimony informative sites; "*aristata* recovered" and "*asiatica* recovered" indicates whether the clades were recovered in the respective phylogeny.

Barcode	**Length (bp)**	**Var**	**% Var**	**PIS**	**% PSI**	***aristata* recovered**	***asiatica* recovered**
ITS2 (nuclear)	560	45	8.04	24	4.29	no	Yes
matK	**1530**	**39**	**2.55**	**18**	**1.18**	**yes**	**Yes**
ndhF (partial)	802	40	4.99	23	2.87	no	Yes
ndhI-ndhG	**501**	**48**	**9.58**	**18**	**3.59**	**no**	**Yes**
rbcL	1452	32	2.20	21	1.45	no	Yes
rbcL-atpB	770	32	4.16	19	2.47	no	Yes
rbcL-psaI	626	59	9.42	28	4.47	no	Yes
rpl32-ndhF	1119	80	7.15	40	3.57	partly	Yes
SSC_noncoding1	741	52	7.02	29	3.91	partly	No
SSC_noncoding2	**790**	**46**	**5.82**	**27**	**3.42**	**yes**	**Yes**
trnH-psbA	580	43	7.41	24	4.14	no	Yes

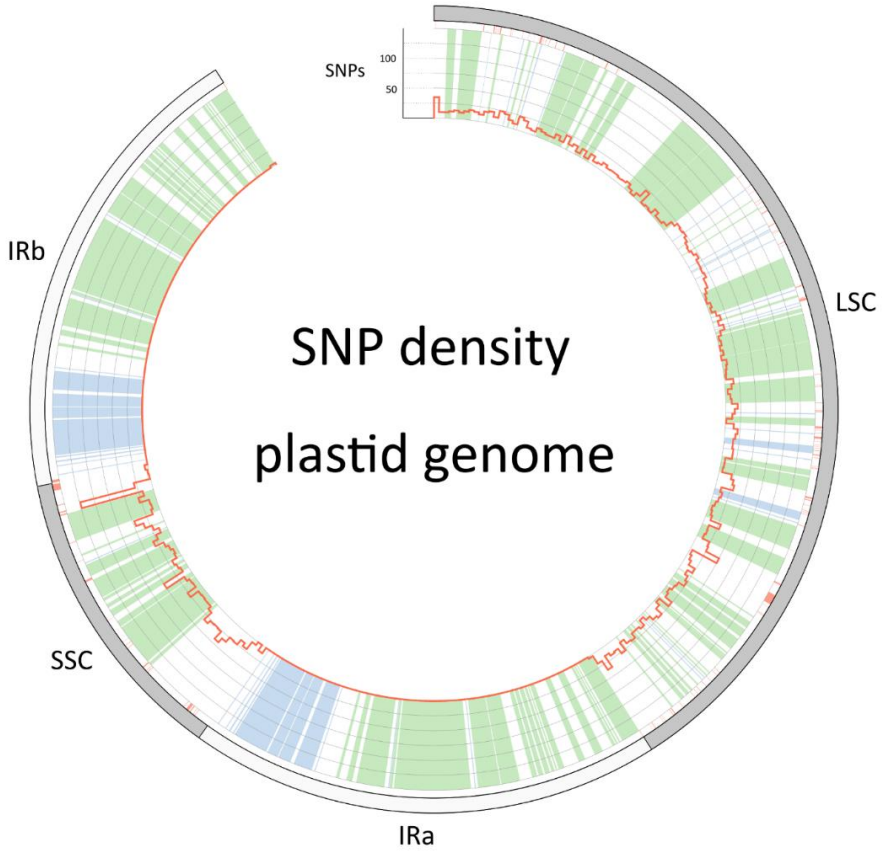


Figure 3-2 SNP density along the plastid genome (red histograms). The outer circle describes the boundaries of the large single copy (LSC), the inverted repeats (IRa and IRb) and the small single copy (SSC). Regions that are coloured green in the inner circle are coding regions, blue are RNA genes (rRNA and tRNA genes) and white is noncoding sequence. Red colour below the outer circle shows regions that have been masked and are thus coded as "N".

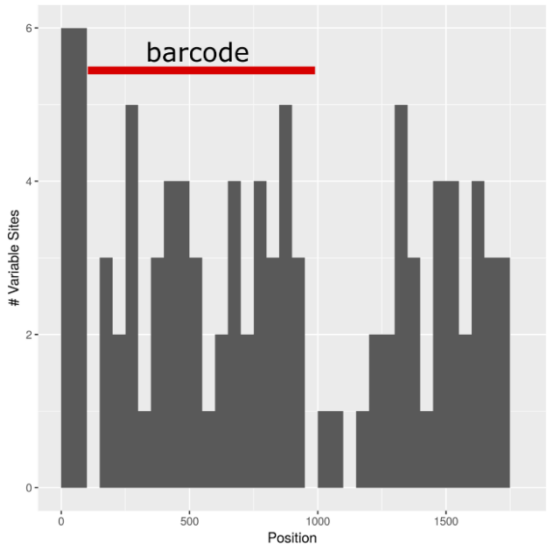


Figure 3-3 Subselection of barcode regions with the SSC_noncoding2 region. The newly determined barcode is marked in red.

None of the individual barcodes retrieved phylogenies with the same topology as the whole plastid phylogeny. Although the *matK* phylogeny is not well resolved overall, species from the *aristata* and *asiatica* groups were recovered. *B. asiatica* is monophyletic in the noncoding SSC_noncoding2 phylogeny, but species from the *aristata* clade are separated into two groups. The same results were retrieved when assuming less complex models of evolution (AF-3). The haplotype networks of these two barcodes show that both groups have distinct haplotypes. The percent variable sites varied between 2.2 in *rbcL* and 9.85 in the intergenic spacer *ndhI-ndhG* (Table 3-2) and the latter was chosen along with *matK* and SSC_noncoding2 as barcodes for further analysis (Figure 3-4).

These three barcodes yielded 133 variable positions in total. Nine positions, including seven containing one SNP variant that was clade-specific, were sufficient to identify seven of the nine groups. Groups 3 and 8 share a barcode, in other words their barcodes are identical. The phylogeny of the concatenated barcodes *matK,* SSC_noncoding2 and *trnI-trnG* barcodes is shown in Figure 3-5. The topology of the tree differs substantially from the total-evidence tree inferred from whole plastid sequences. However, four of the major clades are identified in both trees. The model test results for molecular evolution of the respective barcodes is given in Table …. In order to investigate the influence of the model selection, we have inferred phylogenies with simpler models than the GTRCAT. The clades of interest were also recovered when applying these models of molecular evolution (Appendix Figure 2), suggesting minor influences on the phylogenetic analyses.

Table 3-3 Results from the model test for molecular evolution. The GTR model was only favoured for the barcode ndhI-ndhG. GTM = General Time Reversible Model; TVM = Transversion Model; TIM = Transition Model; TPM = 3-parameter Model.

Barcode	Model	Partition	-lnL	AIC	deltaAIC
ndhI-ndhG	GTR	12345	1012.45	2376.899	0
	TVM	12314	1014.13	2378.263	1.364
	TIM2	10232	1015.328	2378.657	1.7582
SSC_noncoding2	TVM	12314	3486.321	7322.642	0
	GTR	12345	3486.284	7324.567	1.9251
	TPM1uf	12210	3491.467	7328.935	6.2929
matK	TPM1uf	12210	2415.091	5176.183	0
	TIM1	12230	2414.665	5177.331	1.1481
	TVM	12314	2415.027	5180.054	3.8709

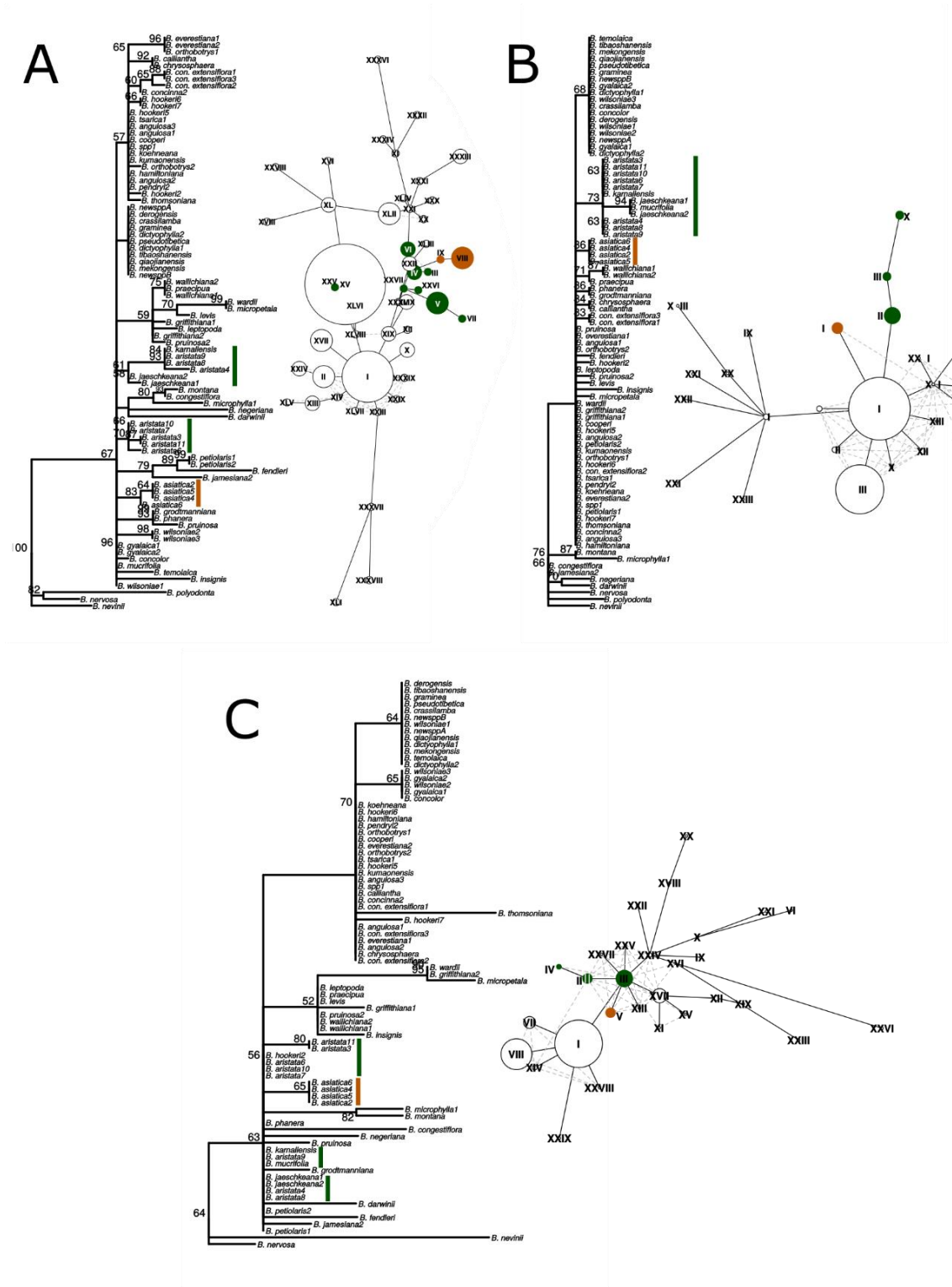


Figure 3-4 Maximum likelihood phylogenies and haplotype networks of individual barcodes. The Roman numerals indicate different haplotypes and the size of the circles corresponds to the number of samples sharing this haplotype. *A*: SSC_noncoding2, *B*: *matK*, *C*: *ndhI-trnG*.

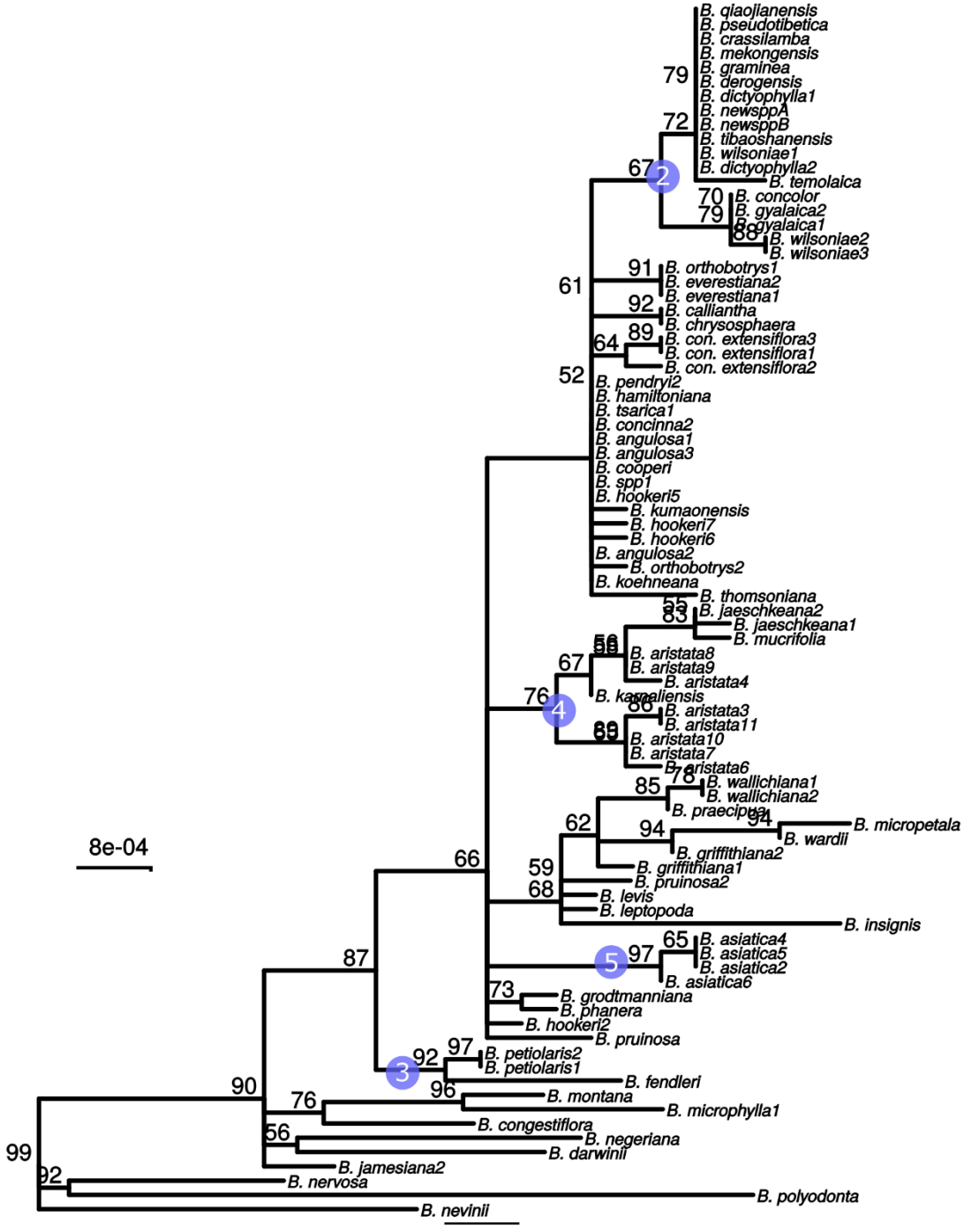


Figure 3-5 Maximum likelihood tree from the concatenated barcodes *matK*, SSC_noncoding2 and *ndhI-ndhG*. Nodes with bootstrap support <50 were collapsed to polytomies. Bootstrap values beween 50 – 99 are shown above branches. No number indicates a bootstrap value of 100. Numbered circles indicate groups that were recovered in the whole plastid phylogeny (see. Fig 3-1).

3.3.3 Testing barcodes

The minimal barcode consists of nine positions and includes barcodes unique to seven groups. No private SNPs were identified for groups 3, 6 and 8. No individual barcode for groups 6 and 8 could be constructed (Table 3-3). The barcodes were evaluated with the test data set (Table 3-3). The commercial samples Market1, Market2 shared the unique barcode of *Mahonia* samples. The sample Market11 shared the barcode with *B. asiatica* samples. The artificial mixtures were identified as comprising *B. asiatica* and *B. aristata.*

Table 3-4 *Top*: Matrix of informative barcode positions. The positions are relative to the consensus of the multiple sequence alignments of each barcode. "SA clade" stands for South American clade. *Bottom*: Results of the test samples. Market1, Market2 and Market11 are commercial samples and Mixture1 and Mixture2 are *in silico* mixtures. Numbers below multiple base calls represent the ratio of nucleotides in the mapping.

	matK				***ndhI-ndhG***			**SSC_ noncoding2**	
Position (bp)	755	857	976	1428	151	182	326	47	700
clade. 1	A	G	G	G	C	A	C	A	G
clade. 2	A	G	G	A	C	A	C	A	A
clade. 3	A	G	G	G	A	A	C	A	A
***aristata* – clade (4)**	C	A	G	G	C	A	C	A	A
***asiatica* – clade (5)**	A	G	G	G	C	C	C	A	A
clade. 6	A	G	G	G	C	A	C	A	A
clade. 7	A	G	A	G	C	A	C	A	A
SA clade (8)	A	G	G	G	A	A	C	A	A
***Mahonia* – clade (9)**	A	G	G	G	A	A	A	C	A
	Test Samples								
Market1	A	G	G	G	A	A	N	C	A
Market2	A	G	G	G	A	A	N	C	A
Market11	A	G	G	G	C	C	C	A	A
Mixture1	A C 1:2	A G 1:2.4	G	G	C	A C 1:1.6	C	A	A
Mixture2	A C 1:7.3	A G 1:9.3	G	G	C	A C 1:6.9	C	A	A

3.4 Discussion

DNA barcoding for quality assurance and pharmacovigilance has great potential and is likely to be implemented as a routine diagnostic method. In this study, we present an approach for barcoding an evolutionary complex group of species with limited availability of samples and successfully tested these barcodes on commercial samples.

With the emergence of new sequencing technologies, whole plastid sequencing has been proposed as an extension of the current barcoding concept (Coissac et al., 2016). It has been shown that whole plastid sequences increase phylogenetic resolution (Parks et al., 2009) and simultaneously increase the effectiveness of discriminating between species. In this study, we show how whole plastid next-generation sequencing can be used to investigate sequence variability patterns for the construction of informative DNA barcodes. We confirm the difficulty of barcoding *Berberis* species as suggested by Roy et al. (2010), even when whole plastid sequences are used for comparison. Although the sampling was limited, with only a few of the species represented with multiple samples, the low resolution of the plastid phylogeny at shallow phylogenetic levels and the presence of polyphyletic species (e.g. *B. aristata*) indicates evolutionary rather than methodological reasons for the failure of barcoding this genus to species level (Mutanen et al., 2016). DNA barcoding is challenging in groups where frequent hybridization occurs in conjunction with plastid capture or where lineage sorting has not yet been completed (Fazekas et al., 2009). Hybridisation in genus *Berberis* (Adhikari et al., 2012) could account for low barcoding success to species level. Evidence for chloroplast capture in *Berberis* is reported in Chapter 2 (Figure 2-8). Whole clade shifts from ancient hybridization events may influence sampling strategy since the species of interest have different sister clades in the nuclear phylogeny. Although some recent chloroplast capture events could be identified, this

process seems to have low influence on barcoding success of *Berberis.* A dramatic case of recent chloroplast capture and failure of barcoding is reported in genus *Salix* (Salicaceae) where haplotypes are shared even between subgenera and where dominant haplotypes are identified (Percy et al., 2014). Low resolution among closely related species, as reported in the whole plastid phylogeny, points towards lack of ancestral polymorphism or incomplete lineage sorting (Naciri and Linder, 2015). Finding suitable species-level barcodes for genera with low resolution, such as *Berberis*, may be possible by incorporating multispecies coalescent approaches (Degnan and Rosenberg, 2009). However, in the case of *Berberis*, where a target species is polyphyletic, a typological rather than an evolutionary approach is needed.

The case of barcoding medicinal *Berberis* species provides an example of how barcoding for regulatory purposes in an evolutionary complex group can be approached. Phylogenies can be essential for formulating adequate barcoding hypotheses; the whole plastid phylogeny reveals that at least three species are nested in the clade with the main species in focus. The polyphyly of *B. aristata* indicates that universal barcodes are unlikely to delineate these species. Furthermore, several clades show low resolution at terminal branches (e.g. Clade 1 + Clade 2, Figure 3-1). We have therefore adapted our classification scheme and defined meaningful operational phylogenetic units (OPUs) that do not correspond to existing species limits, numbered here in the phylogeny (Figure 3-1). The barcodes presented in this study derive from an integrative approach based on the interpretation of a whole plastid phylogeny, coupled with the detection of diagnostic nucleotides in relatively short barcodes for well-supported groups.

Distance-based and phylogenetic methods rely on a reference database, ideally containing all species of the group of interest with several individuals per species

(Meyer and Paulay, 2005; Raja et al., 2017). The dataset used in this study is therefore neither suitable for distance-based methods nor for pure phylogenetic barcoding methods, since only a fraction of the species was available, many only represented with one member (Figure 3-1). The barcode presented in this study is based on diagnostic nucleotides for monophyletic groups of species, referred to here as OPUs. Similar to morphological classification of species, diagnostic methods provide a set of unique characters to assign specimens to species or species groups (Little and Stevenson, 2007). The method has been implemented in various analysis tools (Sarkar et al., 2008; Weitschek et al., 2013), mainly for specimen identification. Some of the algorithms use logic mining techniques (Bertolazzi et al., 2009). Logic mining for DNA barcoding refers to a two-step process, where the barcode is first reduced to a set of very informative nucleotides and second a logic mining method is applied, where a set of formulas for separating the species are defined. More recent approaches, such as BLOG 2.0 (Weitschek et al., 2013), provide a diagnostic, character-based approach for species identification that are based on supervised machine learning. Character-based approaches circumvents analytical issues such as the nearest-neighbour problem in distance-based methods (DeSalle et al., 2005). The *in silico* mixtures presented in this study derive from samples that were used for producing the DNA barcode and are therefore not true test samples. However, the analysis shows the simplicity of analyzing mixed samples based on diagnostic nucleotides when shotgun sequencing data is available.

We believe that this approach is the way forward for regulatory purposes since the barcodes we present are intuitively understandable. DNA barcoding is beyond doubt a powerful method for specimen identification, but its implementation as a routine process for quality assurance (Sgamma et al., 2017) and pharmacovigilance (de Boer et

al., 2015) should consider the ease of application. Neither phylogenetic nor distance methods are appropriate, since they depend on large databases, sophisticated tools and lack objective criteria. For this reason, the British Pharmacopoeia (BP) approach is to present a sequence which samples must match for authentication. Pharmacopoeias ensure the safe use of pharmaceuticals by defining certain quality standards and DNA barcodes have recently been published in the BP for the first time (British Pharmacopoeia Commission, 2017). The question "does this sample belong to species XY?" is addressed by comparison to the pharmacopoeial sequence, since methods based on diagnostic nucleotides provide an easy and straight-forward way to answer the question. Identifying such sequences for inclusion in a pharmacopeia is the challenge addressed by this study. The whole plastid approach described here could become a model that can be applied to species that are difficult to resolve. Success depends on devising a sampling strategy that includes species that are closely related to the target species. Furthermore, the inclusion of distantly related, congeneric species increases the confidence in detected synapomorphic nucleotide polymorphisms.

Chapter 4 Perspectives on global trade and on the regulation of medicinal plants revealed by DNA barcoding

4.1 Introduction

Herbal medicines are a significant part of many traditional medicinal systems and are often partly integrated into mainstream healthcare systems. There is a significant market for so-called complementary medicine (CM) – healthcare practices not originating from a country's own tradition or from conventional medicine and not fully integrated in the country's medicinal system (WHO 2014). For example, in Britain, over a quarter of the population uses CM at least once a year (Thomas et al., 2001). The use of the term CM as a category of medicinal healthcare practices reflects the integration and commercialization of traditional medicinal systems on a global scale, and indicate an increasing popularity of healthcare systems like Traditional Chinese Medicine (TCM) or Ayurveda (Saks, 2008). The assimilation of healthcare practices including the use of herbal medicines is a dynamic process of interchange between local and global pharmacopoeias (Leonti and Casu, 2013). The globalization of herbal medicines creates a tension between the needs of regulatory agencies and pharmaceutical companies, and the systems from which the medicines are drawn. Complex trade networks form the link between the origins of the plant medicines and their users.

Herbal monographs describe medicinal species, providing tools for their identification and quality assurance. Pharmacopoeias are produced by public health authorities, such as the British Pharmacopoeia Commission, whose interest is to protect public health by assuring the correct composition and preparation of drugs (Marini-Bettolo, 1975). When pharmacopoeias incorporate herbal medicines, standards for raw

materials or a list of potential adulterants are documented in monographs and are essential in the process of quality and safety control undertaken by manufacturers and producers. The preparation of monographs requires accurate specimen selection so that methods for determining species according to the reference material are valid. This specimen selection process consists at first of an accurate translation from traditional knowledge to scientific terminology, for example, matching vernacular names to Linnaean taxonomy. Within the very first publications of state-produced pharmacopoeias, medicinal plants were described following Linnaean binomial nomenclature. For example, in the first publication of the British Pharmacopoeia (BP) in 1884, the pharmaceutical drug *Cocculus Indicus* was described as the fruit of *Anamirta cocculus*, an Indian plant from the family Menispermaceae. This example illustrates that regulation often acts at the level of the species, and that herbal medicines are defined according to Linnaean binomial nomenclature. Since the initial publication of the BP, several analytical tests for medicinal plants have been established. DNA barcoding methods are the most recent of these, and were first incorporated in the 2017 publication.

Local medicinal plant collectors are often at the starting point of medicinal plant trade (Olsen and Larsen, 2003), so plants may be expected to enter global trade under local, vernacular names. Although folk nomenclature is similar to the Linnaean nomenclature in that it adheres to a hierarchical classification system (Berlin et al., 1973), it is well reported that taxa in folk nomenclature do not necessarily match taxa recognized in the Linnaean system. In the Ayurvedic Pharmacopoeia for example, an estimate of 20,000 Sanskrit names are attributed to approximately 1,750 biological species and, similarly, many scientific taxa share the same vernacular name (Payyappallimana, 2008). This discrepancy may pose substantial problems for

establishing appropriate reference standards for pharmacopoeias. Global trade chains of medicinal plants are usually a convoluted network of actors and economic agents, involving thousands of harvesters, traders and manufacturers (Olsen and Bhattarai, 2005). Despite reports that the international trade of herbal medicines increased substantially in the two countries with the largest herbal medicine export volume, India (Scindia, 2010) and China (Liu et al., 2009), relatively little is known about the complexities of international trade chains. Particularly, the attached implications for regulatory bodies, for the pharmaceutical industry and for public health are unknown. DNA barcoding offers the possibility to identify trade samples at different points in a supply chain, allowing for an investigation into whether species composition reflects the ethnotaxonomy of local markets.

In this study, we analyse globally- and locally-traded samples that are putatively of two species: *Berberis aristata* DC. and *Phyllanthus amarus* Schumach. & Thonn. Both are used in the Ayurvedic system of medicine (Ayurvedic Pharmacopoeia of India, 2001; Patel et al., 2011). In that system, *B. aristata* is referred to as *Daruharidra* (दारुहरिद्र), among other names (Ayurvedic Pharmacopoeia of India, 2001) and *P. amarus* as *Bhoomyaamalakee* (भूम्यामलकी; Patel et al., 2011). However, both species have numerous vernacular names on the Indian subcontinent (http://www.medicinalplants.in/; last accessed 15/08/2017), and the names may be applied to these and to other closely related species. The samples were analysed using two different next-generation sequencing approaches to DNA barcoding methods, a target enrichment approach with phylogenetic inference for *B. aristata* samples, and a DNA metabarcoding approach (Taberlet and Coissac, 2012; Yu et al., 2012) to identify the entity of species in potentially mixed samples for *P. amarus*. The aim was to identify species in global trade, and interpret findings in terms of the nature of species

adulteration and substitution, of local versus global markets and of the translation of vernacular names to Linnaean taxonomy.

4.2 Material and methods

4.2.1 Sampling

For the analysis of *Berberis*, 16 samples were purchased, two of which were from local markets in Kathmandu, specifically herbal shops used by locals. Six samples from India were purchased via the Internet and ten samples were bought in the UK (see Table 4-1 for further details). The material purchased in the UK were business to business samples, destined for the UK market place.

Table 4-1 *Berberis* trade samples.

Sample	Form	Company	Place of Purchase
Market 1	Stem/Bark/Root	UK_1	UK
Market 2	Stem/Bark/Root	UK_1	UK
Market 3	Stem/Bark/Root	UK_2	UK
Market 4	Stem/Bark/Root	UK_2	UK
Market 5	Stem/Bark/Root	UK_2	UK
Market 6	Stem/Bark/Root	UK_2	UK
Market 7	Stem/Bark/Root	UK_2	UK
Market 8	Powder	UK_3	UK
Market 9	Stem/Bark/Root	KTM_1	Nepal, Kathmandu
Market 10	Stem/Bark/Root	KTM_2	Nepal, Kathmandu
Market 11	Powder	India_1	India, Rajasthan (Internet)
Market 12	Powder	India_2	India, Mumbai (Internet)
Market 13	Powder	India_2	India, Mumbai (Internet)
Market 14	Powder	India_3	India, Uttarakand (Internet)
Market 15	Powder	India_4	India, Surat (Internet)
Market 16	Stem/Bark/Root	India_5	India, unknown (Internet)

For the investigation of *Phyllantus*, ten globally-traded products from four companies in the UK were sampled. From each, three representative sub-samples were taken (Table 4-2).

Table 4-2 *Phyllanthus* trade samples

Sample	Form	Company	Place of Purchase
Product 1	Twigs	UK_4	UK
Product 2	Whole plant	UK_5	UK
Product 3	Whole plant	UK_6	UK
Product 4	Whole plant	UK_6	UK
Product 5	Whole plant	UK_7	UK
Product 6	Whole plant	UK_7	UK
Product 7	Whole plant	UK_7	UK
Product 8	Whole plant	UK_8	UK
Product 9	Whole plant	UK_8	UK
Product 10	Whole plant	UK_8	UK

4.2.2 *Berberis* phylogeny

The traditional barcode regions *ndhF* or ITS have limited discriminatory power in genus *Berberis* (Adhikari et al., 2015). Previous analysis revealed that the targeting of 396 nuclear DNA regions resulted in a well-resolved phylogeny of the genus *Berberis* (Chapter 2). This method was therefore used in the analysis of these 16 market samples. The DNA marker development, library preparation, sequencing and marker assembly are described in Chapter 2. Altogether, 43 accessions of *Berberis* and 16 market samples were included in the analysis. *Berberis* samples were mainly from simple-leaved Himalayan species of the group Septentrionales. Furthermore, three exemplars of the compound-leaved *Mahonia* were included. Revised generic limits now recognize *Mahonia* as *Berberis* (Mabberley, 2008; Marroquin and Laferriere, 1997), but we refer

to *Mahonia* here, following accepted names in the Plant List (www.theplantlist.org; last accessed 15/08/2017). The naming of *Berberis* samples used in the phylogeny are the same as in Chapter 2 (for specimen information, see Appendix table AT-1). The phylogeny was reconstructed with the concatenated DNA marker alignments and maximum likelihood tree estimation was performed using 100 rapid bootstrap replicates and under the GTRGAMMA substation model in RAxML 8.2.9 (Stamatakis, 2014).

4.2.3 *Phyllanthus*: DNA metabarcoding

The dataset was generated as described in Sgamma et al. (2017). Modification to the data analysis is made here, and the data are analysed and interpreted for the first time in the context of local versus global markets and of the translation of vernacular names. In short, each of the ten products (P1 – P10) was sampled multiple times, resulting in 43 samples. DNA from each subsample was extracted and the regions *trnH-psbA*, ITS2, *trnL-F* and *rbcL* were amplified and sequenced on an Illumina MiSeq sequencer. The bioinformatics pipeline applied was as described in Sgamma et al. (2017) except that, in order to aid interpretation, numbers of reads matching a taxonomic category were summed across subsamples. The resulting species abundance table is graphically displayed as a heatmap (Figure 4-2).

4.3 Results

Overall, the phylogenetic placement of traded *Berberis* samples suggests that market samples are either placed in the clade with *B. asiatica* or in the clade with *B. aristata* and *B. karnaliensis* specimens. Based on this result, we identify the two clades as the "asiatica" and "aristata" clades (Figure 4-1). However, two market samples are apparently *Mahonia* spp.

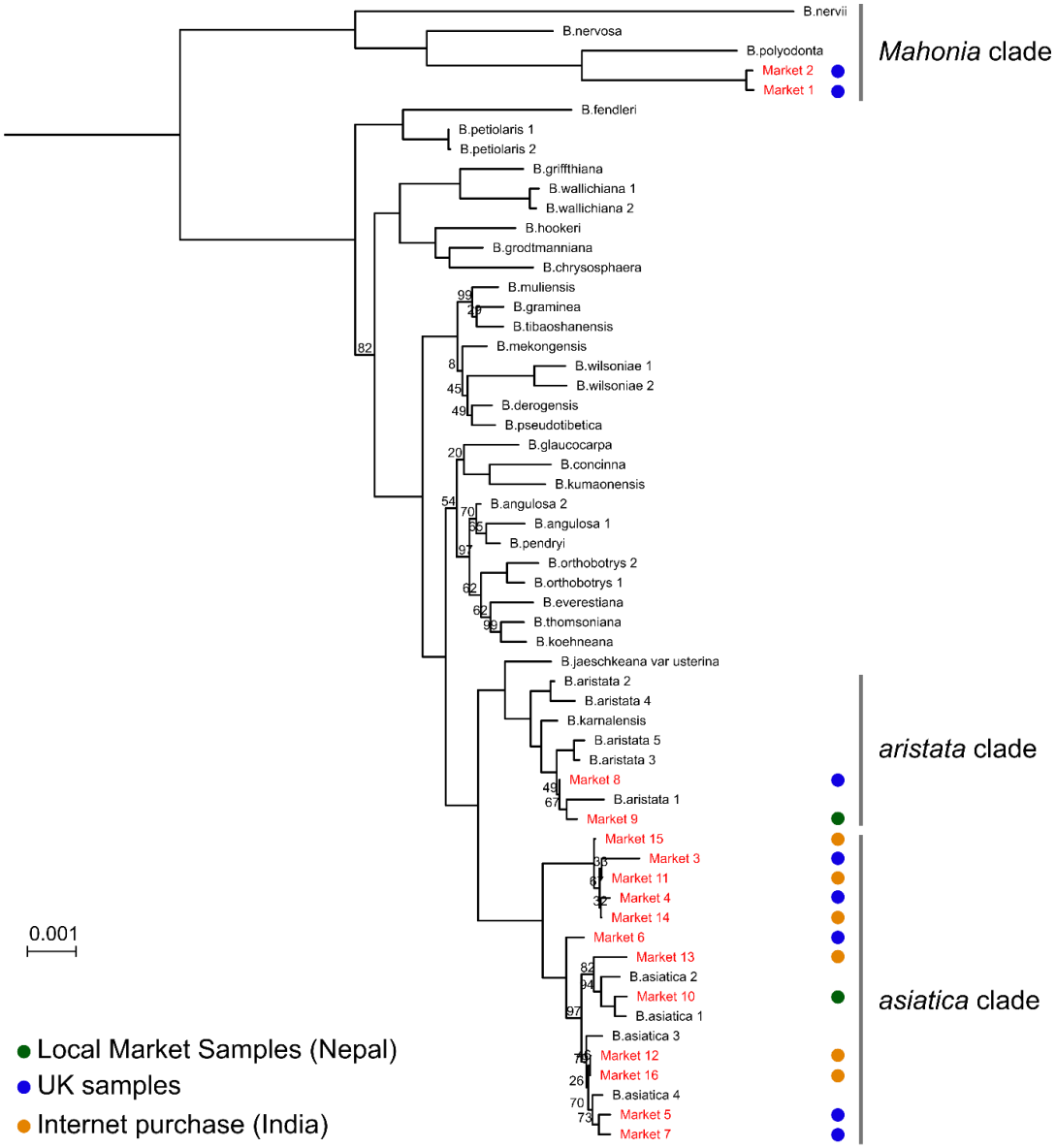


Figure 4-1 Phylogeny of 43 *Berberis* species and 16 market samples. The labels of market samples are coloured in red. Only bootstrap values < 100 are shown (numbers above branches). Coloured dots represent where the samples have been bought. In the case of UK samples, the provenance of the raw material is unknown and were purchased in the UK.

Identifications of samples sold as *Phyllanthus amarus* are shown in Figure 4-2. Four of the products consisted only of species of *Phyllanthus*, while the other six included species from other plant families, and even organisms from other major groups such as fungi and bacteria. Three species of *Phyllanthus* were identified in these samples, *P. amarus*, *P. debilis* and *P. maderaspatensis*. Product P1 consists mostly of

P. maderaspatensis and P2 and P7 are products showing admixture with either *P. maderaspatensis* or *P. debilis*. Products P4 – P6 could be considered as homogenous *P. amarus* samples since there are reads which match to this species but not to others. Matches to genus *Phyllanthus* are indicative of sequences that do not discriminate between the species of *Phyllanthus* in the database.

Although P3 shows a response to Cyperae, this may be explained by a potentially unintentional contamination in the product. The only product where no reads from the genus *Phyllanthus* were identified is P8, suggesting a bad quality sample. Overall, reads that indicate contamination or substitution were from Acanthaceae and Amaranthaceae, *Amaranthus*, *Corchorus* (Malvaceae), *Sida spinosa* (Malvaceae), *Indigofera suffruticosa* (Leguminosae), Malvaceae, and *Dactyloctenium aegypticum* (Poaceae).

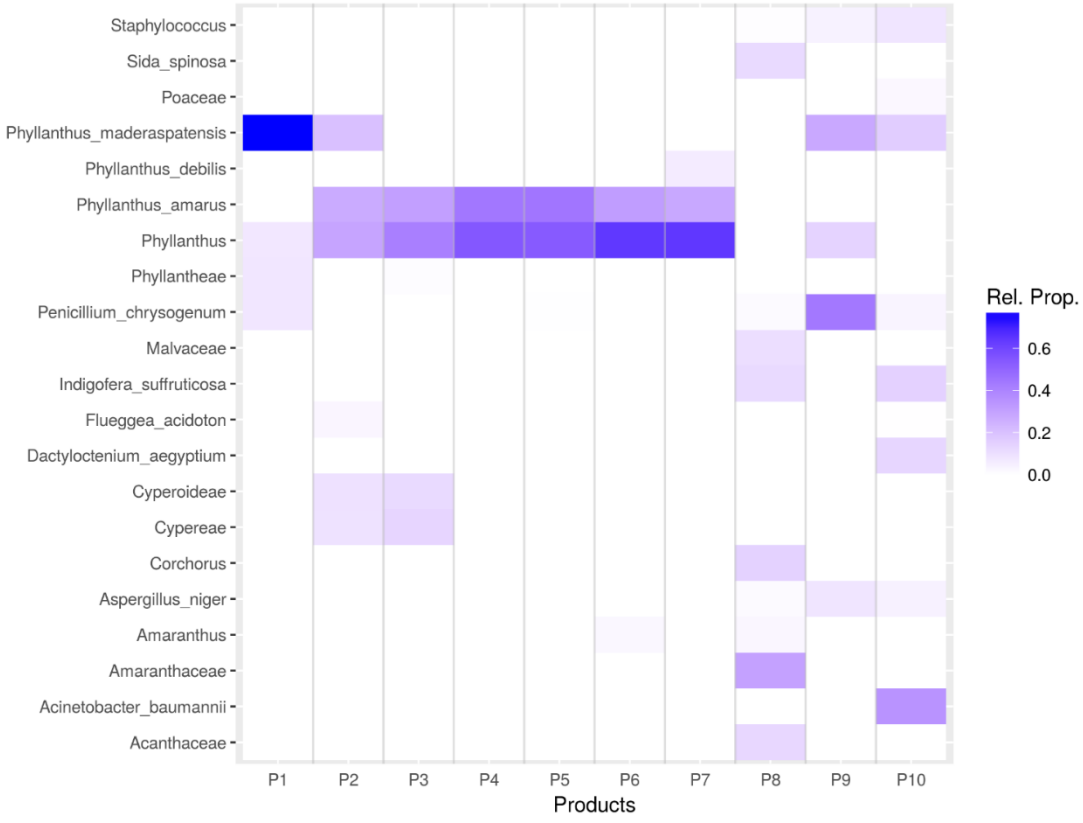


Figure 4-2 Heatmap of species identified. The relative abundance of reads per product mapping to species are represented with the intensity of colour.

4.4 Discussion

4.4.1 Species composition of globally traded products

The analysis of traded material exposes that products marketed as *Phyllanthus* and *Berberis* contain several congeneric species. The DNA metabarcoding analysis reveals that several products contained more than one species of *Phyllanthus*, which suggests species admixture either at the harvesting and/or bulking stage. The analysis also shows that the overall quality between products varies dramatically from homogenous *P. amarus* products (P4 – P7) to a product without any evidence for *Phyllanthus* species (P 10). The phylogenetic approach for analyzing traded *Berberis* species shows that mainly *B. asiatica* is in trade. The two local samples were either *B. aristata* or *B. asiatica*, which indicates that both samples are sold and used in a local context. While most of the samples are either *B. asiatica* or *B. aristata*, we found *Mahonia* in the samples from one one company (UK_1).

In the wider context of medicinal plant use and trade, the use of congenerics is well-described. Generic complexes, the inclusion of several botanical species under one vernacular name (Berlin et al., 1973), are a global phenomenon. Linares and Bye (1987) documented a series of generic complexes in markets from Mexico and the adjacent United States and identified a so-called label species in each complex. The label species is the most prevalent species in trade and could therefore be used for labelling or naming the species complex. The characteristics of the label species is that this species is of high value and is traded well beyond its natural occurrence and is substituted occasionally by local plants which themselves are not traded in quantity beyond the limits of their distribution. While the complexes reported by Linares and Bye (1987) often included species from several genera and even families, there are well-known

examples of generic complexes among closely related species, for example members of the genus *Salvia* subgenus *Calosphace* (Jenks and Kim, 2013). By applying the concept of a label species as either the most widely distributed or most commonly identified species (Ouarghidi et al., 2012), *P. amarus* appears to be the label species of *Phyllanthus* and *B. asiatica* of *Berberis*.

4.4.2 Global trade mirrors local markets

For both medicinal complexes studied in this paper, the results of local market surveys are available in the literature for comparison. In only two out of ten medicinal plant markets that sold *Daruharidra*, the species identified was *B. aristata*, which is, according to official pharmacopeias, the correct species. On the other hand, most other markets sold mainly *B. asiatica* (Srivastava and Rawat, 2013). Furthermore, ethnobotanical field studies and market inventories list either *B. aristata* (Acharya and Rokaya, 2005; Subedi and Panderey, 2011; Tiwari et al., 2004), *B. asiatica* (Joshi and Joshi, 2000; Kunwar et al., 2013; Uprety et al., 2010) or both (Humagain and Shrestha, 2009). In a study where *Phyllanthus* market samples were studied, 19 out of 25 samples contained almost purely *P. amarus* and, in the remaining six samples, five other *Phyllanthus* species were determined, including *P. maderaspatensis* and *P. debilis*. These species were also identified in our study in globally traded samples. The high congruence of species found in local and global trade exemplifies that the latter is highly dependent on local structures. This finding has major implications for regulation and for the development of medicinal plant markets. Furthermore, it has been observed in local trade networks in different parts of the world (Mander, 1998; Olsen and Bhattarai, 2005) that potentially thousands of local harvesters and traders form part of the value chain of medicinal plants. Regulation may affect the livelihood strategies of

these economic agents in local trade systems by restricting and devaluing potentially equivalent species.

4.4.3 Generic complexes and concepts of substitution

Species adulteration and substitution in herbal medicines is a widely documented phenomenon (Srirama et al., 2017), and a series of severe cases of poisoning through deliberate or unintentional adulteration and substitution have been reported, such as that of infants following consumption of adulterated star anise tea (Ize-Ludlow, 2004). Adulteration of herbal medicines describes accidental or intentional variation in identity, strength or purity of herbal remedies. Substitution is a special case of adulteration where substances are replaced by another, potentially less expensive substance (Foster, 2011).

In practice, the terms ‘substitution’ and ‘adulteration’ in herbal medicines seem to embrace two concepts. On the one hand, the terms are used to describe products of reduced quality that potentially threaten the health of consumers for various reasons, such as the mislabelling of ingredients, the existence of unlabelled filler species, contamination with other plant species or the addition of pharmaceuticals or chemicals. (Coghlan et al., 2015, 2012; Newmaster et al., 2013; Posadzki et al., 2013). On the other hand, the terms are commonly applied in scientific literature when traded medicinal plants are not congruent with published species in pharmacopoeias (e.g. Srirama et al., 2017). In that sense, our analysis of *Berberis* samples confirms the previously reported species substitution for *B. aristata* with *B. asiatica.* (Srirama et al., 2017; Srivastava and Rawat, 2013). However, we propose an alternative interpretation of our results: traditional herbal medicinal systems such as Ayurveda undergo a process of modernization through standardization (Banerjee, 2008), which is a central principle

within modern pharmaceutics. This process is also referred to as the pharmaceuticalization of Ayurveda (Banerjee, 2008). An example of this process is the translation of traditional preparation techniques to mass production processes accompanied by standardized quality controls. This stands in contrast to traditional use, where, even in written medicinal systems like Ayurveda, norms of preparation of herbal medicines are more varied (Banerjee, 2008). We argue that the mechanisms of translating vernacular names to scientific taxa in official pharmacopoeias, such as the Ayurvedic Pharmacopoeia or the British Pharmacopoeia, can be understood as an attempt to modernize and standardize traditional knowledge. The central question is whether *Daruharidra* is directly translatable to only *B. aristata* or whether it is a generic complex, consisting of several species of Himalayan members of genus *Berberis.*

The evidence presented in this paper, in conjunction with local market studies and ethnobotanical field studies, favours the view that species from the *aristata* and *asiatica* clade are equivalent in local use and in international trade. Consequently, we conclude that species from both clades should be recognized as members of the generic complex *Daruharidra*. Whether *Mahonia* forms part of the *Daruharidra* generic complex is unclear. Although no *Mahonia* species are recognized in the Ayurvedic Pharmacopoeia of India (Welfare Ministry of Health and Family, 2010), there are reports of local uses of *M. nepaulensis* in Nepal (Shrestha and Dhillion, 2003; Uprety et al., 2012), where ophthalmological use is congruent with the therapeutic application of *Berberis* species. However, other therapeutic applications differ. Nepal is a major exporter of medicinal plants to the Ayurvedic industry in India (Olsen, 1998) and *Berberis* is traded from Nepal to India. In Nepali language, the vernacular name for *Mahonia* is *Jamanemandro* (जमानेमान्द्रो) and most *Berberis* species share the common

name *Chutro* (चुत्रो). Given the evidence that local uses for *Mahonia* differ in some aspects from the use of *Berberis,* in morphological distinctiveness and in different vernacular names, we favour the view that *Mahonia* is a substitute for *B. aristata* and *B. asiatica.* Substitution can be legitimate practice in cases where the substituent has similar traditional uses. For example, *Heterotheca inuloides* is reported to be an appropriate substitute for *Arnica montanta* (Gafner and Applequist, 2016). DNA barcoding and other species identification techniques in conjunction with ethnobotanical field and market studies provide a starting point for pharmacognostic research. In the case of *Mahonia* spp., further research is needed to determine differences in phytochemistry and clinical efficacy.

The identification of the substitution or the adulteration of globally-traded herbal medicines that are used in a transcultural context is dependent on interdisciplinary research. Local and global market studies, ethnobotanical field studies and interpretation of traditional knowledge, together with accurate species identification, are fundamental for an adequate regulatory framework.

Chapter 5 Impact of targeting paralogues on phylogenomic inference

5.1 Introduction

A basic requirement for phylogenetic inference is the use of orthologous DNA sequences. In the era of next-generation sequencing (NGS), where hundreds or thousands of genomic loci are compared, a major concern is that misleading phylogenetic signals are introduced by unidentified paralogues (Philippe et al., 2011). In the absence of assembled whole genome sequences for non-model organisms, and given the relatively high financial and labour investments needed for producing those for large-scale organismic studies, researchers use different techniques to target phylogenetically informative genetic loci for phylogenomic inference (Cronn et al., 2012). One widely used method in phylogenomics is target enrichment via hybridization capture (e.g. Mandel et al., 2014; Schmickl et al., 2015; Stephens et al., 2015; Weitemier et al., 2014). Hybridisation probes (hereafter referred to as probes) are usually RNA oligonucleotides that bind to complementary DNA sequences of interest. Probes are designed by mapping sequence data from one or more transcriptomes of a target species or a close relative to scaffolds from a draft genome assembly (Schmickl et al., 2015; Weitemier et al., 2014). Another approach is to compare transcriptome data from a target species to published orthologues sequence databases (e.g. Wanke et al., 2016, Mandel et al., 2014), such as the Putative Orthologous Groups Data Base (http://cas-pogs.uoregon.edu/; last accessed 15/08/2017).

Whatever means are employed to design probes for hybridization capture, it is not possible to completely avoid targeting multi-copy regions if complete, assembled genomes are unavailable. For example, when a transcriptome is mapped to a draft

genome, putative single copy regions are filtered by selecting sequences which map only once to the target. However, since only parts of the genome are known, any putative single copy region may in fact be represented by several copies in the genome. Furthermore, because a small number of species are used for marker development and capture is extended to a large number of species, gene duplication events may have occurred outside of the set of species used for probe development. In these scenarios, capture of paralogues can occur because the hybridization probes allow for some degree of mismatch. In that case, the resulting DNA reads will therefore stem from several copies within a genome. The presence of reads derived from multi-copy regions might falsely attribute variation between paralogues to within-region (or allelic) polymorphism when the DNA sequences are assembled. Read assembly errors in the presence of unidentified paralogy are therefore perceived as a potential source of error (e.g. Faircloth, 2015; Johnson et al., 2016; Nicholls et al., 2015; Prum et al., 2015).

Given the challenges of excluding the capture of paralogous loci *a priori*, researchers have put substantial effort into devising analyses to eliminate read assembly errors due to paralogy. Several methods have been proposed for minimizing the effect of paralogous read mapping to the reference. For example, in analysis pipelines where *de novo* assembly of reads from a targeted locus is performed, paralogy is assumed when more than one assembled contig matches equally well to the reference (e.g. Faircloth, 2015; Heyduk et al., 2016). In pipelines where only read mapping has been performed, increased read mapping stringency has been proposed as a way to reduce mapping of relatively distant paralogues to the reference. Furthermore, the analysis of levels of variation within locus alignments across all samples, or the investigation of levels of coverage have been proposed (Nicholls et al., 2015).

We use a simulated *in silico* target enrichment study of genus *Arabidopsis* and investigate the effect of having reference sequences that derive from single-copy genes and paralogous gene clusters. We selected 500 nuclear loci that are known to be single-copy orthologues (hereafter referred to as orthologous markers) and 666 loci from paralogous gene families (referred to as paralogous markers) and used them as reference loci (referred to as markers). In particular, the following hypotheses are tested:

1. A method used to filter loci which are putatively contaminated with paralogous copies relies on sequencing coverage. The assumption is that, if several paralogous copies of the targeted sequence are present in the genome, the average sequencing coverage should be higher in comparison to true orthologous loci (Nicholls et al., 2015). We test this hypothesis by comparing coverage patterns between orthologues and paralogues.
2. The second method is based on the prediction that, if reads derive from several paralogous copies, a higher density of SNPs will be observed in these loci (Zhou and Holliday, 2012). These different SNP densities should lead to increased summed branch lengths in gene trees. We evaluate this assumption by calculating the summed branch length of the single gene trees.
3. The method described in Chapter 2 is based on the assumptions that, if reads derive from several paralogous copies, the sequence divergence of the two alleles per enriched marker is higher than if the reads derive from only a single-copy gene. Similarly, in the case that these alleles are true allelic variants, their phylogenetic distance within all allelic variants from a given marker should be smaller than if the putative allelic variants represent paralogous copies. We apply the method developed for filtering markers applied to the *Berberis* dataset

(Chapter 2) to the *Arabidopsis* dataset. We evaluate whether the mean sequence similarity and the mean phylogenetic distance between pairs of alleles can be used to distinguish orthologous and paralogous loci in *Arabidopsis*.

4. Phylogenies derived from markers with paralogous copies in the genome should be incongruent with phylogenies derived from orthologous markers. We infer maximum likelihood phylogenies of concatenated orthologous and paralogous markers and compare their topologies. The estimated tree from orthologous markers is considered the true tree.

5.2. Materials and methods

5.2.1 Sampling

Shotgun sequencing data from members of genus *Arabidopsis* were downloaded from The Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra; last accessed 07/08/2017). In total, shotgun sequencing data from 18 samples were included in this study, of which six belong to *A. thaliana* (Table 5-1). Only diploid species were included.

5.2.3 Marker assembly and allele reconstruction

Raw reads were quality trimmed using Trimmomatic v. 0.33 (Bolger et al., 2014) and reads shorter than 80 bp were discarded. The filtered reads were then mapped to the reference markers with Bowtie2 with default options (Langmead and Salzberg, 2012). The resulting BAM files were sorted and phased with SAMtools (Li et al., 2009). Phasing allows the reconstruction of two alleles based on read alignments in the BAM files (He et al., 2010). The phased BAM files were then used as an input to call variants with bcftools and transform each sequence alignment to fastq files with

'vcfutils vcf2fastq' (Danecek et al., 2011). Fastq sequences were converted to fasta format with seqkt (https://github.com/lh3/seqtk; last accessed 07/08/2017). The consensus sequences of each allele pair were reconstructed using EMBOSS (Rice et al., 2000). Thus, our resulting dataset comprised two alleles (allele 1 and allele 2) and a consensus sequence of each allele pair for 1,166 markers from 18 samples.

Table 5-1 *Arabidopsis* samples used in this study. The number of reads comprises forward and reverse reads.

Sample	Species		subsp.	Sequence Read Archive (SRA)	Number of reads
A_aren	*Arabidopsis*	*arenosa*		SRR4128971	367,133,398
				SRR4128972	259,277,800
				SRR4128973	291,971,246
				SRR4128974	284,998,284
A_cebe	*Arabidopsis*	*cebennensis*		SRR2040777	52,939,764
A_croa_1	*Arabidopsis*	*croatica*		SRR2040778	40,637,296
A_croa_2	*Arabidopsis*	*croatica*		SRR2040779	48,297,660
A_hall	*Arabidopsis*	*halleri*	*halleri*	ERR1760144	221,844,272
				ERR1760145	316,000,000
				ERR1760146	310,713,334
				ERR1760147	326,626,720
A_lyra	*Arabidopsis*	*lyrata*		SRR5003828	429,883,364
A_negl_ro	*Arabidopsis*	*neglecta*	*robusta*	SRR2040831	104,821,966
A_neg_1	*Arabidopsis*	*neglecta*		SRR3111444	61,238,438
A_neg_2	*Arabidopsis*	*neglecta*		SRR3111445	33,800,834
A_neg_3	*Arabidopsis*	*neglecta*		SRR3111446	65,145,234
A_petr	*Arabidopsis*	*petrogena*		SRR2040833	86,147,774
A_thal_1	*Arabidopsis*	*thaliana*		SRR2626429	320,224,586
A_thal_2	*Arabidopsis*	*thaliana*		SRR3166543	324,725,120
A_thal_3	*Arabidopsis*	*thaliana*		SRR4136216	404,278,916
A_thal_4	*Arabidopsis*	*thaliana*		SRR4136238	385,596,896
A_thal_5	*Arabidopsis*	*thaliana*		SRR4136242	425,561,556
A_thal_6	*Arabidopsis*	*thaliana*		SRR4146470	446,791,922
A_unez	*Arabidopsis*	*unezawana*		SRR2040810	48,670,190

The average coverage for each marker per sample was calculated with SAMtools using the 'samtools depth' command. Since the coverage within samples fluctuated substantially, the largest outliers (n=66) were removed for scaling reasons before producing the boxplots.

The assumption of the phasing algorithm implemented in SAMtools is that the input data stems from diploid organisms, since the algorithm tries to find only two copies. The sequence similarity between each pair of sister alleles was calculated with a custom python script and averaged for each marker. The sequence similarity score between two alleles was calculated as

Score = m/(t-g-N);

where m is the number of nucleotide matches (without gaps); t is the length of the alignment; g is the number of gaps; and N is the number of columns where data was missing in at least one sequence. Since missing data and gaps are often prevalent in target enrichment data sets, a mismatch penalty would be inappropriate for assessing the sequence similarity. The resulting data set contained the marker ID, average sequence similarity and the standard deviation of the sequence similarity. The mean phylogenetic distance was calculated as described in section 2.2.3.4.

5.2.5 Tree reconstruction

Three different datasets were built: first, the orthologous marker dataset, where alignments were of orthologous markers; second, the paralogous marker dataset, where only alignments from paralogous markers were included and finally, the combined dataset, consisting of all markers. The following analysis was repeated with each dataset. First, maximum likelihood trees from the individual gene alignments were estimated using RAxML v. 8.2.9 (Stamatakis, 2014) with 100 rapid bootstrap

replicates. The summed branch lengths for each single gene tree were calculated using a custom bash script. The resulting trees served as the input trees for inferring the majority consensus extended trees where gene support frequencies (GSF) are calculated. Alignments of all markers of a given dataset were concatenated using phyutility v.2.2.6 (Smith and Dunn, 2008). The concatenated datasets were used to estimate the maximum likelihood tree in RAxML v. 8.2.9 (Stamatakis, 2014) with 100 rapid bootstrap replicates with partitions for each marker. The resulting phylogenies were inspected by eye for topological differences.

5.3 Results

The average coverage between orthologues and paralogues is shown in Figure 5-1. There is no apparent difference in coverage between orthologues and paralogues. However, in some cases, the paralogues have lower coverage. Similarly, no differences between the summed branch lengths for gene trees derived from orthologues or paralogues were apparent (Figure 5-2), suggesting that similar levels of variation are present in both datasets. The data generated in applying the pipeline devised for the *Berberis* dataset and described in Chapter 2 are shown in Figure 5-3. Figure 5-3A shows that orthologues and paralogues cannot be distinguished in terms of the sequence divergence between putative alleles as calculated using a pairwise similarity approach. Neither can they be distinguished using phylogenetic distances between putative allelic pairs derived from phasing (Figure 5-3B).

The maximum likelihood trees resulting from analyses of concatenated datasets are shown in Figure 5-4. Comparison of topologies suggests that they are congruent whether alignments derived from orthologous markers, from markers known to include

paralogous copies, or from both combined. However, gene support frequencies within *Arabidopsis thaliana* were lower in the phylogeny inferred from paralogous markers.

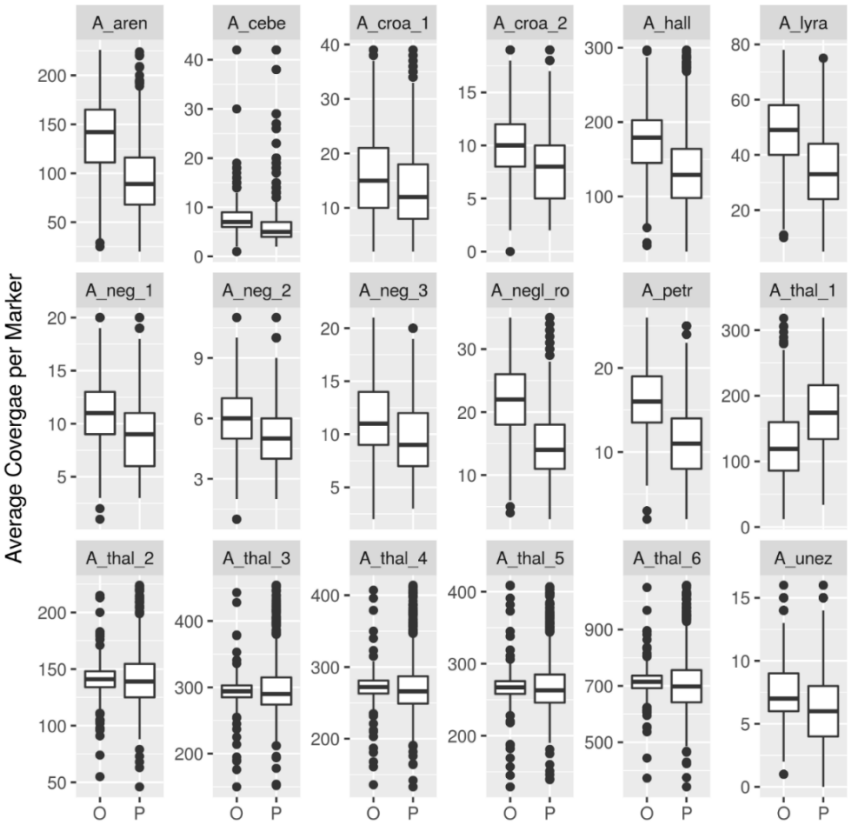


Figure 5-1 The boxplots describe the average coverage of orthologous (left) and paralogous markers (right) per sample. For reasons of scaling the individual plots, the highest 66 data points per sample were removed prior to producing the boxplots (final data set: n = 1,100 markers per sample).

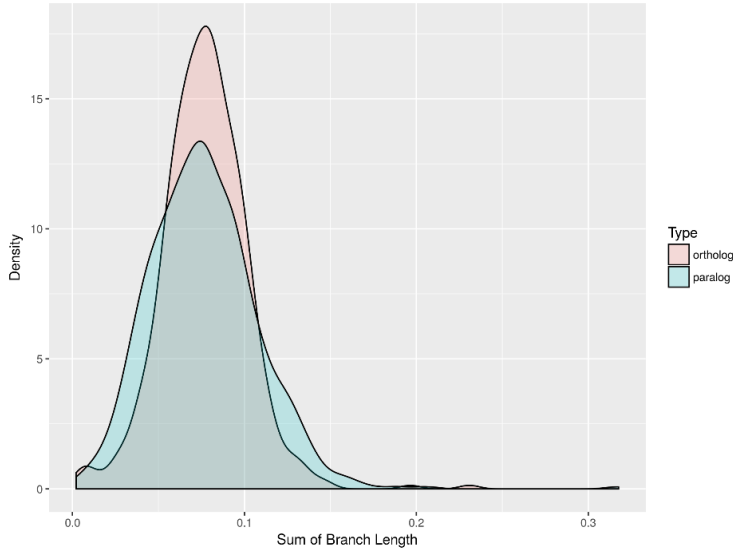


Figure 5-2 Density plot of sums of branch length.

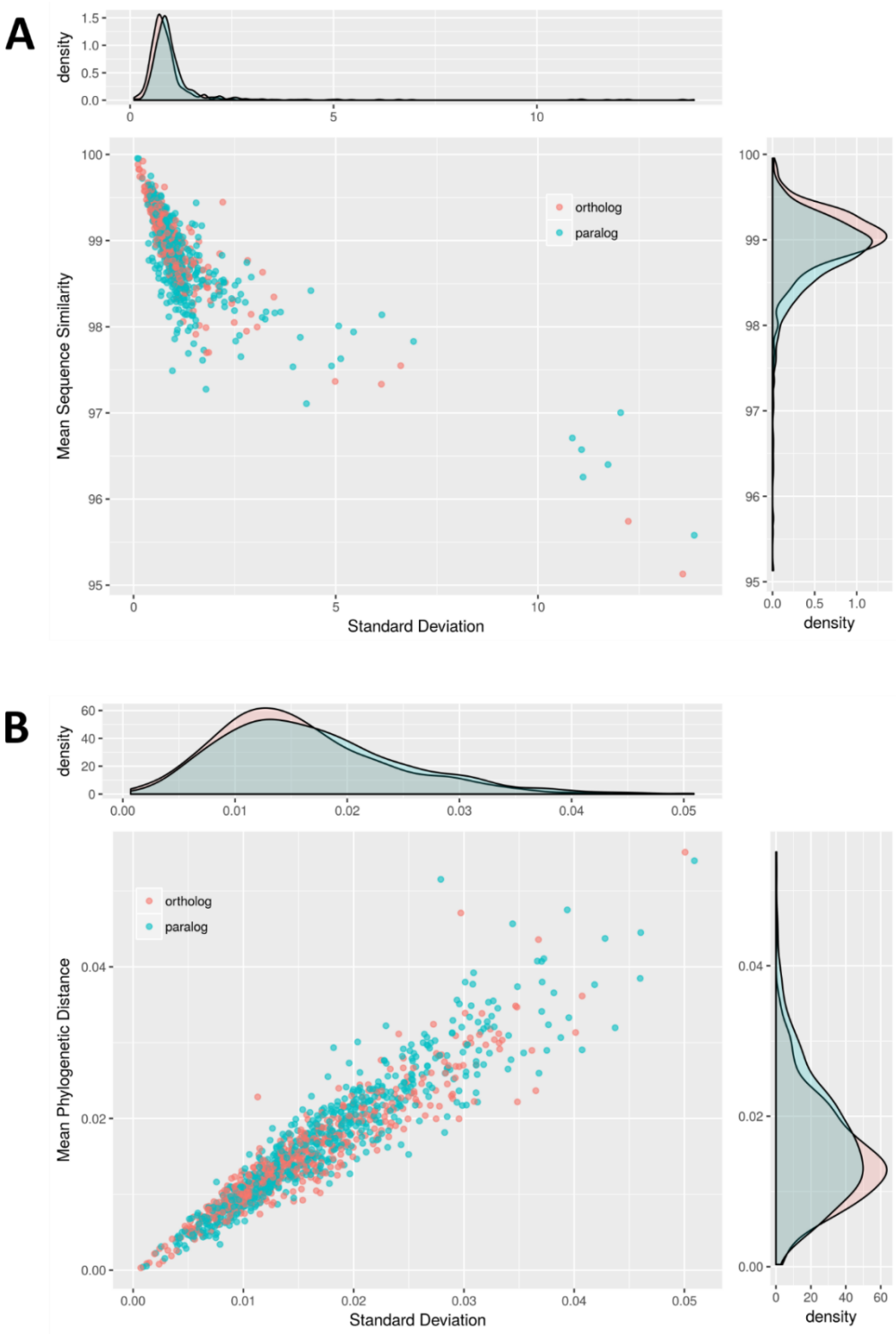


Figure 5-3 *A*: Mean and standard deviation of sequence similarity of each marker, calculated from the pairwise distances between pairs of alleles. *B*: Mean and standard deviation of phylogenetic distance of pairwise alleles of each marker. Density plots on top and right of the scatter plots describe the distribution of points along the respective axis.

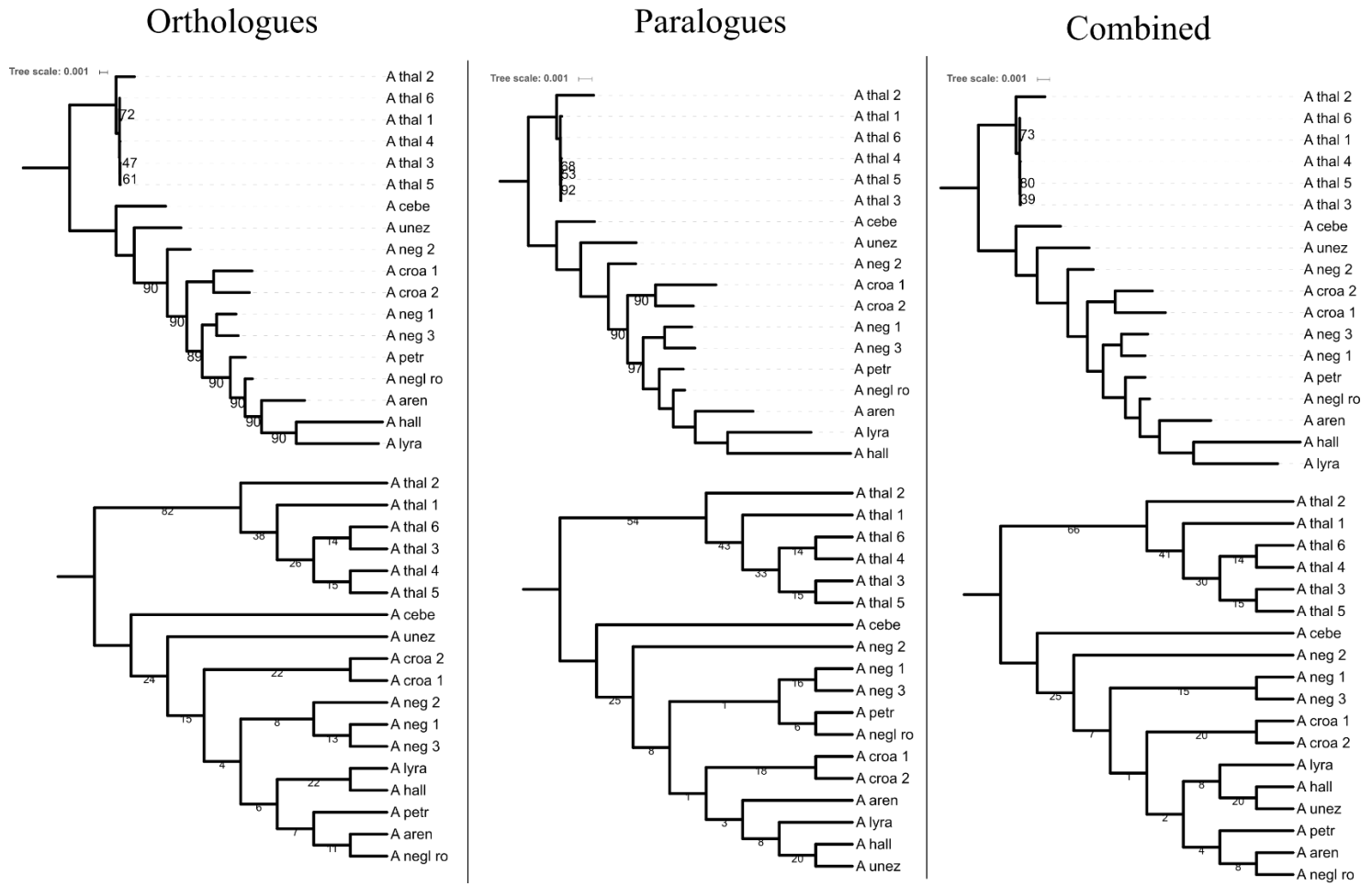


Figure 5-4 Phylogenetic trees of concatenated orthologous, paralogous and the combined marker dataset. Top row: Maximum likelihood tree with 100 bootstrap replicates. Numbers below branches are shown when the bootstrap support is lower than 100. Bottom row: Majority rule extended consensus tree. Branch lengths are ignored. Numbers above branches are gene support frequencies (GSF) in percent.

5.4 Discussion

In this study, we analysed the effect of assembling genes for phylogenomic inference from reference sequences that are either known to be single-copy genes or belong to known paralogous clusters. Furthermore, we tested whether some of the methods used for identifying assemblies that are contaminated with reads from paralogous genes are reproducible in *Arabidopsis*.

The paralogy problem has concerned phylogeneticists since the advent of molecular systematics (Doyle, 1992). When phylogenetic reconstruction depended on the generation of data from one or few loci, mistaken orthology could result in the reconstruction of topologies fundamentally incongruent with the true species phylogeny. In this context, mistaken orthology was the result of direct comparison of an orthologous sequence with a paralogous one. The resulting topologies might be identified as problematic if, for example, one gene tree conflicted with high node support with another. Spurious long branches might also highlight the inclusion of paralogous sequences (Struck, 2014). In the context of phylogenomics, the effect of paralogues on tree inference has been demonstrated in several cases (Struck, 2013), and Philippe et al. (2011) list the incorrect identification of orthologues as a source of non-phylogenetic signal. The starting points of these studies are fully reconstructed gene sequences, where the sequences are each found in the organism, but the relationship between the sequences is not necessarily one of orthology. In this study, the focus is on the implications for phylogenetic reconstruction of the failure during assembly to discriminate paralogues. The assembly of genes from sequencing reads is a crucial step and the effect of contamination on these assemblies with reads from untargeted loci is

perceived as significant (e.g. Nicholls et al., 2015; Prum et al., 2015; Chapter 2), since it may lead to false positive base calls.

Here, we use four measures (coverage, summed branch length of gene trees, allelic distances and phylogenetic infernece) to determine the effect of inclusion of known paralogues in terms of sequences assembled. None of the four measures identify any effect of inclusion of known paralogues in a phylogenomic dataset.
Coverage patterns often fluctuate considerably from the theoretical expectations and are influenced by the DNA sequence and other biases introduced in the lab work (Sims et al., 2014). Nevertheless, the average coverage of a targeted gene with cognate paralogues should, in theory, be higher because the mapped reads derive from two different regions. The coverage patterns reported here exhibited no difference between the orthology and paralogy datasets. This could be explained by the reads from only one paralogous copy mapping to the reference, suggesting that sequence divergence within paralogue clusters is sufficiently high for copies to be discriminated between by the read mapping algorithm. Notably, the sequence data we explored are shotgun sequence datasets, not captured regions. An empirical study to investigate the different coverage for known paralogues versus known orthologues, perhaps using *Arabidopsis*, might resolve the coverage question.

We have shown that the distribution of the sums of branch lengths of gene trees derived from the orthology and paralogy datasets are similar (Figure 5-2). A target enrichment study on *Populus trichocarpa,* where there is good evidence for an ancient-duplication event (Tuskan and Torr, 2007), separated the genes into two groups, as we do here. The first group consisted of genes with retained paralogues and the second group genes without paralogues. The analysis of SNP density revealed that the number of putative SNPs was not significantly higher in the paralogues group (Zhou and

Holliday, 2012). These data from *Populus trichocarpa* suggest that, where orthologues and paralogues are retained in the genome, contaminant reads do not map to the initial reference. In our study, paralogous sequences are from recent duplications, and in this case, though paralogous copies map to the reference, read assembly errors do not influence outcomes. Whether there is a hierarchical level at which duplications map to the reference but do influence outcomes may become apparent as more systems are investigated.

Applying the pipeline described in Chapter 2 to the *Arabidopsis* dataset showed that the orthologues and paralogues were not distinguishable using allelic comparisons. The hypothesis that phylogenetic distance and sequence similarity between alleles are different in the orthology and paralogy datasets could not be confirmed. In the case of the orthology data set, phasing must reconstruct alleles since there are no paralogues for these regions. The alleles that were recovered in the paralogy dataset may be true allelic variants, rather than copies derived from different genomic regions, since they have the same characteristics as the alleles of orthologues. The phylogenies from all three data sets were almost identical (Figure 5-4). However, longer terminal branch lengths in some samples in the ML tree derived from paralogues may indicate a higher level of variability in the paralogy dataset. The results from the majority rule consensus tree which shows how many gene trees support each split in the phylogeny indicates that the all three datasets support the same topology, but that the gene support frequency (GSF) is higher in the ortholog dataset compared to the paralog dataset. However, the MRC trees that were built from individual genes confirm findings that gene trees often favour alternative topologies to the respective total-evidence tree (Salichos and Rokas, 2013). Only the *Arabidopsis thaliana* clade exhibits a GSF that is higher than 50%. In conclusion, the ML and the MRC approaches suggest that the difference between the

datasets is minimal which indicates that the effect of contamination is minimal, if present (Figure 5-4).

From the evidence generated by this study, we conclude that read mapping from cognate paralogues in *Arabidopsis* has little to no effect on phylogenomic inference. That at least SNP density is unaffected by the inclusion of paralogues in *Populus trichocarpa* is an indication that similar results can be expected in other species. However, the *Populus trichocarpa* dataset includes only this species, and the *Arabidopsis* dataset contains only congeneric species; it may be inappropriate to extrapolate our findings to analyses when different genera or families are compared. It is also possible that the results will be different if the mapping stringency is reduced. Read mapping algorithms perform alignments of reads to the reference where mismatch-penalties are applied (e.g. Bowtie 2; Langmead and Salzberg, 2012). Lowering those penalties may result in the mapping of more divergent sequences.

Our results suggest that efforts to devise analyses to eliminate read assembly errors due to paralogy may not contribute to improved phylogenetic inference. In the context of phylogenomic read assemby, paralogy may not present significant challenges. The focus on excluding paralogy when assembling sequences may be misplaced, since targeting sequences from previously unidentified paralogous clusters through hybridization capture is not comparable to including entire paralogous sequences in an alignment when reconstructing species trees. There is still a need to be alert to the possibility of hidden paralogy when different copies are present or lost in different species since, in this case, orthologues would be compared with paralogues. Paralogy of this kind is best identified using the gene tree approaches that are already well-established.

Chapter 6 General discussion

6.1 Summary of findings

This thesis investigates genomic approaches to evolution and to DNA barcoding. In Chapter 2, I show that target enrichment of hundreds of nuclear loci is a significant improvement to Sanger sequencing-based approaches for phylogenetic inference. The data enabled me to ask specific evolutionary questions and I found that genus *Berberis* exhibits strong phylogeographic patterns. The phylogenetic trees shed light on the impact of the different uplift histories of the Himalayas and the Hengduan Mountains. Recent research suggests that *in situ* speciation in the Hengduan Mountains is an important factor shaping the floral composition of this mountain system (Xing and Ree, 2017). The study of *Berberis* confirms the importance of *in situ* speciation in the Hengduan Mountains. Furthermore, the analyses reveal that the Qinghai-Tibetan Plateau is likely to act as a high-elevation bridge between the Himalayas and the Hengduan Mountains. In summary, this chapter provides an example of how the floras of these mountain systems are related.

Chapters 3 and 4 are dedicated to the authentication of herbal medicines using DNA barcoding and phylogenetics. In Chapter 3, I show how next-generation sequencing of plastid genomes can be employed for designing informative barcodes, even in groups where the taxonomy is still under development and where recent evolutionary processes result in low genetic variation between species. The use of operational phylogenetic units for barcoding of herbal medicines provides a conceptual approach that may be employed in other DNA barcoding studies. DNA barcoding is becoming a routine tool for the authentication of herbal medicines and I provide DNA

barcodes based on diagnostic characters to facilitate application for regulatory purposes. In Chapter 4, I show that genomic tools are powerful for identifying specimens of traded herbal medicinal products. The metagenomics approach applied to the *Phyllanthus* dataset has been proven to be effective for analyzing mixed samples (Coghlan et al., 2015, 2012; Raclariu et al., 2017; Sgamma et al., 2017). In line with previous work, I confirm the potential of this approach for routine quality control of herbals. The phylogenetic approach applied to traded *Berberis* samples provides valuable knowledge about global trade. The analysis of traded samples reveals that global trade chains for natural commodities are complex and are highly dependent on local, potentially ancient economic structures. These findings highlight that results obtained from DNA barcoding allow questions beyond species identification to be asked, and are emerging as a valuable tool for market investigations.

Chapter 5 describes investigations that were conceptualized after having devised a pipeline for filtering potentially contaminated read alignments in the *Berberis* dataset. Several different approaches have been taken to address this possible problem (e.g. Nicholls et al., 2015). However, none of these pipelines have actually been tested on data where information about single-copy genes and paralogous clusters were available. I have shown that contamination of reads from unidentified paralogous copies is minimal. This finding is likely to be applicable to groups other than genus *Arabidopsis.* Future target-enrichment studies may use this study as reference to decipher the potential impact of reads from different paralogous loci on DNA assembly.

6.2 The future of phylogenomics

The transition from using few genes to genome-scale data in phylogenetics allows for new approaches for studying the evolution of non-model organisms. The

study presented in Chapter 2 shows that phylogenomics approaches outperform phylogenetic inference when only a few genes are used, as shown in Adhikari et al. (2015). At the beginning of the phylogenomics era, the increased amount of data, in conjunction with certain analytical methods (e.g. concatenation of genes), painted the picture of fully resolved phylogenies (e.g. Regier et al., 2008; Smith et al., 2011; Zhou et al., 2012). However, several studies showed discordance between gene trees and species trees (e.g. Kubatko and Degnan, 2007; Salichos and Rokas, 2013), highlighting that different parts of the genome have different evolutionary histories. The introduction of the multispecies coalescent method (Degnan and Rosenberg, 2009) is a significant improvement for phylogenomic inference using genome-scale data, since likelihood-based reconstruction of species phylogenies with concatenated data can be statistically inconsistent (Kubatko and Degnan, 2007; Roch and Steel, 2015). Species tree estimation is difficult in presence of gene tree conflict. This discordance may be the result of ILS which is most probable in closely related taxa or when ancient rapid radiations occurred (Degnan and Rosenberg, 2009). ILS is mathematically modelled by the MSC and takes into account gene tree discordance by treating each gene as an independent trial of the coalescence process in a phylogeny. In contrast, concatenated datasets consider the same history for all genes and do therefore not allow for genealogical independence of different genes (Edwards, 2009). The emergence of the MSC exemplifies that efforts in developing phylogenetic theory are necessary for improving species tree estimation.

The enriched genomic sequences used in this study were treated as "anonymous", meaning that the physiological function that those regions are responsible for remained unknown. However, they were specifically designed for genus *Berberis.* Several tools and databases exist to assign functions to a sequence of coding

DNA (e.g. BLAST2GO; Conesa et al., 2005). Applying these tools may lead to new ways of studying evolution by specifically targeting genes that are likely beneficial for adaptation or *de novo* evolution of genes (Pease et al., 2016). In the case of *Berberis,* it would be interesting to study how the Berberine biosynthetic pathway is conserved among different species or what genetic factors determine specific traits, such as either being evergreen or deciduous. Combining functional properties of genes with evolutionary theory could identify evolutionary mechanisms in unprecedented detail.

If the aim is to study deep phylogenetic relationships, rather than the mode of species evolution at shallow phylogenetic levels, it is crucial to target genes that are ubiquitous across a wide range of species. In a recent study, a set of DNA hybridization probes targeting ~500 loci was developed, designed using genomic resources from 43 angiosperm species (Buddenhagen et al., 2016). The study shows the applicability of this bait set for resolving deep, intermediate and shallow angiosperm relationships. Such resources may be used for improving the resolution of deep angiosperm phylogenetic relationships. Furthermore, in-solution hybridization techniques are particularly useful when DNA quality is impaired (e.g. Stenzel et al., 2009), as is the case in herbarium specimens (Särkinen et al., 2012). Hybridisation probes as presented by Buddenhagen et al. (2016) may improve sequencing of such specimens.

In terms of data processing of NGS reads to phylogenies, I have presented a pipeline for read assembly in Chapter 2. This pipeline included a step for filtering loci that are likely to be contaminated with reads from unidentified paralogous clusters. However, in Chapter 5, I show that loci from paralogous clusters perform almost equally well as single-copy genes for phylogenomic inference, making the loci filtering step redundant. The process from raw next-generation reads to phylogenies is still under development and researchers have not yet agreed on common practice. However, such

pipelines are emerging (Faircloth, 2015; Johnson et al., 2016) and, in the future, NGS data processing for phylogenomic inference will be a routine task.

6.3 The future of medicinal plant barcoding

It is important to remember that companies manufacturing herbal medicinal products need to follow guidelines published in pharmacopoeias. Three different approaches to DNA identification of specimens are presented in this thesis. The first approach is presented in Chapter 3 and is based on diagnostic characters in a DNA sequence. As stated in the discussion of that chapter, I am convinced that the approach used has the advantage of being easily implemented in a regulatory context.

The second approach presented is based on phylogenetic placement using hundreds of nuclear DNA sequences. Phylogenetic methods for DNA identification have the disadvantage of being computationally intensive and requiring knowledge about phylogenetics (Casiraghi et al., 2010). Furthermore, the study design presented in this thesis is highly specific to *Berberis* and the applicability of the hybridization capture probes to distantly related species is unlikely to produce consistent results. A bait set that targets nuclear genes across all angiosperms has recently been developed (Buddenhagen et al., 2016) and may be used as a standardized process for specimen identification. Using hundreds of nuclear genes could certainly increase the resolution of the DNA barcodes and circumvents known issues of using plastid DNA. However, an important factor determining the applicability of DNA barcoding as a routine diagnostic tool is the presence of a database for those respective loci, which would need to be produced for each group of species under study. Furthermore, data processing and specimen identification are labour-intensive and demand specialized skills. Nevertheless, hybridization capture has significant advantages when using degraded

DNA, as exemplified by its application in retrieval and sequencing of ancient DNA (e.g. Kistler et al., 2014; Stenzel et al., 2009).

The third approach is based on DNA metabarcoding (Taberlet and Coissac, 2012; Yu et al., 2012) of herbal mixtures. DNA metabarcoding is based on amplifying common DNA barcodes through PCR and sequencing the PCR products using next-generation sequencing technologies. Through parallel sequencing, the diversity of the PCR fragments is retained, which contrasts Sanger sequencing where only one sequence per PCR product is produced. Through BLAST search (Altschul et al., 1990) against a database, the fragments are then assigned to a specific taxonomic rank (Huson et al., 2007). This approach has huge potential for the herbal medicines industry since it not only authenticates the presence of a species, but also identifies contaminations with other species (Coghlan et al., 2015, 2012; Raclariu et al., 2017). Furthermore, it may be used for relative quantification of different ingredients. Nevertheless, significant advances in benchmarking this approach are needed. Standard DNA barcodes that are currently in use are either located in the plastid genome or in the rDNA, both existing in multiple copies within a cell. If relative quantification is the goal of this approach, assessing the variation of genome copy numbers, the impact of PCR success and sources of errors introduced through PCR or sequencing need to be benchmarked. This issue has been recognized by the DNA metagenomics community which mainly analyses microbial communities and was recently addressed in an excellent study by Amore et al. (2016). The authors produced artificial microbial mock communities that were sequenced using different library preparation settings (e.g. number of PCR cycles) and sequencers. In the context of DNA metabarcoding of herbal medicines, analysing artificial mixtures will increase confidence in the method by defining thresholds for species identification.

I have described advantages and disadvantages of distance-based, phylogenetic and diagnostic character-based identification methods in Chapter 3. According to a study with simulated DNA barcodes of closely related species, diagnostic methods consistently outperform other methods (van Velzen et al., 2012). Several tools are available for identification of diagnostic nucleotides in a given set of barcodes (e.g. Sarkar et al., 2008; Weitschek et al., 2013). These methods use hierarchical classification systems and can be seen as a classification chart with "if-then rules" (Weitschek et al., 2013). The proposition made to use whole plastid sequences as DNA barcodes (Coissac et al., 2016; Kane et al., 2012) may require the use of sophisticated algorithms for classification. Essentially, all barcoding methods share the same supervised learning paradigm, where a set of sequences with known class are used as the training set and a set of unknown sequences are attributed to these known classes (Fiscon et al., 2016). The increasing amount of sequence data available demands highly effective classification systems. Supervised classification using machine learning algorithms such as Naïve Bayes seem promising (Weitschek et al., 2014).

6.4 Research questions emerging from this study

The sequence data used in the studies presented in this PhD are only partly explored and I believe that sequencing data may be used to target other research questions. Shotgun sequencing was applied to numerous samples in this study. Although the number of sequencing reads is not enough for whole genome assembly, the data may be used, for example, to investigate differences in biosynthetic pathways between species. *Berberis* species produce benzylisoquinoline alkaloids, such as Berberine, mainly as a response to pathogenic attack (Dittrich and Kutchan, 1991). Little is known as to which extent such important pathways are conserved among

species. Exploring the sequencing data by mapping reads to specifically selected genes may give additional insights into the evolution of the species or, alternatively, into the evolution of important biosynthetic pathways.

The phylogeographic study of *Berberis* in the Himalayas and Hengduan Mountains should ideally be complemented with a calibrated tree. Wang et al. (2012) identified a two-phase growth of the Hengduan Mountains: the first occurring between 30-25 Myr ago and the second between 10-15 Myr ago. According to Xing and Ree (2017), high rates of *in situ* diversification coincide with the second period of fast orogeny. A dated phylogeny could be used to test the hypothesis of whether the diversification of deciduous *Berberis* species in the Hengduan Mountains is related to the second pulse of rapid exhumation in the Hengduan Mountains. Furthermore, it would be interesting to calculate diversification rates in the two mountain systems.

Berberis s.s. is diverse in South America and in the Sino-Himalayan region. This distribution pattern provides an excellent opportunity for studying the rate of evolution through time and across two unconnected regions, and would further give insights into the role of orogeny for species diversification. As identified by Adhikari et al., (2015), there is still no clear evidence of how the distribution pattern of simple-leaved *Berberis* emerged. The competing hypotheses are either long-distance dispersal or Cretaceous vicariance between Africa and South America (Kim et al., 2004). Testing of these hypotheses would require sampling African taxa (Adhikari et al., 2015). A collaborative study with dense sampling of taxa in South America, Africa and Eurasia using the laboratory methods presented in this study would give interesting and unprecedented insights into the mode and tempo of species diversification of an antitropically distributed genus.

Bibliography

Acharya, K.P., Rokaya, M.B., 2005. Ethnobotanical Survey of Medicinal Plants Traded in the Streets of Kathmandu Valley. Sci. World 3, 44–48.

Acosta, C.M., Premoli, A.C., 2010. Evidence of chloroplast capture in South American Nothofagus (subgenus Nothofagus, Nothofagaceae). Mol. Phylogenet. Evol. 54, 235–242.

Adhikari, B. et al., 2015. Systematics and biogeography of Berberis s . l . inferred from nuclear ITS and chloroplast ndhF gene sequences 64, 39–48.

Adhikari, B. et al., 2012. A Revision of Berberis S.S. (Berberidaceae) in Nepal. Edinburgh J. Bot. 69, 447–522.

Ahrendt, L.W.A., 1961. Berberis and Mahonia: A Taxonomic Revision. Bot. J. Linn. Soc. 57, 1–410.

Altschul, S.F. et al., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Andrews, S., 2010. FastQC: A quality control tool for high throughput sequence data.

Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796–815.

Arnot, D.E., Roper, C., Bayoumi, R.A.L., 1993. Digital codes from hypervariable tandemly repeated DNA sequences in the Plasmodium falciparum circumsporozoite gene can genetically barcode isolates. Mol. Biochem. Parasitol. 61, 15–24.

Austerlitz, F. et al., 2009. DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. BMC Bioinformatics 10 (Suppl 1), S10.

Ayurvedic Pharmacopoeia of India, 2001. Ayurvedic Pharmacopoeia of India. Government of India, Ministry of Health and Family Welfare, Department of Health, New Delhi.

Baird, N.A. et al., 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3, 1–7.

Banerjee, M., 2008. Ayurveda in Modern India. Standardisation and Pharmaceuticalization, in: Wujastyk, D., Smith, F.M. (Eds.), Modern and Global

Ayurveda. Pluralism and Paradigms. State University of New York Press, Albany, NY, pp. 201–214.

Bankevich, A. et al., 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comput. Biol. 19, 455–477.

Berlin, B., Breedlove, D.E., Raven, P.H., 1973. General principles of classification and nomenclature in folk biology. Am. Anthropol. 75, 214–242.

Bertolazzi, P., Felici, G., Weitschek, E., 2009. Learning to classify species with barcodes. BMC Bioinformatics 10, 1–12.

Bevan, M., Walsh, S., 2005. The Arabidopsis genome : A foundation for plant research 1632–1642.

Blaxter, M., 2003. Counting angels with DNA. Science 421, 9–11.

Blumberg, B.S. et al., 1989. Hepatitis B virus and hepatocellular carcinoma--treatment of HBV carriers with Phyllanthus amarus. Cancer Detect. Prev. 14, 195–201.

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

Braukmann, T.W.A. et al., 2017. Testing the efficacy of DNA barcodes for identifying the vascular plants of Canada. PLoS One 12, 1–19.

British Pharmacopoeia, 2016. Medicines and Healthcare Regulatory Agency (MHRA), London.

British Pharmacopoeia Commission, 2017. Deoxyribonucleic acid (DNA) based identification techniques for herbal drugs, in: British Pharmacopoeia Appendix XI V. London: TSO.

Buddenhagen, C. et al., 2016. Anchored Phylogenomics of Angiosperms I: Assessing the Robustness of Phylogenetic Estimates. doi:http://dx.doi.org/10.1101/086298 (availble at bioRxiv.org)

Casiraghi, M. et al., 2010. DNA barcoding: A six-question tour to improve users' awareness about the method. Brief. Bioinform. 11, 440–453.

CBOL Plant Working Group et al., 2009. A DNA barcode for land plants. Proc. Natl. Acad. Sci. U. S. A. 106, 12794–7.

Chase, M.W. et al., 2007. A Proposal for a Standardised Protocol to Barcode All Land Plants. Taxon 56, 295–299.

Chikhi, R., Medvedev, P., 2014. Informed and automated k-mer size selection for genome assembly. Bioinformatics 30, 31–37.

Cho, Y. et al., 2004. Mitochondrial substitution rates are extraordinarily elevated and variable in a genus of flowering plants. Proc. Natl. Acad. Sci. U. S. A. 101, 17741–17746.

Cho, Y. et al., 1998. Explosive invasion of plant mitochondria by a group I intron. Proc. Natl. Acad. Sci. U. S. A. 95, 14244–14249.

Clare, E.L. et al., 2007. DNA barcoding of Neotropical bats: species identification and discovery within Guyana. Mol. Ecol. Notes 7, 184–190.

Coghlan, M.L. et al., 2015. Combined DNA, toxicological and heavy metal analyses provides an auditing toolkit to improve pharmacovigilance of traditional Chinese medicine (TCM). Sci. Rep. 5, 17475.

Coghlan, M.L. et al., 2012. Deep sequencing of plant and animal DNA contained within traditional Chinese medicines reveals legality issues and health safety concerns. PLoS Genet. 8 (4), 1 - 11.

Coissac, E. et al., 2016. From barcodes to genomes: extending the concept of DNA barcoding. Mol. Ecol. 25, 1423 – 1428.

Conesa, A. et al., 2005. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674–3676.

Cronn, R. et al., 2012. Targeted enrichment strategies for next-generation plant biology. Am. J. Bot. 99, 291–311.

Cronn, R. et al., 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res. 36, e122. doi:10.1093/nar/gkn502

D'Amore, R. et al., 2016. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. BMC Genomics 17, 55.

Danecek, P. et al., 2011. The variant call format and VCFtools. Bioinformatics 27, 2156–2158.

de Boer, H.J., Ichim, M.C., Newmaster, S.G., 2015. DNA Barcoding and Pharmacovigilance of Herbal Medicines. Drug Saf. 38, 611–620.

De Sousa, F. et al., 2014. Phylogenetic properties of 50 nuclear loci in Medicago (Leguminosae) generated using multiplexed sequence capture and next-generation sequencing. PLoS One 9, 1–16.

Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24, 332–340.

Delsuc, F., Brinkmann, H., Philippe, H., 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6, 361–375.

DeSalle, R., 2006. Species discovery versus species identification in DNA barcoding efforts: Response to Rubinoff. Conserv. Biol. 20, 1545–1547.

DeSalle, R., Egan, M.G., Siddall, M., 2005. The unholy trinity: taxonomy, species delimitation and DNA barcoding. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 360, 1905–16.

Directive 2001/83/EC, 2001. Off. J. Eur. Union L 311, 67–128.

Directive 2004/83/EC, 2004. Off. J. Eur. Union L 136, 85–90.

Dittrich, H., Kutchan, T.M., 1991. Molecular cloning, expression, and induction of berberine bridge enzyme, an enzyme essential to the formation of benzophenanthridine alkaloids in the response of plants to pathogenic attack. Proc. Natl. Acad. Sci. U. S. A. 88, 9969–9973.

Doyle, J.J., 1992. Gene Trees and Species Trees : Molecular Systematics as One-Character Taxonomy. Syst. Bot. 17, 144–163.

Doyle, J.J., Doyle, J.L., 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. 19, 11–15.

Ebersbach, J. et al., 2017. In and out of the Qinghai-Tibet Plateau: divergence time estimation and historical biogeography of the large arctic-alpine genus Saxifraga L. J. Biogeogr. 44, 900–910.

Edwards, S. V., 2009. Is a new and general theory of molecular systematics emerging? Evolution (N. Y). 63, 1–19.

Edwards, S. V. et al., 2016. Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. Mol. Phylogenet. Evol. 94, 447–462.

Eggert, L.S., Rasner, C. a, Woodruff, D.S., 2002. The evolution and phylogeography of the African elephant inferred from mitochondrial DNA sequence and nuclear microsatellite markers. Proc. Biol. Sci. 269, 1993–2006.

Eisen, J.A., Fraser, C.M., 2003. Phylogenomics: Intersection of evolution and genomics. Science 300, 1706–7.

Ekblom, R., Galindo, J., 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity 107, 1–15.

Ernst, M. et al., 2016. Evolutionary prediction of medicinal properties in the genus Euphorbia L. Sci. Rep. 6, 30531.

European Medicines Agency, 2006. Guideline on Good Agricultural and Collection Practice (GACP) for Starting Materials of Herbal Origin.

Faircloth, B.C., 2015. PHYLUCE is a software package for the analysis of conserved genomic loci.

Favre, A. et al., 2015. The role of the uplift of the Qinghai-Tibetan Plateau for the evolution of Tibetan biotas. Biol. Rev. Camb. Philos. Soc. 90, 236–253.

Fazekas, A.J. et al., 2009. Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? Mol. Ecol. Resour. 9 (Suppl s1), 130–9.

Fiscon, G. et al., 2016. MISSEL: a method to identify a large number of small species-specific genomic subsequences and its application to viruses classification. BioData Min. 9, 38.

Floyd, R. et al., 2002. Molecular barcodes for soil nematode identification. Mol. Ecol. 11, 839–50.

Folk, R.A., Mandel, J.R., Freudenstein, J. V, 2016. Ancestral Gene Flow and Parallel Organellar Genome Capture Result in Extreme Phylogenomic Discord in a Lineage of Angiosperms. Syst. Biol. 0, 1–18.

Foster, S., 2011. A Brief History of Adulteration of Herbs, Spices, and Botanical Drugs. HerbalGram 42–57.

Gafner, S., Applequist, W., 2016. on Adulteration of Arnica montana. Bot. Adulterants Bull. 1–5.

Goldstein, P.Z., DeSalle, R., 2005. Phylogenetic Species, Nested Hierarchies, and Character Fixation. Cladistics 16, 364–384.

Gurevich, A. et al., 2013. QUAST: Quality assessment tool for genome assemblies. Bioinformatics 29, 1072–1075.

Harber, J., 2017a. Berberis zhaotongensis. A new species in sect. Wallichianae Berberidaceae. Curtis's Bot. 34, 98–104.

Harber, J., 2017b. Berberis bowashanensis. Curtis's Bot. 34, 105–110.

Harvey, M.G. et al., 2016. Sequence Capture versus Restriction Site Associated DNA Sequencing for Shallow Systematics. Syst. Biol. 65, 910–924.

He, D., Choi, A., Pipatsrisawat, K., 2010. Optimal algorithms for haplotype assembly from whole-genome sequence data. Bioinformatics 26, i183-90.

Hebert, P.D.N. et al., 2004. Identification of birds through DNA barcodes. PLoS Biol. 2, 1657-1663.

Hebert, P.D.N. et al., 2003. Biological identifications through DNA barcodes. Proc. Biol. Sci. 270, 313–21.

Heyduk, K. et al., 2016. Phylogenomic analyses of species relationships in the genus Sabal (Arecaceae) using targeted sequence capture. Biol. J. Linn. Soc. 117, 106–120.

Hollingsworth, P.M., Graham, S.W., Little, D.P., 2011. Choosing and using a plant DNA barcode. PLoS One 6, 1-13.

Hughes, C.E., Atchison, G.W., 2015. The ubiquity of alpine plant radiations: From the Andes to the Hengduan Mountains. New Phytol. 207, 275–282.

Humagain, K., Shrestha, K.K., 2009. Medicinal plants in Rasuwa district , central Nepal : trade and livelihood. Bot. Orient. 6, 39–46.

Huson, D. et al., 2007. MEGAN analysis of metagenome data. Genome Res. 17, 377–386.

Ize-Ludlow, D., 2004. Neurotoxicities in Infants Seen With the Consumption of Star Anise Tea. Pediatrics 114, e653–e656.

Janzen, D.H., 2004. Now is the time. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 359, 731–2.

Jenks, A.A., Kim, S.C., 2013. Medicinal plant complexes of Salvia subgenus Calosphace: An ethnobotanical study of new world sages. J. Ethnopharmacol. 146, 214–224.

Johnson, M.G. et al., 2016. HybPiper: Extracting Coding Sequence and Introns for Phylogenetics from High-Throughput Sequencing Reads Using Target Enrichment. Appl. Plant Sci. 4, 1600016.

Joshi, A. R., Joshi, K., 2000. Indigenous knowledge and uses of medicinal plants by local communities of the Kali Gandaki Watershed Area, Nepal. J. Ethnopharmacol. 73, 175–83.

Kane, N. et al., 2012. Ultra-barcoding in cacao (Theobroma spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. Am. J. Bot. 99, 320–9.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.

Kaul, S. et al., 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796–815.

Kent, W.J., 2002. BLAT—The BLAST-Like Alignment Tool. Genome Res. 12, 656–664.

Kim, Y.-D., Kim, S.-H., Landrum, L.R., 2004. Taxonomic and phytogeographic implications from ITS phylogeny in Berberis (Berberidaceae). J. Plant Res. 117, 175–182.

Kistler, L. et al., 2014. Transoceanic drift and the domestication of African bottle gourds in the Americas. Proc. Natl. Acad. Sci. U. S. A. 111, 1–5.

Kong, W. et al., 2004. Berberine is a novel cholesterol-lowering drug working through a unique mechanism distinct from statins. Nat. Med. 10, 1344–1351.

Kress, W.J. et al., 2005. Use of DNA barcodes to identify flowering plants. Proc. Natl. Acad. Sci. U. S. A. 102, 8369–74.

Krzywinski, M. et al., 2009. Circos: an information aesthetic for comparative genomics. Genome Res. doi: 10.1101/gr.092759.109

Kubatko, L.S., Degnan, J.H., 2007. Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence. Syst. Biol. 56, 17–24.

Kunwar, R.M. et al., 2013. Medicinal plants, traditional medicine, markets and management in far-west Nepal. J. Ethnobiol. Ethnomed. 9, 24.

Landis, M.J. et al., 2013. Bayesian analysis of biogeography when the number of areas is large. Syst. Biol. 62, 789–804.

Landrum, L.R., 1999. Revision of Berberis (Berberidaceae) in Chile and Adjacent Southern Argentina. Ann. Missouri Bot. Gard. 86, 793–834.

Lanfear, R. et al., 2016. PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. Mol. Biol. Evol. 34, 772-773.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat Meth 9, 357–359.

Lee, E. et al., 2009. Apollo: A community resource for genome annotation editing. Bioinformatics 25, 1836–1837.

Lemmon, A.R., Emme, S.A., Lemmon, E.M., 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61, 727–744.

Leonti, M., Casu, L., 2013. Traditional medicines and globalization: current and future perspectives in ethnopharmacology. Front. Pharmacol. 4, 92.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079.

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

Li, W., Godzik, A., 2006. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659.

Li, Y.L. et al., 2010. The fossil record of Berberis (Berberidaceae) from the Palaeocene of NE China and interpretations of the evolution and phytogeography of the genus. Rev. Palaeobot. Palynol. 160, 10–31.

Linares, E., Bye, R., 1987. A study of four medicinal plants complexes of Mexico and adjacent United States. J. Ethnopharmacol. 19, 153–183.

Little, D.P., Stevenson, D.W., 2007. A comparison of algorithms for the identification of specimens using DNA barcodes: Examples from gymnosperms. Cladistics 23, 1–21.

Liu, C. et al., 2012. CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. BMC Genomics 13, 715.

Liu, L. et al., 2015. Estimating phylogenetic trees from genome-scale data. Ann. N. Y. Acad. Sci. 1360, 36–53.

Liu, X. et al., 2009. The current global status of Chinese materia medica. Phyther. Res. 23, 1493–1495.

Long, Q. et al., 2013. Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. Nat. Genet. 45, 884–890.

Luo, R. et al., 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1, 18.

Ma, J. et al., 2013. The complete chloroplast genome sequence of Mahonia bealei (Berberidaceae) reveals a significant expansion of the inverted repeat and phylogenetic relationship with other angiosperms. Gene 528, 120–131.

Mabberley, D.J., 2008. Mabberley's Plant-Book, 3rd ed. Cambridge University Press, New York.

Mallet, J., Willmott, K., 2003. Taxonomy: renaissance or Tower of Babel? Trends Ecol. Evol. 18, 57–59.

Manandhar, N.P., 2002. Plants and People of Nepal. Timber Press, Inc., Portland.

Mandel, J.R. et al., 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. Appl. Plant Sci. 2, 1300085.

Mander, M., 1998. Marketing of indigenous medicinal plants in South Africa - a case study in Kwazulu-Natal, FAO - Food and Agriculture Organization of the United Nations. Rome.

Marini-Bettolo, G., 1975. Pharmacopoeia as a pharmaceutical code for public health authorities. Ann 1st Super Sanita 11, 254–268.

Marroquin, J.S., Laferriere, J.E., 1997. Transfer of Specific and Infraspecific Taxa from Mahonia to Berberis. J. Arizona-Nevada Acad. Sci. 30, 53–55.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10–12.

Matzke, N.J., 2014. Model selection in historical biogeography reveals that founder-event speciation is a crucial process in island clades. Syst. Biol. 63, 951–970.

McCormack, J.E. et al., 2013. A Phylogeny of Birds Based on Over 1,500 Loci Collected by Target Enrichment and High-Throughput Sequencing. PLoS One 8, e54848.

McKenna, A. et al., 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303.

McVay, J.D., Carstens, B.C., 2013. Phylogenetic Model Choice: Justifying a Species Tree or Concatenation Analysis. Phylogenetics Evol. Biol. 1, 1–8.

Meyer, C.P., Paulay, G., 2005. DNA barcoding: error rates based on comprehensive sampling. PLoS Biol. 3, e422.

Meyer, M., Kircher, M., 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc. doi:10.1101/pdb.prot5448.

Miehe, G., Weidinger, J.T., 2015. Himalayan Landforms and Processes, in: Miehe, G., Pendry, C. (Eds.), Nepal. An Introduction to the Natural History, Ecology and Human Environment of the Himalayas. Royal Botanic Garden Edinburgh, Edinburgh, pp. 103–124.

Miller, M.A., Pfeiffer, W., Schwartz, T., 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees, in: Proceedings of the Gateway Computing Environments Workshop (GCE). New Orleans, LA, pp. 1–8.

Mirarab, S. et al., 2014. ASTRAL: Genome-scale coalescent-based species tree estimation. Bioinformatics 30, 541–548.

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

Moritz, C., Cicero, C., 2004. DNA barcoding: Promise and pitfalls. PLoS Biol. 2, 1529-1531.

Mulch, A., Chamberlain, C.P., 2006. Earth science: The rise and growth of Tibet. Nature 439, 670–671.

Mutanen, M. et al., 2016. Species-level para- and polyphyly in DNA barcode gene trees: Strong operational bias in European Lepidoptera. Syst. Biol. 65, 1024–1040.

Myers, N. et al., 2000. Biodiversity hotspots for conservation priorities. Nature 403, 853–858.

Naciri, Y., Linder, H.P., 2015. Species delimitation and relationships: The dance of the seven veils. Taxon 64, 3–16.

Nanney, D.L., 1982. Genes and phenes in Tetrahymena. Bioscience 32, 783–788.

Newmaster, S.G. et al., 2013. DNA barcoding detects contamination and substitution in North American herbal products. BMC Med. 11, 222.

Nicholls, J.A. et al., 2015. Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus Inga (Leguminosae: Mimosoideae). Front. Plant Sci. 6, 1-20

Novikova, P.Y. et al., 2016. Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. Nat. Genet. 48, 1077–1082.

Olsen, C.S., 1998. The Trade in Medicinal and Aromatic Plants from Central Nepal to Northern India. Econ. Bot. 52, 279–292.

Olsen, C.S., Bhattarai, N., 2005. A Typology of Economic Agents in the Himalayan Plant Trade. Mt. Res. Dev. 25, 37–43.

Olsen, C.S., Larsen, H.O., 2003. Alpine Medicinal Plant Trade and Himalayan Mountain Livelihood Strategies. Geogr. J. 169, 243–254.

Ouarghidi, A. et al., 2012. Species substitution in medicinal roots and possible implications for toxicity in Morocco. Econ. Bot. 66, 370–382.

Page, A.J. et al., 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb. Genomics 2, 38190.

Pagel, M., Meade, A., 2006. Bayesian Analysis of Correlated Evolution of Discrete Characters by Reversible-Jump Markov Chain Monte Carlo. Am. Nat. 167, 808–825.

Paradis, E., 2010. Pegas: An R package for population genetics with an integrated-modular approach. Bioinformatics 26, 419–420.

Paradis, E., Claude, J., Strimmer, K., 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20, 289–290.

Parks, M., Cronn, R., Liston, A., 2012. Separating the wheat from the chaff: mitigating the effects of noise in a plastome phylogenomic data set from Pinus L. (Pinaceae). BMC Evol. Biol. 12, 100.

Parks, M., Cronn, R., Liston, A., 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biol. 7, 84.

Parmentier, I. et al., 2013. How Effective Are DNA Barcodes in the Identification of African Rainforest Trees? PLoS One 8, e54921.

Patel, J.R. et al., 2011. Phyllanthus amarus: Ethnomedicinal uses, phytochemistry and pharmacology: A review. J. Ethnopharmacol. 138, 286–313.

Payyappallimana, U., 2008. Ayurvedic Pharmacopoeia Databases in the Context of the Revitalization of Traditional Medicine, in: Wujastyk, D., Smith, F.M. (Eds.), Modern and Global Ayurveda. Pluralism and Paradigms. pp. 139–155.

Pease, J.B. et al., 2016. Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. PLoS Biol. 14, 1–24.

Percy, D.M. et al., 2014. Understanding the spectacular failure of DNA barcoding in willows (Salix): Does this result from a trans-specific selective sweep? Mol. Ecol. 4737–4756.

Philippe, H. et al., 2011. Resolving difficult phylogenetic questions: Why more sequences are not enough. PLoS Biol. 9, e1000602.

Posadzki, P., Watson, L., Ernst, E., 2013. Contamination and adulteration of herbal medicinal products (HMPs): An overview of systematic reviews. Eur. J. Clin. Pharmacol. 69, 295–307.

Prum, R.O. et al., 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526, 569–573.

Quinlan, A.R., Hall, I.M., 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842.

Raclariu, A.C. et al., 2017. Comparative authentication of Hypericum perforatum herbal products using DNA metabarcoding, TLC and HPLC-MS. Sci. Rep. 7, 1291.

Raja, H.A. et al., 2017. DNA barcoding for identification of consumer-relevant mushrooms: A partial solution for product certification? Food Chem. 214, 383–392.

Rambaut, A. et al., 2014. Tracer v1.6.

Ree, R.H. et al., 2005. a Likelihood Framework for Inferring the Evolution of Geographic Range on Phylogenetic Trees. Evolution (N. Y). 59, 2299–2311.

Regier, J.C. et al., 2008. Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. Syst. Biol. 57, 920–938.

Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 16, 276–277.

Rieseberg, L.H., Brouillet, L., 1994. Are Many Plant Species Paraphyletic? Taxo 43, 21–32.

Rieseberg, L.H., Soltis, D.E., 1991. Phylogenetic consequences of cytoplasmic gene flow in plants. Evol. trends Plants 5, 65–84.

Robinson, M.M., Zhang, X., 2011. Traditional Medicines : Global Situation , Issues and Challenges, The World Medicines Situation. Geneva.

Roch, S., Steel, M., 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theor. Popul. Biol. 100, 56–62.

Rokas, A. et al., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425, 798–804.

Ronquist, F., 1997. Dispersal-vicariance analysis: a new approach to the quantification of historical biogeography. Sys Biol 46, 195–203.

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.

Rounsaville, T., Ranney, T., 2010. Ploidy levels and genome sizes of Berberis L. and Mahonia Nutt. species, hybrids, and cultivars. HortScience 45, 1029–1033.

Roy, S. et al., 2010. Universal plant DNA barcode loci may not work in complex groups: a case study with Indian berberis species. PLoS One 5, e13674.

Royden, L.H., Burchfiel, B.C., Hilst, R.D. Van Der, 2008. The Geological Evolution of the Tibetan Plateau. Science 321, 1054–1058.

Saks, M., 2008. Plural Medicine an East-West Dialogue, in: Wujastyk, D., Smith, F.M. (Eds.), Modern and Global Ayurveda. Pluralism and Paradigms. State University of New York Press, Albany, NY, pp. 29–41.

Salichos, L., Rokas, A., 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497, 327–331.

Sarkar, I.N., Planet, P.J., Desalle, R., 2008. CAOS software for use in character-based DNA barcoding. Mol. Ecol. Resour. 8, 1256–1259.

Särkinen, T. et al., 2012. How to open the treasure chest? Optimising DNA extraction from herbarium specimens. PLoS One 7, e43808.

Saslis-Lagoudakis, C.H. et al., 2012. Phylogenies reveal predictive power of traditional medicine in bioprospecting. Proc. Natl. Acad. Sci. U. S. A. 109, 15835–40.

Schliep, K.P., 2011. phangorn: phylogenetic analysis in R. Bioinformatics 27, 592–593.

Schmickl, R. et al., 2015. Phylogenetic marker development for target enrichment from transcriptome and genome skim data: The pipeline and its application in southern African Oxalis (Oxalidaceae). Mol. Ecol. Resour. doi: 10.1111/1755-0998.12487

Scindia, J.M., 2010. India's herbal product exports rising at a compounded annual rate of 16.8 per cent news. [WWW Document]. URL http://www.domain-b.com/industry/Healthcare/20101213_herbal_product.html

Searle, M.P., 2011. Geological evolution of the Karakoram Ranges. Ital. J. Geosci. 130, 147–159.

Seberg, O., Petersen, G., 2009. How many loci does it take to DNA barcode a crocus? PLoS One 4, 2–7.

Sgamma, T. et al., 2017. DNA Barcoding for Industrial Quality Assurance. Planta Med. doi: https://doi.org/10.1055/s-0043-113448

Shrestha, P.M., Dhillion, S.S., 2003. Medicinal plant diversity and use in the highlands of Dolakha district, Nepal. J. Ethnopharmacol. 86, 81–96.

Sims, D. et al., 2014. Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. 15, 121–32.

Smith, M.A. et al., 2006. DNA barcodes reveal cryptic host-specificity within the presumed polyphagous members of a genus of parasitoid flies (Diptera: Tachinidae). Proc. Natl. Acad. Sci. U. S. A. 103, 3657–62.

Smith, M.A., Poyarkov, N. a, Hebert, P.D.N., 2008. DNA BARCODING: CO1 DNA barcoding amphibians: take the chance, meet the challenge. Mol. Ecol. Resour. 8, 235–46.

Smith, S.A., Dunn, C.W., 2008. Phyutility: A phyloinformatics tool for trees, alignments and molecular data. Bioinformatics 24, 715–716.

Smith, S. a. et al., 2011. Resolving the evolutionary relationships of molluscs with phylogenomic tools. Nature 480, 364–367.

Srirama, R. et al., 2017. Species Adulteration in the Herbal Trade: Causes, Consequences and Mitigation. Drug Saf. 1–11.

Srivastava, S., Rawat, a K.S., 2013. Quality evaluation of ayurvedic crude drug daruharidra, its allied species, and commercial samples from herbal drug markets of India. Evid. Based. Complement. Alternat. Med. 2013, 472973.

Stamatakis, A., 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313.

Stegemann, S. et al., 2012. Horizontal transfer of chloroplast genomes between plant species. Proc. Natl. Acad. Sci. 109, 2434–2438.

Stenzel, U. et al., 2009. Five Neandertal mtDNA Genomes. Science 325, 318–321.

Stephens, J.D. et al., 2015. Species tree estimation of diploid Helianthus (Asteraceae) using target enrichment. Am. J. Bot. 102, 30602.

Straub, S.C.K. et al., 2012. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. Am. J. Bot. 99, 349–64.

Straub, S.C.K. et al., 2011. Building a model: developing genomic resources for common milkweed (Asclepias syriaca) with low coverage genome sequencing. BMC Genomics 12, 211.

Struck, T.H., 2014. TreSpEx—Detection of Misleading Signal in Phylogenetic Reconstructions Based on Tree Information. Evol. Bioinforma. 52–67.

Struck, T.H., 2013. The Impact of Paralogy on Phylogenomic Studies - A Case Study on Annelid Relationships. PLoS One 8.

Subedi, B.B.P., Panderey, S.S., 2011. Cross-border NTFP value chains: Nepal - India (No. 64), INBAR Working Paper.

Sukumaran, J., Holder, M., 2015. SumTrees: Phylogenetic Tree Summarization. Available at https://github.com/jeetsukumaran/DendroPy.

Taberlet, P., Coissac, E., 2012. Towards next-generation biodiversity assessment using DNA metabarcoding 33, 2045–2050.

Tautz, D. et al., 2003. A plea for DNA taxonomy. Trends Ecol. Evol. 18, 70–74.

Tautz, D., Ellegren, H., Weigel, D., 2010. Next generation molecular ecology. Mol. Ecol. 19 (Suppl. 1), 1–3.

Taylor, H.R., Harris, W.E., 2012. An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. Mol. Ecol. Resour. 12, 377–88.

Thomas, K.J., Nicholl, J.P., Coleman, P., 2001. Use and expenditure on complementary medicine in England: a population based survey. Complement. Ther. Med. 9, 2–11.

Tiwari, N.N., Poudel, R.C., Uprety, Y., 2004. Study on Domestic Market of Medicinal and Aromatic Plants (MAPs) in Kathmandu Valley.

Tsitrone, A., Kirkpatrick, M., Levin, D. a, 2003. A model for chloroplast capture. Evolution 57, 1776–1782.

Tuskan, G.A., Torr, P., 2007. The Genome of Black Cottonwood ,. Science (80-.). 1596, 1596–1605.

Uprety, Y. et al., 2012. Diversity of use and local knowledge of wild edible plant resources in Nepal. J. Ethnobiol. Ethnomed. 8, 16.

Uprety, Y. et al., 2010. Indigenous use and bio-efficacy of medicinal plants in the Rasuwa District, Central Nepal. J. Ethnobiol. Ethnomed. 6, 3.

Van der Auwera, G.A. et al., 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.[GATK]. Curr Protoc Bioinforma. 11. doi: 10.1002/0471250953.bi1110s43

van Velzen, R. et al., 2012. DNA barcoding of recently diverged species: relative performance of matching methods. PLoS One 7, e30490.

Varshney, R.K. et al., 2009. Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol. 27, 522–30.

Vaughn, J.N. et al., 2014. Whole Plastome Sequences from Five Ginger Species Facilitate Marker Development and Define Limits to Barcode Methodology. PLoS One 9, e108581.

Vlietinck, A., Pieters, L., Apers, S., 2009. Legal requirements for the quality of herbal substances and herbal preparations for the manufacturing of herbal medicinal products in the European union. Planta Med. 75, 683–8.

Wang, C. et al., 2008. Constraints on the early uplift history of the Tibetan Plateau. Proc. Natl. Acad. Sci. 105, 4987–4992.

Wang, E. et al., 2012. Two-phase growth of high topography in eastern Tibet during the Cenozoic. Nat. Geosci. 5, 640–645.

Wang, W. et al., 2007. Phylogenetic and Biogeographic Diversification of Berberidaceae in the Northern Hemisphere. Syst. Bot. 32, 731–742.

Wanke, S. et al., 2017. Recalcitrant deep and shallow nodes in Aristolochia (Aristolochiaceae) illuminated using anchored hybrid enrichment. Mol. Phylogenet. Evol.

Ward, R.D. et al., 2005. DNA barcoding Australia's fish species. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 360, 1847–57.

Weitemier, K. et al., 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics 2, 3–9.

Weitschek, E. et al., 2014. Supervised DNA Barcodes species classification: analysis, comparisons and results. BioData Min. 7, 4.

Weitschek, E. et al., 2013. BLOG 2.0: A software system for character-based species classification with DNA Barcode sequences. What it does, how to use it. Mol. Ecol. Resour. 13, 1043–1046.

Whittall, J.B. et al., 2010. Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. Mol. Ecol. 19 (Suppl.1), 100–14.

WHO, 2014. WHO Traditional Medicine Strategy 2014 - 2023.

WHO, 2002. WHO Traditional Medicine Strategy 2002 - 2005.

Wolfe, K.H., Li, W.-H., Sharp, P.M., 1987. Rates of Nucleotide Substitution Vary Greatly among Plant Mitochondrial, Chloroplast, and Nuclear DNAs. Proc. Natl. Acad. Sci. U. S. A. 84, 9054–9058.

Wyman, S.K., Jansen, R.K., Boore, J.L., 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20, 3252–3255.

Xing, Y., Ree, R.H., 2017. Uplift-driven diversification in the Hengduan Mountains, a temperate biodiversity hotspot, Proceedings of the National Academy of Sciences.

Yin, J., Xing, H., Ye, J., 2008. Efficacy of berberine in patients with type 2 diabetes mellitus. Metabolism. 57, 712–717.

Yu, D.W. et al., 2012. Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods Ecol. Evol. 3, 613–623.

Zhao, D. et al., 2017. De novo genome assembly of Camptotheca acuminata, a natural source of the anti-cancer compound camptothecin. Gigascience. doi: https://doi.org/10.1093/gigascience/gix065

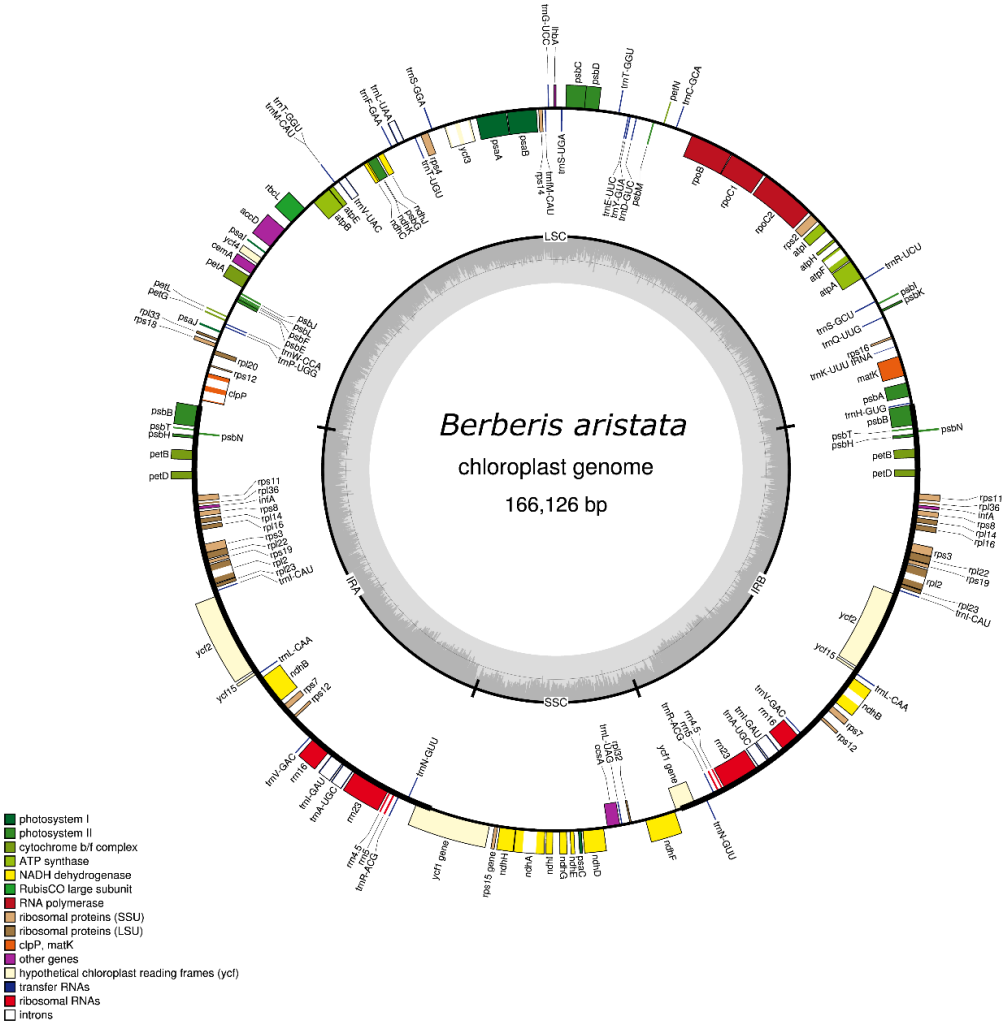
Zhou, L., Holliday, J.A., 2012. Targeted enrichment of the black cottonwood (Populus trichocarpa) gene space using sequence capture. BMC Genomics 13, 703.

Zhou, X. et al., 2012. Phylogenomic analysis resolves the interordinal relationships and rapid diversification of the laurasiatherian mammals. Syst. Biol. 61, 150–164.

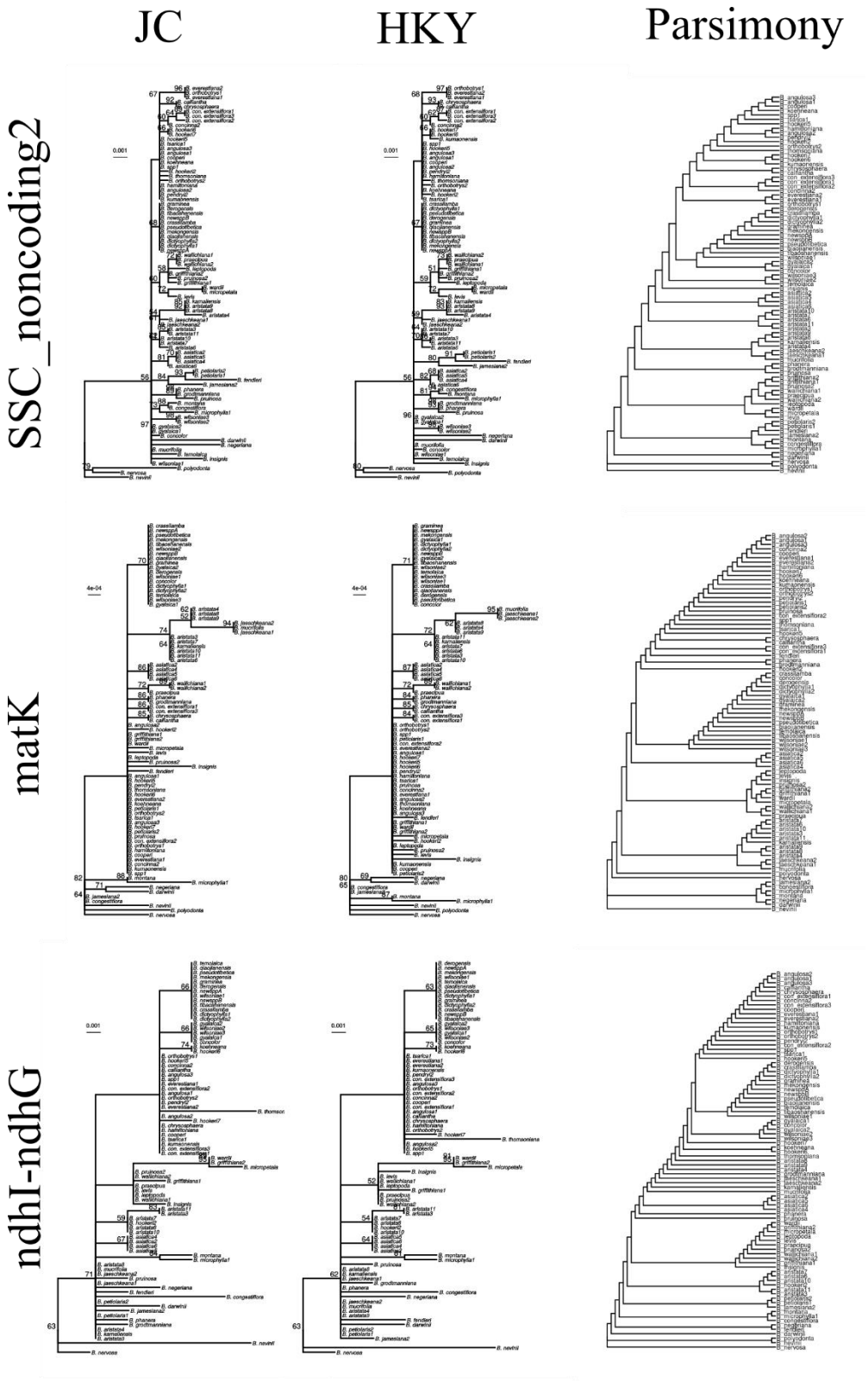
Appendices

Appendix figures

Appendix Figure AF-1 Consensus network from a sample of 984 trees from the Bayesian analysis. The network is a 3D representation drawn in 2D. No conflicting splits were detected.



Appendix Figure AF-2 Gene map of the plastid genome of *Berberis aristata*. Genes on the outside of the circle are transcribed clockwise and genes on the inside anti-clockwise. The dark grey histograms in the inner circle show the GC content.



Appendix Figure AF-3 Phylogenies of the selected barcodes *ndhI-ndhG*, *mat*K and SSC_noncoding2 under different models of evolution. The aristata and asiatica clades were both recovered, leading to the same conclusion as under the GTRCAT model.

Appendix tables

Appendix Table AT-1 Table with specimen information.

Sample	Species	Locality	Lat.	Long	Collector(s)	Coll. Date	Voucher	Comments
B_angulosa1	*B. angulosa* Wall. ex Hook.f. & Thomson	Nepal, Illam District	27.11	87.99	Adhikari, B. et al.	14-Jun-07	LKSRB71	
B_angulosa2	*B. angulosa* Wall. ex Hook.f. & Thomson	Nepal, Rasuwa District	28.21	85.57	Adhikari, B.	03-Aug-07	BL244	
B_angulosa3	*B. angulosa* Wall. ex Hook.f. & Thomson	Bhutan, Haa	27.27	89.17	Di McNab	01-Jul-05	AS97	Cultivated (J. Harber Coll.)
B_angulosa4	*B. angulosa* Wall. ex Hook.f. & Thomson	Nepal, Bimtang	28.64	84.47	N/A	13-Aug-08	20815195	
B_aristata1	*B. aristata* DC.	Nepal, Makwanpur District	27.59	85.77	Adhikari, B. et al.	01-Sep-14	Col_35.5	
B_aristata10	*Berberis aristata* DC.	Nepal, Dhankuta District	27.04918	87.35425	Adhikari, B. et al.	01-Aug-14	WP21.1	
B_aristata11	*Berberis aristata* DC.	Nepal, Gandaki District	28.39255	83.77315	Adhikari, B.	5 October 2006	EA109	
B_aristata2	*B. aristata* DC.	Nepal, Doti District	29.29	81.01	N/A	29-Jun-09	Bhatjang20915004	
B_aristata3	*B. aristata* DC.	Nepal, Dhankuta District	27.05	87.35	Adhikari, B. et al.	01-Sep-14	WP21.5	
B_aristata4	*B. aristata* DC.	N/A	N/A	N/A	N/A	N/A	1260210	
B_aristata5	*B. aristata* DC.	Nepal, Hile	27.04	87.32	Adhikari, B. et al.	01-Sep-14	WP18.3	
B_aristata6	*Berberis aristata* DC.	Nepal, Koshii District	27.04918	87.35425	Adhikari, B. et al.	01-Aug-14	WP32.5	
B_aristata7	*Berberis aristata* DC.	Nepal, Koshii District	27.04048	87.31713	Adhikari, B. et al.	01-Aug-14	WP18.2	
B_aristata8	*Berberis aristata* DC.	Nepal, Dhawalagiri District	28.66222	83.59472	Adhikari, B.	17 August 2007	EA243	
B_aristata9	*Berberis aristata* DC.	Nepal, Dhawalagiri District	28.66028	83.59389	Adhikari, B.	17 August 2007	EA249	
B_asiatica1	*B. asiatica* Roxb. ex DC.	Nepal, Mustang District	28.59	83.65	Adhikari, B.	17-Aug-07	EA254	
B_asiatica2	*B. asiatica* Roxb. ex DC.	Nepal, Makwanpur District	27.58	85.16	Adhikari, B. et al.	25-Aug-17	Coll_7.1	
B_asiatica3	*B. asiatica* Roxb. ex DC.	Nepal, Doti District	29.32	81.02	N/A	30-Jun-09	20915008	
B_asiatica4	*B. asiatica* Roxb. ex DC.	India, no further details	N/A	N/A	C. Chadwell	N/A	AS82	Cultivated (J. Harber Coll.)
B_asiatica5	*Berberis asiatica* Roxb. ex DC.	Nepal, Narayani Zone	27.6541	85.09973	Adhikari, B. et al.	01-Aug-14	Coll_38.1	
B_asiatica6	*Berberis asiatica* Roxb. ex DC.	Nepal, Bagmati Zone	27.77278	85.43166	Adhikari, B. et al.	02-Sep-14	SB1	
B_calliantha	*B. calliantha* Mulligan	China, Tibet	28.91	89.61	F. Kingdon-Ward, Ex Hillier	21-Nov-24	AS38	Cultivated (J. Harber Coll.)
B_chilensis	*B. chilensis* Gillet	Región VII	N/A	N/A	Gardner et al.	22-Jan-90	19900509	Cultivated (RBGE)
B_chrysosphaera	*B. chrysosphaera* Mulligan	China, Tibet	28.65	97.46	F. Kingdon-Ward, Ex Hillier	10-Dec-33	AS39	Cultivated (J. Harber Coll.)
B_con_extensiflora1	*B. concinna* var. *extensiflora* Ahrendt	Nepal, Manang District	28.61	84.47	N/A	14-Aug-08	20812277	
B_con_extensiflora2	*B. concinna* var. *extensiflora* Ahrendt	Nepal, Myagdi District	28.4	83.69	N/A	04-Oct-06	EA104	
B_con_extensiflora3	*B. concinna* var. *extensiflora* Ahrendt	Nepal	N/A	N/A	C. Chadwell	N/A	AS74	Cultivated (J. Harber Coll.)
B_concinna	*B. concinna* Hook.f.	Nepal, Rasuwa District	28.1	85.38	Adhikari, B.	21-May-08	GB10	
B_concinna2	*Berberis concinna* Hook.f.	India, Sikkim	27.83472	88.69944	T D. Atkinson	05-Jul-05	AS102	
B_concolor	*B. concolor* W. W. Smith	China, Yunnan	28.47	98.91	D. E. Boufford et al.	20-Aug-13	43135	
B_congestiflora	*B. congestiflora* Gay	Chile, Región IX	N/A	N/A	Gardner et al.	19-Feb-88	1988.0916	Cultivated (RBGE)
B_cooperi	*B. cooperi* Ahrendt	Bhutan, Timphu	27.47	89.64	J. F. Harber s.n.	01-Aug-97	AS9	Cultivated (J. Harber Coll.)
B_crassilamba	*B. crassilamba* C. Y. Wu ex S. Y. Bao	China, Yunnan	27.61	99.89	D. E. Boufford et al.	04-Sep-13	43437	
B_darwinii	*B. darwinii* Hook.	Argentina : Prov. Río Negro	N/A	N/A	Unknown	N/A	1987.2408	Cultivated (RBGE)
B_derogensis	*B. derogensis* T. S. Ying	China, Sichuan	29.09	99.38	D. E. Boufford et al.	22-Aug-13	43164	
B_dictyophylla1	*B. dictyophylla* Franch.	China, Yunnan	27.89	99.68	B & S Wynn-Jones	17-Sep-00	AS93	Cultivated (J. Harber Coll.)
B_dictyophylla2	*B. dictyophylla* Franch.	China, Yunnan	25.94	100.4	Z. W. Liu s.n.	N/A	AS100	Cultivated (J. Harber Coll.)
B_empetrifolia	B. empetrifolia Lam.	Argentina, Tierra del Fuego	N/A	N/A	N/A	N/A	1976.1088A	Cultivated (RBGE)
B_everestiana1	*B. everestiana* var. *ventosa* Ahrendt	Nepal, Solu Khumbu District	27.86	86.64	N/A	23-Sep-05	DNEP3BY156	
B_everestiana2	*B. koehneana* C. K. Schneid.	Nepal, Mustang District	28.82	83.86	Adhikari, B.	16-Aug-07	EA217	
B_fendleri	*B. fendleri* A.Gray	N/A	N/A	N/A	N/A	N/A	N/A_2	Cultivated (RBGE)
B_glaucocarpa	*B. glaucocarpa* Stapf	Nepal, Doti District	29.35	81.06	N/A	01-Jul-09	20918011	
B_graminea	*B. graminea* Ahrendt	China, Sichuan	28.12	101.18	D. E. Boufford et al.	06-Sep-13	43466	
B_griffithiana1	B. *griffithiana* C.K.Schneid.	India, Arunchal Pradesh	27.58	91.88	SF 06008	24-Nov-06	AS55	Cultivated (J. Harber Coll.)
B_griffithiana2	B. *griffithiana* C.K.Schneid.	India, Arunchal Pradesh	27.33	92.31	A Clark 5260	01-Oct-04	AS54	Cultivated (J. Harber Coll.)
B_grodtmanniana	*B. grodtmanniana* C. K. Schneider	China, Sichuan	27.69	101.22	D. E. Boufford et al.	06-Sep-13	43471	
B_gyalaica1	*Berberis gyalaica* Ahrendt ex F.Br.	China, Tibet	29.65056	94.36	W. Bentall	27-Jun-05	WB	
B_gyalaica2	*Berberis gyalaica* Ahrendt ex F.Br.	China, Tibet	28.97444	93.69472	W. Bentall	NA	AS6	Cultivated (J. Harber Coll.)

Table Appendix AT 1 (continued)

Sample	Species	Locality	Lat.	Long	Collector(s)	Coll. Date	Voucher	Comments
B_hamiltoniana	*Berberis hamiltoniana* Ahrendt	Nepal, Bajhang District	29.61553	81.00556	Adhikari, B.	NA	20915095	
B_hamiltoniana1	*B. hamiltoniana* Ahrendt	Nepal, Humla District	29.98	81.81	N/A	21-Jun-08	JRSB162	
B_hamiltoniana2	*B. hamiltoniana* Ahrendt	Nepal, Bajhang District	29.62	81.01	N/A	13-Jul-09	20915095	
B_hookeri1	*B. hookeri* Lem.	Nepal, Panchthar District	27.11	87.94	Adhikari, B. et al.	08-Jun-07	LKSRB12	
B_hookeri2	*B. hookeri* Lem.	Nepal, Khumbu District	27.76	86.71	N/A	29-Sep-05	DNEP3BY213	
B_hookeri3	*B. hookeri* Lem.	Bhutan	27.42	90.21	J. F. Harber	01-Aug-97	AS29	Cultivated (J. Harber Coll.)
B_hookeri4	*B. hookeri* Lem.	Bhutan	N/A	N/A	Ruth Liddington	20-Jun-05	AS63	Cultivated (RBGE)
B_hookeri5	*Berberis wallichiana* DC.	Nepal, Panchthar District	27.10263	87.96897	Adhikari, B. et al.	8 June 2007	LKSRB28	
B_hookeri6	*Berberis hookeri* Lem.	Nepal, Myagdi District	28.4014	83.70257	Adhikari, B.	4 October 2006	EA106	
B_hookeri7	*Berberis hookeri* Lem.	Nepal, Myagdi District	28.40443	83.69923	Adhikari, B.	13 July 2009	Bajhang0920915095	
B_insignis	Berberis *insignis* Hook.f. & Thomson	Nepal, Illam District	27.06317	88.01702	Adhikari, B. et al.	16 June 2007	LKSRB144	
B_jaeschkeana1	*B. jaeschkeana* var. *usteriana* C.K.Schneid.	Nepal, Jumla District	29.32	82.18	N/A	03-Jun-08	JRSA12	
B_jaeschkeana2	*Berberis jaeschkeana* var. *usteriana* C.K.Schneid.	Nepal, Mustang District	28.71222	83.55889	Adhikari, B.	17 August 2007	EA238	
B_jamesiana2	*B. jamesiana* Forrest & W. W. Smith	China, Yunnan	26.11	100.17	D. E. Boufford et al.	14-Sep-13	43530	
B_karnalensis	*B. karnaliensis* Bh.Adhikari	Nepal, Jumla District	29.3	82.18	N/A	03-Jun-08	JRSA5	
B_koehneana	*B. koehneana* C. K. Schneid.	Nepal, Mustang District	28.68	83.6	N/A	30-Sep-06	EA56	
B_kumaonensis	*B. kumaonensis* C. K. Schneid.	Nepal, Doti District	29.38	81.12	N/A	02-Jul-09	20915029	
B_leptopoda	*B. leptopoda* Ahrendt	India, Arunchal Pradesh	28.57	95.06	K. Rushforth		AS103	Cultivated (J. Harber Coll.)
B_levis	*B. levis* Franch.	China, Yunnan	25.96	100.39	D. E. Boufford et al.	15-Sep-13	43557	
B_mekongensis	*B. mekongensis* W. W. Smith	China, Yunnan	28.33	99.12	D. E. Boufford et al.	19-Aug-13	43131	
B_micropetala	*B. micropetala* C.K.Schneid.	India, Manipur	24.67	93.92	N. Macer	04-Jul-05	AS104	Cultivated (J. Harber Coll.)
B_microphylla1	B. *microphylla* G.Forst.	N/A	N/A	N/A	N/A	N/A	1961.063803	Cultivated (RBGE)
B_microphylla2	B. *microphylla* G.Forst.	Chile, Región XI (Aisén)	N/A	N/A	Beavis, Derek S.	21-Mar-92	1992.2583	Cultivated (RBGE)
B_montana	*B. montana* Gay	Chile : Región X	N/A	N/A	Gardner et al.	15-Jun-05	1993.2827B	Cultivated (RBGE)
B_mucrifolia	*Berberis mucrifolia* Ahrendt	Nepal, Mustang District	28.71194	83.55889	Adhikari, B.	Nov 2009		
B_negeriana	B. *negeriana* Tischler	Chile, Región VIII	N/A	N/A	Hechenleitner Vega	11-Mar-04	200404971	Cultivated (RBGE)
B_nervosa	B. *nervosa* Pursh	Canada, British Columbia	N/A	N/A	Halliwell, Brian	23-Aug-78	1978.2559	Cultivated (RBGE)
B_nevinii	*B. nevinii* A. Gray.	N/A	N/A	N/A	Unknown	Unknown	HC1066	Cultivated (Rancho Santa Ana Botanical
B_newsppA	*Berberis* new_speciesA	China Yunnan	27.53	99.64	D. E. Boufford et al.	31-Aug-13	43334	
B_newsppB	*Berberis* new_speciesB	China Yunnan	28.57	99.83	D. E. Boufford et al.	31-Aug-13	43304	
B_orthobotrys1	*B. orthobotrys* var. *rubicunda* Ahrendt	Nepal, Rasuwa District	28.21	85.53	Adhikari, B.	03-Aug-07	BL239	
B_orthobotrys2	*B. orthobotrys* var. *rubicunda* Ahrendt	Nepal, Khumbu District	27.79	86.71	N/A	12-Sep-05	DNEP3BY22	
B_pendryi	*B. pendryi* Bh.Adhikari	Nepal, Mustang District	28.82	83.87	Adhikari, B.	16-Aug-07	EA25	
B_pendryi2	*Berberis pendryi* Bh.Adhikari	Nepal, Mustang District	28.81694	83.87	Adhikari, B.	16 August 2007	EA29	
B_petiolaris1	*B. petiolaris* Wall. ex G. Don	Nepal, Mugu District	29.65	82.11	N/A	12-Jun-08	JRSA122	
B_petiolaris2	*B. petiolaris* Wall. ex G. Don	Nepal, Mugu District	29.65	82.11	N/A	12-Jun-08	JRSA122	Technical Replicate
B_phanera	*B. phanera* C.K. Schneider	China, Sichuan	28.12	101.18	D. E. Boufford et al.	06-Sep-13	43465	
B_polyodonta	*B. polyodonta* Fedde	China Yunnan	N/A	N/A	Lijiang et al.	12-Jun-05	1991.1138	Cultivated (RBGE)
B_praecipua	B. *praecipua* C.K.Schneid.	Bhutan	27.32	89.55	Ruth Liddington	20-Jun-05	AS64	Cultivated (J. Harber Coll.)
B_pruinosa	*B. pruinosa* Franch.	China, Yunnan	27.46	99.9	D. E. Boufford et al.	04-Sep-13	43442	
B_pruinosa2	*Berberis pruinosa* Franch.	China, Yunnan	26.11111	99.95083	A. Clark	NA	AS106	Cultivated (J. Harber Coll.)
B_pseudotibetica	*B. pseudotibetica* C. Y. Wu	China, Yunnan	28.29	99.16	D. E. Boufford et al.	19-Aug-13	43134	
B_qiaojianensis	*B. qiaojianensis* S. Y. Bao	China, Yunnan	26.19	103.27	D. E. Boufford et al.	19-Sep-13	43528	
B_rotundifolia	*B. rotundifolia* Poepp. & Endl.	Chile	N/A	N/A	Hechenleitner Vega	26-Jun-05	20080789 C	Cultivated (RBGE)
B_spp1	*Berberis* spp.	Nepal, Panchthar District	27.10389	87.9475	Adhikari, B. et al.	8 June 2007	LKRSB17	
B_temolaica	*Berberis telomaica* Ahrendt	China, Tibet	29.2169	94.21528	A. Clark	NA	AS67	Cultivated (J. Harber Coll.)

Table Appendix AT 1 (continued)

Sample	Species	Locality	Lat.	Long	Collector(s)	Coll. Date	Voucher	Comments
B_thomsoniana	*Berberis thomsoniana* C.K.Schneid.	Nepal, Myagdi District	28.40217	83.70247	Adhikari, B.	3 October 2006	EA101	
B_thomsoniana1	*B. thomsoniana* C.K.Schneid.	Nepal, Panchthar District	27.1	87.95	Adhikari, B. et al.	08-Jun-07	LKSRB17	
B_thomsoniana2	*B. thomsoniana* C.K.Schneid.	Nepal, Jumla District	29.37	82.15	N/A	05-Jun-08	JRSA49	
B_thomsoniana3	*B. thomsoniana* C.K.Schneid.	Nepal, Rasuwa District	28.1	85.36	Adhikari, B.	21-May-08	GB14	
B_thomsoniana4	*B. thomsoniana* C.K.Schneid.	Nepal, Myagdi District	28.4	83.7	N/A	03-Oct-06	EA101	
B_tibaoshanensis	*B. tibaoshanensis* S. Y. Bao	China, Yunnan	27.61	99.89	D. E. Boufford et al.	04-Sep-13	43436	
B_tsarica	*B. tsarica* Ahrendt	Nepal, Solu Khumbu District	27.94	86.61	N/A	20-Sep-05	DNEP3BY132	
B_tsarica1	*Berberis tsarica* Ahrendt	Nepal, Khumbu District	27.94111	86.61	Adhikari, B. et al.	20 September 2005	DNEP3BY132	
B_wallichiana1	*B. wallichiana* DC.	Nepal, Panchthar District	27.1	87.97	Adhikari, B. et al.	08-Jun-07	LKSRB28	
B_wallichiana2	*B. wallichiana* DC.	Nepal, Rasuwa District	28.17	85.36	Adhikari, B.	02-Aug-07	BL220	
B_wallichiana3	*B. wallichiana* DC.	Nepal	N/A	N/A	Chadwell C.	N/A	JH2	Cultivated (J. Harber Coll.)
B_wardii	*Berberis wardii* C.K.Schneid	India, Assam	26.00472	94.99806	F. Kingdon-Ward	NA	AS66	Cultivated (J. Harber Coll.)
B_wilsoniae1	*B. wilsoniae* Hemsley	China, Yunnan	27.61	99.72	D. E. Boufford et al.	31-Aug-13	43337	
B_wilsoniae2	*B. wilsoniae* Hemsley	China, Yunnan	24.96	102.66	Z. W Liu	N/A	AS99	Cultivated (J. Harber Coll.)
B_wilsoniae3	*B. wilsoniae* Hemsley	China, Yunnan	29.99	101.95	X. H. Li	05-Jul-05	AS98	Cultivated (J. Harber Coll.)
B_wilsoniae4	*B. wilsoniae* Hemsley	China, Yunnan	27.34	103.72	A. Clark	17-Jun-05	AS12	Cultivated (J. Harber Coll.)

Appendix Table AT-2 Sequencing information. The seqeuencing strategy describes whether the sample was target enriched (TE), shotgun sequenced (SG) or borth (TE + SG). Numbers in the row "Capture" indicates which samples were pooled in the hybridization capture. Furthermore, the average coverage and standard deviation (Stdev) are displayed.

		Sequcning	Plastid		Coverage Nuclear		Coverage Plastid	
Sample	Sequencer	strategy	Phylo	Capture	Average	Stdev	Average	Stdev
B_angulosa1	NextSeq	TE+SG	yes	6	371.8	380.8	353.7	155.7
B_angulosa2	NextSeq	TE+SG	yes	5	447.0	436.0	373.3	130.4
B_angulosa3	NextSeq	TE+SG	yes	4	198.5	411.7	144.4	39.4
B_angulosa4	MiSeq	TE	no	2	584.5	530.5	-	-
B_aristata1	NextSeq	TE	no	8	47.3	59.8	-	-
B_aristata2	NextSeq	TE	no	8	86.2	110.1	-	-
B_aristata3	NextSeq	TE	yes	4	215.6	278.5	209.1	94.0
B_aristata4	NextSeq	TE	yes	5	61.5	79.2	20.6	22.1
B_aristata5	NextSeq	TE	no	8	42.9	61.5	-	-
B_asiatica1	MiSeq	TE	no	1	348.3	418.5	-	-
B_asiatica2	NextSeq	TE	yes	4	389.3	389.0	135.3	110.2
B_asiatica3	NextSeq	TE	no	4	296.8	291.2	-	-
B_asiatica4	NextSeq	TE+SG	yes	4	267.3	468.3	360.3	86.1
B_calliantha	NextSeq	TE+SG	yes	4	452.2	589.8	83.7	35.3
B_chilensis	NextSeq	TE+SG	no	6	61.2	125.8	-	-
B_chrysosphaera	NextSeq	TE+SG	yes	4	437.5	521.4	46.5	21.4
B_con_extensiflora1	NextSeq	TE+SG	yes	5	236.0	250.7	68.7	49.2
B_con_extensiflora2	NextSeq	TE+SG	yes	7	122.2	144.5	-	-
B_concinna	NextSeq	TE	no	4	311.1	311.1	-	-
B_con_extensiflora3	NextSeq	TE+SG	yes	4	243.7	483.6	235.5	52.1
B_concolor	MiSeq	TE+SG	yes	3	148.3	398.4	-	-
B_congestiflora	NextSeq	TE+SG	no	6	276.3	483.8	-	-
B_cooperi	NextSeq	TE+SG	yes	5	98.2	190.2	172.9	40.9
B_crassilamba	NextSeq	TE+SG	yes	5	555.4	655.8	188.5	64.0
B_darwinii	NextSeq	TE+SG	no	6	78.9	163.0	24.9	22.3
B_derogensis	NextSeq	TE+SG	yes	7	146.3	472.2	36.7	34.1
B_dictyophylla1	NextSeq	TE+SG	yes	4	370.3	496.1	98.4	30.7
B_dictyophylla2	NextSeq	TE	yes	4	198.8	370.1	133.8	27.4
B_empetrifolia	NextSeq	TE+SG	no	6	20.1	50.2	-	-
B_everestiana1	NextSeq	TE+SG	yes	8	67.9	153.1	99.7	64.8
B_everestiana2	NextSeq	TE+SG	yes	4	371.9	355.1	60.1	21.8
B_fendleri	NextSeq	TE+SG	yes	6	212.0	393.8	43.7	31.0
B_glaucocarpa	NextSeq	TE+SG	yes	7	111.8	157.6	164.8	63.9
B_graminea	NextSeq	TE+SG	yes	8	258.9	306.6	133.5	70.7
B_griffithiana1	NextSeq	TE+SG	yes	4	424.8	618.9	69.3	31.5
B_griffithiana2	NextSeq	TE	yes	5	278.4	420.2	57.7	28.7
B_grodtmanniana	MiSeq	TE+SG	yes	3	194.8	682.6	112.3	66.1
B_hamiltoniana	MiSeq	TE	yes	2	121.0	7.9	25.4	16.9
B_hookeri1	NextSeq	TE	no	4	195.5	342.5	-	-
B_hookeri2	NextSeq	TE	yes	5	354.5	402.2	16.2	14.4
B_hookeri3	NextSeq	TE	no	4	340.2	492.4	-	-
B_hookeri4	NextSeq	TE	no	5	318.9	460.3	-	-
B_hookeri5	NextSeq	TE+SG	no	5	399.4	421.4	42.5	25.1
B_jaeschkeana1	MiSeq	TE+SG	yes	3	134.0	250.6	72.2	26.5

Appendix Table AT-2 (continued)

Sample	Sequencer	Sequcning strategy	Plastid Phylo	Capture	Coverage Nuclear		Coverage Plastid	
					Average	Stdev	Average	Stdev
B_jamesiana2	MiSeq	TE+SG	yes	3	123.1	279.8	97.2	67.3
B_karnalensis	NextSeq	TE+SG	yes	8	93.9	131.0	16.3	17.3
B_koehneana	NextSeq	TE+SG	yes	5	513.0	506.3	32.0	24.9
B_kumaonensis	NextSeq	TE+SG	yes	7	46.7	93.0	53.2	27.8
B_leptopoda	NextSeq	TE+SG	yes	4	457.1	600.2	112.9	48.3
B_levis	NextSeq	TE+SG	yes	5	343.1	377.9	27.2	13.8
B_mekongensis	NextSeq	TE+SG	yes	7	127.2	166.3	106.9	47.1
B_micropetala	NextSeq	TE+SG	yes	5	264.6	378.8	55.9	27.5
B_microphylla1	NextSeq	TE+SG	no	6	286.7	476.3	-	-
B_microphylla2	NextSeq	TE+SG	no	6	19.3	67.0	-	-
B_montana	NextSeq	TE+SG	no	6	168.5	376.6	196.3	88.6
B_newsppA	MiSeq	TE+SG	yes	3	137.2	206.7	415.5	165.6
B_newsppB	NextSeq	TE+SG	yes	4	415.8	451.5	149.0	72.0
B_negeriana	NextSeq	TE+SG	no	6	255.4	411.8	142.9	74.2
B_nervosa	NextSeq	TE+SG	no	6	355.5	501.3	87.5	54.6
B_nevinii	NextSeq	TE+SG	no	7	241.4	423.1	-	-
B_orthobotrys1	NextSeq	TE	yes	5	305.3	268.8	186.3	55.8
B_orthobotrys2	NextSeq	TE+SG	yes	8	35.2	47.3	16.9	10.7
B_pendryi	NextSeq	TE+SG	no	8	44.5	58.7	-	-
B_petiolaris1	NextSeq	TE+SG	yes	4	170.4	237.5	24.3	14.0
B_petiolaris2	NextSeq	TE	yes	8	340.7	431.4	80.1	31.7
B_phanera	NextSeq	TE+SG	yes	7	197.5	273.3	310.1	113.8
B_polyodonta	NextSeq	TE+SG	no	6	463.1	695.0	664.4	239.9
B_praecipua	NextSeq	TE+SG	yes	5	327.7	555.5	249.1	83.8
B_pruinosa	NextSeq	TE+SG	yes	4	266.8	351.2	16.1	16.1
B_pseudotibetica	NextSeq	TE+SG	yes	7	101.8	121.6	43.2	32.6
B_qiaojianensis	NextSeq	TE+SG	yes	7	113.8	170.7	361.7	124.4
B_rotundifolia	NextSeq	TE+SG	no	6	90.7	292.6	-	-
B_spp3	NextSeq	TE+SG	no	6	145.5	234.9	-	-
B_spp1	NextSeq	TE+SG	yes	6	121.3	177.7	36.3	33.6
B_spp2	NextSeq	TE+SG	no	4	23.5	40.2	-	-
B_thomsoniana	NextSeq	TE+SG	yes	5	346.4	316.7	50.3	23.3
B_tibaoshanensis	NextSeq	TE+SG	yes	7	112.7	160.0	368.1	121.8
B_tsarica1	NextSeq	TE+SG	yes	5	161.3	234.3	91.8	34.0
B_tsarica2	MiSeq	TE+SG	no	3	55.0	124.4	-	-
B_wallichiana1	NextSeq	TE+SG	yes	7	149.8	325.6	177.9	82.3
B_wallichiana2	NextSeq	TE	yes	5	371.2	610.7	140.2	71.7
B_wilsoniae1	MiSeq	TE+SG	yes	3	139.5	172.7	131.8	62.7
B_wilsoniae2	NextSeq	TE	yes	5	327.9	365.6	90.8	38.9
B_wilsoniae3	NextSeq	TE+SG	yes	5	268.0	360.1	46.1	27.3
B_wilsoniae4	NextSeq	TE	no	5	369.1	537.7	-	-