THE UNIVERSITY *of* EDINBURGH

CHARACTERIZING THE GENETIC BASIS OF PLANT SPECIES DIFFERENCES AND ITS APPLICATION TO SPECIES DISCRIMINATION

Wu Huang

Thesis submitted for the Doctor of Philosophy

The University of Edinburgh

Royal Botanic Garden Edinburgh

2022

Declaration

I hereby declare that this thesis has been composed by myself. Any included research is my own work, except where indicated throughout the thesis and summarised and clearly identified on the declarations page of the thesis. All work contained herein has not been submitted for any other degree. I have acknowledged the nature and extent of work carried out in collaboration with others.

Signature of candidate:

Wu Huang

Abstract

Telling species apart plays a key role in understanding global biodiversity, monitoring change, and managing biodiversity. However, species discrimination is often difficult due to the sheer volume of species on earth and the complexity of the nature of species. This led to the development of DNA barcoding which uses standardised regions of DNA for species identification. The standard plant DNA barcodes are based on small regions from the plastid genome and ribosomal DNA. The approach works well in some plant groups, but does not provide unique species-level resolution in many others.

In this thesis, I explore the potential for using data from the nuclear genome for improving and enhancing species discrimination in plants. Access to the nuclear genome via high-throughput sequencing now enables the generation of large amounts of sequence data from multiple unlinked loci. Such data offer the opportunity for designing the next generation of plant barcoding approaches based on a detailed understanding of genomic differences between species. Various studies have shown the ability of multiple unlinked nuclear markers to provide high discriminatory power in many plant groups, separating species, and infra-specific taxa. But there has not yet been an overview and synthesis of exactly how powerful these approaches can be, and how best to guide future efforts in building plant identification tools.

In this thesis, I provide a first overview of how nuclear genomes perform in telling species apart. Overall I tackled the following questions 1) what is the proportion of species that are distinguishable with nuclear markers? 2) what is the nature of the inter-specific differences and what are the attributes of loci that are the most informative in telling species apart? And 3) how many markers are needed and what markers are needed to maximise the species identification success?

To answer those questions, I first outlined the conceptual issues to address in assembling and analysing appropriate multi-locus nuclear sequence datasets. I then developed a new pipeline called NucBarcoder which supports workflows and analysis using multi-locus nuclear sequence data for species discrimination. I then tested this workflow on a case study, consisting of a dataset of sequence data from 810 nuclear genes from 453 individuals from 133 *Inga* species including 69 species which were represented by multiple-sampled individuals. Of the 69 species with multiple individuals sampled, 45 resolved as monophyletic (65%). The density of species-specific SNPs for each *Inga* species ranged from 0 to 1503 per megabase. Compared to the full dataset of 810 genes and 205,871 SNPs, subsampling analysis revealed that a random selection of 70 genes or 2500 SNPs, or a combination of 9 ‘best performing’ genes could achieve levels of species discrimination success similar to the full dataset. I found a positive correlation ($r = 0.42$) between the number of species distinguished and the nucleotide diversity of the genes used for species discrimination.

To search for broader generalisations, I then compiled data from 149 different genera to assess the proportion of plant species that resolve as monophyletic. I then selected 29 genera with suitable available data for further study and calculated the abundance and density of species-specific SNPs (SSSNPs), and the proportion of species that can be distinguished by different subsets of the data and also by targeting the best-performing gene regions. Finally, I evaluated the characteristics of the best-performing gene regions in terms of levels of nucleotide diversity and density of SSSNPs. In the

analysis of 149 genera, overall, of the 1,701 multiple-sampled species evaluated 1,206 resolved as monophyletic (71%). At the level of individual genera, 37 of the 149 genera (25.8%) had 100% of species resolved as monophyletic, and 75 (50.3%) genera had at least 75% of the species resolving as monophyletic. Among the 29 genera examined in more detail, the density of SSSNPs of all species ranged from 0 to 27,262 per Mb, with a median density of 323 SSSNPs per Mb (a median density of one SSSNP every 3,098 bp). Of the total of 460 species from 29 genera assessed, 411 species (89.3%) had at least one SSSNP. Resampling of these datasets showed that with around 3,000 SNPs, almost all genera have asymptoted in their levels of species discrimination, with 21/23 genera (91%) having >85% of their distinguishable species (e.g. those told apart in the full data set) distinguished with 3,000 randomly selected SNPs. Furthermore, in a detailed investigation of six genera, there are clearly some loci that are much better than others in telling species apart. Between one and nine pre-selected genes were able to recover equivalent levels of species discrimination compared to several hundred genes from the full datasets. A closer investigation of the attributes of the best-performing genes showed some positive correlations between the number of species resolved as monophyletic and the nucleotide diversity of a given gene, although this relationship was not clear cut, and the genes that give the highest species resolution are not always the most variable genes.

These findings provide some key general information on the proportions of plant species that are resolvable using multi-locus nuclear sequence data from plants and the nature of the sequence variation between plant species. In the final chapter of the thesis, I summarise these findings and identify a set of priority research and infrastructure needs to take forward the development and use of multi-locus nuclear DNA barcoding of plants.

Acknowledgment

I am grateful that this Ph.D. has been a happy and enjoyable journey. Lots of people have contributed to building this nourishing environment where I grow intellectually, mentally, and physically.

To my supervisors Pete and Alex, you have been the best mentors I could ever have wanted. The knowledge you imparted, the freedom you gave, and the support you provided have helped me to flourish, be independent, and feel profoundly secure. I bless the day I met you at IBC in 2017. I was lost and carrying the thoughts that my future would probably be doomed. But you led me into this interesting world of DNA barcoding and biodiversity genomics where I am developing a career.

Pete, you not only guided me in science but also demonstrated to me the best example of effective leadership. You listen and cheer me up in all circumstances which kept me happy and light-hearted. These feelings are especially priceless in the time of COVID-19. You support me in every possible aspect unconditionally so that I feel safe to explore the world because I know you'll be there. And you are utterly understanding and considerate and taught me to respect myself and my emotions, and that fosters a radical acceptance of my true self.

Alex, you are my model for developing a career with passion and love. Your love of plants and your dedication to science has encouraged me to pursue my interests. The inclusivity of the Twyford lab gave me a sense of belonging throughout my Ph.D. - I can't imagine how lonely I would be without the welcoming vibe from this cohort. You are also the assurance that I'd get support when Pete's schedule gets tight, and this adds to the sense of security throughout the four years.

Now I have come to another crossroads in my life, I still don't know what the right path is, but I no longer feel helpless and lost, as I know, I will meet people as great as both of my supervisors, and they will guide me through the next stages of life as you have done. The best thing is, my inner self has grown strong enough to be my own guide.

My love also goes to the Royal Botanic Garden Edinburgh and the University of Edinburgh which provide resource to build my academic and personal skills, a platform to develop myself, and a safe harbour that wherever I go, I'll carry the identity of being a student in Edinburgh for the rest of my life and I'm proud of it. Also my deepest thanks to my funder the Darwin Trust of Edinburgh who saw the potential in me and made my wonderful journey as a Ph.D. student possible.

For people who would never read my thesis, let alone this acknowledgment, but are all important parts to my successful Ph.D. and a happy life, this is a reminder to give them hugs and say thank you. My friends Alejandra, Zhao Ning, Apple and Tom, Hanna, Huo Yu, Joan and Rebecca, Ding, Meng and Rowan, David and Xin, Nahuel and Nadine, Wenyue, Alex, Aliz, Andrea and Yunyu, Heather, and my cohort in RBGE and Twyford lab, Natalia, Cynthia, Gustavo, Pakkapol, Flavia, Zhifang, Pengcheng, Madhavi, Hanna, Tibo, and Surabhi, Lucia, Emily, Hannes, Oriane, Mauricio, Mario, and Max.

I am happy that I am writing this at home in China, and I'll keep giving as much company as I can to my mom, dad, and grandma.

I am indebted to those who have contributed sequence data that were included in the thesis which are the foundations of my Ph.D. Thanks to the generosity of not only sharing the data but addressing any questions I had, however naïve and stupid, to help me move forward. Here is a dedication to my collaborators and data providers:

Many thanks to the following people from the Royal Botanic Garden Edinburgh and the University of Edinburgh for providing the first few datasets that enabled me to develop the pipeline NucBarcoder:

- Alex Twyford, Mario Duran, and Andrew Hudson for the *Antirrhinum* dataset
- Toby Pennington, Catherine Kidner, and Kyle Dexter for the *Inga* dataset
- Alex Twyford, Hannes Becher, and Yanqian Ding for the *Euphrasia* datasets
- Oriane Loiseau for the *Geonoma* dataset
- Natalia Contreras-Ortiz for the *Lupinus* dataset

And to people from Kunming Institute of Botany who generously shared unpublished data:

- De-zhu Li, Lian-ming Gao, Shuang-xiu Xu, and Han-tao Qin for the *Taxus* and *Camellia* datasets

And thanks to people from the Royal Botanic Garden Kew for providing datasets from various study systems:

- Alexandre Antonelli, Christine Bacon for the *Attalea* and *Syagrus* datasets
- Rowan Schley for the *Brownea* datasets

Many thanks to the Soltis lab from the University of Florida

- Pamela Soltis, Douglas Soltis, Gaynor Shelly for the *Diapensia* dataset
- Andre A. Naranjo for the *Dicerandra* dataset
- Jacob B Landis for the *Linanthus* and *Leptosiphon* datasets
- Joanna Jantzen for the *Tibouchina* dataset

And finally, I am grateful to have this hardworking and supportive community who made their data available to the public, especially those who have responded to my inquiries and generously offer help when I was in need.

- Andrew Hipp, Jeannine Cavender-Bares for the *Quercus* dataset
- Natascha Dorothea Wagner for the *Salix* dataset
- Morrgan Gostel for the *Commiphora* dataset
- Elliot Gardner for the *Artocarpus* dataset
- Khalid Sedeek for the *Ophrys* dataset
- Camille Christe for the *Capurodendron* dataset
- Qiu-yun Xiang, Kira Lindelof for the *Cornus* dataset
- Zhi-Yuan Du for the *Aesculus* dataset
- Madeline Chase for the *Mimulus* dataset

- Yuan-Yuan Feng for the *Tsuga* dataset
- Jess Stephens for the *Sarracenia* dataset
- Jeffrey Rose for the *Polemonium* dataset
- Xiu-qun Liu for the *Vitis* dataset
- Mario Fernández-Mazuecos for the *Linaria* dataset

Preface

The thesis is structured into six chapters. These consist of

Chapter 1: A general introduction to the topic of telling plant species apart, the strengths and limitations of existing DNA barcodes, and the potential for the use of sequence data from the nuclear genome to improve levels of species discrimination.

Chapter 2: A consideration of some key issues and challenges that need to be addressed in developing methods for using multi-locus sequence data for plant species discrimination

Chapter 3: A technical chapter summarising a newly developed pipeline (NucBarcoder) for supporting the analysis of telling plant species apart with multi-locus DNA sequences

Chapter 4: A case study focusing on a complex and species rich genus (*Inga*), utilising the NucBarcoder pipeline to test how multi-locus nuclear sequence data can be used to tell plant species apart

Chapter 5: The main analysis in the thesis consisting of a meta-analysis of multiple studies, evaluating the extent of plant species monophyly from multi-locus sequence data, the distribution of species specific SNPs, and exploring the minimal amount of data necessary to give maximal species discrimination.

Chapter 6: A summary chapter and brief forward looking perspective identifying future research priorities.

From these six chapters I plan to submit three or four papers. These will be:

1) A technical paper outlining the scripts and protocols used for assessing the signal in multi-locus nuclear sequence data from plants for species discrimination. This paper is likely to be submitted to a methodological journal such as Protocols.io, and will be based heavily on the existing text in Chapter 3. The co-authors will be Alex Twyford and Pete Hollingsworth
2) The main meta-analysis paper, giving an overview of the headline findings of the proportion of plant species that resolve as monophyletic, the distribution of species specific SNPs among plant species, and the minimum amounts of data required to tell plant species apart. This paper is based on the analyses described in Chapter 5 and is currently at an advanced draft stage with submission planned for PNAS. The co-authors will be Alex Twyford and Pete Hollingsworth, along with the collaborators who have provided datasets used in that paper. A list of collaborators could be found in the acknowledgement.
3) The next paper(s) will either be a focal paper exploring the issues of species discrimination in detail, focusing on the *Inga* case study, and/or a more general perspective, drawing together strands from Chapters 1, 2 and 6 to outline a road-map for developing next generation barcodes for plants.

Table of Content

Chapter 1 Introduction

An estimated 25% of species on earth are threatened with extinction due to large-scale environmental change and human pressures on the natural environment (IPBES, 2019). To address this global biodiversity crisis, there is a pressing need to enhance our understanding of biodiversity and monitor biodiversity change, and to guide interventions to address its decline. Species are a fundamental component of biodiversity, yet they can be difficult to tell apart due to intrinsic morphological and genetic complexities (Mallet et al., 2016). Improved methods of telling plant species apart are therefore important to assist efforts to respond to the biodiversity crisis. The specific focus of this project is to get a better understanding of the nature of the genomic differences between plant species, to guide the improvement of DNA-based methods for telling plant species apart.

1.1. Species concept

At the outset of a thesis aiming to understand the genomic nature of plant species and to improve methods of telling species apart, it is important, first of all, to consider and define what a species is. There is an extremely rich literature on species concepts (De Queiroz, 2007), with many hundreds of papers written, and many ongoing and unresolved disagreements. Some examples of different species concepts are outlined below.

The biological species concept is defined as natural mating among groups of organisms that produces viable and fertile offspring (Gregor, 1940, Mayr, 1942, Dobzhansky, 1950). As one of the oldest, most popular and well-accepted species concept, this definition emphasises genetic compatibility rather than morphological conformity. The biological species concept focuses on biological interactions and natural processes rather than being merely a set of instructions for the demarcation of species, and it is on this that a species concept must be based according to Mayr (Mayr, 1942). One obvious limitation of this concept is that it does not take into account organisms which reproduce by autogamy or asexual reproduction, and in practice observing the mating behaviour and reproductive outcomes of many species is operationally impractical for many taxonomic groups, especially plants.

The ecological species concept was introduced based on the occupation of a distinct geographical range and ecological niche (Van Valen, 1976). Adapting to a set of environmental components offers a good example to answer the Darwinian 'why' question as to why species exist. Two main challenges to the ecological species concept are firstly that widespread species can occupy heterogeneous niches, yet clearly represent one species based on other criteria. Another challenge that refutes this concept is the opposite to the first point – it is about when a single set of offspring from the same parents differentiates and demonstrates the feature as multiple species should do. An example for this is the trophic species of cichlids (Meyer, 1990).

According to Nixon and Wheeler (Nixon et al., 1990) the phylogenetic species concept defines a species as "the smallest aggregation of populations (sexual) or lineages (asexual) diagnosable by a unique combination of character states in comparable individuals". It was originally based on the clustering of similar morphological characters. However, with the availability of multi-locus or even whole-genome sequence data, and to suit the purposes of phylogeny reconstruction, cladistic classification, and the study of evolutionary process, the phylogenetic species concept was then supplemented with several alternative properties with different emphases. These properties are: 1) Monophyletic – a species should consist of an ancestor and all of its descendants (Donoghue, 1985). This is under the premise that the species has been established long enough to accumulate abundant mutations. It also relates to the mode of speciation events which will be discussed later. 2) Diagnosable (qualitative, fixed differences) – certain loci with critical mutations associated with adaptivity, sexual compatibility, and reproductive functions being involved during the speciation process, i.e. speciation genes, and fixed post-speciation events (Nixon et al., 1990). 3) Hennigian – The ancestor becomes extinct when lineage splits (Hennig, 1979). 4) Exclusive coalescence of alleles - many alleles within a species should coalesce to a single ancestor within that species.

In practice, for the vast majority of species, the operational species concept used is a typological concept, which is based on taxonomists identifying morphological discontinuities. Based on those morphological discontinuities, taxonomists assign a species name and develop a description and diagnosis associated with that name, and the approach is very much focused on describing biological discontinuities, rather than focusing on evolutionary processes that have led to those discontinuities.

1.2. Approaches for telling species apart

As noted above, the vast majority of species have been delimited and are identified using morphological characters. Since Linnaeus's first treatise on species description (Linné, 1753), the long history of this morphology-based identification has drawn on (and supports the development of) an abundance of historical records (such as herbarium, museum, and garden collections). The well-established standards of morphological descriptions, identification keys and guides, and voucher specimens in museums and herbaria has created an effective, widely used, and accessible system. However, the approach does have limitations. First of all, where the differences between species are subtle, recognising diagnostic characters requires high-levels of taxonomic expertise and the task of telling species apart is often slow and demanding. Secondly, in order to identify species using morphology, it is often necessary to have access to adult fertile specimens; e.g. the specimen must be intact and in its full-grown adult shape, preferably with flowers and fruits for plants. Thus this method requires access to optimal material which is not always available. This is particularly true for studies looking at ecological processes (e.g. recruitment/pollination/diet) or for forensic/authentication studies where the material available is often juvenile or fragmentary or processed (Ogden et al., 2009, Hollingsworth et al., 2016).

Other popular approaches for telling species apart include various protein-based or chemical methods and these continue to be deployed. A recent study successfully identified the moth species of origin of wild silk used in antiquity by examining unique peptides using protein mass spectrometry (Lee et al., 2022). Another successful study examined the peak patterns from gas-liquid chromatography to successfully distinguish two honey bee species (Lavine & Carlson, 1987). This use of chemical fingerprinting, or 'chemocoding' has been promoted as a robust way for distinguishing morphologically similar species at a single site and for identifying widespread species across continental-scale ranges (Endara et al., 2018).

Such approaches are widespread, and build on a rich history of using protein polymorphism in biosystematics, or secondary compounds in chemotaxonomy (Soltis et al., 1989). These protein-based or chemical approaches provide another set of characters that are argued to be less susceptible to plastic responses to the environment than morphological-based methods (Lee et al., 2022). However, these approaches are based on the downstream expression products of genes and they can be susceptible to tissue-specific profiles and to environmentally induced variation. The resulting data are also fundamentally complex in nature, often consisting of different ratios of product presence, or combinations of products, making analysis complex across taxa and across analytical platforms.

Recognising the challenges using morphological or chemical methods of species discrimination, there is an enormous opportunity and an enormous globally distributed effort to use DNA-based methods to improve species discrimination. DNA is universal across eukaryotic organisms and stable enough such that genetic information can be recovered from a wide range of specimens and samples, including museum specimens and even from fossil specimens. The basic property of a universal genetic code, and four discrete character states (A, C, G, T) which do not show developmental or environmental plasticity make DNA a prime target for a global species identification system (Hebert et al., 2003).

1.3. Speciation modes and mechanisms

A prerequisite to thinking about how best to deploy DNA-based methods to tell plant species apart is considering, firstly, the evolutionary processes that lead to plant speciation, and secondly, the genetic and sequence diversity associated with these processes. In this section, I summarise some key modes and mechanisms of speciation and the expected phylogenetic patterns that arise from these processes (i.e., species monophyly, paraphyly, and polyphyly).

1.3.1. Three major geographical scales of speciation

Speciation categories were originally defined by the proportion of geographic range overlap of diverging populations (Fitzpatrick et al., 2009), namely allopatric speciation, sympatric speciation, and parapatric speciation.

Allopatric speciation is the classic speciation by isolation model. For populations that are geographically isolated, the physical separation restricts gene flow between populations, leading to independent evolutionary trajectories. Over time, genetic divergence accumulates due to various evolutionary processes such as genetic drift, mutation, and natural selection acting independently in each isolated population.

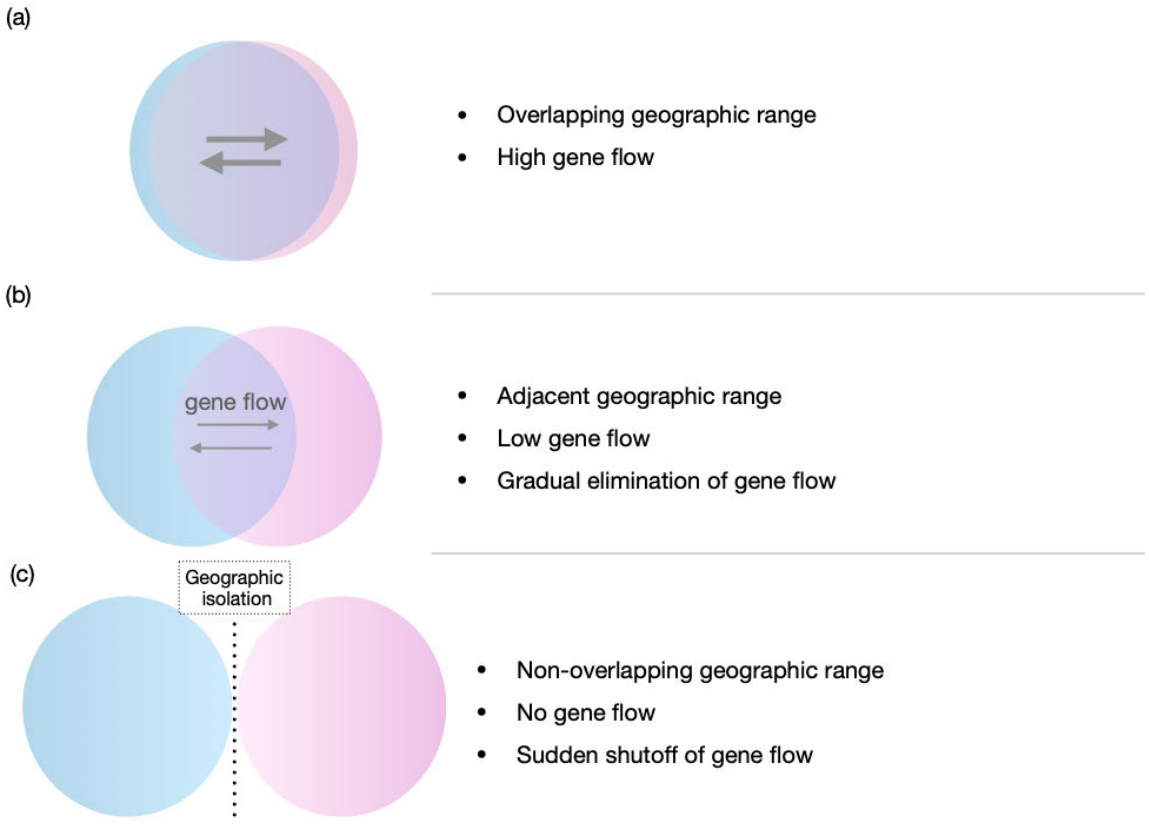


Figure 1.1. Speciation modes with different levels of geographic overlap and gene flow; (a) sympatric speciation. (b) parapatric speciation. (c) allopatric speciation. Pink and Blue circles represent two diverging populations, with different levels (thickness) of initial gene flow (arrows) displayed.

Parapatric speciation refers to divergence across a species range in the face of partial restrictions on gene flow. Some level of geographical restriction to gene flow, can promote divergence when (semi-isolated) populations experience different ecological conditions. Thus differentiation and speciation can occur where selective forces promote divergence, and the homogenising influence of gene flow is comparatively weak compared to the forces of divergence (Gavrilets, 2004).

Sympatric speciation is defined as a situation where populations co-occur in their range, and previously experienced a high level of reciprocal gene flow (Kondrashov et al., 1986, Kawecki, 2004). The divergence of two populations happens due to an initial mutation that promotes divergence, and may trigger a cascade of further changes which leads to an elimination of gene flow (Kawecki, 2004).

1.3.2. Species diversification mechanisms in relate to patterns of DNA divergence

Within these geographical contexts there are different diversification mechanisms including speciation due to reproductive incompatibilities that arise following divergence in isolation, and/or due following adaptations to different environmental conditions, as well as sometimes more abrupt forms of speciation promoted by hybridisation, chromosomal/ploidy changes, and speciation due to mating system shifts (Rieseberg et al., 2007). These leads to different patterns of DNA divergence which account for the ability to distinguish species.

Patterns of DNA divergence in allopatric speciation typically involve the accumulation of fixed differences or unique alleles in each isolated population. As the populations experience different selective pressures and genetic drift, genetic differences accumulate, leading to distinctive genomic signatures. The accumulation of fixed differences or unique alleles in allopatric populations provides clear genetic boundaries between species. These differences can be detected by DNA sequencing and genotyping, and can be used to identify and distinguish species.

In sympatric speciation, genetic divergence occurs despite the absence of physical barriers. DNA divergence in sympatric speciation can be more nuanced compared to allopatric speciation. It may involve subtle genetic variations, such as allele frequency differences or changes in gene expression patterns, rather than fixed differences. This can make it challenging to distinguish species solely based on traditional DNA sequence analysis.

In the parapatric speciation mode, populations occupy distinct ecological niches or habitats that are adjacent to each other, allowing for some gene exchange, albeit at reduced rates compared to sympatric speciation. Patterns of DNA divergence in parapatric speciation can exhibit a gradient of genetic variation along the geographic range of the species. The ability to distinguish parapatric species based on DNA markers can be challenging due to the continuum of genetic variation and the potential for shared genetic variants between neighbouring populations. But on the continuum genetic variation spectrum, fixed differences or unique alleles can be found which could be used for species identification.

1.4. Species and phylogeny

An alternative way of thinking about the nature of plant species is through the optic of phylogenetics. Hennig firstly distinguished two kinds of groups that he called mono- and paraphyletic groups, and postulated that 'only monophyletic groups are natural groups' (Hennig, 1979). The claim was challenged (Nixon et al., 1990), given the wide range of mechanisms by which plant species might arise, and in turn lead to species that are either monophyletic, paraphyletic, or polyphyletic. These phylogenetic relationships among species will be influenced by different modes and timing of speciation.

The classic model of 'speciation in allopatry' due to range-splitting vicariance will typically generate reciprocally monophyletic daughter species. The amount of time required for two species to resolve as monophyletic will depend on (a) the size of the populations, with larger populations taking longer to become reciprocally monophyletic compared to smaller populations; (b) the region of DNA being examined to test for monophyly, the slower the mutation rate of the region concerned, the longer the time before sequences will resolve as monophyletic; (c) the generation time of the species in question, with longer generation times, leading to longer time to reciprocal monophyly, and (d) the extent of residual gene flow between diverging lineages, with ongoing gene flow retarding the transition to reciprocal monophyly (Avise, 2004).The reciprocally monophyletic daughter species are easy to identify as separate monophyletic clades on the phylogenetic trees.

The circumstances under which we expect species to resolve as paraphyletic include when one species 'buds-off' from another. This may occur if populations become isolated at the margins of a species range, resulting in a derivative species nested within the variation of its progenitor species. The state of paraphyly itself may be transitory, as overtime gene flow within the progenitor species, and divergence from the derivative species should lead to a transition to reciprocal monophyly (Coyne et al., 2004) Likewise, when a species is generated from hybridization, the hybrids are usually nested between the two parental species, thus rendering both of the parental species paraphyletic. In this context, though only the daughter species forms a monophyletic clade on the phylogeny, the mother species is also identifiable to be resolved as paraphyletic clade immediate close to the daughter species.

The circumstances under which we expect species to resolve as polyphyletic include repeated transitions from the same starting genetic material such as repeated allopolyploid hybridisation from the same parental species, or independent ecological adaptations giving rise to morphologically and ecologically similar entities (Avise, 2004). In the case of independent origins of allopolyploidy, many taxonomists will treat the resulting polyphyletic entities as members of the same species (as they may in practice be indistinguishable). In the case of multiple (parallel) evolution of similar morphologies and ecologies, the different monophyletic units within a polyphyletic 'species' complex may in practice be reclassified as different species, reflecting their different evolutionary origins. This is the most intricate scenario trying to identify species based on the phylogeny. The best hope is one or several genomic regions that account for the species boundary, albeit rampant genetic exchanges happening at the rest of the genome, could be identified and thus for the use of species identification.

1.5. Species identification using DNA sequence

In light of the aforementioned evolutionary processes and how they relate to the genomic patterns of differences between species, I now take a closer look at the methods that have been developed for telling species apart using DNA sequences, focusing on the widely applied approach of DNA barcoding.

1.5.1. DNA barcoding

DNA barcoding aims to provide a minimal diagnostic region of the genome and works on the principle of using short standardised regions of DNA to tell wild species apart (Hebert et al., 2003). There are two major ways to identify species with DNA sequence from DNA barcodes – Distance-based and phylogeny-based methods.

Distance-based methods focus on finding a barcoding gap between the inter-specific nucleotide distances and intra-specific nucleotide distances (Čandek et al., 2015). It is based on the premise that the nucleotide distances among individuals from the same species should be smaller than those between individuals from different species. This difference between inter-specific and intra-specific nucleotide distances is called the barcoding gap. The barcoding gap has been implemented as a fixed threshold of sequence divergence to define species boundaries. In many taxa, there is a clear and unambiguous barcoding gap which can be used to assign unknown specimens to known species, as demonstrated by the use of DNA barcoding to study North American birds (Hebert et al., 2004). In this study, the inter-specific sequence distances in the cytochrome oxidase 1 DNA barcode (CO1) was 19–24 times greater in magnitude than the intra-specific differences (7.05%–7.93% versus 0.27%–0.43%, respectively). With this clear barcode gap, the authors concluded that most North American bird species can be discriminated using DNA barcode-based molecular diagnosis (Hebert et al., 2004).

Using a fixed threshold DNA-barcoding gap has been criticised as closely related species may have inter-specific genetic distances that fall below a divergence threshold that works for more divergent taxa (Moritz et al., 2004). In general, a priori determined divergence thresholds are not widely accepted, as even where the target species are ‘good’ species, the levels of divergence between species is expected to vary in different taxonomic groups, and with different speciation histories (based on factors such as effective population sizes and species divergence times (Hickerson et al., 2006, Rannala, 2015)). Thus it is often difficult to establish an a priori threshold (Goldstein et al., 2000, DeSalle et al., 2005). The establishment of the automated barcode gap discovery (ABGD) program (Puillandre et al., 2012) assists the estimation of an a priori barcoding gap. he Barcode Index Number System (Ratnasingham et al., 2013), offers a more sophisticated approach for taxon discrimination based on sequence similarities and flexible thresholds, and this has been widely used for arthropod DNA barcoding (Hebert et al., 2016). More simple approaches for using distance-based methods for barcode identification include firstly checking which species are readily distinguishable (e.g. cases where species in a reference library have smaller intra-specific sequence distances than they have inter-specific sequence distances to any other species), and then comparing the distances of unknown

specimens to those in the reference library using approaches such as BLAST (Altschul et al., 1990).

Another widely used approach for DNA barcoding involves tree-based methods. The approach involves building phylogenetic trees, and separating species based on their assignment into discrete monophyletic groups on the tree. This approach, of course assumes that the species in question is monophyletic in the first place, and rapid species diagnostics is more problematic for species whose underlying species tree does not involve monophyly. A desirable attribute of this approach, however, is its simplicity. A set of sequences can be aligned, and a tree computed and the resulting clusters examined to see which species any unknown samples cluster with, and hence which species it is inferred to belong to.

1.5.2. The DNA regions used as standard DNA barcodes in major eukaryotic groups.

DNA barcoding in animals uses sequence from the mitochondrial cytochrome oxidase 1 (CO1) gene region (Folmer et al., 1994, Hebert et al., 2003). This c. 650 base pairs of a mitochondrial DNA region is maternally inherited, and typically discriminates a large proportion of known animal species, with only occasional failure at the species level due to closely related species pairs, or a slower mitochondrial evolutionary rate in a fraction of animal groups (McFadden et al., 2011, Vargas et al., 2012, DeBiasse et al., 2014).

Compared to animals, studies show that in fungi, CO1 is prone to having multiple introns and is difficult to amplify with universal primers (Dentinger et al., 2011). Instead, the internal transcribed spacer (ITS) region of nuclear ribosomal DNA was selected as the standard DNA barcode marker for fungi and has been widely adopted and used by mycologists, and supplemented with clade-specific markers as required (Schoch et al., 2012).

In plants DNA barcoding has proven more challenging. Various regions have been proposed with the key features being ease of recovery of the barcode region (e.g. universality of primers), the quality of the resulting sequence traces (to avoid ambiguity of sequence reads), and the discriminatory power of the barcode region (CBOL, 2009). The standard plant barcode consists of two plastid regions *rbcL* (Chase et al., 1993) and *matK (*Hilu *et al.,* 1997, CBOL, 2009*).* This combination of loci tells many plant species apart, but also in many cases provides resolution only to a group of related species instead of unique species identity (Hollingsworth et al., 2011). The standard DNA barcode is usually augmented or used in various combinations with plastid spacer regions such as *trnH-psbA* (Pang et al., 2012), or the *trnL* intron (Taberlet et al., 2007). Another core standard DNA barcode for plants is the internal transcribed spacer region of nuclear ribosomal DNA (nrDNA ITS) or just a subset of this region (namely ITS2) (Yao et al., 2010, China Plant Barcoding Group, 2011, Hollingsworth, 2011,). These different plant barcoding regions have different strengths and weaknesses, and some regions or combination of regions have been proven to be successful in some plant groups but not in others, as in the case of *Cymbidium*, where 68% of the samples could be correctly assigned to species by *rbcL* and *matK*, whereas multiple copies of ITS and ITS2 meant they were not useful for species identification

in this group (Zhang et al., 2022). An overriding conclusion from the plant barcoding literature is that all of the current 'standard' DNA barcodes for plants result in discrimination success that is in general, lower, compared with that in animals (Blaxter, 2016, Hollingsworth et al., 2016, Page, 2016).

1.5.3. Application of DNA barcoding.

The use of DNA barcoding as a tool for species identification, has applications in numerous fields. One of the main applications is new species discovery. Thus when a well-sampled reference library is available, if DNA barcodes from unknown specimens do not group with any specimens of known species, this indicates a potential discovery of a new species (Bell et al., 2012). DNA barcoding can also offer a reliable way of spotting adulterants in traded natural products such as food, medicine, and timber (Hu et al., 2021), and the ability to detect alien invasive species in ecological surveys (Van De Wiel et al., 2009, Ghahramanzadeh et al., 2013, Xu et al., 2018, Madden et al., 2019). A growing interest in DNA barcoding is the use of large-scale reference libraries to support biomonitoring studies using bulk sample meta-barcoding to track patterns of species diversity and dynamics across time and space (Andersen et al., 2012, Sickel et al., 2015, Pandit et al., 2021).

1.5.4. The global DNA barcoding infra-structure

The Consortium for the Barcode Of Life (CBOL, http://barcoding.si.edu) was launched in May 2004 with the aim of supporting global efforts to build a complete barcode library of all eukaryotic life. Its role was subsequently overtaken by iBOL, the International Barcode of Life (www.iBOL.org). iBOL is a global initiative led from the University of Guelph in Canada, and serves to coordinate international DNA barcoding efforts. It consists of three temporally phased programs. The first phase, Barcode 500K was completed in 2015, and established barcode records for 500K taxa. The second and current program called BIOSCAN was launched in 2019 and focuses on (a) greatly accelerating species discovery and reference specimen barcoding, (b) accelerating the study of species interactions, by multi-amplicon barcoding of specimens to reveal their associated symbiome, and (c) accelerating studies of species dynamics by bulk sample metabarcoding across major ecoregions, all with the collective aim of laying the foundations for a global biomonitoring system for the final program of iBOL due to launch in 2027 (https://ibol.org/programs/program-overview/).

The iBOL program has given rise to a series of associated national and international initiatives, such as large-scale projects like Biodiversity Genomics Europe (https://biodiversitygenomics.eu) which focuses on building a continental barcoding network for Europe, as well as national reference library and biomonitoring projects like Barcode UK (Jones et al., 2021), the Norwegian Barcode of Life (https://www.norbol.org), Arise-Biodiversity in Netherland (https://www.arise-biodiversity.nl), and the Panama Barcode of Life (Kress *et al.,* 2009).

To date, the iBOL associated data repository, Barcode of Life Data System (BOLD), has archived 11,587,000 barcodes, representing approximately 339,000 formally described species, including 72,000 plant species (accessed online December 2022).

As noted previously, in addition to the Linnean binomial nomenclature system, BOLD incorporates a Barcode Index Number (BIN) system which automates the delineation of molecular operational taxonomic units (OTU) as proxies for animal species. This BIN system allows for the classification of species based on sequence data, prior to a more taxonomic formal description. The BIN system does not work for plants, however, due to the lower resolving power of plant DNA barcodes.

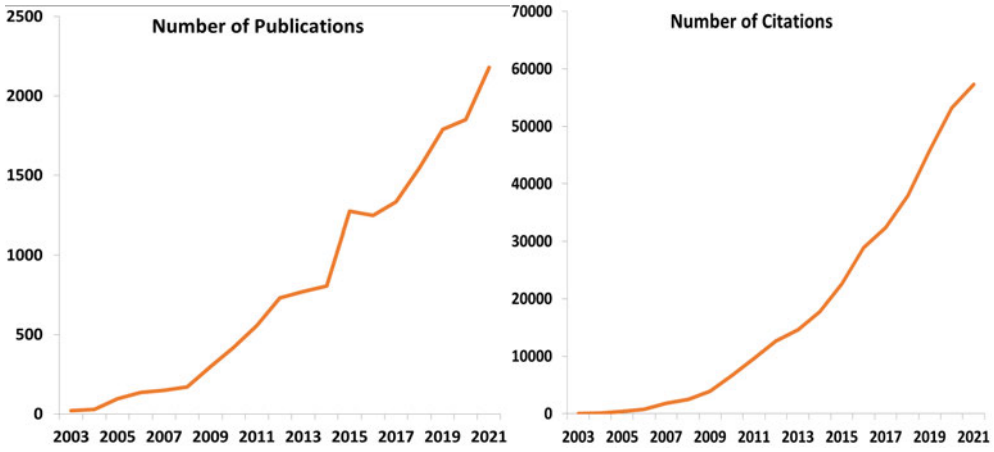


Figure 1.2. The growth of DNA barcoding influence (data provided by Paul Hebert, University of Guelph)

1.6. Efforts to improve DNA barcode resolution in plants by augmenting the standard DNA barcodes

An obvious starting point for increasing species discrimination in plants is to add more loci. The standard plant barcodes (*rbcL, matK, trnH-psbA, trnL*, ITS (or ITS2)) can be used in combination, and /or further supplemented with additional coding regions (e.g. *rpoB, rpoC1, ndhH*), and/or intergenic spacer regions (*atpF-atpH, psbK-psbI, ndhF-rpl32, rpl32-ccsA, psbK-psbI, petA-psbJ*) from the plastid genome (Yan et al., 2015, Mu et al., 2021, Zheng et al., 2021, Torke et al., 2022), or going one step further to sequence the entire plastome which is referred to by some authors as a 'super-barcode' (Comer et al., 2015, Bohmann et al., 2020, Simmonds et al., 2021, Su et al., 2021, Fu et al., 2022), or adding additional nuclear ribosomal DNA regions such as the external transcribed spacers (ETS, Liu et al., 2022).

When more plastid or ribosomal data are added, studies have shown different levels of improvements in discrimination success, but overall the gains have been modest. A small-scale study on a species complex *Mukdenia rossii* and *M. acanthifolia* shows that the complete plastome resolves individuals from two species, while the ITS + ETS combination failed (Liu et al., 2022). In a taxonomical complex genus *Corydalis*, five nuclear and chloroplast DNA regions, ITS, ITS2*, matK, rbcL*, and psbA-*trnH*, were preliminarily assessed, and the combination of ITS *+ matK* (69.6%) provided the highest species resolution among all single barcodes and their combinations (Ren et al., 2019). In a well-studied group *Stipa*, evaluation of tens of commonly used loci shows that no single locus or combination of loci could tell 70% of the species apart. The complete plastome sequences, a.k.a. super-barcodes, appeared to be more effective with a higher nucleotide variation, but still failed to reach a desired species

resolution in this feather grass genus (Krawczyk et al., 2018). An extreme case in *Schima* demonstrated a growth of discrimination success from 0% (using the standard DNA barcodes) to 27% when adding the whole plastome sequence to the pool. But the final discrimination rate is still too low to be confidently used in future species identification (Yu et al., 2022).

Several other studies have tried to expand from standard DNA barcodes to multiple plastid regions or the whole plastome. A phylogeographic analysis on 78 plastid protein-coding sequence (CDS) loci, assembled based on whole-plastome data, of the genus *Polygonatum* distributed on the Himalaya-Hengduan Mountains (HHM), successfully clustered specimens from the two main subsections (Xia et al., 2022). At the species level, the plastome data from the olive genus Olea was reported to be able to cluster individuals from the same species together (Dong et al., 2021). The success was also demonstrated at a small scale in the genus *Catalpa* (Dong et al., 2022). However, the information from the whole plastid genome provide is not always sufficient, as in the case of *Lauraceae*, where a study sequenced 191 plastid genomes from 131 species from 25 genera, and the plastome data overall were only able to discriminate ~60% of the species, with this representing a modest improvement from 40–50% discrimination success with the standard plant DNA barcodes (Liu et al., 2021). A newly published study sequencing the whole plastome of *Cymbidium* showing that the identification rate increased from 58% to 68% compared to the standard DNA barcodes (Zhang et al., 2022). Likewise, in the genus *Rhododendron* (Fu et al., 2022), there was a modest increase in discriminatory power from 33% using a combination of the standard barcodes (*ITS+matK+rbcL+trnH-psbA)* to 55% using plastid genome sequences and nrDNA arrays.

It is thus clear that simply adding more data from standard barcoding regions does not solve the challenge of substantially increasing discriminatory power for plant species. Expressed another way, sequencing the whole plastome doesn't tackle the fundamental problem. This is because the uniparentally inherited plastid genome of plants often doesn't track species boundaries (Hollingsworth et al., 2016). As a single non-recombining entity, the plastid genome is susceptible to incomplete lineage sorting (ILS) due to its smaller effective population size (*Ne*). Lateral transfer (Mallet et al., 2016) and plastid capture (du Preez et al., 2018) events also confound species inference using plastid genome sequence. In addition, the predominant maternal inheritance (e.g. seed dispersal) (du Preez et al., 2018) of plastids in plants is another intrinsic limitation, as new mutations can be slow to spread throughout a species range, due to the much lower dispersal of seeds compared to pollen in plants (Hollingsworth et al., 2011). Thus adding additional plastid sequence data may simply improve the precision in the failure for plastid DNA to tell species apart rather than making material gains compared to the resolving power of standard plastid barcode regions.

A similar issue occurs with adding more ribosomal data to ITS, e.g. not just sequencing the internal transcribed spacer regions, but instead sequencing the entire 18S-5.8S-26S array with the internal and external transcribed spacers and the associated intergenic spacer. The entire ribosomal repeat array does provide additional characters, but the regions are all fundamentally linked together and like the plastid genome may not track species boundaries. In addition, nuclear ribosomal DNA is present in multiple copies per cell, with the number of copies ranging from a few hundred to thousands, with no clear correlation with the genome size (Prokopowich et

al., 2003). These ribosomal DNA repeats are subject to concerted evolution, which can result in gene tree and species tree discordance, and where concerted evolution is incomplete can lead to multiple different variants recovered from a single individual adding considerable complexity to interpretation (Hollingsworth et al., 2011).

Collectively, it is clear that sequencing entire plastid genomes and/or ribosomal arrays will not fundamentally result in achieving universal (or near universal) species-level discrimination of plants.

1.7. The need to develop multi-locus nuclear DNA barcodes for plants

The failure of the existing barcodes (standard barcodes, or entire plastomes/rDNA arrays) in telling species apart has been reported in many plant groups. Examples include the genus *Inga* (Dexter et al., 2017), *Calligonum* (Li et al., 2014), *Schima* (Yu et al., 2022), feature grass *Stipa* (Krawczyk et al., 2018), European bladderworts *Utricularia* (Astuti et al., 2019), British eyebrights *Euphrasia* (Yeo, 1968, French et al., 2007, Wang et al., 2018), and on a broader scale, in the family Bromeliaceae (Maia et al., 2012) and Lauraceae (Liu et al., 2021). Similar failure has also been reported among distant taxa occurring at specific geographic locations which is the case for a range of recently radiating plant groups in Hawaii (Stallman et al., 2019). A particularly illustrative example is in the willow shrub genus *Salix*. Here, in this large genus of approximately 450 species, a single widespread identical DNA barcode sequence was recovered from 337 individuals representing 53 species, with hybridisation followed by selective sweeps identified as a likely mechanism for the barcode sharing (Percy et al., 2014).

These frequent cases where DNA barcoding does not provide species-level resolution in plants, have been attributed to several related mechanisms: 1) incomplete lineage sorting of ancestral polymorphism. 2) hybridization leading to the transfer of sequence variants between species resulting in identical barcode sequences becoming shared among species within a genus 3) selective sweeps also resulting in a particular barcode sequence becoming abundant in multiple species, or 4) cases where species are young and/or mutation rates are slow, where there may not have been enough time for species-specific mutations to arise in the barcoding regions (Twyford, 2014).

To address these limitations of DNA barcodes for plants, accessing multiple independent loci is required. To gain as much information from independently inherited loci as possible, the exploration of the nuclear genomes is urgently needed, as multiple studies have suggested (Hollingsworth et al., 2016, Liu et al., 2021, Zhang et al., 2022). Accessing the nuclear genome avoids the vagaries of making biological inference from individual linkage groups like plastid or nrDNA. Turning to the nuclear genome will give access to large numbers of variable markers (nuclear genomes are larger than plastid genomes) and independent data points (individual linkage groups resulting from sexual inheritance of the nuclear genomes and recombination) (Hollingsworth et al., 2016).

1.7.1. Sequencing and bioinformatic advances enable accessing the nuclear genome of plants at a large-scale

Next-generation sequencing (NGS) platforms offer transformative opportunities for accessing the nuclear genome in large-scale fashion in plants. Sequencing plant nuclear genomes has been hampered by their complexity, e.g., large genome size and frequent polyploidy (Levin, 2020), and the resulting costs of achieving appropriate sequence coverage and the informatics challenges of assembly, annotation, and identification of orthologous loci. However, the shift from Sanger sequencing to the new wave of sequencing platforms has led to a significant increase in the tractability of studies of complex genomes (Heather et al., 2016). Key developments include a

series of massively parallel short-read sequencing machines (Lander et al., 2001, Li et al., 2010) (Illumina platform is dominant) and the real-time long-read sequencing platforms from Oxford Nanopore (Akeson et al., 2012, Goodwin et al., 2015) (ONT) and Pacific Biosystems (Eid et al., 2009, English et al., 2012) (PacBio).

Box1. Overview of major post-Sanger sequencing technologies

Illumina: The platform detects fluorescent signals while synthesizing complementary strands of the target DNA templates. By adding only one fluorescently labelled deoxynucleotide (dNTP) to the template on each cycle, the imaging system can determine which of the four nucleotides was incorporated by the colour released. Reads generated are highly accurate with limited length which ranges from 100 bp to 300 bp, but the highly parallel nature of the approach generates up to 6000 gigabases of data from a single run on the NovaSeq 6000 machine.

PacBio: This approach detects optical signals emitted by fluorescently labelled nucleotides when incorporated into a single DNA molecule guided by an engineered polymerase. The single molecule approach bypasses the PCR stage in library production(Levene et al., 2003). The reads are relatively long (~15 kb) but originally had a high error rate (10 – 15%). With the implementation of circular consensus sequencing (CCS) mode on PacBio long-read systems, this produces highly accurate long reads, or HiFi reads (Wenger et al., 2019), which reach an accuracy of 99.9%, and are on par with Illumina short reads. The PacBio Sequel II machine can produce tens of gigabases of data per run.

ONT: This rapidly progressing method recognizes the electronic patterns of ion flow when a single-stranded DNA is driven through a narrow nucleopore protein channel. Sequence read length can exceed that of PacBio, with ~ 4 Mb single reads reported recently. However, because the molecule movement is hard to control, the data generated reach an error rate as high as 30%. Sequencing a higher coverage of the target regions can largely reduce the error, i.e. reaching 99.999% accuracy with 100× data. The portable MinION machines produce 50 gigabases of sequence data per flow cell run (https://nanoporetech.com/products/minion).

The decreasing cost and increasing throughput of next-generation DNA sequencing platforms have facilitated a myriad of whole-genome sequencing projects and general surveys of whole-genome nucleotide variation. Initial whole-genome sequencing projects usually focused on model organisms or species of economic importance (Fleischmann et al., 1995, Lander et al., 2001, Yu et al., 2002) and there are ongoing targeted efforts to sequence representatives of major clades of life (Jaillon et al., 2007). Finally, as the costs continue to fall, there are an increasing number of projects gathering entire genome sequences from multiple species across genera or families (Sun et al., 2017, Wei et al., 2021). In parallel to whole-genome sequencing projects, a range of studies have been undertaken with less comprehensive genome coverage including genome skimming (Dodsworth, 2015), transcriptome sequencing (Wang et al., 2009), and reduced representation sequencing techniques such as RAD sequencing (Davey et al., 2010, Davey et al., 2011, Peterson et al., 2012), genotyping-by-sequencing (GBS) (Elshire et al., 2011, Deschamps et al., 2012) and target capture (Weitemier et al., 2014). These types of projects gather a subset of the whole-genome per individual, and have the advantage of being able to sequence more individuals/more species given a set budget.

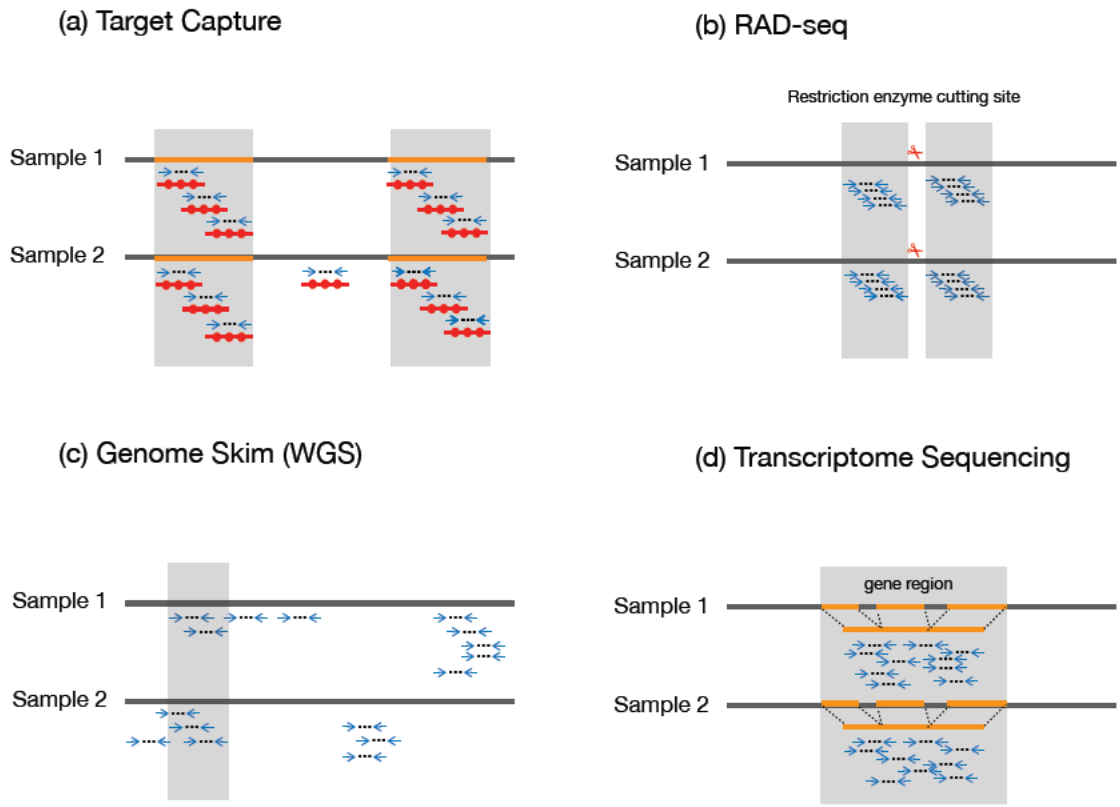


Figure 1.3. Diagram of popular reduced representation sequencing approaches. (a) Target capture methods 'capture' fragmented DNA from the homologous regions (grey shading) of the genome using designed short oligonucleotide probes (lined red dots). (b) RAD-seq targets regions flanked (grey shading) by chosen restriction enzyme cutting sites (red scissors). (c) Genome skim sequences recover high copy number organelle genomes, along with relatively random fragments of nuclear genome regions with low sequence coverage. Homologous sequences are a collection of overlapped assemblies recovered by valid coverage of reads (grey shading). (d) Transcriptome sequencing involves sequencing the expressed total RNA in an organism, that are often exons (orange bars) of gene regions (grey shading) expressed at the time. All of the above methods result in a subset of the nuclear genome being sequenced via a NGS platform (paired-end sequencing illustrated).

Aside from the improvement of the sequencing platforms, the dramatic improvement of bioinformatics tools has also facilitated the exploration of the nuclear genome. For example, clustering millions of reads and assembling them into contigs used to take days with heavy computational cost. With a surge of bioinformatics advancement during the past 40 years, now assembling a human genome with informatic pipelines such as wtdgb2 (Ruan et al., 2020) only takes less than 20 hours on a personal affordable machine (32 CPUs).

1.7.2. Feasibility of a systematic evaluation of variation in the nuclear genomes of plants to inform the development of multi-locus DNA barcodes

The reduced sequencing cost and the easy use of the bioinformatics tools allow the application of genomics to large-scale evolutionary or ecological studies. Notably, the Earth BioGenome Project (Lewin et al., 2018) (https://www.earthbiogenome.org/), illustrates a global effort to efficiently sequence the genomes of all known species, and to use genomics to help discover the remaining 80 to 90 percent of species that are currently hidden from science. The data produced from this giant project could be used in various analyses, which include building phylogenetic trees, examining population genetic diversity, and understanding functional diversity (e.g. by sequencing species/variants with atypical morphologies or ecologies). Other important major projects include One Thousand Plant Transcriptomes Initiative (OneKP or 1KP, (One Thousand Plant Transcriptomes, 2019) which provided transcriptome data from a diverse sample set of 1000 phylogenetically disparate plant species, greatly enhancing knowledge of sequence variation among coding regions in plants. Building on this resource, the Plant and Fungal Tree of Life Project (PAFTOL) developed the Angiosperm 353 target capture probe set (Baker et al., 2021, Baker et al., 2022) which is being used to recover data from 353 nuclear loci across a wide diversity of angiosperm species. In addition to these large-scale infrastructure projects, there is an abundance of interesting small-scale studies targeting finer geographic and evolutionary scopes that are benefiting from the advancement of sequencing and bioinformatic technologies. In biogeography-centred studies, genomics tools can give better understandings of the local biome assemblage, e.g. the rainforest tree communities across the Amazon basin (Dexter et al., 2017). For taxonomists whose main foci are a certain family, genus, or species, genomic data can improve the resolution of phylogenies of these groups of their taxonomic levels. Examples include a better resolved phylogeny of the carrot family (Clarkson et al., 2021), a higher-resolution phylogeny of the genus *Salix* (Wagner et al., 2018, He et al., 2020, Sanderson et al., 2020), or a detailed history of how the varieties of citrus cultivars evolved with generations of breeding, selection, and cultivation (Wu et al., 2018). Numerous studies that have focused on hybridisation and speciation processes have also resulted in a better understanding of which genomic regions might shape speciation and species differences such as in the genera *Mimulus* (Chase et al., 2017) and *Antirrhinum* (Otero et al., 2021, Durán-Castillo et al., 2022).

Collectively these large-scale and smaller-scale projects are providing data which greatly increase the ease of access and understanding of the sequence complexity of the nuclear genomes of plants. **This then leads to the prospect of a systematic**

evaluation of the patterns of sequence variation among plant species by reusing and reanalysing the sequence data from various published studies. The aforementioned studies, along with others, have produced large quantities of DNA sequence data that are available on public data repositories. These data should allow an evaluation of the proportion of plant species that resolve as monophyletic using large amounts of nuclear sequence data, and also an understanding of the nature of genomic differences between species (e.g. the frequency and distribution of taxon-specific nucleotide substitutions). Such analyses offer the potential to enhance our conceptual understanding of what constitutes a plant species. Furthermore, improving understanding of the nature of genetic differences between plant species is a critically important step in developing improved DNA-based methods for telling plant species apart. Thus better understanding of the nature of sequence differences between plant species is informative about the nature of plant species themselves, and the type of assay that would be required to routinely improve discriminatory power, beyond that achieved with standard plastid or nrDNA barcodes, and the extended 'super-barcode' approach.

At a technical level any future multi-locus nuclear barcoding approach will need to recover sufficient amounts of data to give resolution among closely related species, it will need to have sufficient universality to provide comparative data across a wide phylogenetic spread of plant diversity enabling the use in floristic applications, and it will also have to accommodate the complexity of plant genomes to enable orthologous sequences to be compared, minimising problems caused by paralogy and the repetitive nature of plant genomes (Fitzek et al., 2018, Sahlin et al., 2021).

Of the currently used approaches for routine recovery of sequences from the nuclear genomes of plants from large sample sets, some methods can be discarded as future candidate barcoding approaches. RAD-seq targets orthologues regions via shared restriction enzyme cut sites, and it is a very cost-effective method of generating large amounts of data. However, the method is notorious for its high levels of missing data because whether a locus could be sequenced depends largely on the enzyme digest success; furthermore the lack of conserved cut sites across a very broad taxonomic scope limits its application only to closely related taxa. Likewise, transcriptome sequencing is a widely used tool for the analysis of gene expression, marker discovery and comparative evolution (Cloonan et al., 2008, Wang et al., 2009, Stark et al., 2019). The main benefit of transcriptomics is that it focuses NGS onto a homologous proportion of the genome, which case is also conserved due to selective constraints. However, the requirement for high-quality fresh material, and the tissue-specific nature of expressed sequences, rules transcriptomics out as a tool for universal plant barcoding.

In contrast, both shotgun sequencing (e.g. genome skims) and target capture offer potential for the next-generation of nuclear DNA barcodes in plants (Hollingsworth et al., 2016). Target capture involves the designing of baits (short oligonucleotide probes, figure 1.3. (a)) to capture homologous DNA regions. The approach can be targeted to gene regions with a predisposition to be single copy across wide sample sets, and can be tailored to include generic and taxon-specific probes (Baker et al. 2021). It is highly scalable as the candidate bait set can be designed one-off with a reasonable cost for subsequent multiple re-uses. Genome skimming requires the least pre-sequencing effort and cost but also results in the most erratic data recovery. Beyond high copy

number regions (e.g. plastome, rDNA), the reads produced by this method are usually spread randomly across the whole-genome, thus making it difficult to recover common orthologous regions from multiple samples when sequence coverage is shallow. However, as sequencing costs continue to fall, it is possible that genome skims can routinely have sufficient sequencing depth to allow routine recovery of multiple orthologous regions enabling effective species discrimination. Both target capture and genome skimming approaches work well with preserved herbarium specimens (Alsos et al., 2009, Forrest et al., 2019), giving access to well-identified sample sets, and making them well-suited for the development of DNA barcode reference libraries.

1.8. Aim of this study

My project aims to undertake a synthetic evaluation of the genetic differences between plant species. The ultimate aim is to better understand plant species, the nature of sequence differences between them, and to provide this information to facilitate the development of a next-generation of high-resolution plant DNA barcodes. I do not focus on the design of a new DNA barcoding system for plants. Rather my focus is developing informatic methods and understanding the nature of the differences between plant species, as a critical enabling step prior to the design of 'plant barcoding 2.0'.

More specifically, in this thesis I:

1) Outline the key issues that need consideration in the development of a nuclear DNA barcodes for plants, with a particular focus on the types of data and analytical steps required (Chapter 2)
2) Develop scripts and pipelines to analyse nuclear sequence data from plants to better understanding the genomic nature of differences between plant species, and the efficacy of nuclear sequence data in telling plant species apart (Chapter 3)
3) Apply the scripts I have developed to a test case, using nuclear sequence data to evaluate the nature of inter-specific differences in the challenging case of the highly-diverse and recently-radiated neotropical tree genus *Inga* (Chapter 4)
4) Extend the application of my methodology to a meta-analysis, mining the available datasets from public repositories and from collaborators to undertake a synthetic evaluation of the genomic nature of the differences between plant species, based on currently available data (Chapter 5)
5) I finish with some general conclusions on future prospects for telling plant species apart with DNA from the nuclear genome and identify key next steps towards the practical development of multi-locus nuclear DNA barcodes for plants (Chapter 6).

1.9. Reference

Akeson, M., Branton, D., Church, G., & Deamer David, W. (2012). Characterization of individual polymer molecules based on monomer-interface interactions. (Patent)

Alsos, I. G., Alm, T., Normand, S., & Brochmann, C. (2009). Past and future range shifts and loss of diversity in dwarf willow (*Salix herbacea* L.) inferred from genetics, fossils and modelling. *Global Ecology and Biogeography,* 18(2), 223-239. doi:10.1111/j.1466-8238.2008.00439.x

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology,* 215(3), 403-410. doi:10.1016/S0022-2836(05)80360-2

Andersen, K., Bird, K. L., Rasmussen, M., Haile, J., Breuning-Madsen, H., KjÆR, K. H., Willerslev, E. (2012). Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. *Molecular Ecology,* 21(8), 1966-1979. doi:10.1111/j.1365-294X.2011.05261.x

Astuti, G., Petroni, G., Adamec, L., Miranda, V. F. O., & Peruzzi, L. (2019). DNA barcoding approach fails to discriminate Central European bladderworts (*Utricularia*, Lentibulariaceae), but provides insights concerning their evolution. *Plant Biosystems - An International Journal Dealing with all Aspects of Plant Biology,* 154(3), 326-336. doi:10.1080/11263504.2019.1610112

Avise, J. C. (2004). *Molecular markers, natural history, and evolution* (Second edition. ed.). Sunderland, Mass: Sinauer Associates.

Baker, W. J., Bailey, P., Barber, V., Barker, A., Bellot, S., Bishop, D.,... Forest, F. (2022). A comprehensive phylogenomic platform for exploring the angiosperm tree of life. *Systematic Biology,* 71(2), 301-319. doi:10.1093/sysbio/syab035

Baker, W. J., Dodsworth, S., Forest, F., Graham, S. W., Johnson, M. G., McDonnell, A.,... Wickett, N. J. (2021). Exploring Angiosperms353: An open, community toolkit for collaborative phylogenomic research on flowering plants. *American Journal of Botany,* 108(7), 1059-1065. doi:10.1002/ajb2.1703

Bell, D., Long, D. G., Forrest, A. D., Hollingsworth, M. L., Blom, H. H., & Hollingsworth, P. M. (2012). DNA barcoding of European *Herbertus* (Marchantiopsida, Herbertaceae) and the discovery and description of a new species. *Molecular Ecology Resources,* 12(1), 36-47. doi:10.1111/j.1755-0998.2011.03053.x

Blaxter, M. (2016). Imagining Sisyphus happy: DNA barcoding and the unnamed majority. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 371(1702). doi:10.1098/rstb.2015.0329

Bohmann, K., Mirarab, S., Bafna, V., & Gilbert, M. T. P. (2020). Beyond DNA barcoding: The unrealized potential of genome skim data in sample identification. *Molecular Ecology*. doi:10.1111/mec.15507

Čandek, K., & Kuntner, M. (2015). DNA barcoding gap: reliable species identification over morphological and geographical scales. *Molecular Ecology Resources,* 15(2), 268-277. doi:10.1111/1755-0998.12304

CBOL. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences,* 106(31), 12794-12797. doi:10.1073/pnas.0905845106

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A., Tang, H., Wang, X.,... Wincker, P. (2014). Plant genetics. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science,* 345(6199), 950-953. doi:10.1126/science.1253435

Chase, M. A., Stankowski, S., & Streisfeld, M. A. (2017). Genomewide variation provides insight into evolutionary relationships in a monkeyflower species complex (*Mimulus* sect. *Diplacus*). *American Journal of Botany,* 104(10), 1510-1521. doi:10.3732/ajb.1700234

Chase, M. W., Soltis, D. E., Olmstead, R. G., Morgan, D., Les, D. H., Mishler, B. D.,... Albert, V. A. (1993). Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. *Annals of the Missouri Botanical Garden,* 80(3), 528-580. doi:10.2307/2399846

Chen, Z. J., Sreedasyam, A., Ando, A., Song, Q., De Santiago, L. M., Hulse-Kemp, A. M.,... Schmutz, J. (2020). Genomic diversifications of five Gossypium allopolyploid species and their impact on cotton improvement. *Nature Genetics,* 52(5), 525-533. doi:10.1038/s41588-020-0614-5

China Plant, B. O. L. G., Li, D. Z., Gao, L. M., Li, H. T., Wang, H., Ge, X. J.,... Duan, G. W. (2011). Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences,* 108(49), 19641-19646. doi:10.1073/pnas.1104551108

Clarkson, J. J., Zuntini, A. R., Maurin, O., Downie, S. R., Plunkett, G. M., Nicolas, A. N.,... Baker, W. J. (2021). A higher-level nuclear phylogenomic study of the carrot family (Apiaceae). *American Journal of Botany,* 108(7), 1252-1269. doi:10.1002/ajb2.1701

Cloonan, N., Forrest, A. R., Kolle, G., Gardiner, B. B., Faulkner, G. J., Brown, M. K.,... Grimmond, S. M. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods,* 5(7), 613-619. doi:10.1038/nmeth.1223

Comer, J. R., Zomlefer, W. B., Barrett, C. F., Davis, J. I., Stevenson, D. W., Heyduk, K., & Leebens-Mack, J. H. (2015). Resolving relationships within the palm subfamily *Arecoideae* (Arecaceae) using plastid sequences derived from next-generation sequencing. *American Journal of Botany*, 102(6), 888-899. doi:10.3732/ajb.1500057

Coyne, J. A., & Orr, H. A. (2004). *Speciation.* Sunderland, Mass: Sinauer Associates.

Davey, J. W., & Blaxter, M. L. (2010). RADSeq: next-generation population genetics. *Brief Funct Genomics,* 9(5-6), 416-423. doi:10.1093/bfgp/elq031

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics,* 12(7), 499-510. doi:10.1038/nrg3012

De Queiroz, K. (2007). Species concepts and species delimitation. *Systematic Biology,* 56(6), 879-886. doi:10.1080/10635150701701083

DeBiasse, M. B., Nelson, B. J., & Hellberg, M. E. (2014). Evaluating summary statistics used to test for incomplete lineage sorting: mito-nuclear discordance in the reef sponge *Callyspongia vaginalis*. *Molecular Ecology,* 23(1), 225-238. doi:10.1111/mec.12584

Dentinger, B. T. M., Didukh, M. Y., & Moncalvo, J.-M. (2011). Comparing COI and ITS as DNA barcode markers for mushrooms and allies (*Agaricomycotina*). *PLoS One,* 6(9), e25081-e25081. doi:10.1371/journal.pone.0025081

DeSalle, R., Egan, M. G., & Siddall, M. (2005). The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 360(1462), 1905-1916. doi:10.1098/rstb.2005.1722

Deschamps, S., Llaca, V., & May, G. D. (2012). Genotyping-by-Sequencing in plants. *Biology,* 1(3), 460-483. doi:10.3390/biology1030460

Dexter, K. G., Lavin, M., Torke, B. M., Twyford, A. D., Kursar, T. A., Coley, P. D.,... Pennington, R. T. (2017). Dispersal assembly of rainforest tree communities across the Amazon basin. *Proceedings of the National Academy of Sciences,* 114(10), 2645-2650. doi:10.1073/pnas.1613655114

Dobzhansky, T. (1950). Mendelian populations and their evolution. *The American naturalist,* 84(819), 401-418. doi:10.1086/281638

Dodsworth, S. (2015). Genome skimming for next-generation biodiversity analysis. *Trends in Plant Science,* 20(9), 525-527. doi:10.1016/j.tplants.2015.06.012

Dong, W., Liu, Y., Li, E., Xu, C., Sun, J., Li, W.,... Suo, Z. (2022). Phylogenomics and biogeography of *Catalpa* (Bignoniaceae) reveal incomplete lineage sorting and three dispersal events. *Molecular Phylogenetics and Evolution,* 166, 107330. doi:10.1016/j.ympev.2021.107330

Dong, W. P., Sun, J. H., Liu, Y. L., Xu, C., Wang, Y. H., Suo, Z. L.,... Wen, J. (2021). Phylogenomic relationships and species identification of the olive genus *Olea* (Oleaceae). *Journal of Systematics and Evolution*. doi:10.1111/jse.12802

Donoghue, M. J. (1985). A critique of the biological species concept and recommendations for a phylogenetic alternative. *The Bryologist,* 88(3), 172-181. doi:10.2307/3243026

du Preez, B., Dreyer, L. L., Schmickl, R., Suda, J., & Oberlander, K. C. (2018). Plastid capture and resultant fitness costs of hybridization in the Hirta clade of southern African *Oxalis*. *South African journal of botany*, 118, 329-341. doi:10.1016/j.sajb.2017.06.010

Durán-Castillo, M., Hudson, A., Wilson, Y., Field, D. L., & Twyford, A. D. (2022). A phylogeny of Antirrhinum reveals parallel evolution of alpine morphology. *New Phytologist,* 233(3), 1426-1439. doi:10.1111/nph.17581

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G.,... Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science,* 323(5910), 133-138. doi:10.1126/science.1162986

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One,* 6(5), e19379. doi:10.1371/journal.pone.0019379

Endara, M. J., Coley, P. D., Wiggins, N. L., Forrister, D. L., Younkin, G. C., Nicholls, J. A.,... Kursar, T. A. (2018). Chemocoding as an identification tool where morphological- and DNA-based methods fall short: *Inga* as a case study. *New Phytologist,* 218(2), 847-858. doi:10.1111/nph.15020

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J.,... Gibbs, R. A. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One*, 7(11), e47768. doi:10.1371/journal.pone.0047768

Ercisli, S., Ipek, A., & Barut, E. (2011). SSR Marker-based DNA fingerprinting and cultivar identification of Olives (*Olea europaea*). *Biochemical Genetics,* 49(9-10), 555-561. doi:10.1007/s10528-011-9430-z

Fitzek, E., Delcamp, A., Guichoux, E., Hahn, M., Lobdell, M., & Hipp, A. L. (2018). A nuclear DNA barcode for eastern North American oaks and application to a study of hybridization in an Arboretum setting. *Ecology and Evolution,* 8(11), 5837-5851. doi:10.1002/ece3.4122

Fitzpatrick, B. M., Fordyce, J. A., & Gavrilets, S. (2009). Pattern, process and geographic modes of speciation. *Journal of Evolutionary Biology,* 22(11), 2342-2347. doi:10.1111/j.1420-9101.2009.01833.x

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R.,... Venter, J. C. (1995). Whole-genome random sequencing and assembly of *Haemophilus-Influenzae* Rd. *Science,* 269(5223), 496-512. doi:DOI 10.1126/science.7542800

Folmer, O., Black, M., Wr, H., Lutz, R., & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial Cytochrome C oxidase subunit I from diverse metazoan invertebrates. *Molecular marine biology and biotechnology,* 3, 294-299.

French, G. C., Hollingsworth, P. M., Silverside, A. J., & Ennos, R. A. (2007). Genetics, taxonomy and the conservation of British *Euphrasia*. *Conservation Genetics,* 9(6), 1547-1562. doi:10.1007/s10592-007-9494-9

Forrest, L. L., Hart, M. L., Hughes, M., Wilson, H. P., Chung, K.-F., Tseng, Y.-H., & Kidner, C. A. (2019). The limits of Hyb-Seq for herbarium specimens: impact of preservation techniques. *Frontiers in Ecology and Evolution,* 7. doi:10.3389/fevo.2019.00439

Fu, C. N., Mo, Z. Q., Yang, J. B., Cai, J., Ye, L. J., Zou, J. Y.,... Gao, L. M. (2022). Testing genome skimming for species discrimination in the large and taxonomically difficult genus *Rhododendron*. *Molecular Ecology Resources,* 22(1), 404-414. doi:10.1111/1755-0998.13479

Futuyma, D. J., & Mayer, G. C. (1980). Non-allopatric speciation in animals. *Systematic Biology,* 29(3), 254-271.

Gavrilets, S. (2004). Genetic theories of allopatric and parapatric speciation. *Adaptive Speciation*, 112-139.

Ghahramanzadeh, R., Esselink, G., Kodde, L. P., Duistermaat, H., Valkenburg, J. L. C. H., Marashi, S. H.,... Wiel, C. C. M. (2013). Efficient distinction of invasive aquatic plant species from non-invasive related species using DNA barcoding. *Molecular Ecology Resources,* 13(1), 21-31. doi:10.1111/1755-0998.12020

Gholave, A., Pawar, K., Yadav, S., Bapat, V., & Jadhav, J. (2017). Reconstruction of molecular phylogeny of closely related *Amorphophallus* species of India using plastid DNA marker and fingerprinting approaches. *Physiology and Molecular Biology of Plants,* 23(1), 155-167. doi:10.1007/s12298-016-0400-0

Ghosh, S., Majumder, P. B., & Sen Mandi, S. (2011). Species-specific AFLP markers for identification of *Zingiber officinale, Z. montanum* and *Z. zerumbet* (Zingiberaceae). *Genetics and molecular research*, 10(1), 218-229. doi:10.4238/vol10-1gmr1154

Goldstein, P. Z., Desalle, R., Amato, G., & Vogler, A. P. (2000). Conservation genetics at the species boundary. *Conservation Biology,* 14(1), 120-131. doi:10.1046/j.1523-1739.2000.98122.x

Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M. C., & McCombie, W. R. (2015). Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Research,* 25(11), 1750-1756. doi:10.1101/gr.191395.115

Gregor, J. W. (1940). The new systematics. *Nature,* 146(3689), 42-43. doi:10.1038/146042a0

He, L., Wagner, N. D., & Hörandl, E. (2020). Restriction-site associated DNA sequencing data reveal a radiation of willow species (*Salix* L., Salicaceae) in the Hengduan Mountains and adjacent areas. *Journal of Systematics and Evolution,* 59(1), 44-57. doi:10.1111/jse.12593

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics,* 107(1), 1-8. doi:10.1016/j.ygeno.2015.11.003

Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences,* 270(1512), 313-321. doi:10.1098/rspb.2002.2218

Hebert, P. D. N., Stoeckle, M. Y., Zemlak, T. S., & Francis, C. M. (2004). Identification of Birds through DNA Barcodes. *PLOS Biology,* 2(10), e312-e312. doi:10.1371/journal.pbio.0020312

Hebert, P. D., Ratnasingham, S., Zakharov, E. V., Telfer, A. C., Levesque-Beaudin, V., Milton, M. A.,... deWaard, J. R. (2016). Counting animal species with DNA barcodes: Canadian insects. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 371(1702). doi:10.1098/rstb.2015.0333

Hennig, W. (1979). *Phylogenetic systematics* Urbana: University of Illinois Press.

Hickerson, M. J., Meyer, C. P., Moritz, C., & Hedin, M. (2006). DNA barcoding will often fail to discover new animal species over broad parameter space. *Systematic Biology,* 55(5), 729-739. doi:10.1080/10635150600969898

Hilu, K. W., & Liang, g. (1997). The *matK* gene: sequence variation and application in plant systematics. *American Journal of Botany,* 84(6), 830-839. doi:10.2307/2445819

Hollingsworth, P. M. (2011). Refining the DNA barcode for land plants. *Proceedings of the National Academy of Sciences,* 108(49), 19451-19452. doi:10.1073/pnas.1116812108

Hollingsworth, P. M., Li, D. Z., van der Bank, M., & Twyford, A. D. (2016). Telling plant species apart with DNA: from barcodes to genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1702). doi:10.1098/rstb.2015.0338

Hu, J. L., Ci, X. Q., Liu, Z. F., Dormontt, E. E., Conran, J. G., Lowe, A. J., & Li, J. (2021). Assessing candidate DNA barcodes for Chinese and internationally traded timber species. *Molecular Ecology Resources*. doi:10.1111/1755-0998.13546

IPBES, B., E. S., Settele, J., Díaz, S., Ngo, H. T. (eds). (2019). Global assessment report of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. *Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*, 1144. doi:10.5281/zenodo.3553579

Jaillon, O., Aury, J. M., Noel, B., Policriti, A., Clepet, C., Casagrande, A.,... French-Italian Public Consortium for Grapevine Genome, C. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature,* 449(7161), 463-467. doi:10.1038/nature06148

Joanna, W., Iveta, M., Rebecca, E. M., Stefano, C., & Michaël, B. (2018). New diagnostic SNP molecular markers for the *Mytilus* species complex. *PLoS One,* 13(7), e0200654. doi:10.1371/journal.pone.0200654

Johnson, M. G., Pokorny, L., Dodsworth, S., Botigué, L. R., Cowan, R. S., Devault, A.,... Renner, S. (2018). A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Systematic Biology*. doi:10.1093/sysbio/syy086

Jones, L., Twyford, A. D., Ford, C. R., Rich, T. C. G., Davies, H., Forrest, L. L.,... de Vere, N. (2021). Barcode UK: A complete DNA barcoding resource for the flowering plants and conifers of the United Kingdom. *Molecular Ecology Resources,* 21(6), 2050-2062. doi:10.1111/1755-0998.13388

Kawecki, T. J. (2004). Genetic theories of sympatric speciation. In *Adaptive Speciation* (pp. 36-53).

Kondrashov, A. S., & Mina, M. V. (1986). Sympatric speciation: when is it possible? *Biological Journal of the Linnean Society,* 27(3), 201-223. doi:10.1111/j.1095-8312.1986.tb01734.x

Krawczyk, K., Nobis, M., Myszczynski, K., Klichowska, E., & Sawicki, J. (2018). Plastid super-barcodes as a tool for species discrimination in feather grasses (Poaceae: *Stipa*). *Scientific Reports,* 8(1), 1924. doi:10.1038/s41598-018-20399-w

Kress, W. J., Erickson, D. L., Jones, F. A., Swenson, N. G., Perez, R., Sanjur, O., & Bermingham, E. (2009). Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. *Proceedings of the National Academy of Sciences - PNAS,* 106(44), 18621-18626. doi:10.1073/pnas.0909820106

Kress, W. J., Garcia-Robledo, C., Uriarte, M., & Erickson, D. L. (2015). DNA barcodes for ecology, evolution, and conservation. *Trends Ecology and Evolution,* 30(1), 25-35. doi:10.1016/j.tree.2014.10.008

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J.,... International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature,* 409(6822), 860-921. doi:10.1038/35057062

Lavine, B., & Carlson, D. A. (1987). Species identification through chemical analysis. *Analytical Chemistry,* 59(6).

Lee, B., Pires, E., Pollard, A. M., & McCullagh, J. S. O. (2022). Species identification of silks by protein mass spectrometry reveals evidence of wild silk use in antiquity. *Scientific Reports,* 12(1), 4579. doi:10.1038/s41598-022-08167-3

Leffler, E. M., Bullaughey, K., Matute, D. R., Meyer, W. K., Segurel, L., Venkat, A.,... Przeworski, M. (2012). Revisiting an old riddle: what determines genetic diversity levels within species? *PLOS Biology,* 10(9), e1001388. doi:10.1371/journal.pbio.1001388

Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G., & Webb, W. W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* 299(5607), 682.

Levin, D. A. (2020). Has the polyploid wave ebbed? *Frontiers in Plant Science,* 11, 251. doi:10.3389/fpls.2020.00251

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A.,... Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences - PNAS,* 115(17), 4325-4333. doi:10.1073/pnas.1720115115

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z.,... Wang, J. (2010). *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Research,* 20(2), 265-272. doi:10.1101/gr.097261.109

Li, Y., Feng, Y., Wang, X.-Y., Liu, B., & Lv, G.-H. (2014). Failure of DNA barcoding in discriminating *Calligonum* species. *Nordic Journal of Botany,* 32(4), 511-517. doi:10.1111/njb.00423

Linné, C. v. (1753). *Species plantarum* (Vol. 2). Stockholm, Sweden: Laurentius Salvius.

Liu, L. X., Deng, P., Chen, M. Z., Yu, L. M., Lee, J., Jiang, W. M.,... Li, P. (2022). Systematics of *Mukdenia* and *Oresitrophe* (Saxifragaceae): Insights from genome skimming data. *Journal of Systematics and Evolution*. doi:10.1111/jse.12833

Liu, Z. F., Ma, H., Ci, X.-Q., Li, L., Song, Y., Liu, B.,... Li, J. (2021). Can plastid genome sequencing be used for species identification in the Lauraceae? *Botanical Journal of the Linnean Society,* 197(1), 1-14. doi:10.1093/botlinnean/boab018

Madden, M. J. L., Young, R. G., Brown, J. W., Miller, S. E., Frewin, A. J., & Hanner, R. H. (2019). Using DNA barcoding to improve invasive pest identification at U.S. ports-of-entry. *PLoS One,* 14(9), e0222291. doi:10.1371/journal.pone.0222291

Maia, V. H., Mata, C. S. d., Franco, L. O., Cardoso, M. A., Cardoso, S. R. S., Hemerly, A. S., & Ferreira, P. C. G. (2012). DNA barcoding Bromeliaceae: achievements and pitfalls (DNA barcoding in Bromeliaceae). *PLoS One,* 7(1), e29877. doi:10.1371/journal.pone.0029877

Mallet, J., Besansky, N., & Hahn, M. W. (2016). How reticulated are species? *Bioessays,* 38(2), 140-149. doi:10.1002/bies.201500149

Mayr, E. (1942). *Systematics and the origin of species from the viewpoint of a zoologist*. New York: Columbia University Press.

McCann, J., Jang, T. S., Macas, J., Schneeweiss, G. M., Matzke, N. J., Novak, P.,... Weiss-Schneeweiss, H. (2018). Dating the species network: allopolyploidy and repetitive DNA evolution in American daisies (*Melampodium* sect. *Melampodium*, Asteraceae). *Systematic Biology,* 67(6), 1010-1024. doi:10.1093/sysbio/syy024

McFadden, C. S., Benayahu, Y., Pante, E., Thoma, J. N., Nevarez, P. A., & France, S. C. (2011). Limitations of mitochondrial gene barcoding in *Octocorallia*. *Molecular Ecology Resources,* 11(1), 19-31. doi:10.1111/j.1755-0998.2010.02875.x

Meyer, A. (1990). Ecological and evolutionary consequences of the trophic polymorphism in *Cichlasoma citrinellum* (*Pisces*: *Cichlidae*). *Biological Journal of the Linnean Society,* 39(3), 279-299. doi:10.1111/j.1095-8312.1990.tb00517.x

Moritz, C., & Cicero, C. (2004). DNA barcoding: promise and pitfalls. *PLOS Biology,* 2(10), e354. doi:10.1371/journal.pbio.0020354

Mu, Y. H., Yu, J. R., Cao, T., Wang, X. H., & Yuan, H. S. (2021). Multi-gene phylogeny and taxonomy of *Hydnellum* (Bankeraceae, Basidiomycota) from China. *Journal of Fungi,* 7(10). doi:10.3390/jof7100818

Nixon, K. C., & Wheeler, Q. D. (1990). An amplification of the phylogenetic species concept. *Cladistics,* 6(3), 211-223. doi:10.1111/j.1096-0031.1990.tb00541.x

Ogden, R. (2011). Unlocking the potential of genomic technologies for wildlife forensics. *Molecular Ecology Resources,* 11, 109-116. doi:10.1111/j.1755-0998.2010.02954.x

Ogden, R., Dawnay, N., & McEwing, R. (2009). Wildlife DNA forensics—bridging the gap between conservation genetics and law enforcement. *Endangered Species Research,* 9, 179-195. doi:10.3354/esr00144

One Thousand Plant Transcriptomes, I. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature,* 574(7780), 679-685. doi:10.1038/s41586-019-1693-2

Otero, A., Fernandez-Mazuecos, M., & Vargas, P. (2021). Evolution in the model genus *Antirrhinum* based on phylogenomics of topotypic material. *Frontiers in Plant Science,* 12, 631178. doi:10.3389/fpls.2021.631178

Page, R. D. (2016). DNA barcoding and taxonomy: dark taxa and dark texts. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 371(1702). doi:10.1098/rstb.2015.0334

Pandit, R., Travadi, T., Sharma, S., Joshi, C., & Joshi, M. (2021). DNA meta-barcoding using *rbcL* based mini-barcode revealed presence of unspecified plant species in Ayurvedic polyherbal formulations. *Phytochemical analysis,* 32(5), 804-810. doi:10.1002/pca.3026

Pang, X., Liu, C., Shi, L., Liu, R., Liang, D., Li, H.,... Chen, S. (2012). Utility of the *trnH – psbA* intergenic spacer region and its combinations as plant DNA barcodes: A meta-analysis (utility of *trnH-psbA* and its combinations). *PLoS One,* 7(11), e48833. doi:10.1371/journal.pone.0048833

Percy, D. M., Argus, G. W., Cronk, Q. C., Fazekas, A. J., Kesanakurti, P. R., Burgess, K. S.,... Graham, S. W. (2014). Understanding the spectacular failure of DNA barcoding in willows (*Salix*): does this result from a trans-specific selective sweep? *Molecular Ecology,* 23(19), 4737-4756. doi:10.1111/mec.12837

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS One,* 7(5), e37135. doi:10.1371/journal.pone.0037135

Prokopowich, C. D., Gregory, T. R., & Crease, T. J. (2003). correlation between rDNA copy number and genome size in eukaryotes. *Genome,* 46(1), 48-50. doi:10.1139/g02-103

Puillandre, N., Lambert, A., Brouillet, S., & Achaz, G. (2012). ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology,* 21(8), 1864-1877. doi:10.1111/j.1365-294X.2011.05239.x

Purvis, A., & Hector, A. (2000). Getting the measure of biodiversity. *Nature,* 405(6783), 212-219. doi:10.1038/35012221

Qiu, T., Liu, Z., & Liu, B. (2020). The effects of hybridization and genome doubling in plant evolution via allopolyploidy. *Molecular biology reports,* 47(7), 5549-5558. doi:10.1007/s11033-020-05597-y

Rannala, B. (2015). The art and science of species delimitation. *Current zoology,* 61(5), 846-853. doi:10.1093/czoolo/61.5.846

Ratnasingham, S., & Hebert, P. D. N. (2013). A DNA-based registry for all animal species: The Barcode Index Number (BIN) system. *PLoS One,* 8(7), e66213-e66213. doi:10.1371/journal.pone.0066213

Ren, F. M., Wang, Y. W., Xu, Z. C., Li, Y., Xin, T. Y., Zhou, J. G.,... Song, J. Y. (2019). DNA barcoding of *Corydalis*, the most taxonomically complicated genus of Papaveraceae. *Ecology and Evolution,* 9(4), 1934-1945. doi:10.1002/ece3.4886

Rieseberg, L. H., & Willis, J. H. (2007). Plant speciation. *Science,* 317(5840), 910-914. doi:10.1126/science.1137729

Ruan, J., & Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nature Methods,* 17(2), 155-158. doi:10.1038/s41592-019-0669-3

Sahlin, K., Lim, M. C. W., & Prost, S. (2021). NGSpeciesID: DNA barcode and amplicon consensus generation from long-read sequencing data. *Ecology and Evolution,* 11(3), 1392-1398. doi:10.1002/ece3.7146

Sanderson, B. J., DiFazio, S. P., Cronk, Q. C. B., Ma, T., & Olson, M. S. (2020). A targeted sequence capture array for phylogenetics and population genomics in the Salicaceae. *Applications in Plant Sciences,* 8(10), e11394. doi:10.1002/aps3.11394

Sarmashghi, S., Bohmann, K., MT, P. G., Bafna, V., & Mirarab, S. (2019). Skmer: assembly-free and alignment-free sample identification using genome skims. *Genome Biology,* 20(1), 34. doi:10.1186/s13059-019-1632-4

Satturu, V., Rani, D., Gattu, S., Md, J., Mulinti, S., Nagireddy, R.,... Yanda, R. (2018). DNA fingerprinting for identification of rice varieties and seed genetic purity assessment. *Agricultural Research,* 7(4), 379-390. doi:10.1007/s40003-018-0324-8

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A.,... Consortium, F. B. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America,* 109(16), 6241-6246. doi:10.1073/pnas.1117018109

Sickel, W., Ankenbrand, M. J., Grimmer, G., Holzschuh, A., Härtel, S., Lanzen, J.,... Keller, A. (2015). Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-indexing approach. *BMC ecology,* 15(1), 20-20. doi:10.1186/s12898-015-0051-y

Simmonds, S. E., Smith, J. F., Davidson, C., & Buerki, S. (2021). Phylogenetics and comparative plastome genomics of two of the largest genera of angiosperms, *Piper* and *Peperomia* (Piperaceae). *Molecular Phylogenetics and Evolution,* 163, 107229. doi:10.1016/j.ympev.2021.107229

Soltis, D. E., & Soltis, P. S. (1989). *Isozymes in plant biology/edited by Douglas E. Soltis and Pamela S. Soltis ; introduction by G.L. Stebbins*. Portland, Or: Dioscorides Press.

Stallman, J. K., Funk, V. A., Price, J. P., & Knope, M. L. (2019). DNA barcodes fail to accurately differentiate species in Hawaiian plant lineages. *Botanical Journal of the Linnean Society,* 190(4), 374-388. doi:10.1093/botlinnean/boz024

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Review Genetics*. doi:10.1038/s41576-019-0150-2

Steindor, M., Nkwouano, V., Stefanski, A., Stuehler, K., Ioerger, T. R., Bogumil, D.,... Kalscheuer, R. (2019). A proteomics approach for the identification of species-specific immunogenic proteins in the *Mycobacterium abscessus* complex. *Microbes and Infection,* 21(3-4), 154-162. doi:10.1016/j.micinf.2018.10.006

Su, H. J., Liang, S. L., & Nickrent, D. L. (2021). Plastome variation and phylogeny of *Taxillus* (Loranthaceae). *PLoS One,* 16(8), e0256345. doi:10.1371/journal.pone.0256345

Suissa, J. S., Kinosian, S. P., Schafran, P. W., Bolin, J. F., Taylor, W. C., & Zimmer, E. A. (2022). Homoploid hybrids, allopolyploids, and high ploidy levels characterize the evolutionary history of a western North American quillwort (*Isoetes*) complex. *Molecular Phylogenetics and Evolution,* 166, 107332. doi:10.1016/j.ympev.2021.107332

Sun, C., Hu, Z., Zheng, T., Lu, K., Zhao, Y., Wang, W.,... Wei, C. (2017). RPAN: Rice pan-genome browser for ~3000 rice genomes. *Nucleic acids research,* 45(2), 597-605. doi:10.1093/nar/gkw958

Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A.,... Willerslev, E. (2007). Power and limitations of the chloroplast *trnL* (UAA) intron for plant DNA barcoding. *Nucleic acids research,* 35(3), e14. doi:10.1093/nar/gkl938

Torke, B. M., Cardoso, D., Chang, H., Li, S. J., Niu, M., Pennington, R. T.,... Chung, K. F. (2022). A dated molecular phylogeny and biogeographical analysis reveals the evolutionary history of the trans-pacifically disjunct tropical tree genus *Ormosia*

(Fabaceae). *Molecular Phylogenetics and Evolution,* 166, 107329. doi:10.1016/j.ympev.2021.107329

Twyford, A. D. (2014). Testing evolutionary hypotheses for DNA barcoding failure in willows. *Molecular Ecology,* 23(19), 4674-4676. doi:10.1111/mec.12892

Urbanelli, S., Della Rosa, V., Punelli, F., Porretta, D., Reverberi, M., Fabbri, A. A., & Fanelli, C. (2007). DNA-fingerprinting (AFLP and RFLP) for genotypic identification in species of the *Pleurotus eryngii* complex. *Applied microbiology and biotechnology,* 74(3), 592-600. doi:10.1007/s00253-006-0684-z

Van De Wiel, C. C. M., Van Der Schoot, J., Van Valkenburg, J. L. C. H., Duistermaat, H., & Smulders, M. J. M. (2009). DNA barcoding discriminates the noxious invasive plant species, floating pennywort (*Hydrocotyle ranunculoides* L.f.), from non-invasive relatives. *Molecular Ecology Resources,* 9(4), 1086-1091. doi:10.1111/j.1755-0998.2009.02547.x

Van Valen, L. (1976). Ecological species, multispecies, and oaks. *Taxon,* 25(2/3), 233-239. doi:10.2307/1219444

Vannarattanarat, S., Zieritz, A., Kanchanaketu, T., Kovitvadhi, U., Kovitvadhi, S., & Hongtrakul, V. (2014). Molecular identification of the economically important freshwater mussels (Mollusca–Bivalvia–Unionoida) of Thailand: developing species-specific markers from AFLPs. *Animal Genetics,* 45(2), 235-239. doi:10.1111/age.12115

Vargas, S., Schuster, A., Sacher, K., Büttner, G., Schätzle, S., Läuchli, B.,... Wörheide, G. (2012). Barcoding sponges: an overview based on comprehensive sampling. *PLoS One,* 7(7), e39345-e39345. doi:10.1371/journal.pone.0039345

Wagner, N. D., Gramlich, S., & Horandl, E. (2018). RAD sequencing resolved phylogenetic relationships in European shrub willows (*Salix* L. subg. *Chamaetia* and subg. *Vetrix*) and revealed multiple evolution of dwarf shrubs. *Ecology and Evolution,* 8(16), 8243-8255. doi:10.1002/ece3.4360

Wang, X., Gussarova, G., Ruhsam, M., de Vere, N., Metherell, C., Hollingsworth, P. M., & Twyford, A. D. (2018). DNA barcoding a taxonomically complex hemiparasitic genus reveals deep divergence between ploidy levels but lack of species-level resolution. *AoB Plants,* 10(3), ply026. doi:10.1093/aobpla/ply026

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Review Genetics,* 10(1), 57-63. doi:10.1038/nrg2484

Waser, N. M., & Campbell, D. R. (2004). Ecological speciation in flowering plants. In *Adaptive Speciation* (pp. 264-277).

Weitemier, K., Straub, S. C., Cronn, R. C., Fishbein, M., Schmickl, R., McDonnell, A., & Liston, A. (2014). Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences,* 2(9). doi:10.3732/apps.1400042

Wei, T., van Treuren, R., Liu, X., Zhang, Z., Chen, J., Liu, Y.,... Liu, H. (2021). Whole-genome resequencing of 445 *Lactuca* accessions reveals the domestication history of cultivated lettuce. *Nature Genetics*. doi:10.1038/s41588-021-00831-0

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T.,... Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology,* 37(10), 1155-1162. doi:10.1038/s41587-019-0217-9

Wu, G. A., Terol, J., Ibanez, V., Lopez-Garcia, A., Perez-Roman, E., Borreda, C.,... Talon, M. (2018). Genomics of the origin and evolution of *Citrus*. *Nature,* 554(7692), 311-316. doi:10.1038/nature25447

Xia, M., Liu, Y., Liu, J., Chen, D., Shi, Y., Chen, Z.,... Qiu, Y. (2022). Out of the Himalaya-Hengduan Mountains: Phylogenomics, biogeography and diversification of Polygonatum Mill. (Asparagaceae) in the Northern Hemisphere. *Molecular Phylogenetics and Evolution,* 169, 107431. doi:10.1016/j.ympev.2022.107431

Xu, S. Z., Li, Z. Y., & Jin, X. H. (2018). DNA barcoding of invasive plants in China: A resource for identifying invasive plants. *Molecular Ecology Resources,* 18(1), 128-136. doi:10.1111/1755-0998.12715

Yan, L. J., Liu, J., Moller, M., Zhang, L., Zhang, X. M., Li, D. Z., & Gao, L. M. (2015). DNA barcoding of *Rhododendron* (Ericaceae), the largest Chinese plant genus in biodiversity hotspots of the Himalaya-Hengduan Mountains. *Molecular Ecology Resources,* 15(4), 932-944. doi:10.1111/1755-0998.12353

Yao, H., Song, J., Liu, C., Luo, K., Han, J., Li, Y.,... Chen, S. (2010). Use of ITS2 region as the universal DNA barcode for plants and animals. *PLoS One,* 5(10). doi:10.1371/journal.pone.0013102

Yeo, D., Srivathsan, A., & Meier, R. (2020). Longer is not always better: optimizing barcode length for large-scale species discovery and identification. *Systematic Biology,* 69(5), 999-1015. doi:10.1093/sysbio/syaa014

Yeo, P. F. (1968). The evolutionary significance of the speciation of *Euphrasia* in Europe. *Evolution,* 22(4), 736-747. doi:10.1111/j.1558-5646.1968.tb03473.x

Yu, X. Q., Jiang, Y. Z., Folk, R. A., Zhao, J. L., Fu, C. N., Fang, L.,... Yang, S. X. (2022). Species discrimination in *Schima* (Theaceae): Next-generation super-barcodes meet evolutionary complexity. *Molecular Ecology Resources*. doi:10.1111/1755-0998.13683

Yu, J., Hu, S., Wang, J., & Li, S. (2002). A draft sequence of the rice genome (*Oryza sativa* L. *ssp*. *indica*). *Science,* 296(5565), 79-92. doi:10.1126/science.1068037

Zhang, L., Huang, Y. W., Huang, J. L., Ya, J. D., Zhe, M. Q., Zeng, C. X.,... Yang, J. B. (2022). DNA barcoding of *Cymbidium* by genome skimming: call for next-generation nuclear barcodes. *Molecular Ecology Resources*. doi:10.1111/1755-0998.13719

Zheng, H. Y., Guo, X. L., Price, M., He, X. J., & Zhou, S. D. (2021). Effects of mountain uplift and climatic oscillations on phylogeography and species divergence of *Chamaesium* (Apiaceae). *Frontiers in Plant Science,* 12, 673200. doi:10.3389/fpls.2021.673200

Chapter 2 Conceptual issues that require consideration prior to assessing and using nuclear DNA for plant barcoding

2.1. Summary

This chapter is an overview of the issues that need to be worked through in the use of nuclear DNA to tell species apart. I firstly briefly outline the strengths and weaknesses of the standard DNA barcoding approach in plants to contextualise the pressing need for adding nuclear DNA data. I then highlight some conceptual issues associated with evaluating nuclear DNA sequence data for DNA barcoding, including taxon and genomic sampling, and handling analytical complexity associated with heterozygosity, paralogy, and repetitive sequences. Building on these technical considerations for evaluating existing datasets to understand discriminatory power and the nature of genomic differences between plant species, I then outline a simple research roadmap to support the development of a nuclear DNA barcode.

2.2. Introduction

DNA barcoding involves the sequencing of standardised DNA regions to tell species apart (Hebert et al., 2003). The approach works extremely well in animals, but in plants, DNA barcoding is more challenging. Although many species can be distinguished using standard plant barcodes, there are also many cases where plant barcodes do not provide adequate species level discrimination (Hollingsworth et al., 2016).

2.3. Strengths and limitations of the existing approach to plant DNA barcoding with plastid regions and nrDNA ITS

Genomic regions are chosen as DNA barcodes if they fit three essential principles of DNA barcoding, i.e., standardisation, minimalism, and scalability (Hebert et al., 2003). The most frequently used DNA barcodes for plants, i.e. the core DNA barcodes, are the plastid gene regions *matK* and *rbcL*, plus the internal transcribed spacer (ITS) and ITS2 from nuclear ribosomal DNA (nrDNA) (Hollingsworth et al., 2011). In addition, several other regions, typically plastid genes, introns or spacers are often used as supplementary barcodes (Table 2.1).

Table 2.1. A list of frequently used genomic regions for plant DNA barcoding

Locus	Med an amp con ength (base pa rs) of barcode reg on	Genom c reg on	Reference
matK	889	P ast d gene	(Johnson et a 1994 Ho ngsworth et a 2011)
rbcLa	654	P ast d gene	(Be et a 2017)
rpoB	548	P ast d gene	(Drancourt et a 2002)
rpoC1	616	P ast d gene	(P on et a 2007)
trnH-psbA	509	P ast d ntergen c spacer	(Loera-Sanchez et a 2020)
atpF atpH	669	P ast d ntergen c spacer	(Renaud et a 2008)
psbK psbI	468	P ast d ntergen c spacer	(Suzuk et a 2014)
trnL-F	994	P ast d ntron and ntergen c spacer	(Chen et a 2013)
ITS	705	R bosoma DNA nterna transcr bed spacer	(CBOL et a 2011)
ITS2	494	R bosoma DNA nterna transcr bed spacer	(Yao et a 2010)

2.3.1. The strengths of plastid regions and nrDNA as DNA barcodes

Plastid and nrDNA are technically easy to sequence due to their biological features. They both have high copy-number per cell. The number of plastids per cell ranges from one in unicellular algae to several hundred in wheat mesophyll cells (Kubinova et al., 2014). The number of rDNA copies ranges from a few hundred to thousands in plant genomes and does not correlate with the genome size (Prokopowich et al., 2003). The multi-copy nature of these regions leads to ease of PCR amplification with ready-to-use primers. These regions are also well-characterised due to decades of use in plant systematic studies.

Compared to the whole genome, the size of DNA barcodes are in a different order of magnitude. The plant genome size ranges from dozens of megabases (~ 63.6 Mb for *Genlisea aurea*) to hundreds of gigabases (~ 150 Gb for *Paris japonica*) (Pellicer et al., 2018). By comparison, DNA barcodes are much smaller, ranging from hundreds to a few thousand base pairs, and hence they are well suited for routine recovery and simple data analyses, with minimal demands on data storage and analytical computing power.

The core DNA barcodes have well-established reference databases. Abundant records can be found in the two most important reference databases for DNA barcoding, the Barcode of Life Data Systems (BOLD) (Ratnasingham et al., 2007) and NCBI GenBank. To date, the BOLD system encompasses DNA barcodes for over 72K plant species (retrieved in December 2022), and there are over 1 million sequences for the core plant DNA barcodes collectively in GenBank (255,443 *matK*, 290,191 *rbcL*, 489,472 ITS & ITS2, retrieved 10 Nov 2022). Regional reference libraries are emerging across continents (Kress et al., 2009, Xu et al., 2018) and bioinformatics pipelines to assist establishing these regional reference libraries are also available (Liu et al., 2021).

The consensus of sequencing core plant barcodes has led to a comparable dataset among different working groups through time. Moreover, the data is also compatible with the popularity of the next generation sequencing, because these core plant barcodes are also recoverable from genome skimming (Dodsworth, 2015, Twyford et al., 2017), target capture (Weitemier et al., 2014, Kozarewa et al., 2015, Stephens et al., 2015, du Preez et al., 2018), and even RAD-seq data (Lin et al., 2019, Scharmann et al., 2021).

2.3.2. The weaknesses of plastid regions and nrDNA as DNA barcodes

Levels of resolution using standard plant barcodes are too low in many cases to discriminate among species and provide adequate discriminatory power. The diversity of the plant world reflects a complex set of evolutionary processes including polyploidization (Zenil-Ferguson et al., 2019, Debray et al., 2022), frequent hybridization and introgression (Soltis et al., 2009, Scharmann et al., 2021), and recent diversification of many plant groups (Magallon et al., 2001, Yardeni et al., 2021). The standard core plant DNA barcodes gave an average discrimination success of 72% based on a study on 907 samples, representing 445 angiosperm, 38 gymnosperm, and 67 cryptogam species (CBOL, 2009). However, the literature is replete with examples of plants sharing barcodes among related species. For example, species-

level resolution can be as low as 10 ~ 20 % using standard plant DNA barcodes as shown in genera such as *Rhododendron* (Fu et al., 2022), *Salix* (Percy et al., 2014), and *Euphrasia* (Wang et al., 2018), or in the floras of specific geographic locations such as young, oceanic islands because of rapid speciation, high incidence of hybridization and polyploidy (Stallman et al., 2019). Therefore, pilot studies, careful project design, and critical use of plant barcoding are essential to avoid disappointing and/or uninformative results (Hollingsworth et al., 2016).

As one of the core plant DNA barcodes, *matK* can be difficult to PCR amplify using a universal primer set due to practical issues such as primer mismatches. Using the best currently available 'universal' primer pair (472F/1248R) results in PCR amplification success of ca. 93.1% in angiosperms (Yu et al., 2011). In gymnosperm, the best 'universal' primer pair (Gym_F1A/Gym_R1A) reaches a 94.7% success PCR amplification rate (Li et al., 2011). Other studies, however, have reported lower rates of success, promoting the design of complex multiplex PCRs or time consuming taxon-specific primer approaches (Heckenhauer, Barfuss, & Samuel, 2016; Jones et al., 2021).Furthermore, *matK* is still not recoverable from some bryophyte and fern groups with available primer sets. Ferns in particular represent a challenge for *matK* recovery as genome rearrangements mean that the gene is not flanked by conserved *trnK* exons in some clades (Kuo et al., 2011), creating additional difficulties in generating full-length *matK* sequences from which to design primers for the barcode region.

The use of nuclear ITS often increases levels of resolution beyond those of plastid markers but within limits. An assessment of species discrimination success and sequence quality based on 3,011 individuals representing 765 species shows that ITS discriminated 67% of samples, while *rbcL* only had a discrimination rate of 26% (China Plant Barcoding of Life Group et al., 2011). In 7% of individuals and 9% of species of all 6286 samples assessed, multiple copies within individuals were detected. The polymorphisms between repeats creates challenges in sequencing and/or analyses. In addition, interspecific barcode sharing either through lack of divergence or hybridization is also not uncommon in ITS datasets (China Plant Barcoding of Life Group et al., 2011).

The intrinsic limitation of using plastid and ITS barcodes for species identification is that the organelle genomes and rDNA haplotypes often do not track species boundaries, and in plants, numerous studies have reported phylogenetic discordance between nrDNA and plastid sequence data (Mu et al., 2020, Stull et al., 2020). These basic limitations are common to extended plant barcodes which use genome skimming to go beyond standard DNA barcodes and obtain complete plastid genomes and complete rDNAs. These approaches can lead to modest gains in resolution but do not address the fundamental problem of limited species identification ability of plastid and rDNA loci (Fu et al., 2022). On the one hand, the limited size of the plastid genome and rDNA constrains the amount of information they carry. This is demonstrated by a phylogeny for diploid *Helianthus* built by using whole plastome data. The whole plastome alignments show a low level of polymorphism and result in a large polytomy for the majority of species and very little species resolution (Stephens et al., 2015). Another challenge is that the plastid genome and rDNA do not always track species boundaries, and the evolutionary history for these loci can be very different from the histories revealed by multiple loci from the nuclear genome. Cytonuclear discordance (between plastid and nuclear gene trees) has been well-documented in

recent radiated groups, such as southern African *Oxalis*, where there was cytonuclear discordance between the species tree built by using 727 low copy number genes (LCN) and the phylogeny built by the whole plastome (du Preez et al., 2018). Likewise various studies of allopolyploids have illuminated the complex dynamics of rDNA evolution, including complex within individual polymorphism (Devos, Oh, Raspé, Jacquemart, & Manos, 2005), or conversely elimination and/or marked asymmetry in the maintenance of different rDNA types due to concerted evolution following hybridisation (Kovarik et al., 2005). These complex dynamics of rDNA evolution hamper its universal utility as a plant barcode.

Overall, the limitations described above of standard barcodes (from plastid regions and/or ITS), or extended barcodes (plastid genomes and rDNA arrays) have triggered interest in exploiting the nuclear genome as a source of markers for plant species discrimination.

2.4. Rationale for a nuclear DNA barcode for plants

Various studies have shown the ability of multiple unlinked nuclear markers to provide high discriminatory power in many plant groups, separating species, and infra-specific taxa (Urbanelli et al., 2007, Gholave et al., 2017, Bi et al., 2021, Hua et al., 2022). This includes previous studies using amplified fragment length polymorphisms (AFLPs), simple sequence repeats (SSRs), and even allozymes. For example, to produce DNA fingerprints for the identification of species used as medicines or in agriculture, multiple studies have developed dozens of AFLP and SSR marker primers for genera such as *Swertia* (Lin et al., 2019), *Zingiber* (Ghosh et al., 2011), *Olea* (Ercisli et al., 2011), and *Oryza* (Satturu et al., 2018), and dozens of ad hoc primer pairs were selected to produce specific fingerprints of each focal species. However, these data are intrinsically poorly suited for comparisons and reuse because each study designs specific primers to amplify the most suited genomic regions for that specific group, and the approaches are too specific to apply to organisms other than their original targets. Similar patterns apply to many other unlinked genetic marker studies, most data are designed for specific use and are hard to compare and scale across groups.

Access to the nuclear genome via high-throughput sequencing now enables the generation of large amounts of intrinsically more comparable sequence data (as opposed to fragment length data) and this is transforming understanding of the nuclear genomes of plants. Chapter 1 summarises the major sequencing techniques and platforms to sequence entire or partial nuclear genomes at a reasonable cost, including genome skimming (Dodsworth, 2015), transcriptome sequencing (Wang et al., 2009), and reduced representation sequencing techniques such as RAD sequencing (Davey et al., 2010, Davey et al., 2011, Peterson et al., 2012), GBS (Elshire et al., 2011, Deschamps et al., 2012) and target capture (Weitemier et al., 2014). The rapid progress in increasing throughput while decreasing the cost of sequencing enables numerous studies to access the nuclear genomes of multiple individuals and populations from a diverse set of plant groups.

With nuclear sequence data increasingly available for individuals and populations, the use of nuclear sequence data for species discovery and specimen identification is becoming realisable. The increasing density of reference genome availability offers the promise of a comprehensive inventory of inter-specific differences. Likewise, the widespread availability of reduced representation datasets gives insights into the frequency distribution of taxonomically informative characters in the nuclear genome (and the patterns of nucleotide variation among species). Such data offer the opportunity for designing the next generation of plant barcoding approaches based on a detailed understanding of genomic differences between species. In light of high density availability of nuclear sequence data, identification of genetic changes associated with specialized traits in specific lineages, and searching for variable loci which are diagnostic at a family, genus, species, variety, population, and even individual levels will become possible. This could ultimately lead to transformative gains in discriminatory power, and realisation of the goal of routine and automated identification of plants, plant parts and plant products at the species level.

2.5. Technical considerations for evaluating existing datasets to understand discriminatory power and the nature of genomic differences between plant species

In this thesis I undertake a series of analyses of patterns of sequence differences among plant species (Chapters 4 & 5). Before doing this, it is important to evaluate the operational and technical issues which can impact on the analysis and understanding of existing datasets. Gaining a better understanding of the extent of genomic differences between plant species must be done carefully, considering a range of different variables in a given dataset. Here, I consider each of these in turn, explaining their potential impacts on the inference made.

2.5.1. Taxon sampling

Taxon sampling is key to the evaluation of the discriminatory power of species identification methods. Where sample density is low (few species sampled, few individuals per species sampled), there is a greater likelihood of detecting DNA substitutions which appear to be diagnostic, but which are not robust to subsequent sampling.

Sample density per species: Fewer sampled individuals per species increase the chances of detecting false positive species-specific SNPs. This is a critical point in the evaluation of improved methods for DNA barcoding, and the discriminatory power of a given dataset. Adding increased quantities of sequence data will inevitably detect variation between samples, but unless multiple individuals are sampled per species, it is not clear if this variation is informative for species identification (as opposed to being autapomorphic or otherwise uninformative substitutions). For instance, various studies have promoted whole plastid genome sequencing as informative for species discrimination, but are based on only single individuals sampled per species (Nock et al., 2011). Such studies only provide limited insights into species discrimination, as it is not clear if the differences detected between the sequenced plastome will actually correspond to reliable markers tracking species boundaries. Thus as a bare minimum, the presence of at least two sampled individuals per species is necessary to evaluate whether a given dataset is informative for species discrimination Intuitively, the more individuals sampled per species, and the better the coverage of a species range, the more likely the sample is to capture intra-species variation, and hence the greater the confidence in the resulting data. A major consideration is therefore how to partition effort between sampling of individuals to obtain robust insights.

Sample density per genus: The fewer the species sampled per genus, the greater the likelihood of over-estimating species discrimination power. If only few species are sampled per genus, there is a lower likelihood of sampling sister species (which are the most difficult taxa to discriminate). Thus a sparse sampling of species per genus is likely to result in a failure to capture species that share the same nucleotide variations or haplotypes. Where the proportion of species sampled per genus is low, then the confidence that the levels of species discrimination detected are robust is correspondingly low. Of course, if the aim is floristic sampling, focusing on a geographically restricted sample set, then partial sampling of genera may be relevant and acceptable. However, it is important to recognise that the discrimination obtained

at a local floristic scale may not then scale when additional congeneric samples are added.

2.5.2 Taxon attributes

When the efficacy of different data sets is being compared and figures are reported on levels of species discrimination, an important set of variables to consider includes the nature of the taxa that were sampled, as some biological situations are intrinsically more complicated than others to resolve.

Genus size: A basic element that might impact the discriminatory power of a given dataset is genus size. Larger genera containing more species, increases the number of taxa to be distinguished and hence may represent an intrinsically more demanding challenge for discrimination. Thus, different target taxa will give different perceptions on the scale of the challenge. Simplistically (and for illustrative purposes excluding the use of character combinations), a genus of 1000 species will need at least 1000 nucleotide variants to uniquely distinguish every species, while a genus of 10 species will only need 10. A confounding practical factor here is the age / levels of divergence of a given genus, and in practice, there may be greater difficulties distinguishing species in a small genus that has radiated recently, compared to a large genus that has a longer diversification time (and greater time for species to accumulate diagnostic divergent sequences).

Biological complexity: The nature of genetic divergence among species will vary depending on their evolutionary history and dynamics. There is an expectation of systematic differences in the signal between groups which have radiated recently, and/or have a history of hybridisation compared to taxa where species diversity is older and hybridisation is infrequent. With a lower frequency of hybridisation and genetic recombination, the mutation accumulation is at a steady and relatively slow pace, usually resulting in a few new species every few million years (Coyne et al., 2004). Some plant groups, appear inherently prone to hybridise and thus some lineages show highly imbalanced phylogenetic patterns of species richness, e.g *Salix* (Gramlich et al., 2016), Hawaiian silversword alliance (Baldwin et al., 1998), *Tragopogon* goatsbeard flowers (Marques et al., 2019), *Helianthus* sunflowers (Rieseberg et al., 2003), monkey flower *Mimulus aurantiacus* complex (Stankowski et al., 2015), and *Rosa* (Debray et al., 2022). During the hybridisation process, the recombination of old genetic variations from both parental lineages enables rapid speciation and adaptive radiation (Marques et al., 2019). This ‘combinatorial mechanism’ was proposed when case studies of speciation show conflict with standard speciation models. For instance, the monkeyflower *Mimulus guttatus* speciated in the past 150 years as a consequence of a pre-existing hybrid lethality mutation hitchhiking to high frequency in a copper mine population by the tight link to a novel copper-tolerance allele (Wright et al., 2013). The key point here is that some datasets which show poor resolution, may relate to limited power of the sequencing approach, or alternatively, a simply very challenging biological situation to resolve.

2.5.3. Genome sampling

The type of data available (target capture, shotgun sequencing, GBS, RAD) will impact the level of analysis needed and there are also systematic differences in signal among such datatypes.

Coding versus non-coding data: One of the most prominent differences the data brings is whether the sequences are from coding or non-coding regions, with data from coding regions tending to be more conserved than that from non-coding regions. Reduced representation sequence methods like RAD-Seq and GBS can be designed to recover sequences with a bias towards coding regions based on the enzyme choice. For example, Sbf1 is a frequently used rare-cutter enzyme, it's a GC-rich enzyme so it tends to cut in coding regions (Cariou et al., 2013, Herrera et al., 2015). But in general, the distribution of the enzyme cutting-sites are random (Herrera et al., 2015) resulting in frequent recovery of non-coding sequences in enzyme based reduced representation studies. Target capture approaches typically focus on genes with introns and adjacent regions depending on the purpose of designing the baits (Kozarewa et al., 2015). RNA-seq, above all, aiming to sequence the full transcriptome, has the highest proportion of coding regions (Wang et al., 2009). Normally, shotgun sequencing methods such as genome skimming are less biased toward specific regions (although their non-targeted approach naturally leads to the recovery of high-copy number sequences more readily than low-copy number sequences).

Inclusion or exclusion of organelle data: Another difference from different types of data is the exclusion or inclusion of organelle sequence data. Different patterns of taxon-specific substitutions are expected between nuclear and organelle genomes, reflecting their different modes of inheritance and dynamics of evolution. The genome skimming method is prone to capture whole plastome sequences due to their high copy number. The target capture method often has a lower chance of capturing organelle sequences if not designed to, but nevertheless plastid sequences are common as 'by-catch' in target capture studies (Baker et al., 2022; Stull et al., 2013). Other sequencing methods lie in between in terms of the proportion of organelle sequences recovered.

The extent of missing data: Recovery success rate varies among different types of data, and varied success in recovery will lead to varying levels of missing data which may influence species discrimination success. The element of missing data introduced by systematic bias should be distinguished from stochastic information loss. Mutation disruption for RAD-seq and GBS, due to restriction enzymes failing to digest conserved sites, typically happens to samples from a certain collection of taxa, rather than random loss. This type of signal is more evident at deeper phylogenetic scales, thus the missing data problem for RAD-seq and GBS is more prominent with a broader sampling scheme (Harvey et al., 2016). Hybridisation failure for target capture applies when baits fail to recognise the target sequence, and the more divergent the groups, the bigger the chance of missing data in taxa that are divergent from the reference genome on which the baits were designed. Low capture efficiency can be exacerbated by degraded DNA in herbarium samples, which typically increase the proportion of missing data in a dataset (Villaverde et al., 2018).

Linkage of loci: Discrimination statistics associated with independent, random loci provide representation across the genome. In contrast, regions that are in linkage disequilibrium provide correlated information which may influence species

discrimination statistics. In comparing the efficacy of different datasets for species discrimination, understanding the level of independence across the different sequenced loci may be informative in understanding the resulting patterns of species discrimination.

2.5.4. Sequence production

The degree to which species-specific SNPs can be reliably identified is closely associated with sequence quality.

Read length: The lengths of the reads have a direct impact on the accuracy of alignment and clustering. Longer reads are preferable as it reduces the chance of clustering paralogues into chimeric assemblies, which can lead to miscalls. RAD-seq loci are often short, and not always easy to assemble into consensus sequences without a high-quality reference genome. Long-read sequencing such as PacBio and Nanopore technologies usually produce reads with lengths spanning out from repetitive regions and therefore works better when no reference genome is available.

Sequence depth: Another factor that can have an impact on the accuracy of variant calling is sequence depth. Low-coverage sequencing can result in miscalls and polymorphic sites appearing uniform, impacting on accurate detection of species-specific SNPs. It's particularly true for genome skim data where only one or two fold coverage might be obtained for many low copy number nuclear regions. Where sample density per species is low, this may give the misleading impression of there being large numbers of fixed homozygous species-specific SNPs which are actually heterozygous and variable, hence giving a misleading impression of discriminatory power.

2.5.5. Analytical complexity

Heterozygosity: Genome heterozygosity is one of the main complexities in retrieving species-specific SNPs. Heterozygosity is pervasive in land plants as a result of frequent hybridisation (Soltis et al., 2009) and more generally, the presence of genetic variation in populations of outcrossing species (Hamrick & Godt, 1996). Heterozygous loci inadvertently scored as homozygotes may impact the accurate detection of species-specific SNPs. This is most likely to be an issue in reusing datasets initially produced for phylogenetic reconstruction, where heterozygosity is usually masked and the most common variant at a given site is called in the consensus sequence. Most of the current nucleotide substitution models for ancestor state inference are based on homozygotes, with only a few establishing models taking the heterozygosity of given sites into consideration (Schrempf et al., 2016, Schrempf et al., 2019, Minh et al., 2020). With the masked datasets, an SNP could be miscalled as species-specific either when the focal species is actually heterozygous, or when the focal species is homozygous, but other non-focal species actually have the same nucleotide as the focal species but are masked during early analytical steps. In this case, obtaining the raw reads will be conducive to recovering the genuine heterozygosity (and hence taxonomically informative) information.

Orthology and paralogy: Another major challenge for the interpretation of nuclear sequence variation in plants is distinguishing orthologous and paralogous loci. Incorrect assessments of orthology may lead to loci being merged, potentially underestimating the frequency of species-specific SNPs. The pervasiveness of paralogy in plants is a combined result of several reasons. One of them is due to numerous independent whole genome duplication (WGD) events across land plant lineages (Soltis et al., 2015, Levin, 2020, Li et al., 2021). Another is the frequent occurrence of more localised gene duplication and gene family expansion events mediated by transposable elements (Munoz-Lopez et al., 2010), tandem and segmental duplication (Leister, 2004), or other mechanisms (Freeling, 2009). As a result, large portions of plant genome regions are present in multiple copies.

Plant species with large amounts of highly repetitive DNA present in sequence datasets may have few species-specific SNPs detected, due to their being a proportionately low representation of single-copy loci (and ultimately few independent loci being assayed). This is a particular challenge for groups like gymnosperms which typically have more than 90% highly repetitive regions (Luo et al., 2022). Random reduced representation sequencing techniques such as RAD-seq, GBS, and genome skimming might not be suitable in such cases, as the probability of recovering low/single copy loci is low.

2.6. Meta-data and data analysis

The selection of the original standard core organelle barcode markers for plants (*rbcL*, *matK*) was based on a common evaluation and data handling framework which allowed different datasets from different research groups to be co-analysed to inform the establishment of community standards (CBOL Plant Working Group, 2009). To take forward the development of nuclear DNA barcodes for plants, a similar approach needs to be taken which accommodates the additional complexities of working with data from the nuclear genome.

Different datasets of nuclear sequences that are generated for different purposes tend to have different types of meta-data and sample processing histories. At a practical level, there is a basic requirement for standardised sets of meta-data and accurate records of data processing approaches to facilitate use and comparisons among datasets, both in terms of guiding the generation and analysis of new datasets, and as a set of criteria outlining the requirements for re-use of existing datasets. To this end, I have outlined below the key areas to enhance the development of datasets which will be informative for understanding the nature of the differences between plants species and contributing towards the development of nuclear DNA barcoding in plants:

2.6.1. Meta-data

1) Basic information should be recorded in a systematic fashion that describes each dataset including the taxonomy, sampling scheme, a naming record tracking file, and sequence alignments. Exemplary files in required formats are provided in 3.3.2. And a template for essential files can be found at https://github.com/Hazelhuangup/Species_specific_alleles_analysis/tree/main/test_file.

2) A tracking history of how misidentifications are handled should be included. In assessing discriminatory power it is necessary to understand how researchers have handled misidentifications that have come to light during a study. This includes rectifying obvious errors of identification, but not arbitrarily discarding or renaming individual samples just because they don't form the expected monophyletic species clusters. This step is important to avoid under- or over-estimating species discrimination success.

3) A summary of the sample scheme and rationale is important as a guideline to assess the confidence level of species discrimination. As we discussed in section 2.5.1, both the taxon and genome sampling approaches can have an impact on levels of species discrimination. Some basic standardised information can help considerably in contextualising the results from a given dataset (such as the proportion of species in the genus that have been sampled, and the level of confidence in the species identifications used in the final dataset). Likewise recording the rationale behind any laboratory procedures can also help contextualise the findings (e.g. the rationale behind the choice of a certain enzyme to digest the genome or the loci selected in a target capture approach).

4) A detailed bioinformatics record and pipeline annotation should always be connected to the data. This point is a re-emphasis of the FAIR Guiding Principles for scientific data management and stewardship (Wilkinson et al., 2016). To be able to maximize the impact of research the bioinformatic analytical steps and code and software used in the production of the original data should be findable, accessible, interoperable and reusable. Having access to such records will promote the reuse and interpretation of datasets for meta-analyses.

5) Sequence alignments are an important resource for assessing the efficacy of sequence data for species discrimination. The gold standard for processed data is the variant calling format (vcf). A file in vcf format contains but is not limited to information including heterozygosity and allele information, variant calling quality, base quality, site depth, and haplotype information if a reference genome was available. All of the above information can help in understanding the signal in a dataset for telling species apart.

2.7. Conclusion

In this chapter, I have undertaken an evaluation on the range of potential issues to consider in using nuclear DNA to tell species apart. highlight some conceptual issues associated with the future developments of nuclear DNA barcoding, including taxon and genomic sampling density, and handling analytical complexity brought up by heterozygosity, paralogy, and repetitive sequences, and meta-data standards that promote the reuse of datasets for species discrimination studies.

2.8. Reference

Baker, W. J., Bailey, P., Barber, V., Barker, A., Bellot, S., Bishop, D., . . . Forest, F. (2022). A comprehensive phylogenomic platform for exploring the angiosperm tree of life. *Systematic Biology,* 71(2), 301-319. doi:10.1093/sysbio/syab035

Baldwin, B. G., & Sanderson, M. J. (1998). Age and rate of diversification of the Hawaiian silversword alliance (Compositae). *Proceedings of the National Academy of Sciences,* 95(16), 9402-9406. doi:10.1073/pnas.95.16.9402

Bell, K. L., Loeffler, V. M., & Brosi, B. J. (2017). An *rbcL* reference library to aid in the identification of plant species mixtures by DNA metabarcoding. *Applications in Plant Sciences,* 5(3), 1600110-n/a. doi:10.3732/apps.1600110

Bi, D., Chen, D., Khayatnezhad, M., Hashjin, Z., Li, Z., & Ma, Y. (2021). Molecular identification and genetic diversity in *Hypericum* L.: A high value medicinal plant using RAPD markers markers. *Genetika,* 53(1), 393-405. doi:10.2298/gensr2101393b

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., & Abebe, E. (2005). Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 360(1462), 1935-1943. doi:10.1098/rstb.2005.1725

Čandek, K., & Kuntner, M. (2015). DNA barcoding gap: reliable species identification over morphological and geographical scales. *Molecular Ecology Resources,* 15(2), 268-277. doi:10.1111/1755-0998.12304

Cariou, M., Duret, L., & Charlat, S. (2013). Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. *Ecology and Evolution,* 3(4), 846-852. doi:10.1002/ece3.512

CBOL, C. P. W. G., Li, D. Z., Gao, L. M., Li, H. T., Wang, H., Ge, X. J., . . . Duan, G. W. (2011). Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences,* 108(49), 19641-19646. doi:10.1073/pnas.1104551108

CBOL, P. W. G. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences,* 106(31), 12794-12797. doi:10.1073/pnas.0905845106

Chen, C. W., Huang, Y. M., Kuo, L. Y., Nguyen, Q. D., Luu, H. T., Callado, J. R., . . . Chiou, W. L. (2013). trnL-F is a powerful marker for DNA identification of field vittarioid gametophytes (Pteridaceae). *Annual Botany,* 111(4), 663-673. doi:10.1093/aob/mct004

Coyne, J. A., & Orr, H. A. (2004). *Speciation / Jerry A. Coyne, H. Allen Orr*. Sunderland, Mass: Sinauer Associates.

Davey, J. W., & Blaxter, M. L. (2010). RADSeq: next-generation population genetics. *Briefings in Functional Genomics,* 9(5-6), 416-423. doi:10.1093/bfgp/elq031

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics,* 12(7), 499-510. doi:10.1038/nrg3012

Debray, K., Le Paslier, M. C., Berard, A., Thouroude, T., Michel, G., Marie-Magdelaine, J., . . . Malecot, V. (2022). Unveiling the patterns of reticulated evolutionary processes with phylogenomics: hybridization and polyploidy in the genus *Rosa*. *Systematic Biology,* 71(3), 547-569. doi:10.1093/sysbio/syab064

Deschamps, S., Llaca, V., & May, G. D. (2012). Genotyping-by-Sequencing in plants. *Biology,* 1(3), 460-483. doi:10.3390/biology1030460

Devos, N., Oh, S.-H., Raspé, O., Jacquemart, A.-L., & Manos, P. S. (2005). Nuclear ribosomal DNA sequence variation and evolution of spotted marsh-orchids (*Dactylorhiza maculata* group). *Molecular Phylogenetics and Evolution,* 36(3), 568-580. doi:10.1016/j.ympev.2005.04.014

Dodsworth, S. (2015). Genome skimming for next-generation biodiversity analysis. *Trends in Plant Science,* 20(9), 525-527. doi:10.1016/j.tplants.2015.06.012

Dodsworth, S., Pokorny, L., Johnson, M. G., Kim, J. T., Maurin, O., Wickett, N. J., . . . Baker, W. J. (2019). Hyb-Seq for flowering plant systematics. *Trends in Plant Science,* 24(10), 887-891. doi:10.1016/j.tplants.2019.07.011

Donoghue, M. J. (1985). A critique of the biological species concept and recommendations for a phylogenetic alternative. *The Bryologist,* 88(3), 172-181. doi:10.2307/3243026

Drancourt, M., & Raoult, D. (2002). *rpoB* gene sequence-based identification of *Staphylococcus* species. *Journal of Clinical Microbiology,* 40(4), 1333-1338. doi:10.1128/JCM.40.4.1333-1338.2002

du Preez, B., Dreyer, L. L., Schmickl, R., Suda, J., & Oberlander, K. C. (2018). Plastid capture and resultant fitness costs of hybridization in the Hirta clade of southern African *Oxalis*. *South African journal of botany,* 118, 329-341. doi:10.1016/j.sajb.2017.06.010

Dupont, L. M., Linder, H. P., Rommerskirchen, F., & Schefuß, E. (2011). Climate-driven rampant speciation of the Cape flora. *Journal of Biogeography,* 38(6), 1059-1068. doi:10.1111/j.1365-2699.2011.02476.x

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One,* 6(5), e19379. doi:10.1371/journal.pone.0019379

Ercisli, S., Ipek, A., & Barut, E. (2011). SSR Marker-based DNA fingerprinting and cultivar identification of Olives (*Olea europaea*). *Biochemical Genetics,* 49(9-10), 555-561. doi:10.1007/s10528-011-9430-z

Freeling, M. (2009). Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annual Review Plant Biology,* 60, 433-453. doi:10.1146/annurev.arplant.043008.092122

Frodin, D. G. (2004). History and concepts of big plant genera. *Taxon,* 53(3), 753-776. doi:10.2307/4135449

Fu, C. N., Mo, Z. Q., Yang, J. B., Cai, J., Ye, L. J., Zou, J. Y., . . . Gao, L. M. (2022). Testing genome skimming for species discrimination in the large and taxonomically difficult genus *Rhododendron*. *Molecular Ecology Resources,* 22(1), 404-414. doi:10.1111/1755-0998.13479

Germano, J., & Klein, A. S. (1999). Species-specific nuclear and chloroplast single nucleotide polymorphisms to distinguish *Picea glauca, P. mariana* and *P. rubens*. *Theoretical and Applied Genetics,* 99(1-2), 37-49. doi:10.1007/s001220051206

Gholave, A., Pawar, K., Yadav, S., Bapat, V., & Jadhav, J. (2017). Reconstruction of molecular phylogeny of closely related Amorphophallus species of India using plastid DNA marker and fingerprinting approaches. *Physiology and Molecular Biology of Plants,* 23(1), 155-167. doi:10.1007/s12298-016-0400-0

Ghosh, S., Majumder, P. B., & Sen Mandi, S. (2011). Species-specific AFLP markers for identification of *Zingiber officinale, Z. montanum* and *Z. zerumbet* (Zingiberaceae). *Genetics and molecular research,* 10(1), 218-229. doi:10.4238/vol10-1gmr1154

Gloyn, A. L., & McCarthy, M. I. (2010). Variation across the allele frequency spectrum. *Nature Genetics,* 42, 648. doi:10.1038/ng0810-648

Goldman, N., & Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution,* 11(5), 725-736. doi:10.1093/oxfordjournals.molbev.a040153

Gramlich, S., Sagmeister, P., Dullinger, S., Hadacek, F., & Horandl, E. (2016). Evolution in situ: hybrid origin and establishment of willows (*Salix* L.) on alpine glacier forefields. *Heredity,* 116(6), 531-541. doi:10.1038/hdy.2016.14

Hamrick, J. L., & Godt, M. J. W. (1996). Effects of life history traits on genetic diversity in plant species. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences,* 351(1345), 1291-1298. doi:10.1098/rstb.1996.0112

Harvey, M. G., Smith, B. T., Glenn, T. C., Faircloth, B. C., & Brumfield, R. T. (2016). Sequence capture versus restriction site associated DNA sequencing for shallow systematics. *Systematic Biology,* 65(5), 910-924. doi:10.1093/sysbio/syw036

Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences,* 270(1512), 313-321. doi:10.1098/rspb.2002.2218

Heckenhauer, J., Barfuss, M. H. J., & Samuel, R. (2016). Universal multiplexable *matK* primers for DNA barcoding of Angiosperms. *Applications in Plant Sciences,* 4(6), 1500137-n/a. doi:10.3732/apps.1500137

Herrera, S., Reyes-Herrera, P. H., & Shank, T. M. (2015). Predicting RAD-seq marker numbers across the eukaryotic tree of life. *Genome Biology and Evolution,* 7(12), 3207-3225. doi:10.1093/gbe/evv210

Hollingsworth, P. M., Graham, S. W., & Little, D. P. (2011). Choosing and using a plant DNA barcode. *PLoS One,* 6(5), e19254. doi:10.1371/journal.pone.0019254

Hollingsworth, P. M., Li, D. Z., van der Bank, M., & Twyford, A. D. (2016). Telling plant species apart with DNA: from barcodes to genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 371(1702). doi:10.1098/rstb.2015.0338

Hua, Z., Jiang, C., Song, S., Tian, D., Chen, Z., Jin, Y., . . . Yuan, Y. (2022). Accurate identification of taxon-specific molecular markers in plants based on DNA signature sequence. *Molecular Ecology Resources*. doi:10.1111/1755-0998.13697

Johnson, L. A., & Soltis, D. E. (1994). matK DNA sequences and phylogenetic reconstruction in Saxifragaceae s. str. *Systematic Botany,* 19(1). doi:10.2307/2419718

Jones, L., Twyford, A. D., Ford, C. R., Rich, T. C. G., Davies, H., Forrest, L. L., . . . de Vere, N. (2021). Barcode UK: A complete DNA barcoding resource for the flowering plants and conifers of the United Kingdom. *Molecular Ecology Resources,* 21(6), 2050-2062. doi:10.1111/1755-0998.13388

Jordon-Thaden, I. E., Beck, J. B., Rushworth, C. A., Windham, M. D., Diaz, N., Cantley, J. T., . . . Rothfels, C. J. (2020). A basic ddRADseq two-enzyme protocol performs well with herbarium and silica-dried tissues across four genera. *Applications in Plant Sciences,* 8(4), e11344. doi:10.1002/aps3.11344

Kovarik, A., Pires, J. C., Leitch, A. R., Lim, K. Y., Sherwood, A. M., Matyasek, R., . . . Soltis, P. S. (2005). Rapid concerted evolution of nuclear ribosomal dna in two tragopogon

allopolyploids of recent and recurrent origin. *Genetics (Austin),* 169(2), 931-944. doi:10.1534/genetics.104.032839

Kozarewa, I., Armisen, J., Gardner, A. F., Slatko, B. E., & Hendrickson, C. L. (2015). Overview of target enrichment strategies. *Current Protocols in Molecular Biology,* 112, 72121-23. doi:10.1002/0471142727.mb0721s112

Kress, W. J., Erickson, D. L., Jones, F. A., Swenson, N. G., Perez, R., Sanjur, O., & Bermingham, E. (2009). Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. *Proceedings of the National Academy of Sciences - PNAS,* 106(44), 18621-18626. doi:10.1073/pnas.0909820106

Kubinova, Z., Janacek, J., Lhotakova, Z., Kubinova, L., & Albrechtova, J. (2014). Unbiased estimation of chloroplast number in mesophyll cells: advantage of a genuine three-dimensional approach. *Journal of Experimental Botany,* 65(2), 609-620. doi:10.1093/jxb/ert407

Kuo, L. Y., Li, F. W., Chiou, W. L., & Wang, C. N. (2011). First insights into fern *matK* phylogeny. *Molecular Phylogenetics and Evolution* 59(3), 556-566. doi:10.1016/j.ympev.2011.03.010

Leister, D. (2004). Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance gene. *Trends in Genetics,* 20(3), 116-122. doi:10.1016/j.tig.2004.01.007

Levin, D. A. (2020). Has the polyploid wave ebbed? *Frontiers in Plant Science,* 11, 251. doi:10.3389/fpls.2020.00251

Li, Y., Gao, L.-M., Poudel, R. C., Li, D.-Z., & Forrest, A. (2011). High universality of *matK* primers for barcoding gymnosperms. *Journal of Systematics and Evolution,* 49(3), 169-175. doi:10.1111/j.1759-6831.2011.00128.x

Li, Z., McKibben, M. T. W., Finch, G. S., Blischak, P. D., Sutherland, B. L., & Barker, M. S. (2021). Patterns and processes of diploidization in land plants. *Annual Review Plant Biology,* 72, 387-410. doi:10.1146/annurev-arplant-050718-100344

Lin, H. Y., Hao, Y. J., Li, J. H., Fu, C. X., Soltis, P. S., Soltis, D. E., & Zhao, Y. P. (2019). Phylogenomic conflict resulting from ancient introgression following species diversification in *Stewartia* s.l. (Theaceae). *Molecular Phylogenetics and Evolution,* 135, 1-11. doi:10.1016/j.ympev.2019.02.018

Liu, Y., Xu, C., Sun, Y., Chen, X., Dong, W., Yang, X., & Zhou, S. (2021). Method for quick DNA barcode reference library construction. *Ecology and Evolution* 11(17), 11627-11638. doi:10.1002/ece3.7788

Loera-Sanchez, M., Studer, B., & Kolliker, R. (2020). DNA barcode *trnH-psbA* is a promising candidate for efficient identification of forage legumes and grasses. *BMC Research Notes,* 13(1), 35. doi:10.1186/s13104-020-4897-5

Luo, X., Chen, S., & Zhang, Y. (2022). PlantRep: a database of plant repetitive elements. *Plant Cell Report,* 41(4), 1163-1166. doi:10.1007/s00299-021-02817-y

Magallon, S., & Sanderson, M. J. (2001). Absolute diversification rates in angiosperm clades. *Evolution,* 55(9), 1762-1780. doi:10.1111/j.0014-3820.2001.tb00826.x.

Marques, D. A., Meier, J. I., & Seehausen, O. (2019). A combinatorial view on speciation and adaptive radiation. *Trends Ecology and Evolution,* 34(6), 531-544. doi:10.1016/j.tree.2019.02.008

Meier, R., Shiyang, K., Vaidya, G., & Ng, P. K. (2006). DNA barcoding and taxonomy in *Diptera*: a tale of high intraspecific variability and low identification success. *Systematic Biology,* 55(5), 715-728. doi:10.1080/10635150600969864

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution,* 37(5), 1530-1534. doi:10.1093/molbev/msaa015

Mu, X. Y., Tong, L., Sun, M., Zhu, Y. X., Wen, J., Lin, Q. W., & Liu, B. (2020). Phylogeny and divergence time estimation of the walnut family (Juglandaceae) based on nuclear RAD-Seq and chloroplast genome data. *Molecular Phylogenetics and Evolution,* 147, 106802. doi:10.1016/j.ympev.2020.106802

Munoz-Lopez, M., & Garcia-Perez, J. L. (2010). DNA transposons: nature and applications in genomics. *Current Genomics,* 11(2), 115-128. doi:10.2174/138920210790886871

Nock, C. J., Waters, D. L. E., Edwards, M. A., Bowen, S. G., Rice, N., Cordeiro, G. M., & Henry, R. J. (2011). Chloroplast genome sequences from total DNA for plant identification: Chloroplast genome sequences for plant identification. *Plant Biotechnology Journal,* 9(3), 328-333. doi:10.1111/j.1467-7652.2010.00558.x

Pellicer, J., Hidalgo, O., Dodsworth, S., & Leitch, I. J. (2018). Genome size diversity and its impact on the evolution of land plants. *Genes (Basel),* 9(2). doi:10.3390/genes9020088

Percy, D. M., Argus, G. W., Cronk, Q. C., Fazekas, A. J., Kesanakurti, P. R., Burgess, K. S., . . . Graham, S. W. (2014). Understanding the spectacular failure of DNA barcoding in willows (*Salix*): does this result from a trans-specific selective sweep? *Molecular Ecology,* 23(19), 4737-4756. doi:10.1111/mec.12837

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One,* 7(5), e37135. doi:10.1371/journal.pone.0037135

Pillon, Y., Fay, M. F., Hedren, M., Bateman, R. M., Devey, D. S., Shipunov, A. B., . . . Chase, M. W. (2007). Evolution and temporal diversification of western European polyploid species complexes in *Dactylorhiza* (Orchidaceae). *Taxon,* 56(4), 1185-1208. doi:10.2307/25065911

Prokopowich, C. D., Gregory, T. R., & Crease, T. J. (2003). correlation between rDNA copy number and genome size in eukaryotes. *Genome,* 46(1), 48-50. doi:10.1139/g02-103

Puillandre, N., Lambert, A., Brouillet, S., & Achaz, G. (2012). ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology,* 21(8), 1864-1877. doi:10.1111/j.1365-294X.2011.05239.x

Rach, J., Desalle, R., Sarkar, I. N., Schierwater, B., & Hadrys, H. (2008). Character-based DNA barcoding allows discrimination of genera, species and populations in *Odonata*. *Proceedings of the Royal Society B: Biological Sciences,* 275(1632), 237-247. doi:10.1098/rspb.2007.1290

Ratnasingham, S., & Herbert, P. (2007). BOLD: The Barcode of Life Data system. *Molecular Ecology Notes*. doi:10.1111/j.1471-8286.2006.01678.x

Renaud, R. Y. L., Vincent, S., Sylvie, D., Olivier, M., & Michelle van der, B. (2008). A test of *psbK-psbI* and *atpF-atpH* as potential plant DNA barcodes using the flora of the Kruger National Park (South Africa) as a model system. *Nature Precedings.*

Rieseberg, L. H., Raymond, O., Rosenthal, D. M., Lai, Z., Livingstone, K., Nakazato, T., . . . Lexer, C. (2003). Major ecological transitions in wild sunflowers facilitated by hybridization. *Science,* 301(5637), 1211-1216. doi:10.1126/science.1086949

Roux, C., Fraisse, C., Romiguier, J., Anciaux, Y., Galtier, N., & Bierne, N. (2016). Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLOS Biology,* 14(12), e2000234. doi:10.1371/journal.pbio.2000234

Sarkar, I. N., Planet, P. J., Bael, T. E., Stanley, S. E., Siddall, M., DeSalle, R., & Figurski, D. H. (2002). Characteristic attributes in cancer microarrays. *Journal of Biomedical Informatics,* 35(2), 111-122. doi:10.1016/s1532-0464(02)00504-x

Satturu, V., Rani, D., Gattu, S., Md, J., Mulinti, S., Nagireddy, R., . . . Yanda, R. (2018). DNA fingerprinting for identification of rice varieties and seed genetic purity assessment. *Agricultural Research,* 7(4), 379-390. doi:10.1007/s40003-018-0324-8

Scharmann, M., Wistuba, A., & Widmer, A. (2021). Introgression is widespread in the radiation of carnivorous *Nepenthes* pitcher plants. *Molecular Phylogenetics and Evolution,* 163, 107214. doi:10.1016/j.ympev.2021.107214

Schrempf, D., Minh, B. Q., De Maio, N., von Haeseler, A., & Kosiol, C. (2016). Reversible polymorphism-aware phylogenetic models and their application to tree inference. *Journal of Theoretical Biology,* 407, 362-370. doi:10.1016/j.jtbi.2016.07.042

Schrempf, D., Minh, B. Q., von Haeseler, A., & Kosiol, C. (2019). Polymorphism-aware species trees with advanced mutation models, bootstrap, and rate heterogeneity. *Molecular Biology and Evolution,* 36(6), 1294-1301. doi:10.1093/molbev/msz043

Soltis, P. S., Marchant, D. B., Van de Peer, Y., & Soltis, D. E. (2015). Polyploidy and genome evolution in plants. *Current Opinion in Genetics and Development,* 35, 119-125. doi:10.1016/j.gde.2015.11.003

Soltis, P. S., & Soltis, D. E. (2009). The role of hybridization in plant speciation. *Annual Review Plant Biology,* 60, 561-588. doi:10.1146/annurev.arplant.043008.092039

Stallman, J. K., Funk, V. A., Price, J. P., & Knope, M. L. (2019). DNA barcodes fail to accurately differentiate species in Hawaiian plant lineages. *Botanical Journal of the Linnean Society,* 190(4), 374-388. doi:10.1093/botlinnean/boz024

Stankowski, S., & Streisfeld, M. A. (2015). Introgressive hybridization facilitates adaptive divergence in a recent radiation of monkeyflowers. *Proceedings of the Royal Society B: Biological Sciences,* 282(1814). doi:10.1098/rspb.2015.1666

Stephens, J. D., Rogers, W. L., Heyduk, K., Cruse-Sanders, J. M., Determann, R. O., Glenn, T. C., & Malmberg, R. L. (2015). Resolving phylogenetic relationships of the recently radiated carnivorous plant genus *Sarracenia* using target enrichment. *Molecular Phylogenetics and Evolution,* 85, 76-87. doi:10.1016/j.ympev.2015.01.015

Stephens, J. D., Rogers, W. L., Mason, C. M., Donovan, L. A., & Malmberg, R. L. (2015). Species tree estimation of diploid *Helianthus* (Asteraceae) using target enrichment. *American Journal of Botany,* 102(6), 910-920. doi:10.3732/ajb.1500031

Stull, G. W., Moore, M. J., Mandala, V. S., Douglas, N. A., Kates, H.-R., Qi, X., . . . Gitzendanner, M. A. (2013). A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes. *Applications in Plant Sciences,* 1(2), 1200497-n/a. doi:10.3732/apps.1200497

Stull, G. W., Soltis, P. S., Soltis, D. E., Gitzendanner, M. A., & Smith, S. A. (2020). Nuclear phylogenomic analyses of asterids conflict with plastome trees and support novel

relationships among major lineages. *American Journal of Botany,* 107(5), 790-805. doi:10.1002/ajb2.1468

Suzuki, J. Y., Matsumoto, T. K., Keith, L. M., & Myers, R. Y. (2014). The chloroplast *psbK-psbI* intergenic region, a potential genetic marker for broad sectional relationships in *Anthurium*. *HortScience,* 49(10), 1244-1252. doi:10.21273/hortsci.49.10.1244

Twyford, A. D., & Ness, R. W. (2017). Strategies for complete plastid genome sequencing. *Molecular Ecology Resources,* 17(5), 858-868. doi:10.1111/1755-0998.12626

Urbanelli, S., Della Rosa, V., Punelli, F., Porretta, D., Reverberi, M., Fabbri, A. A., & Fanelli, C. (2007). DNA-fingerprinting (AFLP and RFLP) for genotypic identification in species of the *Pleurotus eryngii* complex. *Applied microbiology and biotechnology,* 74(3), 592-600. doi:10.1007/s00253-006-0684-z

Villaverde, T., Pokorny, L., Olsson, S., Rincón-Barrado, M., Johnson, M. G., Gardner, E. M., . . . Sanmartín, I. (2018). Bridging the micro- and macroevolutionary levels in phylogenomics: Hyb-Seq solves relationships from populations to species and above. *New Phytologist,* 220(2), 636-650. doi:10.1111/nph.15312

Wang, X., Gussarova, G., Ruhsam, M., de Vere, N., Metherell, C., Hollingsworth, P. M., & Twyford, A. D. (2018). DNA barcoding a taxonomically complex hemiparasitic genus reveals deep divergence between ploidy levels but lack of species-level resolution. *AoB Plants,* 10(3), ply026. doi:10.1093/aobpla/ply026

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Review Genetics,* 10(1), 57-63. doi:10.1038/nrg2484

Weitemier, K., Straub, S. C., Cronn, R. C., Fishbein, M., Schmickl, R., McDonnell, A., & Liston, A. (2014). Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences,* 2(9). doi:10.3732/apps.1400042

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data,* 3, 160018. doi:10.1038/sdata.2016.18

Wright, K. M., Lloyd, D., Lowry, D. B., Macnair, M. R., & Willis, J. H. (2013). Indirect evolution of hybrid lethality due to linkage with selected locus in *Mimulus guttatus*. *PLOS Biology,* 11(2), e1001497. doi:10.1371/journal.pbio.1001497

Xu, S. Z., Li, Z. Y., & Jin, X. H. (2018). DNA barcoding of invasive plants in China: A resource for identifying invasive plants. *Molecular Ecology Resources,* 18(1), 128-136. doi:10.1111/1755-0998.12715

Yao, H., Song, J., Liu, C., Luo, K., Han, J., Li, Y., . . . Chen, S. (2010). Use of ITS2 region as the universal DNA barcode for plants and animals. *PLoS One,* 5(10). doi:10.1371/journal.pone.0013102

Yardeni, G., Viruel, J., Paris, M., Hess, J., Groot Crego, C., de La Harpe, M., . . . Leroy, T. (2021). Taxon-specific or universal? Using target capture to study the evolutionary history of rapid radiations. *Molecular Ecology Resources*. doi:10.1111/1755-0998.13523

Yu, J., Xue, J.-H., & Zhou, S.-L. (2011). New universal *matK* primers for DNA barcoding angiosperms. *Journal of Systematics and Evolution,* 49(3), 176-181. doi:10.1111/j.1759-6831.2011.00134.x

Zenil-Ferguson, R., Burleigh, J. G., Freyman, W. A., Igic, B., Mayrose, I., & Goldberg, E. E. (2019). Interaction among ploidy, breeding system and lineage diversification. *New Phytologist,* 224(3), 1252-1265. doi:10.1111/nph.16184

Chapter 3 Bioinformatics methods

3.1. Abstract

Acquiring sequence data generated for various purposes and repurposing it to evaluate the nature of genomic differences between plant species requires thorough dataset assessments, efficient data processing, and careful management. In this chapter, I summarise the methodology and bioinformatics pipeline I have developed for assessing the genomic nature of differences between plant species based on datasets mined from multi-repositories and shared by collaborators. The process ranges from searching for appropriate datasets, data collection and filtering, to calculating the ratio of monophyletic versus non-monophyletic species, extracting information on ancestry informative loci, and sub-sampling the data to evaluate the minimum amount of data to achieve maximal species discrimination. I provide examples of the data formats to guide future downloading and acquiring of the datasets, and the links to curated publicly available bioinformatic tools and self-written scripts to allow reproduction and reuse of this data processing pipeline. The approach I have developed is designed for Linux and Mac OS X, and the modules are self-functioning to allow them to be easily embedded in other analytical pipelines.

3.2. Introduction

There is now a significant number of studies that have sequenced multiple loci from the nuclear genomes of plants and these are available in public databases or in private data repositories (if the study hasn't yet been published). Collectively these datasets have great potential for understanding the nature of genomic differences between plant species. To harness this existing information I have developed a set of workflows, with the logic behind my approach being to

- Develop key criteria for selecting suitable datasets focusing only on datasets which have sampled multiple individuals from multiple congeneric species
- Search the literature and public data repositories, and contact collaborators for datasets that match these criteria
- Obtain and organise the selected datasets
- Use these datasets to estimate the proportion of species that resolve as monophyletic units as one measure of species discrimination success
- Establish the frequency distribution of taxonomically informative nucleotide substitutions among taxa to better understand the genomic nature of differences between plant species
- Subsample the data to better understand the effectiveness of smaller numbers of loci in recovering maximal species discrimination, as well as evaluating the attributes of the gene regions that are most effective at telling species apart.

The following sections summarise the data collection and management methods, and the main bioinformatic steps to evaluate the proportion of plant species that are distinguishable by nuclear DNA sequence data, to extract the species diagnostic signals from the sequence data available, and to assess how much data is needed to optimise the discrimination success and to better understand the pattern of sequence variation that makes species distinguishable.

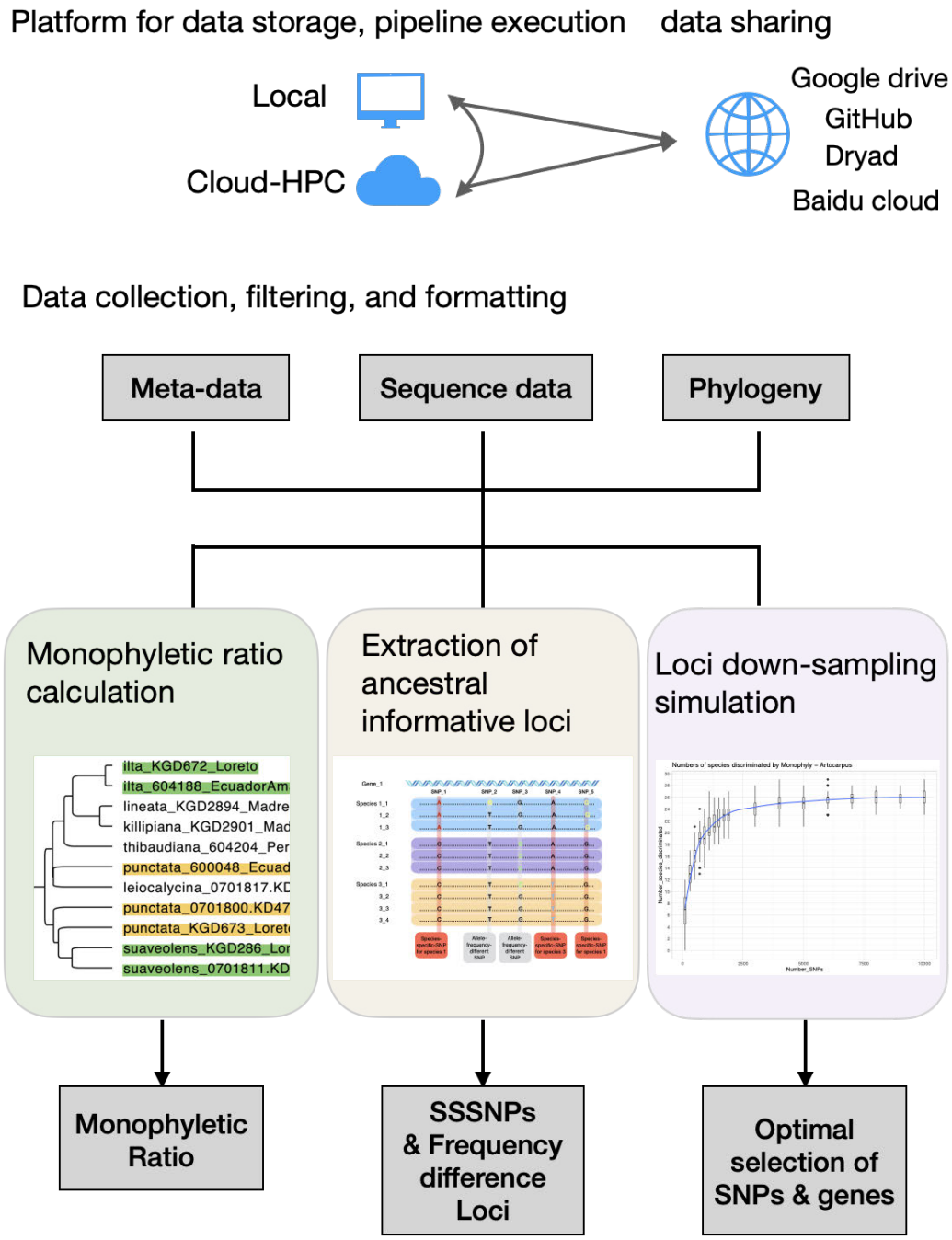


Figure 3.1. Overview of the data management and the bioinformatic pipeline NucBarcoder. The data storage and pipeline execution are performed on both a local computer and a High-Performance Computing system via cloud service. The original data from public and private providers is deposited and downloaded from Google Drive, Dryad, and Baidu cloud. The data underwent format tidying up and filtering locally and was then uploaded to HPC for intensive processing and analytical tasks. The Nucbarcoder pipeline comprises three main parts: Monophyletic ratio calculation, Extraction of ancestry informative loci, and loci down-sampling simulation (sub-sampling of the data). The scripts and software to execute the three steps are synced to GitHub regularly.

3.3. Data collection, filtering, and formatting

3.3.1. Selection of data sets

To compile data suitable for meta-analysis, I first searched journal publications from 2013 onwards for studies that sequenced multiple loci from the nuclear genome and which sampled multiple individuals of multiple congeneric species. The cut-off of 2013 was selected as this reflects the initiation of widespread use of next-generation sequence platforms for recovery of nuclear sequence data from plants. I used ambiguous matching patterns to search in the Web of Science and University of Edinburgh literature search engines. The matching patterns are listed in Table 3.1.

Table 3.1. Advance literature search keywords and matching patterns

“phy ogen*” AND “RAD*" AND “p ants*”
“phy ogen*” AND “GBS" AND “p ants*”
“phy ogen*” AND “genome sk m*" AND “p ants*”
“phy ogen*” AND “transcr ptome*" AND “p ants*”
“phy ogen*” AND “target capture" OR "Hyb-seq" AND “p ants*”
“phy ogen*” AND “WGS" AND “p ants*”
“*RAD*" AND “p ants*” AND "genus"
“GBS" AND “p ants*”AND "genus"
“genome sk m*" AND “p ants*”AND "genus"
“transcr ptome*" AND “p ants*”AND "genus"
“target capture" OR "Hyb-seq" AND “p ants*”AND "genus"
“WGS" AND “p ants*”AND "genus"

Note: a of the advance search patterns cou d be merged nto one query techn ca y (comb n ng by OR and AND) but the search eng nes g ve not as comprehens ve a resu t as search ng separate y

With the full-text publications downloaded, the next step was selecting publications manually. I only kept studies if they satisfied the following criteria.

1) Three or more un-linked nuclear loci were sequenced.
2) More than two species had multiple individuals successfully sequenced and retained.
3) A phylogeny is available either in visual format or in a machine readable text format (newick, phylip, or nexus formats).
4) The species identities on the phylogeny could be interpreted and related to the sampled species.

In addition to mining the published literature, I contacted potential collaborators to request access to unpublished datasets. This involved designing a data request form

(supplementary file S.3.3) to request standardised metadata and agree a data-sharing scheme. The shared data were uploaded to Google Drive (https://www.google.co.uk/intl/en-GB/drive/), and then downloaded and managed on the UK crop diversity bioinformatics HPC platform (https://www.cropdiversity.ac.uk).

Three important files were requested from our collaborators or extracted from online data repositories:

1) Metadata. This includes information that corresponds to the sample IDs in the consensus sequence files which links sequences of individuals to their scientific names (species identities). This information is also useful for tracking changes of names wherever this has been applied, as well as monitoring and understanding inclusion and exclusion of individuals and loci.

Sample ID	Species name
FG_186	*Inga_poeppigiana*
KD_13	*Inga_poeppigiana*
LA_2023	*Inga_poeppigiana*
M46A	*Inga_poeppigiana*

Figure 3.2. Sample IDs corresponding to species' names.

2) Sequence alignment file. This alignment covers the sequence variation in all individuals per dataset, and depending on how the original raw reads were processed, takes the form of either multiple-aligned sequences in .fasta, .phylip, or .nex formats, or SNP matrix in .vcf or .fasta format.

a) fasta format

```
>FG_186
CATTGTTCCTCCATAACACACAGATTTTGGCCGGA
>KD_13
CATTGTTCCTCCATAACANACAAATTTTGGCCGGA
>LA_2023
CATTGTTCCTCCATAACACACAAATTTTGGCCGGA
>M46A
CATTGTTCCTCCATAACACACANATTTTGGCCGGA
```

b) phylip format

```
Samp e  ID  103     CATACATCTTCAGCACTACAGNTATCT
Samp e  ID  1300    CATACATCTTCAGCACTACAGTTATCT
Samp e  ID  32      CATACATCTTCAGCACTACAGTTATCT
Samp e  ID  8577    CATACATCTTCAGCACTACAGTTATCT
Samp e  ID  8578    CATACATCTTCAGCACTACAGTTATCT
Samp e  ID  8579    CATACATCTTCAGCACTACAGTTATCT
```

c) nexus format

```
beg n characters;
        d mens ons nchar=31;
        format datatype=dna m ss ng=? gap=-;
        matr x
        Samp e  ID  1 CCATGACTTGATTAGCATCTGTCAAAATCCC
        Samp e  ID  2 CCATGACTTGATTAGCACCTGTCAAAATCCC
        Samp e  ID  3 CCATGACTTGATTAGCACCTGTCAAAATCCC
        Samp e  ID  4 CCATGACTTGATTAGCACCTGTCAAAATCCC
;
end;
```

d) vcf format

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	samp e 1 sorted bam
Chr1	31150		G	A	999			GT:PL:DP:SP:ADF:ADR:AD	/ :0 0 0:0:0:0 0:0 0:0 0
Chr1	31174		A	T	70			GT:PL:DP:SP:ADF:ADR:AD	/ :0 0 0:0:0:0 0:0 0:0 0
Chr1	252833		G	T A	243			GT:PL:DP:SP:ADF:ADR:AD	0/0:0 9 98 9 98 98:3:0:3 0 0:0 0 0:3 0 0

Figure 3.3. Examples of aligned sequences in a) fasta, b) phylip, and c) nex formats, and SNP matrix in d) vcf format.

3) Phylogenetic trees. Phylogenetic trees were recovered from each study to enable easy estimation of the proportion of species that resolve as monophyletic. The preferred format is for this to be a machine readable text format such as the newick format. However, where only a graphical representation of the tree was available, this was also retained and used, to maximise the number of studies analysed.

a) the phylogenies in newick format

```
((FG  186:0 0004627773 M46A:0 0003625627)100:0 0002685118 (KD  13:0 0003441657 LA  2023:0 0003470779)1
00:0 0003440200)100:0 0013278046)100:0 0003416399
```

b) visualized phylogeny of a)

Figure 3.4. An exemplar phylogeny in a) newick format and b) visualisation. Four samples and the lengths of the branches are included.

The resulting datasets are available on GitHub: https://github.com/Hazelhuangup/Species_specific_alleles_analysis/tree/main/test_file.

3.3.2. Data cleaning and filtering

Data cleaning and filtering were performed at both the individual and locus levels. Upon acquiring each dataset, I examined how taxon re-identifications were performed and any ambiguous samples were removed based on information given in the metadata of collaborators and in publications. Any individuals with unresolved species names (including cf., aff., and other uncategorised naming methods) and hybrids were removed from the matrix. Individuals that were outside of the focal genus of each dataset were removed except for the purpose of acting as outgroups in phylogenetic analyses. In the test run, subspecies were both treated as separate entities and as one species. To achieve conform diagnostic signals at species level, multiple varieties or subspecies in one species were treated as a single taxon at the species level for the main analysis. Individuals with a high proportion of missing data were also removed, the threshold ranges from 75 ~ 80% data presence according to the data quality after manual checking. The threshold is determined by the data quality for the specific genus. If the general data quality is good, samples were removed at a lower missing data threshold. Further details on the removal of individuals and why they were removed from each genus are provided in Table 3.2.

After the filtering of individuals, I undertook locus filtering. For target capture or transcriptome data where genomic segment information is available, loci that were missing from 80% of all individuals were deleted. This also applies to RAD-seq/GBS data where stack or assembly information on each locus was provided. For datasets without genomic segment information, such as vcf and concatenated consensus fasta files, a nucleotide site was removed when over 40% of the individuals are missing. This threshold is determined by the number of individuals with valid data that enables informative phylogenetic relationships (a quartet). When the depth and quality information was accessible, usually in vcf format, nucleotide sites were retained only when the depth is within a reasonable range (lower depth limit of 2 to an upper limit of 2 to 3 times of the average sequencing depth) according to the sequencing depth reported by the collaborators. Higher coverage were treated as repetitive regions and so these regions were removed.

Different clean-up and filtering tools were used for aligned sequence formats and SNP matrices. VCFtools (0.1.17) (Danecek et al., 2011) was used to deal with the vcf format and self-written scripts were used for parsing the aligned sequences files, including format conversion, removing individuals, and removing sites. The main steps are listed in Table 3.2.

Table 3.2. Tools or scripts used for data clean up and filtering

Data format	Too s or scr pts	Usefu parameters and the r funct ons	Usage and descr pt on
vcf	vcftoo s	--remove --m n-meanDP --max-meanDP --m nDP --max-m ss ng --remove- nde s --m n-a e es	Spec fy nd v dua s to remove from the dataset; Remove s tes by read depth and m ss ng data rate; Remove a nsert ons and de et ons; Spec fy to nc ude on y b - a e c s tes
fasta phy p nex	rm s te h gn N rate py	defau t	Remove s tes based on m ss ng data rate >40% (adaptab e);
	fa2phy py nex2fa py phy2fa one ne py formatt ng pyrad a e es2fasta py	defau t	Convert the f e format for spec f c usage
	g ve me mu seq py	--seqf e --seq st	Run the scr pts w th the command g ve me mu seq py -f nput fasta - IDs to keep -o output fasta

3.4. Monophyletic Ratio

The Monophyletic Ratio (MR) was calculated as a simple measure of species discrimination success from different data sets. This is defined as the number of species resolved as monophyletic on a given phylogeny divided by the total number of multiple-sampled species (individuals ≥2 samples per species) from a genus.

3.4.1. Monophyletic Ratio calculation requirements

Monophyletic clades were first counted from published phylogenies based on the data and format available. This involved either manual scoring of species resolving as monophyletic when only graphic representations of trees were available, and automated extraction of the MR for machine readable data (see section 3.4.3). Where phylogenies were not available for a given study, then I built a basic phylogeny using the aligned sequence data. Where sequence alignments were not available, I discarded these datasets due to the time constraints of producing high quality alignments from multiple different sources.

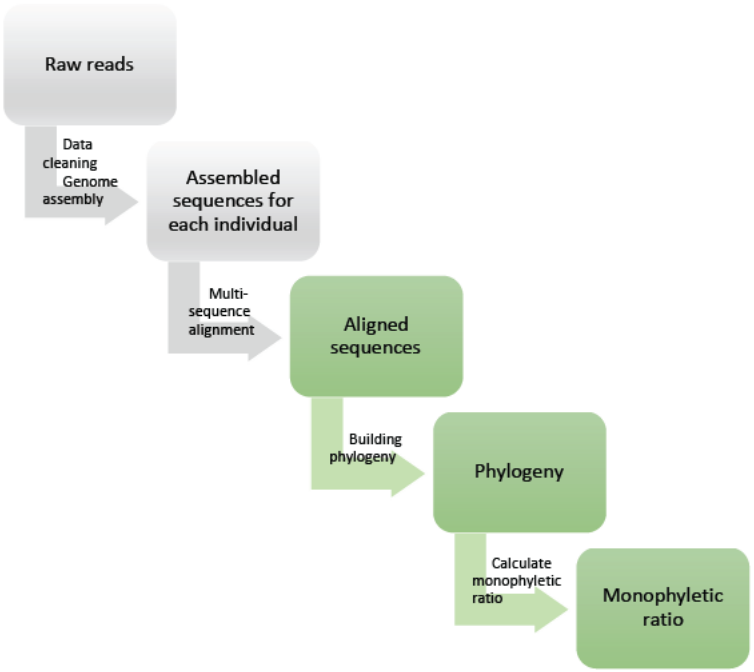


Figure 3.5. Steps to a successful monophyletic ratio calculation. Green colour blocks indicate where I accepted data, and grey blocks shows the stages where I discarded datasets if these were the only data available.

3.4.2. Building phylogenies

I built phylogenies with the aligned sequences as input using IQTREE2 (Minh et al., 2020). IQTREE (Nguyen et al., 2015) is a widely used software package for phylogenetic inference using maximum likelihood. The second version incorporates Polymorphism-aware phylogenetic models (PoMo) to parse the IUPAC Ambiguity Codes (Schrempf et al., 2016), which is useful when the dataset maintains heterozygosity information. I also built quick UPGMA trees using R package ape v.5.0 (Paradis et al., 2019) and phangorn (Schliep, 2011). Multiple software and parameters were tested on five early accessed datasets. For example, multiple nucleotide

substitution models were tested using both IQTREE2 and RAxML NG. Taking the availability of constant sites into consideration, ascertainment bias correction was also tested for SNP-only datasets. For datasets which retain the alternative allele information, the impact of keep ambiguity code was also assessed. The details could be found in table S6. The different substitution models do not or only slightly change the monophyletic ratio. So the later analyses use the default models for the phylogeny building software used.

Table 3.3. Tools or scripts used for building phylogeny

Data format	Too s or scr pts	Parameters	Usage and descr pt on
a gned fasta	IQTREE2	--seqtype -B -m -T -o	Spec fy the type of data as nput e g DNA Spec fy the rep cates for u trafast bootstrap e g 1000 spec fy the subst tut on mode e g HKY spec fy the number of threads for para e comput ng spec fy the dent f er of the outgroup
	ape dna d st R ape d st tree R	defau t	Run the scr pts w th the command Rscr pt ape dna d st R Input fasta Output d st and Rscr pt ape d st tree R Input d st Output tree

The test data for evaluating the tree-building pipeline consisted of 393 individuals with 1,313,489 base pairs per individual. To build a tree using this dataset, IQTREE2 requires a minimum of 7 Gb RAM and 131 hours CPU time. With 64 CPU for parallel computing, it took a total of 2 hours and 34 minutes to finish the task. Building quick trees with ape requires less computing power because it doesn't have a model testing and bootstrap stage. Running the whole dataset to build a UPGMA tree took 34 minutes CPU time.

3.4.3. Calculating the monophyletic ratio

If a phylogeny is in visualized format only, the identification could only be done by manual calculation. Figure 3.6. demonstrates an example where 3 species have multiple sampled individuals, 2 of them are monophyletic on the phylogeny. In this case, the monophyletic ratio is 2/3 = 0.67.

Figure 3.6. Phylogeny of the 11 individuals from the genus *Inga* featuring monophyletic and non-monophyletic taxa. The green highlights individuals from species *Inga ilta* and *I. suaveoolens* which form monophyletic clades, and yellow highlights individuals from *I. puctata* that form a non-monophyletic clade.

With phylogenies provided in newick or other text-based formats, this process could be automated using MonoPhy (Schwery et al., 2016), which is a quick and user-friendly method for assessing the monophyly of taxa in a given phylogeny. MonoPhy builds on the existing packages ape 5.0, phytools (Revell, 2012), phangorn, RColorBrewer (https://CRAN.R-project.org/package=RColorBrewer) and taxize (Chamberlain et al., 2013), and the installation of these packages is also required.

Table 3.4. Tools or scripts used for identifying monophyletic clades

Data format	Too s or scr pts	Parameters	Usage and package requ rements
rooted new ck tree ID correspond ng f e	MonoPhy	defau t	Rscr pt MonoPhy r Input treef e outgroup ID output ape phytoo s phangorn Rco orBrewer and tax ze packages requ red

The output of MonoPhy indicates taxa are either monophyletic, non-monophyletic or monotypic. By counting the number of species that are scored “Yes” for ‘monophyly’, divided by the number of species with ‘tips’ > 1 (= species with multiple-samples), gives the Monophyletic Ratio. For the example shown in Table 3.5, the monophyletic ratio is 6/7 = 0.86. A tutorial on running MonoPhy and how to interpret the results is given at https://cran.r-project.org/web/packages/MonoPhy/vignettes/MonoPhyVignette.html.

Table 3.5. An example of the output of MonoPhy

Spec es name	Monophy y	T ps
Inga acreana	Yes	8
Inga acrocephala	Yes	2
Inga acuminata	No	2
Inga alata	Yes	7
Inga alba	Yes	5
Inga auristellae	Yes	8
Inga bourgonii	Yes	7

Running the main analysis command ‘AssessMonophyly’ on 393 individuals using standard settings is very rapid, and used only 0.19 seconds on a MacBook Pro with 2.2 GHz Intel 6-Core i7 and 16 GB RAM. MonoPhy cannot be run in parallel by default.

3.5. The extraction of taxonomically informative Loci (Species-specific SNPs and Allele-frequency-different SNPs)

To provide a quantification of the distribution of taxonomically informative polymorphism, I extracted information on the frequency distribution of ancestry informative loci that were either fixed (Species-specific) or with marked nucleotide frequency differences.

Specifically, I divided SNPs into two categories: a) species-specific-SNPs which are fixed in all individuals from one species and distinct from all other ingroup species, and b) Frequency different SNPs where SNPs show a significant allele frequency difference in one species in comparison to other groups.

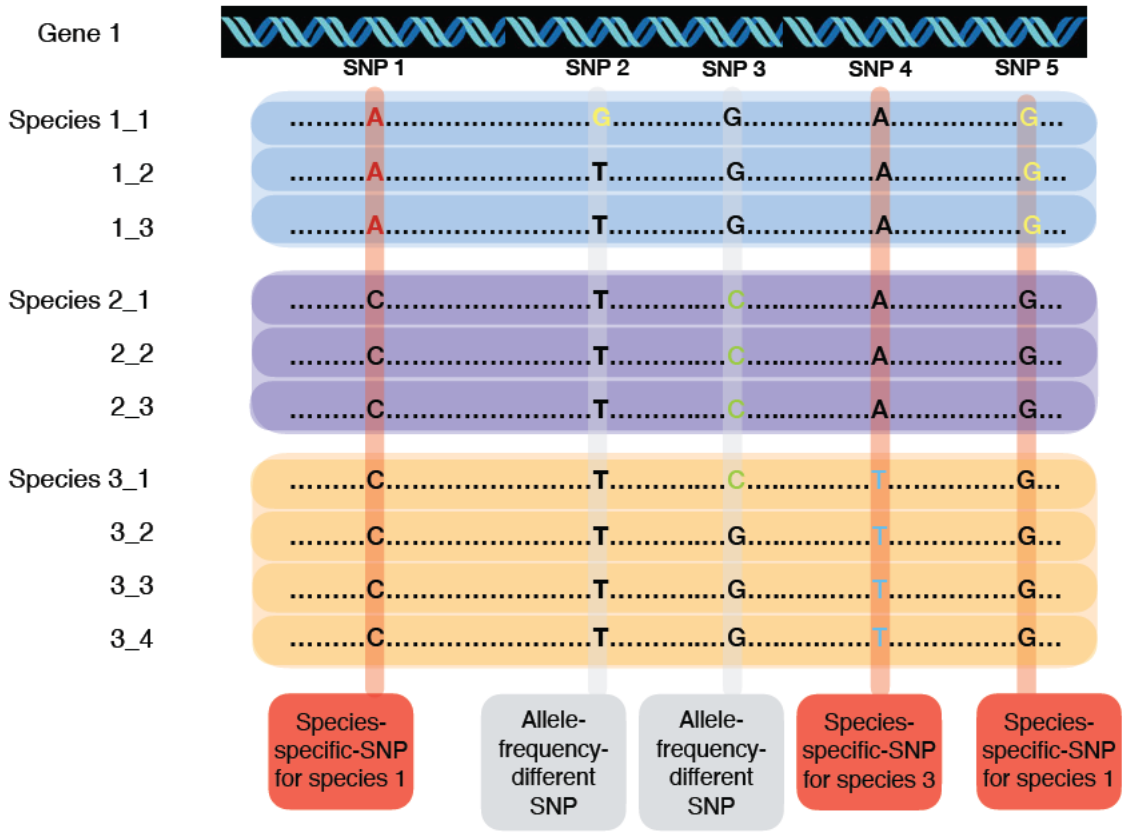


Figure 3.7. Diagram of two categories of polymorphic sites. SNPs (1,4,5) that are highlighted in red are species-specific, SNPs (2,3) that are in grey are SNPs with allele frequency differences.

To extract both types of SNPs from the aligned sequenced files, a python script extract_ancestral_informative_SNPs*.py was developed with the following steps:

1) For each locus, there should be at least 2 individuals for the target species and 4 individuals from another two species whose locus coverages are greater than one.
2) Calculate allele frequency for the target taxon. If the allele frequency (AF) of the major allele (M_A) is higher than 87.5%, progress to stage 3).

3) Calculate allele frequency for the M_A across all other taxa (aggregated). If less than the 10% threshold then proceed to 4).

4) Calculate allele frequency across other species with multi-sampled populations. If the number of individuals in this group is greater than 4 (included), and A_M is lower than 12.5% (at most only 1/8 allele belongs to the minor one), then retain the site. Otherwise, only if the minor alleles are homogeneous in a taxon would this site be retained.

The pipeline also includes following extra calculations steps:

5) Count the number of SSSNPs for each species that has multiple sampled individuals (e.g. the number of SNPs that are fixed and different compared to all other sampled species).

6) Calculate the density of SSSNPs by dividing the number of SSSNPs by the average total nucleotides that do not contain missing data in that species.

Scripts for extracting ancestry informative SNPs are available in two versions, one is for the situation where the input is in vcf format (extract_ancestral_informative_SNPs_vcf.py), and one is for fasta, phylip, and nexus format (extract_ancestral_informative_SNPs_hete.py). Both of these two versions are heterozygous aware, and can accommodate heterozygosity in the designation of SSSNP calling.

Table 3.6. Tools or scripts used for extracting ancestral informative SNPs

Data format	Too s or scr pts	Parameters	Descr pt on
vcf	extract ancestra nfo rmat ve SNPs vcf py	-f --name --se ectedspps --se ectedsamp e	Spec fy the nput f e n vcf/fasta format The vcf f e need GT and DP co umns Spec fy the samp e ID to spec es names correspond ng f e G ve a st of spps names that are mu t p e samp ed G ve a st of samp e names that be ongs to target genus (ID)
fasta phy p and nexus	extract ancestra nfo rmat ve SNPs hete py		

In very large data sets consisting of many hundreds of thousands of nucleotides, one might expect to encounter shared SNPs between any group of samples, purely due to random mutations. To distinguish biological signal of SSSNPs from random sampling artefacts, I developed an approach to test whether the number of SSSNPs is greater than expected due to chance alone based on a random distribution of the data. This was implemented by randomising the label of the sequences using the shuf function in bash command lines and repeating the same analysis counting the distribution of SSSNPs. I then compared the number of SSSNPs extracted from the original datasets versus the randomised label datasets using Wilcoxon signed-rank test (Knapp, 2017) with wilcox.test in R (parameters: paired = FALSE, alternative = “greater”).

To visualise and compare the distribution of SSSNPs across all species in available datasets, I produced a script plotting the density distribution of SSSNPs from all multiple sampled species annotated with whether a given species resolved as monophyletic or not, using ggplot2 embedded in the script Number_of_SSSNPs_vs_Monophyly.R.

Computing requirements for extracting species-specific SNPs and allele-frequency-different SNPs depended largely on the dataset size, i.e. the number of sites to parse, and to some extent on the number of multiple-sampled species. Table 3.7. illustrates the run times for data sets involved in this study.

Table 3.7. The running time for extracting ancestral informative SNPs from different datasets

Dataset	Data s ze (Mb)	Input format	Number of mu t p e-samp es spec es	Runn ng t me
Brownea	1 3	vcf	11	42s
Euphrasia	1 8	vcf	4	22s
Commiphora	1 9	fasta	22	2m n12s
Syagrus	3 5	vcf	17	3m n
Linanthus	4 4	fasta	20	2m n37s
Bee orchid	6	vcf	4	1m n40s
Cornus	9 5	fasta	18	4m n32s
Polemonium	9 9	fasta	12	3m n1s
Aesculus	17	fasta	15	7m n12s
Tsuga	38	fasta	8	8m n35s
Quercus	52	fasta	7	10m n55s
Vitis	76	vcf	8	38m n10s
Antirrhinum	119	vcf	21	3h59m n20s
Capurodendron	148	fasta	20	51m n12s
Artocarpus	157	fasta	42	2h58m ns
Salix	190	vcf	23	2h3m n40s
Linaria	235	fasta	13	54m n12s
Inga	576	fasta	69	16h8m n30s
Geonoma	969	fasta	44	16h12m n18s

3.6. Genomic region down-sampling and the impact on species resolving power

3.6.1 Performance of each locus in telling species apart

To assess the range of variation in species discrimination power among loci within datasets, I compared the discrimination success of individual loci/genes. This step is important in identifying the attributes of loci that are most informative in species discrimination which could help inform the choice of markers for future use. To do this, I built phylogenetic trees based on each single locus and looked at how many species resolved as monophyletic (compared to the equivalent figure using the total dataset). Given the scale of this task, tree-building was restricted to building quick UPGMA trees using ape_dna_dist.R and ape_dist_tree.R, and then quantifying the species resolved as monophyletic using MonoPhy. Descriptions of these scripts can be found in Table 3.3 and Table 3.4.

To assess the relationship between nucleotide diversity of individual loci and their performance in telling species apart, I calculated the average nucleotide diversity of each single locus among all individuals in a genus using nuc.div function in R package pegas (Paradis, 2010) (script calculate_pi.R). I then plotted the number of monophyletic specie resolved by this locus, against the nucleotide diversity, and the density of SSSNPs at this locus, using ggplot2 scripted in NucDiv_DensSSSNPs_No_spps_mono_by_each_gene.R. Regression curves were then fitted for both the series of nucleotide diversity and the density of SSSNPs against the number of species that are resolved as monophyletic using the CORREL function in Excel.

3.6.2. Species discrimination success with down-sampled sequence data

To further explore how much data is needed to tell species apart, I subsampled each dataset to reduce the amount of sequence information and assess the consequent impact on the species discrimination success.

Specifically, I developed a pipeline (DS_snp.sh, available on *GitHub page) to down sample the data at laddered intervals. To model a reduced amount of sequence information, I tested down-sampling in terms of a) the number of SNPs, and b) the number of DNA segments (genes, exon, or assembled RAD-tags) where this information is available. The steps in both regards are the same:

1) Select a specified number of SNPs/DNA segments randomly using python script.
2) Build a fast UPGMA tree using R package ape and phangorn.
3) Identify monophyletic clades using MonoPhy and count the number of monophyletic clades with python.
4) Repeat step 1-3 for 50 times (bootstrap = 50) which is enough to give a clear distribution.
5) Change the number of SNPs/DNA segments specified and repeat steps 1-4.
6) Visualize the result using ggplot2 (Wickham, 2011).

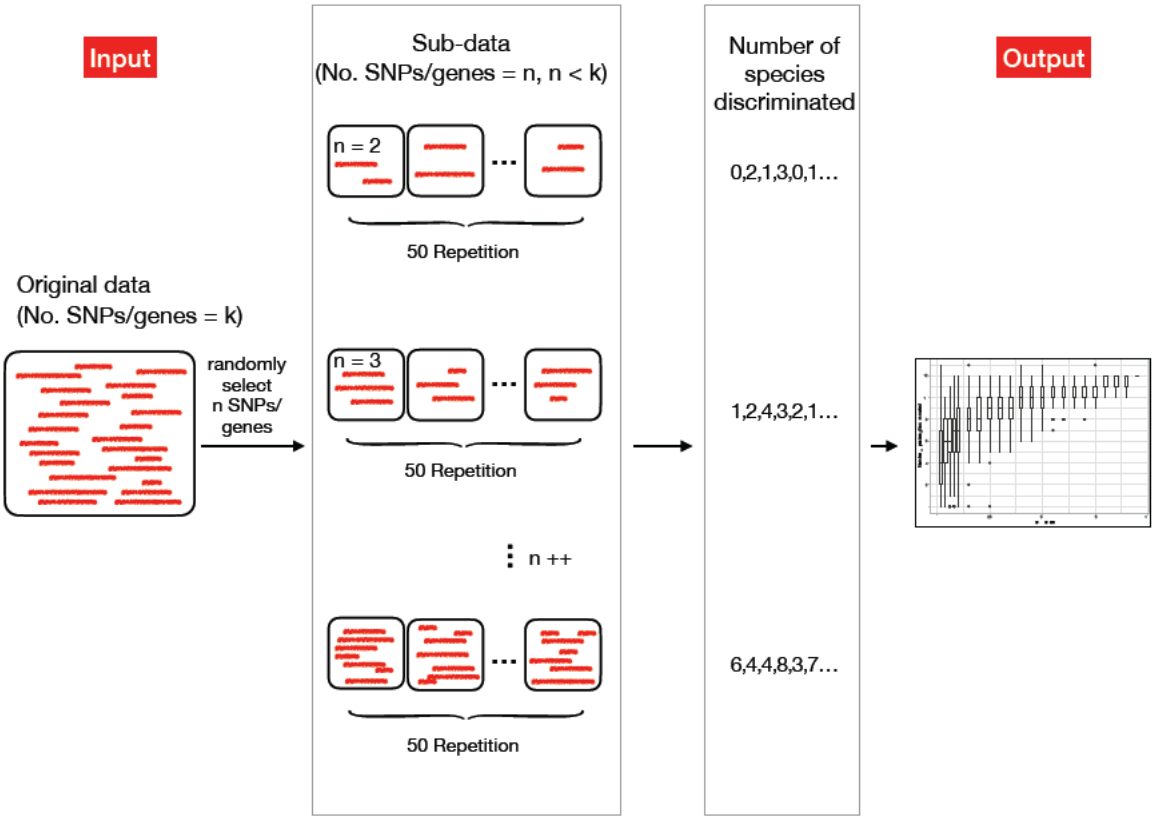


Figure 3.8. Diagram for the main steps of down-sampling simulation. For both SNP and segment-based down-sampling, the input is the original comprehensive collection of SNPs/DNA segments. I then randomly selected a specified number of SNPs/DNA segments starting from 100 SNPs/10 segments, and at each laddered step I increased the dataset size by randomly selecting a further 100 SNPs/10 segments, and at each step I recorded the number of species being told apart. The random sampling was repeated 50 times at each step. The resulting data were used to plot the distribution of the number of species being told apart with a given number of randomly selected SNPs/segments. The x-axis of the output figure is the number of SNPs/DNA segments from 0 to the maximum the dataset allows, and the y-axis is the distribution of the number of species being told apart based on 50 random samples of SNPs/DNA segments.

For datasets in which the performance of individual genes was analysed (see section 3.6.1), I conducted a further analysis focusing on the best performing genes. Here I started with the genes which showed the maximum amount of species discrimination, and sequentially added one gene at a time, selecting at each stage the next best performing locus. At each stage, I recorded the number of genes used as well as the number of species that resolved as monophyletic. This analysis complements the preceding random selection of loci, by instead assessing the minimal number of best-performing loci required to achieve the maximum amount of species discrimination from a given data set.

3.7. Code for Data Processing, Storing, and Plotting

All data used in this study are stored and run at the UK Crop Diversity platform (www.cropdiversity.ac.uk).

Python (version 3+) and bash shell scripts were used to process the data.

The R Statistical Programing Language (version 4.1.0) and the Rstudio integrated development environment (Racine, 2012) were used for most of the plotting along with Excel (version 16.58) charts.

All code scripts, workflow, and an example to run the pipeline are available at https://github.com/Hazelhuangup/Species_specific_alleles_analysis.

3.8. Reference

Chamberlain, S. A., & Szöcs, E. (2013). taxize: taxonomic search and retrieval in R. *F1000 research,* 2, 191. doi:10.12688/f1000research.2-191.v1

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics,* 27(15), 2156-2158. doi:10.1093/bioinformatics/btr330

Eaton, D. A., & Ree, R. H. (2013). Inferring phylogeny and introgression using RADseq data: an example from flowering plants (Pedicularis: Orobanchaceae). *Systematic Biology,* 62(5), 689-706. doi:10.1093/sysbio/syt032

Knapp, H. (2017). Wilcoxon test. United Kingdom: SAGE Publications Ltd.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution,* 37(5), 1530-1534. doi:10.1093/molbev/msaa015

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution,* 32(1), 268-274. doi:10.1093/molbev/msu300

Paradis, E. (2010). pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics,* 26(3), 419-420. doi:10.1093/bioinformatics/btp696

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics,* 35(3), 526-528. doi:10.1093/bioinformatics/bty633

Racine, J. S. (2012). RStudio: A platform-independent IDE for R and Sweave. *Journal of Applied Econometrics,* 27(1), 167-172. doi:10.1002/jae.1278

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology. *Methods in Ecology and Evolution,* 3(2), 217-223. doi:10.1111/j.2041-210X.2011.00169.x

Schliep, K. P. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics,* 27(4), 592-593. doi:10.1093/bioinformatics/btq706

Schrempf, D., Minh, B. Q., De Maio, N., von Haeseler, A., & Kosiol, C. (2016). Reversible polymorphism-aware phylogenetic models and their application to tree inference. *Journal of Theoretical Biology,* 407, 362-370. doi:10.1016/j.jtbi.2016.07.042

Schwery, O., & O'Meara, B. C. (2016). MonoPhy: a simple R package to find and visualize monophyly issues. *PeerJ Computer Science,* 2. doi:10.7717/peerj-cs.56

Web of science. In *WOS*. Philadelphia, Pa: Institute for Scientific Information.

Wickham, H. (2011). ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics,* 3(2), 180-185. doi:10.1002/wics.147

Chapter 4 Characterising the Genetic Bases of Species Differences in the Genus *Inga*

4.1. Abstract

To establish a test case to evaluate the genomic basis of species differences, I have explored the levels of species discrimination, and the underlying signals behind this, in the neotropical genus *Inga*. This dataset consists of target capture sequences from 453 individuals from 133 *Inga* species and an individual from *Zygia mediana* as the outgroup. Among all the species, 69 of them are represented by multiple-sampled individuals. The target capture panel involves sequence data from 810 genes. Following file formatting, I then carried out a unified set of analyses to address the questions 1) what is the proportion of *Inga* species distinguishable with nuclear markers? 2) what is the nature of the inter-specific differences and what are the attributes of loci that are the most informative in telling species apart? And 3) how many markers are needed and what markers are needed to maximise the species identification success? Of the 69 species with multiple individuals sampled, 45 resolved as monophyletic (65%). The density of species-specific SNPs for each *Inga* species ranged from 0 to 1,503 per megabase. Compared to the full dataset of 810 genes and 205,871 SNPs, subsampling analysis revealed that a random selection of 70 genes or 2500 SNPs, or a combination of 9 'best performing' genes could achieve levels of species discrimination success similar to the full dataset. I found a positive correlation ($r = 0.42$) between the number of species distinguished and the nucleotide diversity of the genes used for species discrimination.

4.2. Introduction

Inga is a species-rich (>300 species) neotropical tree genus with a crown age of 2–10 myr (Pennington et al., 1997). The main distribution of this genus is in central and south America centred around Amazonia. *Inga* lineages have radiated rapidly and the genus consists of many closely-related species showing mixed morphological characters (Lavin, 2006). The complexity of morphological characters thus poses many difficulties in morphological identifications, and a recent study showed a 40% error rate of specimen identification in *Inga* using leaf and fruit morphology (Baker et al., 2017). An early DNA barcoding study of *Inga* using seven barcoding loci (*rpoC1, rpoB, rbcL, matK, trnH-psbA, atpF-atpH, psbK-psbI*) resulted in low levels of species discrimination, with many species sharing identical barcodes, and the few multiple sampled species routinely failing to resolve as monophyletic (Hollingsworth et al., 2009). A more recent study sequenced over 6 kb from seven plastid regions and nuclear ITS sequence for 210 individuals from 124 species, and here the whole dataset give around 41% species resolution on phylogeny (Dexter et al., 2017). Chemocoding has been evaluated as an alternative method but does not solve the issue completely (Endara et al., 2018). There is thus a pressing need to explore new approaches to assist *Inga* species identification.

One potential approach for improving levels of species discrimination in *Inga* is to utilise a large-scale targeted enrichment array of nuclear genes that has been designed from transcriptome data from three *Inga* species (Nicholls et al., 2015). Since the publication of this array, a large sample set has been assembled and sequenced by Pennington et al. (Royal Botanic Garden Edinburgh, unpublished data). This dataset includes several hundred nuclear genes targeted from dozens of species with multiple individuals sampled per species, and it includes a low rate of missing data, and careful bioinformatics analysis to produce the consensus sequences for each individual. This dataset from *Inga* serves as a perfect test case for exploring patterns of species discrimination, and road-testing analytical pipelines prior to a more comprehensive meta-analysis.

The aims of this chapter are to use this exemplar dataset from *Inga* to evaluate the proportion of plant species that are distinguishable by nuclear DNA markers, and to evaluate the minimal amount of data needed to tell the maximum number of species apart, and also to explore the underlying signals behind species identification success.

To achieve these aims, I calculated the monophyletic ratio (MR) of *Inga* species in the phylogeny. I then examined the density and frequency distribution of species-specific SNPs (SSSNPs) and compared the density of SSSNPs to species that resolved as monophyletic. I then subsampled the data to evaluate the amount of data required to achieve maximum species discrimination. Finally, I examined the attributes of genes that were particularly useful at species identification, and assessed whether identification success is correlated with a particular level of sequence diversity.

4.3. Materials and Methods

4.3.1. Assembling the target capture dataset for *Inga*

The *Inga* target capture dataset for this study was provided by the *Inga* Working Group in the spring of 2019, and the essential information is presented in Table 4.1. The number of loci retained is that after screening out loci that had poor coverage or that failed the quality tests that are outlined in the original paper (Nicholls et al., 2015).

Table 4.1. Characteristics of the *Inga* target capture dataset

Genus	*Inga*
Number of accepted spec es n th s genus	380
Number of samp ed spec es (nc udes spec es n rev ew)	133
Number of mu t p e samp ed spec es	69
Sequenc ng method	Target Capture
Number of genes targeted	810
Average gene engt (bp)	1622
Tota engt of the a gned sequences (bp)	1 313 489

Three primary data files were produced. Firstly, a name tracking file that consists of a table with the accession identifier, the species identifier, (and optionally, and not used in this case, a sequence label if it is different to the accession identifier). Second, the sequences themselves consisted of a multi-aligned fasta file. The consensus sequences were derived (i.e, only A, T, C, and Gs left) by choosing the major allele where heterozygosity was detected, and the software MAFFT (Katoh et al., 2002) was used to align the consensus sequences for all individuals and output in aligned format. Thirdly, the phylogeny was present in a newick format. An example of the structures of these files is given below.

a)

FG 186 *Inga poeppigiana*

KD 13 *Inga poeppigiana*

LA 2023 *Inga poeppigiana*

M46A *Inga poeppigiana*

b)

```
>FG_186
CATTGTTCCTCCATAACACACAGATTTTGGCCGGA
>KD_13
CATTGTTCCTCCATAACANACAAATTTTGGCCGGA
>LA_2023
CATTGTTCCTCCATAACACACAAATTTTGGCCGGA
>M46A
CATTGTTCCTCCATAACACACANATTTTGGCCGGA
```

c)

((FG 186:0 0004627773 M46A:0 0003625627)100:0 0002685118 (KD 13:0 0003441657 LA 2023:0 0003470779)100:0 0003440200)100:0 0013278046)100:0 0003416399

Figure 4.1. An example of the structure of three primary data files. a) name tracking records (in txt format); b) Aligned sequence of the four samples in fasta format; and c) A subset of *Inga* phylogeny in newick format and visualised version. Four samples and the lengths of the branches are included.

4.3.2. Data analysis

The methodology used for analysing the data is described in detail in Chapter 3. In summary:

To estimate the Monophyly Ratio (MR), I recorded the number of species represented by more than one sampled individual that resolved as monophyletic, as a proportion of the total number of species in the dataset with more than one sampled individual.

To estimate the density and abundance of species-specific SNPs, the number (and density per Mb) of species-specific SNPs (SSSNPs) was extracted from the total dataset with an SSSNP defined as an SNP that had a character state that was fixed present in one species, and which was not present in any other species. I also extracted the number of SNPs that showed a major difference in allele frequencies (variants present at >87.5% frequency in the focal species; present at < 12.5 % in any other species).

To assess the relationship between SSSNP density and species monophyly, species were plotted in order of the number of species-specific SNPs each contained, and this was mapped onto whether the species resolved as monophyletic or not.

To evaluate the minimum amount of data needed for species discrimination (e.g. to assess how efficient the data could be in telling species apart), the data was

subsampled to estimate the minimum number of randomly selected genes or SNPs that recovered the maximum number of separable species. This involved running 50 replicates of 10 genes or 200 SNPs selected at random and generating a UPGMA tree for each replicate and recording the proportion of species that resolved as monophyletic using Monophy (Schwery et al., 2016). This process was repeated by incrementally adding more data. For the dataset consisting of genes as the focal until, this involved increasing the amount of data in steps of 10 genes at each step until 100 genes were reached, then adding 40 genes at each step until 300 genes were reached, and after this, adding 100 genes at each further step. For SNPs, the incremental steps involved a further 200 SNPs at each step until 2000 SNPs, then adding 1000 SNPs at each further step. The analyses were terminated when asymptote in discrimination was reached.

To assess whether a small number of carefully selected loci could result in species discrimination that was equivalent to a large random selection of loci, all loci were placed in rank order of the number of species they successfully resolved, and I recorded the cumulative number of best-performing genes that was required to recover the same species discrimination success as the total dataset.

4.4. Results

4.4.1. Assessing the overall discriminatory power of the target capture array for *Inga*

Of the 69 species with multiple individuals sampled, 45 resolved as monophyletic (65%). The remaining 24 species resolved as either polyphyletic (21 species, 31%), with individuals clearly scattered on the tree, or as paraphyletic (3 species, 4%), with individuals of another species nested within the variation encompassed by the species (Table S8).

4.4.2. Assessing the distribution of taxonomically important SNPs among *Inga* species

The number of SSSNPs for *Inga* ranges from 0 to 1,627 for each species (Figure 4.2), which translates to a density of 0 to 1,503 SSSNPs per megabase (median = 97), which in turn translates to a maximum density of one SSSNP every 665 bp, with a median across all species of one SSSNP every 10,309 bp. The number of SNPs that were not species-specific, but showed a marked frequency difference ranging from 0 to 3,828 for each species, which translates to 0 to 3,635 SNPs per megabase (median = 26). There are thus SNPs showing marked allele frequency differences at a maximum density of one every 275 bp, with a median occurrence across all species of one SNP every 38,461 bp. Figure 4.2 shows the relationship between the abundance of SSSNPs in each species with multiple individuals sampled, and whether that species resolves as monophyletic or not. There is the expected association that species that resolve as monophyletic have a greater number of SSSNPs. However, there are multiple species that do not resolve as monophyletic that still show the presence of SSSNPs.

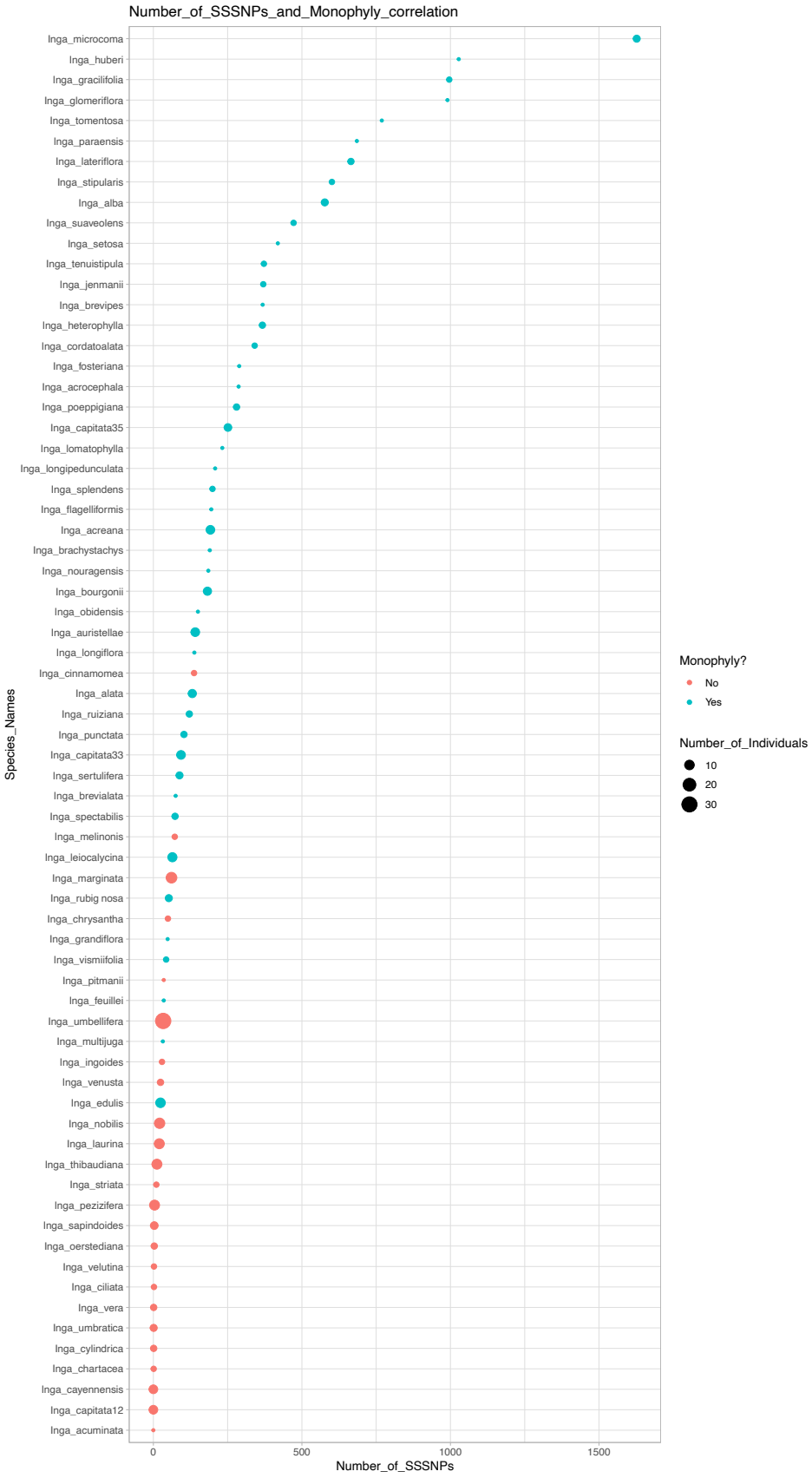


Figure 4.2. Distribution of the number of Species-Specific SNPs for multi-sampled *Inga* species. Red dots represent the species that resolve as non-monophyletic on the phylogeny and blue dots the species that resolve as monophyletic. The size of the dots represents the number of individuals sampled for a given species (the bigger the dot the more sampled individuals)

4.4.3. Assessing the impacts of sub-sampling the data on species discrimination

To explore how reducing the number of genes impacts the species identification success, random subsets of the data were drawn at stepwise intervals, with 50 replicates per interval and the levels of species monophyly recorded. When the analysis was carried out with the resampling unit being at the gene level, there is a marked decrease in levels of discrimination when only 10 genes were sampled (Figure 4.3). As the number of genes increased, there was a steep rise in the number of species discriminated, with an asymptote developing with a median number of 38 species resolved with a randomly selected 70 genes (8% of the 810 gene total data set). The maximum median number of species resolved in random draws from the 810 genes set up to a total of 600 genes, is 39 species (compared to 45 species resolving as monophyletic from the full gene set).

When the resampling analysis was undertaken using randomly selected SNPs, very few species are recovered as monophyletic using 100 randomly selected SNPs (median 10 species), but there is a steep rise in the number of species recovered up until around 1,000 SNPs, with a clear asymptote in species discrimination at around 2,500 SNPs (1.2 % of the total 205,871 SNPs in the dataset) where a median of 43-44 species are resolved as monophyletic (Figure 4.3).

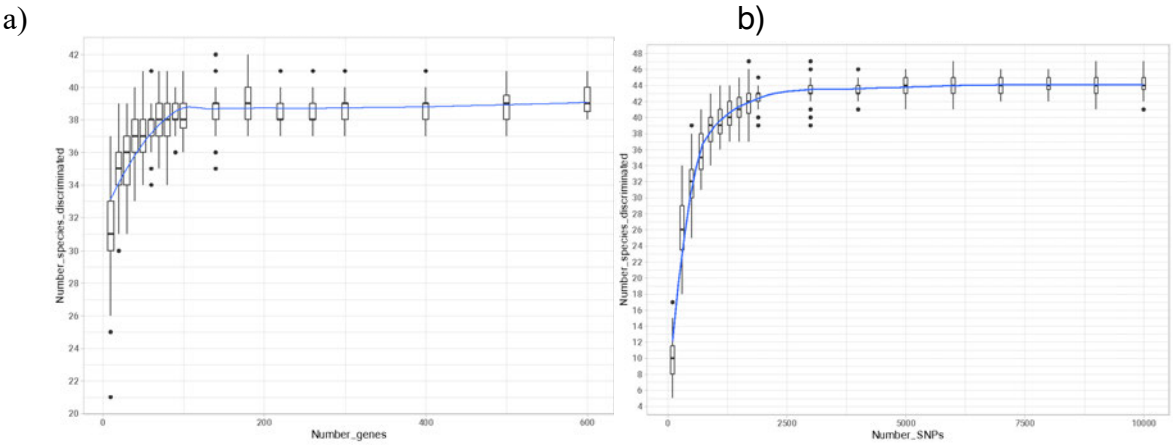


Figure 4.3. The number of species being discriminated with subsets of *Inga* target capture sequence data. The unit for subsampling for a) is the gene, and the size of the subset starts from 10 genes with 10 genes as the step until 100 genes, with 40 genes as the step until 300, and with 100 genes as the step after 300 genes; The unit for subsampling for b) is SNP, and the size of the subset starting from 200 SNPs with 200 SNPs as the step until 2,000 SNPs, then with 1,000 SNPs as the step until 10,000 SNPs. Each boxplot shows the distribution of the number of species being discriminated by a random draw of each number of SNPs/genes with 50 replicates with the median represented by a solid horizontal line.

The best-performing individual genes resolved a maximum of 31 species as monophyletic (compared to 45 species in the total dataset). The frequency distribution of the resolving power of individual genes is shown in Figure 4.4. When the five genes with the greatest resolving power were combined (0.6% of the total dataset), 42 species can be resolved as monophyletic. When this was extended to the nine genes with the greatest resolving power (1% of the total dataset), 44 species resolved as monophyletic. A list of genes and the number of species each of them could tell apart could be found in table S5.

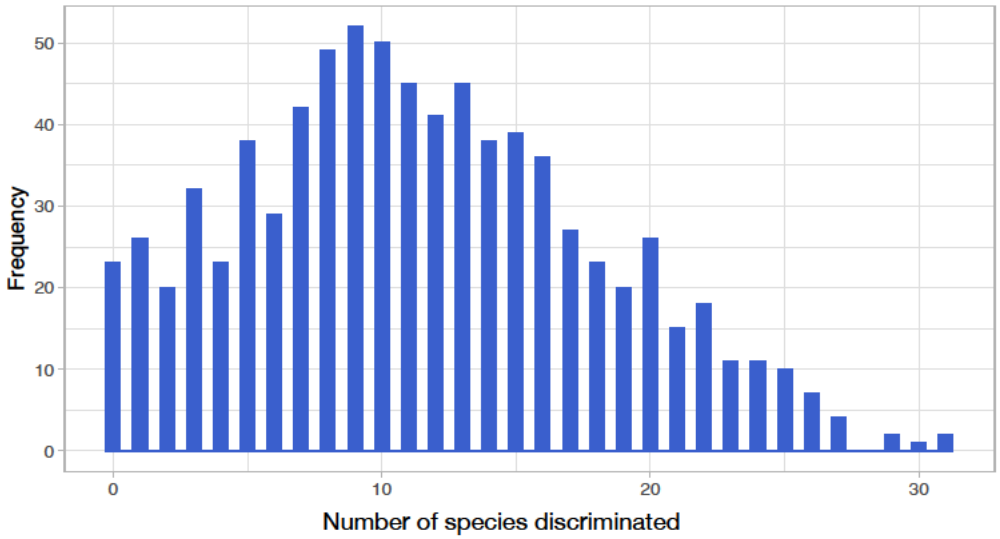


Figure 4.4. Frequency distribution of the number of species resolved as monophyletic by individual genes

4.4.4. Assessing the characteristics of the genes that show the greatest discriminatory power

To evaluate the characteristics of the genes with the greatest discriminatory power, they were plotted in rank order from least to greatest resolving power (equal to the number of species each resolved as monophyletic) in Figure 4.5. In Figure 4.5, the number of species a given gene resolved as monophyletic is positively correlated with the density of SSSNPs at that locus, although there is considerable spread in the data ($r = 0.61$, p-value <0.001). There is also a positive correlation ($r = 0.42$, p-value <0.001) between the number of species distinguished by a given gene and its associated nucleotide diversity though again there is considerable spread in the data, with the most variable loci not being the ones that show the greatest recovery of monophyly. A similar pattern and positive correlation also apply to the relationship between the density of SSSNPs and the nucleotide diversity ($r = 0.49$, p-value <0.001).

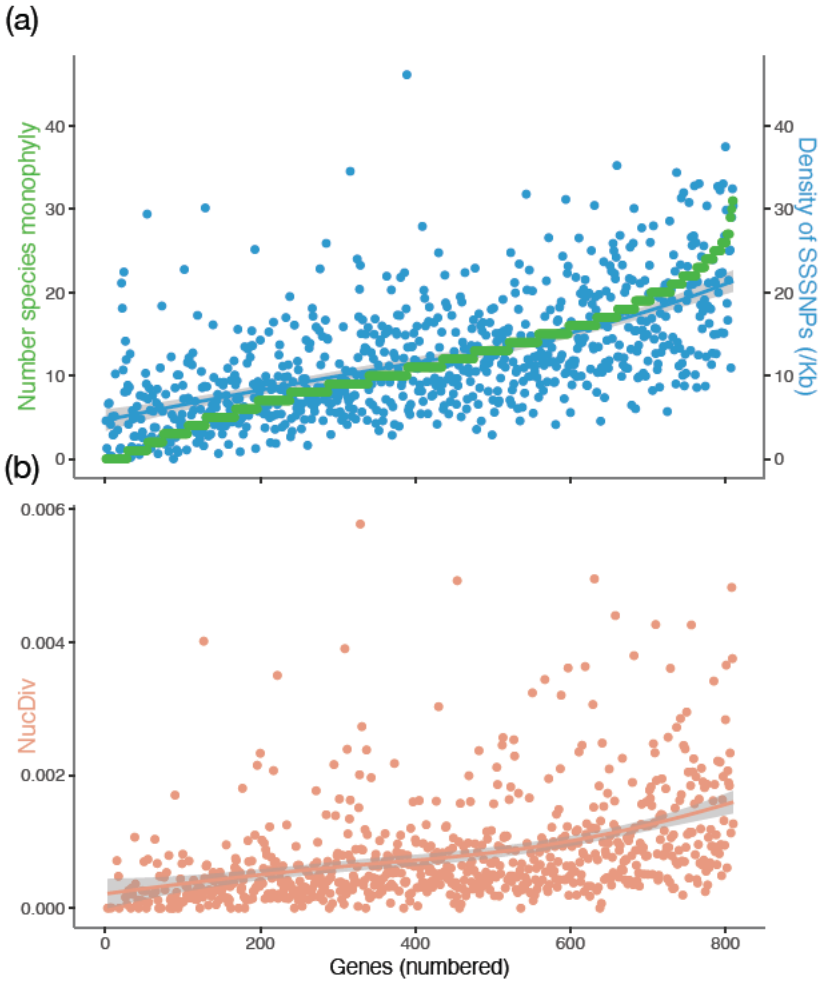


Figure 4.5. The relationship between the resolving power of individual genes, plotted against the (a) density of SSSNPs for each gene, and (b) the nucleotide diversity of those genes. The x-axis plots each of the 810 genes ordered by the number of species they resolve as monophyletic (from lowest to highest shown as the green dots in (a)). The blue dots represent the density of SSSNPs (per kilobases) on each gene, and the peach dots represent the nuclear diversity of this gene across the whole genus.

4.5. Discussion

4.5.1. Overall patterns of species discrimination

Inga is a challenging test case for species discrimination. Its recent radiation (Richardson et al., 2001) results in a species-rich assemblage of tree species with recent common ancestry, and overall, low levels of genetic divergence among species. In addition, the genus contains some species that are widespread and variable with large population sizes, and other species that are more locally distributed and that may have evolved by new taxa, budding off from the range margins of widespread taxa. Collectively these attributes might be predicted to lead to many species resolving as non-monophyletic (Pennington et al., 2016). The Monophyly Ratio from this *Inga* data set is 65%. This is a much greater level of species resolution for multi-sampled species than was detected by (Hollingsworth et al., 2009) using plastid barcoding loci (2/7). It is also much higher than the overall success of species discrimination reported by (Dexter et al., 2017) who recovered 18 species as monophyletic from 44 species (41%) with multiple sampled individuals using seven plastid regions (total 5,916 bp) and ITS sequences (572 bp).

The analyses revealed the expected association between the abundance of species-specific SNPs and species that resolve as monophyletic, with more SSSNPs in species that resolve as monophyletic, than in those that do not. Perhaps more interestingly, are the cases where there are species that do not resolve as monophyletic, which nevertheless possess species-specific SNPs. Examples of this include *Inga cinnamomea* where the species does not resolve as monophyletic, yet the 3 sampled individuals had 137 SSSNPs (130 SSSNPs / Mb). Likewise, *Inga marginata* also did not resolve as monophyletic, and its 13 sample individuals had 61 SSSNPs (57 SSSNPs / Mb). These SNPs can arise due to several reasons. The leading reason could be that the species is newly formed so that monophyly hasn't been achieved on a whole genome scale. These SSSNPs in non-monophyletic species, may be SNPs linked to regions of the genome under selection and thus linked to the cohesiveness of a species (Burri, 2017, Wu, 2001). Alternatively, they may just reflect a stochastic process of allele fixation during the history of species divergence (Kimura, 1962), and the SSSNPs consisting of loci that happen to have become fixed at an earlier stage in the history of the species divergence than many other loci in the genome.

Likewise, the lack of SSSNPs in monophyletic species could be explained from multiple perspectives. Firstly, it is possible that the genomic regions containing SSSNPs are not under strong selective pressure or are not linked to traits that contribute to the cohesiveness of the species (Burri, 2017). In such cases, the genetic variation among the species may be sparsely distributed across the genome, resulting in a lack of distinct SNPs specific to that species with the reduced-representative dataset. Additionally, the lack of SSSNPs in monophyletic species may be a consequence of the formation process of the species, e.g. through combinatorial mechanism (Marques et al., 2019). During combinatorial speciation, genetic variation is derived from the ancestral population through recombination and genetic exchange. Through various mechanisms such as hybridization, introgression, or horizontal gene transfer, species arise through the recombination and reshuffling of existing genetic variation rather than the accumulation of de novo mutations. In this mechanism, new

species can form without the fixation of unique alleles or SNPs specific to those species. Finally, it is important to note that the absence of SSSNPs may also due to a high within-species genetic diversity. There could be variations among different populations so that SNPs identified are not distinctive for the species. The last reason could be addressed by a more developed bioinformatic algorithm which is not touched at the moment.

This presence of species-specific SNPs in species that do not resolve as monophyletic is potentially important from a practical identification perspective. Clearly, monophyly is a somewhat restrictive criterion for 'successful' species discrimination, given a priori expectations that not all 'good' species will resolve as monophyletic (Pennington et al., 2016, Rieseberg et al., 2019). Being able to identify and quantify the abundance of SSSNP in non-monophyletic taxa thus has potential use for increasing the proportion of species that can be identified using sequence data, above and beyond those that resolve as monophyletic.

Of the 69 species from which multiple individuals were sampled, only three species did not resolve as monophyletic, or possess any SSSNPs. These species have no SNPs that showed significant allele frequency differences either. In this case, the complexity of the species is beyond the scope and extra innovative methods are required which could go beyond DNA-based identification.

4.5.2. Establishing the minimal amount of data for species recovery

SSSNPs were present at a variable density in *Inga* species, with a median presence of one SSSNP every 10,309 bp in the 69 multiple sampled species. The sub-sampling analyses conducted show that a random panel of 2500 SNP loci would give as good a level of species discrimination (43/44 species) as the full panel of 810 genes (= 205,871 SNPs in this dataset). When considered at the gene level, the analysis shows that c 70 randomly selected genes give almost as good a level of discrimination as the full data set (recovering on average 38 species as monophyletic). There is an interesting subtle difference in the discrimination power of the individual genes versus the SNP subsampling, with the subsampling asymptoting at slightly high levels of discrimination for SNPs. The cause of this is not immediately obvious, but it may simply be caused by the greater genomic coverage of the subsampled SNP data (which will be distributed across all loci, and hence more independent loci should be included in each replicate), whereas any of the sub-sampled gene replicates will leave entire loci out which could have a stronger disruptive effect on monophyly.

When the best-performing loci were selected (those genes that resolved the maximum amount of species as monophyletic), high levels of species discrimination can be achieved with small amounts of data, with only five loci resolving >>90% of the species as monophyletic compared to the total sample of 810 genes, with 100% monophyly recovery (of the species resolving as monophyletic using 810 genes) being reached with nine loci. Clearly these regions contain strong signals for species discrimination and in *Inga*, would represent prime targets for the development of diagnostic assays.

4.5.3. Characteristics of the loci that showed greatest species discrimination

There is a statistically significant positive correlation between the proportion of species that a given gene resolves as monophyletic and its levels of nucleotide diversity. However, there is considerable variation in the data, and it is clear that many of the most variable loci and not necessarily the most informative or useful loci for species discrimination. What is clear from the analysis (Figure 4.5) is that the genes that show very low levels of diversity, and generally not effective at recovering species as monophyletic, whereas the converse is not necessarily true. This guards against an overly strong focus on simply finding the most variable nuclear regions for targeting for species identification purposes.

It is worthwhile to notify that the nucleotide diversity is a within-species measure originally, and when measured within species, it is correlated with effective population size (*Ne*) (Nei *et al.*, 1979) and might therefore not be predictive of the most informative genes, because the probability of incomplete lineage sorting (ILS) increases with *Ne* (Maddison, 1997). In this thesis scenario, the nucleotide diversity is a within-genus measurement. Similarly, the correlation with *Ne* also applies to this extended definition, i.e. the size of the genus. The lack of predictive ability of the most informative genes could be caused by (ILS).

4.5.4. Implications for developing a nuclear DNA barcoding system for *Inga*

The aim of this study was to gain a general picture of the patterns of variation between species in the genus *Inga* and to road-test analytical pipelines for more general analyses on other datasets. The aim was not to design a barcoding / diagnostic identification panel for *Inga*. Nevertheless, it is worth reflecting on the findings of this chapter on the implications for the design of species identification systems.

At the outset it is worth outlining some caveats. Firstly, although extensive, the dataset used here is not comprehensive for *Inga* species, and any conclusions should be borne with that in mind. Secondly - the future success of DNA-based identification methods for plants will be dependent on their universality. Thus developing an identification system for a given genus does not address the larger challenge of developing routine methodologies for telling all plant species apart. Thirdly, in this particular dataset, heterozygosity was masked during the dataset production stage. This may impact on the resolving power of the data.

With the above caveats in mind, the analyses conducted here are informative about the types of nuclear data that will be required to tell species apart in challenging groups.

Many Inga species are not monophyletic: One simple observation is that only 65% of species resolved as monophyletic. The corollary of this is that 35% of species did not. Thus for genera like *Inga*, a monophyly-based diagnostic approach will leave many species unidentified. This contrasts strongly with the signature of DNA barcode sequences in animal groups such as Lepidoptera where genuine non-monophyly is uncommon (Mutanen et al., 2016). A further qualifier that needs adding is that some non-monophyletic species of *Inga* may be the result of species misidentifications or imperfect taxonomy. As noted by (Mutanen et al., 2016) in their evaluation of this issue in a well-studied sample set of European butterflies, some cases of non-monophyly may be attributable to operational taxonomic problems as opposed to underlying

species biology. On the other hand, there are good theoretical reasons to assume that non-monophyletic species are not uncommon in plants (Rieseberg et al., 2019), and especially in groups like *Inga* (Pennington et al., 2016).

Most examined species had taxon-specific SNPs: The distribution and density of species-specific SNPs is informative for the design of species discrimination assays for *Inga*. The positive observation is that 66/69 species have at least one SSSNP (based on the sample set examined here). This type of information could be used to design an SNP panel to screen for the presence of these SNPs in a cost-effective fashion.

Large numbers of randomly selected nuclear loci are required to tell Inga apart: While the widespread presence of SSSNPs is a positive finding, what is less encouraging in more general terms, considering extrapolation to other groups where there is no a priori data, is the overall density of these SSSNPs. Recovering maximum levels of species discrimination in *Inga* could be done with only a small portion of the 810 genes and 1,313,489 bp of the total data set, but this nevertheless still required 2500 randomly selected SNPs, or c70 different genes. This is a substantial increase in data requirements beyond conventional barcodes of hundreds or a few thousand bps, and such assays would require extensive resources (consumables, informatics, data processing) to operate at scale. This observation is in accord with some previous smaller scale studies, such as that by Ruhsam et al. (2015) who showed that sequencing a small number (11) of randomly selected nuclear genes did not provide high levels of resolution in the *Araucaria* species of New Caledonia.

Some loci are much better than others at species discrimination: One of the most striking findings of this analysis is that there are some gene regions that are particularly good at species discrimination. These are not the most variable loci in the target capture bait set, and instead have intermediate levels of nucleotide variation (with the top five genes for species discrimination having levels of nucleotide diversity of 0.11 – 0.48% (mean 0.27%). In the case of *Inga*, the use of 5-10 best-performing genes for species identification would result in equivalent levels of recovery of monophyletic species as the full panel of 810 genes.

4.5.5. An appraisal of the practical implementation steps required to operationalise a nuclear DNA barcode for *Inga*

In this final section I continue the thought exercise outlined in section 4.5.4. to work through some of the practicalities of moving from knowledge of the patterns of sequence variation among species to how that might translate into a practical assay.

Current advancements in NGS DNA sequencing platforms allow a variety of approaches to routinely recover multi-locus DNA data. Taking the resource availability into consideration, such as access to lab equipment, regionally available sequencing services, and the budget constraints, I have worked through a set of five different potential approaches Table 4.2. The prominent challenges are different for each of the methods. For example, herbarium specimens are usually used in botanical taxonomic studies. The feature of the dried specimens is fragmented and contaminations, making the PCR-based amplicon sequencing and further genome digestion methods (RAD-seq/GBS) challenging. The genome skimming method requires no probe or bait

design, but it does involve library construction. Furthermore, although the reported genome size of *Inga* in the Kew C-value database is relatively compact (one species has been recorded, *Inga dulcis*, 1C = 0.49 pg ≈ 0.49 Gbp (https://cvalues.science.kew.org/search/angiosperm; accessed 29/12/2022) the cost for obtaining adequate coverage of the nuclear genome is non-trivial when scaled over multiple samples. The SNP array and Hyb-seq methods are designed to handle fragmented DNA, and the loci to sequence are predetermined so the cost is not sensitive to genome size variation. The initial laboratory costs of these methods are high including the one-time synthesis of a new set of custom probes (for SNP array) or biotinylated baits (for Hyb-seq). With the service provider, a customized probe set for SNP array costs ~$70 per sample, and ~$99 per sample for Hyb-seq. However, the cost could be reduced by advanced experimental design such as integrating SNPs for multiple species on one array, and pooling libraries of selected 'good' genes for multiple genera. For example, a study successfully reduced the overall cost for Hyb-seq to $22.66 per sample with bulk purchase and frugal experimental design (Hale et.al., 2020).

Table 4.2. Appraisal of implementation approaches and steps to multi-locus DNA barcoding

Multi locus DNA barcoding candidate methods	Minimum information required	Maximum number of species distinguished	Requires high quality starting material	Genome assembly required	nitial laboratory cost	Bioinformatic analytical complexity	Sequencing cost
SNP microarray	random 2000 SNPs / ~188 SSSNPs	44 ± 2 / 66*	No	No	High §	Low	Not applicable
GBS/RAD seq	random 2000 SNPs / ~188 SSSNPs	44 ± 2 / 66*	High molecular weight DNA preferred	Yes	Medium	High	Medium
Genome skimming	random 2000 SNPs / ~188 SSSNPs	44 ± 2 / 66*	No	Yes	Low	High	Medium to high®
PCR based Amplicon Sequencing	random 70 genes $ / 9 good genes	39 ± 2 / 44	High molecular weight DNA preferred	No	Medium	Medium	Low
Hyb seq/Target enrichment	random 70 genes / 9 good genes	39 ± 2 / 44	No	No	High §	Medium	Low

* The confidence in the diagnostic capability of the SSSNPs should be considered and evaluated.

® depending on the genome size.

$ Might incur multiple amplicon probe sets for each sample because commercially viable target regions for amplicon sequencing are usually < 50 genes. (Recommendation from Illumina:

https://emea.illumina.com/techniques/sequencing/dna-sequencing/targeted-resequencing/targeted-panels.html).

§ Cost could be reduced to medium by advanced experimental design.

4.6. Conclusion

This chapter assessed the potential of using nuclear sequence data to tell *Inga* species apart, and also road-tested analytical pipelines for a more general comparison among species. Based on a dataset of 810 genes across 453 individuals from 133 *Inga* species. Among the 69 multiple-sampled species, around 45 of these species resolve as monophyletic on the phylogeny and 66 species have at least one diagnostic SNP. Distinguishing among *Inga* species with randomly selected genes or SNPs can be done using a much smaller fraction of data than the 810 gene set, but it is still a relatively data-intensive task (e.g. 2500 SNPs or 70 different genes). Pre-selecting the best-performing genes in the case of *Inga* leads to substantial improvements in the efficiency of telling species apart, with 5-9 genes giving equivalent levels of species discrimination to the full dataset.

4.7. Reference

Baker, T. R., Pennington, R. T., Dexter, K. G., Fine, P. V. A., Fortune-Hopkins, H., Honorio, E. N., . . . Vasquez, R. (2017). Maximising synergy among tropical plant systematists, ecologists, and evolutionary biologists. *Trends Ecology and Evolution*, 32(4), 258-267. doi:10.1016/j.tree.2017.01.007

Burri, R. Interpreting differentiation landscapes in the light of long-term linked selection. *Evolution Letters*, 2017, 1(3): 118-131.

Dexter, K. G., Lavin, M., Torke, B. M., Twyford, A. D., Kursar, T. A., Coley, P. D., . . . Pennington, R. T. (2017). Dispersal assembly of rain forest tree communities across the Amazon basin. *Proceedings of the National Academy of Sciences*, 114(10), 2645-2650. doi:10.1073/pnas.1613655114

Endara, M. J., Coley, P. D., Wiggins, N. L., Forrister, D. L., Younkin, G. C., Nicholls, J. A., . . . Kursar, T. A. (2018). Chemocoding as an identification tool where morphological- and DNA-based methods fall short: *Inga* as a case study. *The New Phytologist*, 218(2), 847-858. doi:10.1111/nph.15020

Hale, H., Gardner, E. M., Viruel, J., Pokorny, L., & Johnson, M. G. (2020). Strategies for reducing per-sample costs in target capture sequencing for phylogenomics and population genomics in plants. *Applications in Plant Sciences*, 8(4), e11337. doi:10.1002/aps3.11337

Hollingsworth, M. L., Clark, A., Forrest, L. L., Richardson, J., Pennington, R. T., Long, D. G., . . . Hollingsworth, P. M. (2009). Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. *Molecular Ecology Resources*, 9(2), 439-457. doi:10.1111/j.1755-0998.2008.02439.x

Hollingsworth, P. M. (2015). Does complete plastid genome sequencing improve species discrimination and phylogenetic resolution in *Araucaria*? *Molecular Ecology Resources*, 15(5), 1067-1078. doi:10.1111/1755-0998.12375

Katoh, K., Misawa, K., Kuma, K. i., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, 30(14), 3059-3066. doi:10.1093/nar/gkf436

Kimura, M. On the probability of fixation of mutant genes in a population. *Genetics*, 1962, 47: 713-719.

Kozarewa, I., Armisen, J., Gardner, A. F., Slatko, B. E., & Hendrickson, C. L. (2015). Overview of target enrichment strategies. *Current Protocols in Molecular Biology*, 112, 7 21 21-23. doi:10.1002/0471142727.mb0721s112

Lavin, M. (2006). Floristic and geographical stability of discontinuous seasonally dry tropical forests explains patterns of plant phylogeny and endemism. London: *Neotropical Savannas and Seasonally Dry Forests* (pp. 433-447).

Maddison, W. P. Gene trees in species trees. *Systematic Biology*, 1997, 46(3): 523-536.

Marques, D. A., Meier, J. I., et al. A combinatorial view on speciation and adaptive radiation. *Trends Ecology and Evolution*, 2019, 34(6): 531-544.

Mutanen, M., Kivela, S. M., Vos, R. A., Doorenweerd, C., Ratnasingham, S., Hausmann, A., . . . Godfray, H. C. (2016). Species-level para- and polyphyly in DNA barcode gene trees: Strong operational bias in European Lepidoptera. *Systematic Biology*, 65(6), 1024-1040. doi:10.1093/sysbio/syw044

Nei, M. and Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, 1979, 76(10): 5269-5273.

Nicholls, J. A., Pennington, R. T., Koenen, E. J., Hughes, C. E., Hearn, J., Bunnefeld, L., . . . Kidner, C. A. (2015). Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: *Mimosoideae*). *Frontiers in Plant Science*, 6, 710. doi:10.3389/fpls.2015.00710

Pennington, T. D. (1997). The Genus *Inga*: Botany. London: *Royal Botanic Gardens, Kew.*

Pennington, R. T., & Lavin, M. (2016). The contrasting nature of woody plant species in different neotropical forest biomes reflects differences in ecological stability. *The New Phytologist*, 210(1), 25-37. doi:10.1111/nph.13724

Richardson, J. E., Pennington, R. T., Pennington, T. D., & Hollingsworth, P. M. (2001). Rapid diversification of a species-rich genus of neotropical rain forest trees. *Science*, 293(5538), 2242-2245. doi:10.1126/science.1061421

Rieseberg, L. H., & Brouillet, L. (2019). Are many plant species paraphyletic? *Taxon*, 43(1), 21-32. doi:10.2307/1223457

Schwery, O., & O'Meara, B. C. (2016). MonoPhy: a simple R package to find and visualize monophyly issues. *PeerJ Computer Science*, 2. doi:10.7717/peerj-cs.56

Wu, C.-I. The genic view of the process of speciation. *Journal of Evolutionary Biology*, 2001, 14(6): 851-865.

Chapter 5 A meta-analysis on the use of nuclear sequence data for plant species discrimination

5.1. Abstract

The recent accumulation of nuclear DNA sequence data for plants offers the potential to explore the signal in these data for species identification. In this chapter I undertake a meta-analysis using available data to evaluate the patterns of sequence differences between species, and the ease with which species can be discriminated using data from the nuclear genome. I compiled data from 149 different genera to assess the proportion of plant species that resolve as monophyletic. I then selected 29 genera with suitable available data for more detailed analysis. Overall I tackled the following questions (1) what is the proportion of species are distinguishable with nuclear markers? (2) what is the nature of the inter-specific differences and what are the attributes of loci that are the most informative in telling species apart? And (3) how many markers are needed and what markers are needed to maximise the species identification success? In the analysis of 149 genera, overall, of the 1,701 multiple-sampled species evaluated 1,206 resolved as monophyletic (71%). At the level of individual genera, 37 of the 149 genera (25.8%) had 100% of species resolved as monophyletic, and 75 (50.3%) genera had at least 75% of the species resolving as monophyletic. The median percentage of species resolved as monophyletic across all genera was 75%.

To understand the genetic basis of species differences, the abundance of species-specific SNPs (SSSNPs) was characterised in 29 datasets representing 21 plant families scattered across the land plant tree of life. Among these genera, the density of SSSNPs of all species ranges from 0 to 27,262 per Mb, with a median density of 323 SSSNPs per Mb (a median density of one SSSNP every 3,098 bp). In species that resolve as monophyletic, the density of SSSNPs in 90% of monophyletic species ranges from 20 – 5,624 per Mb; in those that do not resolve as monophyletic, 90% of non-monophyletic species have SSSNPs at a density between 0 – 648 SSSNPs per Mb. Of the total of 460 species from 29 genera assessed, 411 species (89.3%) had at least one SSSNP.

When the data were subsampled to evaluate the minimum amounts of data required for species discrimination, there was an asymptote at around 2,500 – 3,000 randomly selected SNPs by which almost all of the species resolved in the full datasets (which ranged from 6,061 – 1,534,400 total number of SNPs) could be distinguished. When the loci were placed in rank order within a selection of six genera to further evaluate the characteristics of the best-performing loci, the single best-performing locus was able to resolve as many or almost as many species as the full dataset (consisting on average of 663 genes, range 360-881) in four of the six genera, with seven and nine genes required respectively in the other two datasets. When I assessed the attributes of these best-performing loci, there is a statistically significant correlation between the density of SSSNPs and the species recovered as monophyletic from different gene regions, and a positive (but weaker) correlation between sequence diversity and discriminatory power of the best-performing loci in four of the six datasets. These findings give a first quantitative assessment using multiple independent nuclear loci

across a wide range of plant groups of the number of species that resolve as monophyletic, the distribution of SSSNPs, and the implications of these data for selecting minimal amounts of sequence data for telling the maximum number of plant species apart.

5.2. Introduction

Since the Linnaean initiation of cataloging the names of over 5,940 plant species during the 18th century (Linné, 1753), and despite over two centuries' endeavours of numerous taxonomists, the volume of species awaiting identification and description remains prodigious -- recent estimates suggest that around 70,000 flowering-plant species await discovery (Bebber et al., 2010). And even for the species that have already been described, it can be difficult to reliably identify plant species, particularly if the material available for identification is sub-optimal in one way or another.

With a growing volume of genomic studies and an expanding repository of DNA sequences, the idea of exploiting the DNA sequence information of standardised genomic regions to represent each species, namely 'DNA barcoding', has gained wide acceptance for species identification. The plastid genes *rbcL* and *matK* were accepted as the standard plant barcode combination for plants (CBOL, 2009). This two marker combination includes a portion of the highly conserved *rbcL* gene that is easy to recover (barcode region length ~ 600 bp) (Chase et al., 1993) along with a more variable portion of the *matK* gene (barcode region length ~ 800 bp) (Dunning et al., 2010). Although these regions have been widely deployed, and augmented with other plastid loci (Hollingsworth et al., 2016), they do not always provide species-level resolution, and there are multiple cases where species can share the same barcodes. Potential reasons for the shared barcodes among species include (a) incomplete lineage sorting of ancestral polymorphism, (b) hybridization leading to the widespread distribution of a certain allele, (c) new mutations followed by selective sweeps (d) the lineage is young and went through a recent rapid radiation (Twyford, 2014).

One drawback of the standard plant DNA barcodes is that both *rbcL* and *matK* are from organelle genomes, and organelle genomes do not necessarily track species boundaries. This is exemplified by the phylogenetic incongruence between nuclear and plastid genes (Stephens et al., 2015, Stephens et al., 2015, Schmickl et al., 2016, Gernandt et al., 2018, Lin et al., 2019, Mu et al., 2020, Scharmann et al., 2021).

The internal transcribed spacers of nuclear ribosomal DNA ITS, or just one of the spacer regions (ITS2) has also been widely used in plant barcoding due to its ability to identify more closely related species than the plastid barcodes (Yao et al., 2010, China Plant BOL Group et al., 2011), and as such it is routinely incorporated into standard barcoding approaches for plants (Hebert et al., 2016). It can often lead to a 10~20% gain in resolution in tested plant groups. However, it can occur in multiple copies in many plant genera that hampers its fully universal use, due to challenges in obtaining clean sequences, and/or difficulties in interpreting the signal from paralogous copies (Hollingsworth, 2011).

The decreasing cost and increasing throughput of the next-generation sequencing techniques has resulted in a large volume of studies obtaining sequences from the nuclear genome of multiple species across different plant families to recover the

evolutionary histories of target groups (Eaton et al., 2013, Nicholls et al., 2015, Kates et al., 2018). Sequencing techniques have been developed to enable sequencing of the entire, or subsets of the nuclear genome in an affordable fashion such as genome skimming (Dodsworth, 2015), RAD-Seq (Miller et al., 2007, Peterson et al., 2012), GBS (Elshire et al., 2011), and target capture (Mamanova et al., 2010, Kozarewa et al., 2015, Hale et al., 2020)

Multiple independent nuclear loci should (conceptually) improve levels of species discrimination, and individual studies have shown this. A recent study in Oak combined the RAD-seq and MassARRAY approaches, and developed an efficient multispecies barcode for this complex tree genus (Fitzek et al., 2018). Another study in *Anacyclus* addressing global medicinal trade routes demonstrated that the target capture data could outperform standard plant barcodes in terms of species resolution (Manzanilla et al., 2022). These examples provide a proof-of-principle for the efficacy of using multiple nuclear regions to assign plant samples to the right species. However, what is lacking is an overview and synthesis of exactly how powerful these approaches can be, and how best to guide future efforts in building plant identification tools.

Outstanding questions include 1) what is the proportion of species distinguishable with nuclear markers? 2) what is the nature of inter-specific genomic differences between plant species and what are the attributes of loci that are the most informative? 3) how many markers and what markers are needed to maximise the species identification success?

The prediction that the proportion of species distinguishable with nuclear markers varies depending on the genetic complexity of species. Species with lower biological complexity are more likely to be distinguishable using nuclear markers. Due to the reason that woody plants have a longer lifespan and are more prone to hybridisation and introgression, the hypotheses of question one is the proportion of species distinguishable is higher in herbaceous than woody plant groups. An additional prediction is that the regions of the genome being sequenced has a significant impact on the proportion of species distinguishable. This leads to the hypothesis that a different level of the proportion of species distinguishable varies among different sequencing methods.

Of the species that are distinguishable with nuclear markers – why are they distinguishable? The predictions are 1) they have lots of fixed nucleotide substitutions which are diagnostic for different species, or 2) they lack fixed differences, but genome-wise they have a high nucleotide diversity so they can be told apart genetically overall – even though there is a shortage of individual gene regions which are uniquely diagnostic. The hypotheses of question two then could be described as species that have more fixed differences (species-specific SNP), have a higher nucleotide diversity, and are prone to be monophyletic.

How many markers and what markers are needed to maximize species identification success? The prediction is that the number of markers required for successful species identification will depend on the level of genetic diversity and the number of species under consideration. A larger number of markers, particularly those with high species-specificity, will increase the accuracy and success of species identification. So the hypothesis for question three is increasing the number of markers will increase the

species identification success, and there are loci that tell more species apart than others.

In this chapter, I compiled nuclear genomic sequences from public repositories and collaborative projects and assessed the proportion of plant species distinguishable by nuclear markers. I undertake a synthetic evaluation of the genetic differences that drive discrimination success and failure. I compare the performances of data from different sequencing techniques and explore the minimum amount of sequence data needed to give an optimal species resolution success.

5.3. Materials and Methods

5.3.1. Assembling published studies for an overview of the extent of plant species monophyly (Dataset 1)

To assemble data for this analysis I searched for publications (from 2013) that sequenced three or more unlinked nuclear loci from at least three individuals from multiple congeneric species. I used the matching pattern “phylogeny*” AND “RAD*” OR “target capture/hyb-seq” OR “genome skimming” OR “Transcript*” on the Web of Science, and selected publications manually. I also obtained unpublished datasets from collaborators. A total of 149 plant groups (Table S1) matched these sampling and sequencing criteria and also included access to a phylogenetic tree where species monophyly could be inferred. They are from a wide range of taxa including 2 moss genera, 3 fern genera, 6 gymnosperms genera, and 138 angiosperm groups. Studies were categorised by sequencing techniques, into 1) Restriction site-associated DNA sequencing (RAD-seq (Baird et al., 2008)) and its derivatives, (e.g., GBS (Elshire et al., 2011), ddRAD-seq (Peterson et al., 2012), 2b-RAD (Wang et al., 2012)); 2) Target Capture (Mamanova et al., 2010, Kozarewa et al., 2015); 3) Genome skimming (Dodsworth, 2015); 4) Transcriptome/exon sequencing (Wang et al., 2009, One Thousand Plant Transcriptomes, 2019).

5.3.2. Assembling datasets to assess genomic differences between plant species

Of the 149 individual datasets used for assessing patterns of monophyly, 29 were suitable for more detailed analysis (Dataset 2). These were selected on the criteria of having (1) Consensus sequences for each individual (either multiple-aligned sequence file in .phylip, .fasta, or .nex formats, or SNP matrix in .vcf or .fasta format), (2) Metadata, including records of changes of names if provided, and the inclusion or exclusion of individuals and loci, (3) a phylogeny. The individual datasets selected are from 21 different families, and also represent a wide range across the tree of life (Figure S5.1).

Table 5.1. Six individual datasets analysed to assess the performance of individual loci for species discrimination (Dataset 4)

Genus	Tota number of oc n the dataset	Locus ength ower quart e (bp)	Locus ength on average (bp)	Locus ength upper quart e (bp)
Artocarpus	517	1755	2302	2615
Capurodendron	615	648	1100	1189
Geonoma	795	2845	3932	4588
Inga	810	1018	1622	1946
Polemonium	360	614	706	750
Tsuga	881	729	1103	1374

Of these 29 individual datasets, a final round of analyses was undertaken on six individual datasets (Dataset 4) to further investigate species discrimination at the level of individual loci. Assessing the performance of each locus in telling species apart is only possible for datasets that have full sequence data on each locus with a low rate of missing data. With these requirements, datasets with only SNP information were excluded, as were those based on RAD sequencing methods as these approaches result in high levels of missing data. The six retained individual datasets (Table 5.1) are from the genera *Inga, Geonoma, Artocarpus, Polemonium,* and *Capurodendron* that were sequenced by target capture, and *Tsuga* based on transcriptome sequencing.

5.3.3. Assembling data sets to evaluate the minimum amounts of data that can recover the maximum amount of species discriminated

Starting with the Dataset 2 described in Section 5.3.2, I retained datasets that were amenable to subsampling SNPs or genes/DNA segments. A total of 23 datasets (Dataset 3) were used for this analysis, and these were selected on the grounds of having enough species that could be told apart by the full dataset, as it is less meaningful to down-sample a dataset that can only tell none or a few species apart in the first place. The list of the 23 genera for Dataset 3 retained for this analysis and their properties can be found in Table 5.4.

A random subset of SNPs/DNA segments was drawn from the whole dataset, repeating the random draw 50 times, and the size of the subset varies from genus to genus according to the data type and single loci length. Normally it starts from 100 SNPs and ends with 10,000 SNPs, or starts from 10 genes until 600 genes.

To evaluate whether there are any genes that give more species resolution information than others I focused on evaluating the individual and cumulative performance of the genes showing highest resolution. To do this, I built a quick tree for every single gene and calculated the number of species each of them can tell apart. To evaluate why some genes/DNA segments are better than others in telling species apart, I then evaluated the nucleotide diversity of each gene to assess the relationship between discriminatory power and nucleotide diversity.

5.3.4. Data analyses summary

The data analyses followed the methods outlined in detail in Chapters 3 and 4. In summary, the analysis aims are to:

- Estimate the proportion of plant species that resolve as monophyletic (the Monophyly Ratio, MR) (Dataset 1, Table S1)
- Calculate the abundance and density of species-specific SNPs (SSSNPs) (Dataset 2, Table S2)
- Assess what proportion of species can be distinguished by subsets of the data by random resampling (Dataset 3, Table 5.4) and also by targeting the best-performing gene regions (Dataset 4, Table 5.3)

- Evaluate the characteristics of the best-performing gene regions in terms of levels of nucleotide diversity and density of SSSNPs (Dataset 4, Table 5.3)

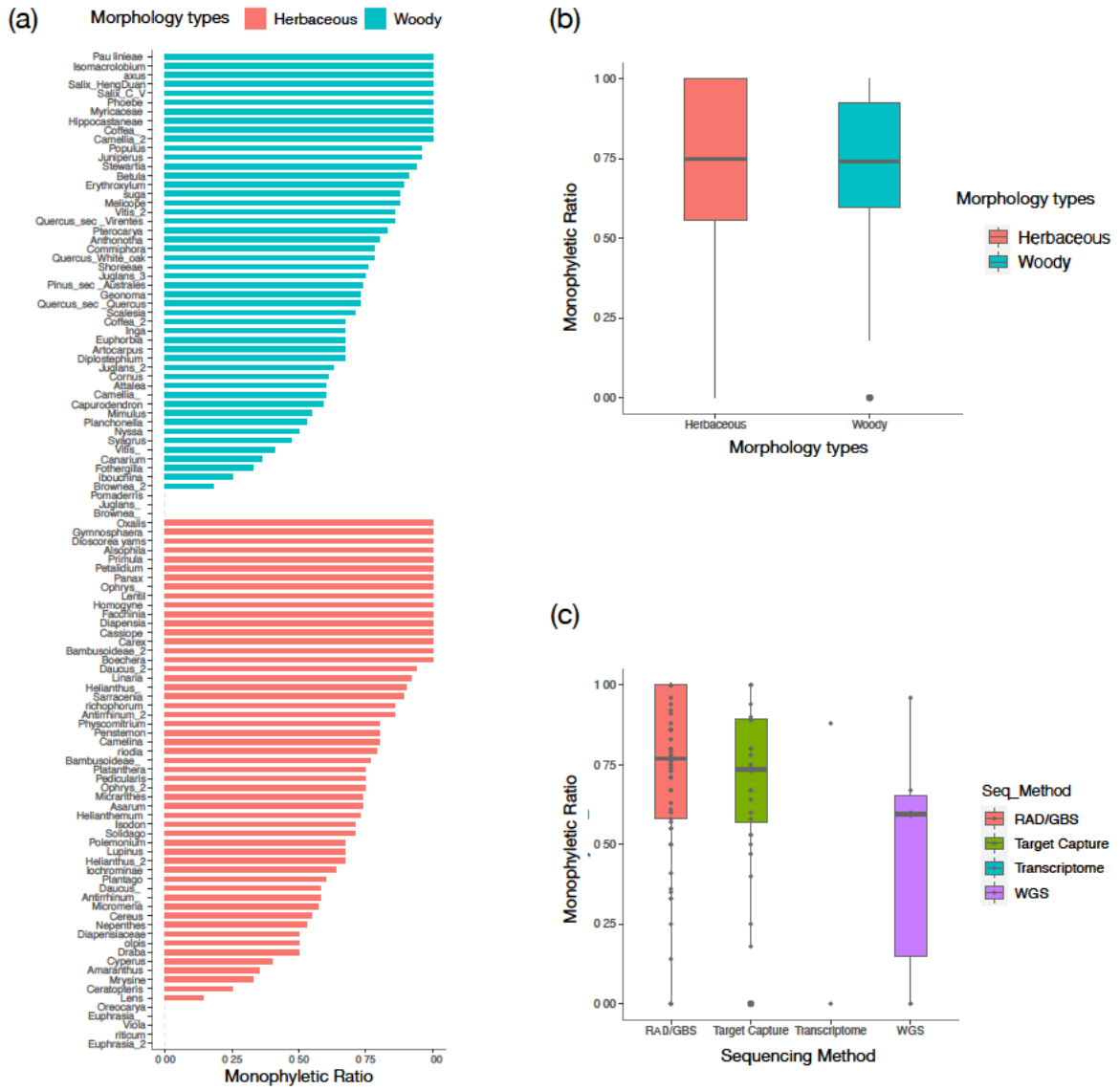
5.4. Results

5.4.1. **How often are plant species monophyletic?**

I calculated the Monophyletic Ratio (MR), defined as the proportion of monophyletic species out of the total number of species with multiple samples, for a total of 149 plant genera (Table S1). This included 138 angiosperm genera as well as 2 mosses, 3 ferns, and 6 gymnosperm genera. The MR here indicates how well the sequence data distinguish species, i.e., a high MR indicates that the sequence data used to build the phylogeny has a high species diagnostic power. Among the 149 genera, 37 (25.8%) resolve 100% of species as monophyletic, while 75 (50.3%) have a MR higher than 75% (Figure 5.1.A). No significant difference in MR was observed between plant groups with different growth strategies ($p = 0.84$, woody vs. herbaceous, Table 5.2, Figure 5.1.B) or between studies using different sequencing techniques such as target capture, genome skimming, transcriptome sequencing, and RAD/GBS (p-values all > 0.1, Table 5.2, Figure 5.1.C); tests conducted using Welch's t-test (Knapp, 2017).

Table 5.2. P-values of the Welsh's t-test between the monophyletic ratio and growth form or sequencing approach

Var ab e pa rs for compar son	P-va ue
Herbaceous vs Woody	0 843
Genome sk mm ng vs RAD GBS	1
Genome sk mm ng vs Target Capture	0 238
Genome sk mm ng vs Transcr ptome	0 397
RAD GBS vs Target Capture	0 238
RAD GBS vs Transcr ptome	0 397
Target Capture vs Transcr ptome	0 543



Figurea 5.1. Monophyletic ratio (MR) of 149 genera/sub-genera. (A) MR of individual taxa (B) Comparison of MR among different growth strategies. (C) Comparison of MR between different sequencing methods. Boxplots show the median, lower, and upper quartiles, with whiskers extending to 1.5 times the interquartile range; the dots in (C) are the MR value for each genus.

5.4.2. Are species-specific SNPs the norm or the exception?

To understand the genetic basis of species differences, I characterised the abundance of species-specific SNPs (SSSNPs) in 29 datasets representing 21 plant families scattered across the land plant tree of life.

Figure 5.2 provides a detailed overview of the distribution of species-specific SNPs for multi-sampled plant species for 29 genera listed. Among these genera, the density of SSSNPs of all species ranges from 0 to 27,262 per Mb, with the median density of SSSNPs being 323 per Mb, translating into a median density of one SSSNP every 3,098 bp. If the analysis is focused on species that resolve as monophyletic, the density of SSSNPs in 90% of monophyletic species ranges from 20 - 5,624 per Mb. In contrast, in the species that do not resolve as monophyletic, 90% of non-monophyletic species have SSSNPs at a density between 0 - 648 SSSNPs per Mb. The density of

SSSNPs between monophyletic species and non-monophyletic species is significantly different (Figure 5.3., Wilcox signed-rank p-value < 0.001). There is also notable variation between genera, with SSSNP densities ranging from 650 - 7,693 SSSNPs per Mb in genera like *Linaria*, to 0 - 4 SSSNPs per Mb in recently diverged taxa like *Ophrys* species. Detailed information for each genus could be found at the appendices (Table S3, Figure S5.34 - S5.62).

To check whether the observed density of SSSNPs was due to genuine biological signal, as opposed to random shared mutations among samples arising due to the sheer size of the datasets, I randomised the taxon assignment of species within genera and conducted the same analysis. The Wilcox signed-rank test shows that the randomise-labelled data results in a significantly smaller number of SSSNPs, usually approaching zero, compared with the original assignment (average p-value < 0.05, Table S4).

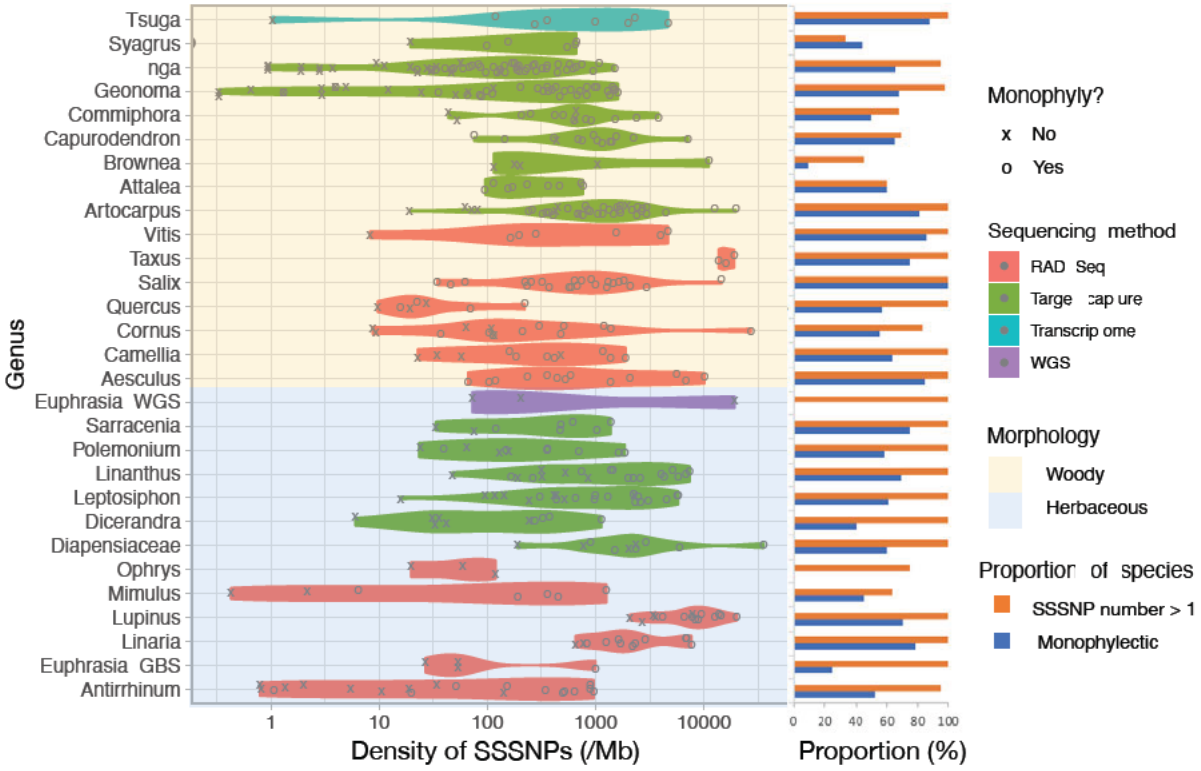


Figure 5.2. Distribution of the Density of Species-Specific SNPs for all genera and the proportion of species that have more than one SSSNPs. The x-axis is standardised by the mean length of a valid sequence of all individuals from each species to eliminate the influence of different lengths of sequence, and is also transformed by log10. The top 16 genera (yellow background) are woody in nature, with the lower genera (blue background) herbaceous. The right panel is the proportion of species that have more than one SSSNPs in each genus (orange bars), and the proportion of species that are monophyletic in that genus (blue bars).

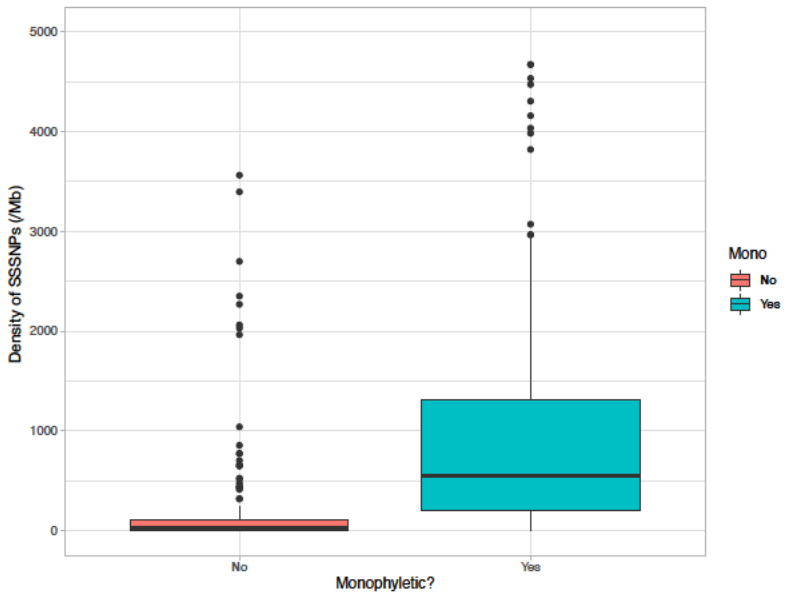


Figure 5.3. The density of SSSNPs comparison between monophyletic species and non-monophyletic species (all genera) (p-value <0.001).

5.4.3. How much data is needed to achieve maximal species discrimination success?

To test the minimal number of loci needed to tell species apart, I down sampled SNP loci in 23 datasets (Table 5.3) and evaluated the point at which optimal species discrimination success was achieved.

Table 5.3. Number of species told apart by full dataset and whether to keep the dataset in down-sampling. N means the dataset is excluded from the analysis, Y means included in this analysis.

Genus	Sequenc ng method	Number of mu t p e-samp ed spec es	Number of monophy et c spec es us ng fu dataset	Se ected for down samp ng?	Note
Leptosiphon	Target Capture	23	14	YES	
Polemonium	Target Capture	12	7	YES	
Artocarpus	Target Capture	42	34	YES	
Capurodendron	Target Capture	20	13	YES	
Geonoma	Target Capture	44	30	YES	
Inga	Target Capture	69	45	YES	
Tsuga	Transcr ptome	8	7	YES	
Aesculus	RAD seq/GBS	13	11	YES	
Antirrhinum	RAD seq/GBS	21	11	YES	
Attalea	Target Capture	15	9	YES	
Camellia	RAD seq/GBS	11	7	YES	
Commiphora	Target Capture	22	11	YES	
Cornus	RAD seq/GBS	18	10	YES	
Dicerandra	Target Capture	10	4	YES	
Linanthus	Target Capture	20	14	YES	
Linaria	RAD seq/GBS	14	11	YES	
Mimulus	RAD seq/GBS	11	5	YES	
Quercus sub Virentes	RAD seq/GBS	7	4	YES	
Salix	RAD seq/GBS	23	23	YES	
Sarracenia	Target Capture	8	6	YES	
Syagrus	Target Capture	18	8	YES	
Taxus	RAD seq/GBS	4	3	YES	
Vitis	RAD seq/GBS	7	6	YES	
Brownea	Target Capture	11	1	No	no spec es cou d be down-samp ed
Diapensiaceae	Target Capture	10	5	No	fam y- eve study
Euphrasia	RAD seq/GBS	4	1	No	no spec es cou d be down-samp ed
Euphrasia WGS	WGS	3	0	No	no spec es cou d be d st ngu shed
Lupinus	RAD seq/GBS	15	10	No	h gh m ss ng data
Ophrys	RAD seq/GBS	4	0	No	no spec es cou d be d st ngu shed

Resampling of these datasets (Figure 5.4, Figure 5.5) showed that the number of species discriminated increases sharply at the start when 100~500 random SNPs are provided. The numbers then reach a plateau when the number of SNPs increases from 500-1,300. At around 3,000 SNPs almost all genera have asymptoted in their levels of species discrimination, and 21/23 genera (91%) have >85% of their distinguishable species distinguished with 3,000 randomly selected SNPs (Table S7). Genera with a small number of multiple-sampled species (such as *Taxus*, *Mimulus*, and *Tsuga*) hit the maximum species discrimination at an earlier stage, i.e., increasing the data volume beyond c. 300 randomly selected SNPs does not increase the number of species resolved as monophyletic. Genera with more than 20 multiple-sampled species, such as *Salix*, *Artocarpus*, *Geonoma*, and *Inga*, show a continued but slower increase from 500-1,000 SNPs. Some genera like *Geonoma, Linaria, and Quercus* continued to slowly increase even after 7,000 SNPs were sampled. The subsampling curve of the individual genera is shown in supplementary Figure S5.2 – S5.24. The comparison of subsampling results between sequencing techniques shows no difference (Figure S5.63).

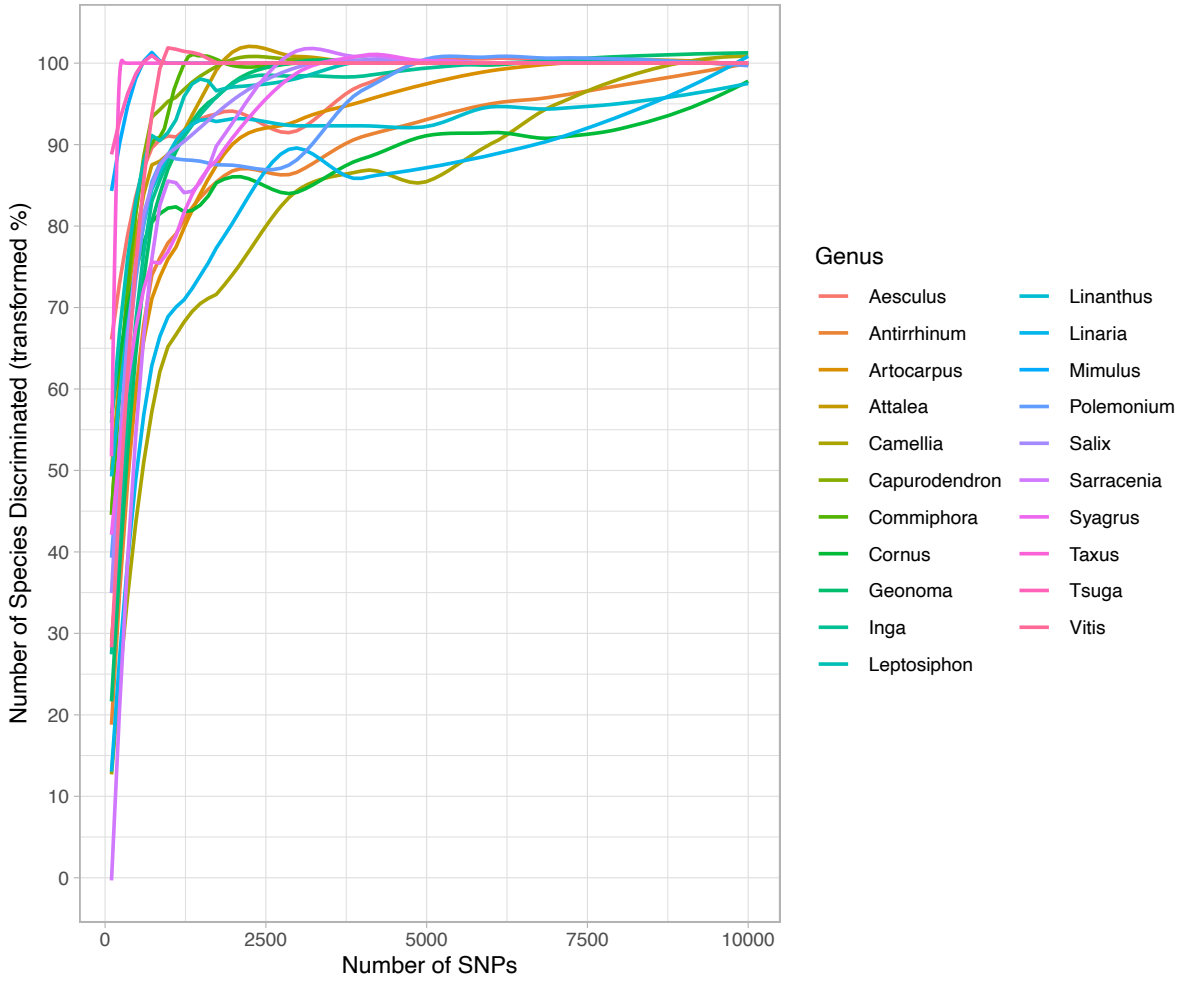


Figure 5.4. The proportion of species discriminated using different numbers of sub-sampled SNPs in each genus. The curve is fitted using the median of each data point (based on 50 replicates per stepwise addition of SNPs). The y-axis is standardised to the maximum number of species that are resolved as monophyletic in each individual dataset using the full original sequence data. The curves thus plot the proportion of the originally discriminated species that are discriminated at increasing random subsamples of SNPs in each genus.

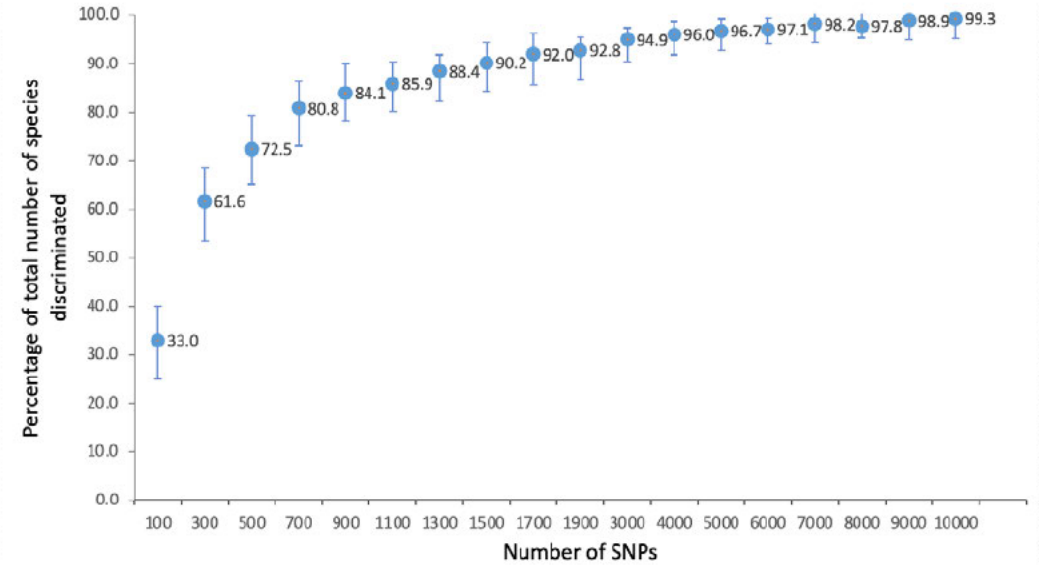


Figure 5.5. The proportion of species discriminated using different numbers of sub-sampled SNPs for all genera congregated. The x-axis is the number of SNPs randomly sampled. The intervals are 200 SNPs from 100 to 1,900 and 1,000 from 3,000 to 10,000. The y-axis is the distribution of the percentage of the number of species discriminated for all genera. The numbers on each datapoint stand for the average of the percentage of the number of species discriminated, with the error bars showing the lower quartile and upper quartile of the percentage of the number of species discriminated for 29 genera.

For the six target capture datasets (Dataset 4, Table 5.4) where individual gene information could be recovered, the subsampling was also done on individual genes (this gene-based analysis was not undertaken for datasets with only small 'loci' like RADseq etc.). The subsampling curves of the individual genera are shown in supplementary Figure S5.25 – S5.33. In five of the 'target capture' genera that had >400 genes samples in their full dataset, four (*Inga, Tsuga, Polemonium, Geonoma*) showed an asymptote in species discrimination at 100 genes or less, whereas in the fifth genus (*Atrocarpus*) there was a slightly more protracted curve with 100 genes on average recovering 26 species as monophyletic, and 200 genes recovering 28 species. The smaller number of genes in the total dataset for a sixth genus (*Commiphora*) did not allow meaningful comparisons to be made.

5.4.4. Do some individual genes show exceptional performance in species discrimination

To assess the performance of single loci in telling species apart, the frequency distribution of the number of species resolved by individual genes was plotted in the six genera from Dataset 4 (Figure 5.6). In five of the six genera, there is an approximation of a normal distribution in the spread of performance of individual loci. In the remaining genus, *Polemonium*, most loci showed low levels of species discrimination with only a few loci being individually able to distinguish more than one species. In all datasets, there are clearly some loci that are much better than others in telling species apart. For example, in the genus *Geonoma*, the locus *LOC105045005* alone resolves 30 species as monophyletic (Table S5. Genes_diagnosability). This is actually bigger a number than that being resolved using all 795 loci (which gave 28 species resolving as monophyletic). In four of the genera, the best single locus can tell a similar amount of species apart as the full dataset Figure 5.6; Table 5.4). Only in *Inga* and *Capurodendron* is there a bigger discrepancy between the efficacy of the best-performing gene and the total dataset. In *Inga*, the best-performing gene distinguished 31 species compared to 45 species from the full dataset; In *Capurodendron*, the best-performing locus distinguished 13 species, compared to 20 species with the full dataset.

Table 5.4. The least number of best performing loci required to match the species discrimination success of the full dataset

Genus	M n mum numbers of good genes for max mum spec es reso ut on	Tota number of oc n the dataset	Tota number of spec es d st ngu shed by severa good oc	Tota number of spec es d st ngu shed by a data (rap d tree)
Artocarpus	1	517	27	29
Capurodendron	7	615	14	13
Geonoma	1	795	30	28
Inga	9	810	44	45
Polemonium	1	360	7	8
Tsuga	1	881	7	7

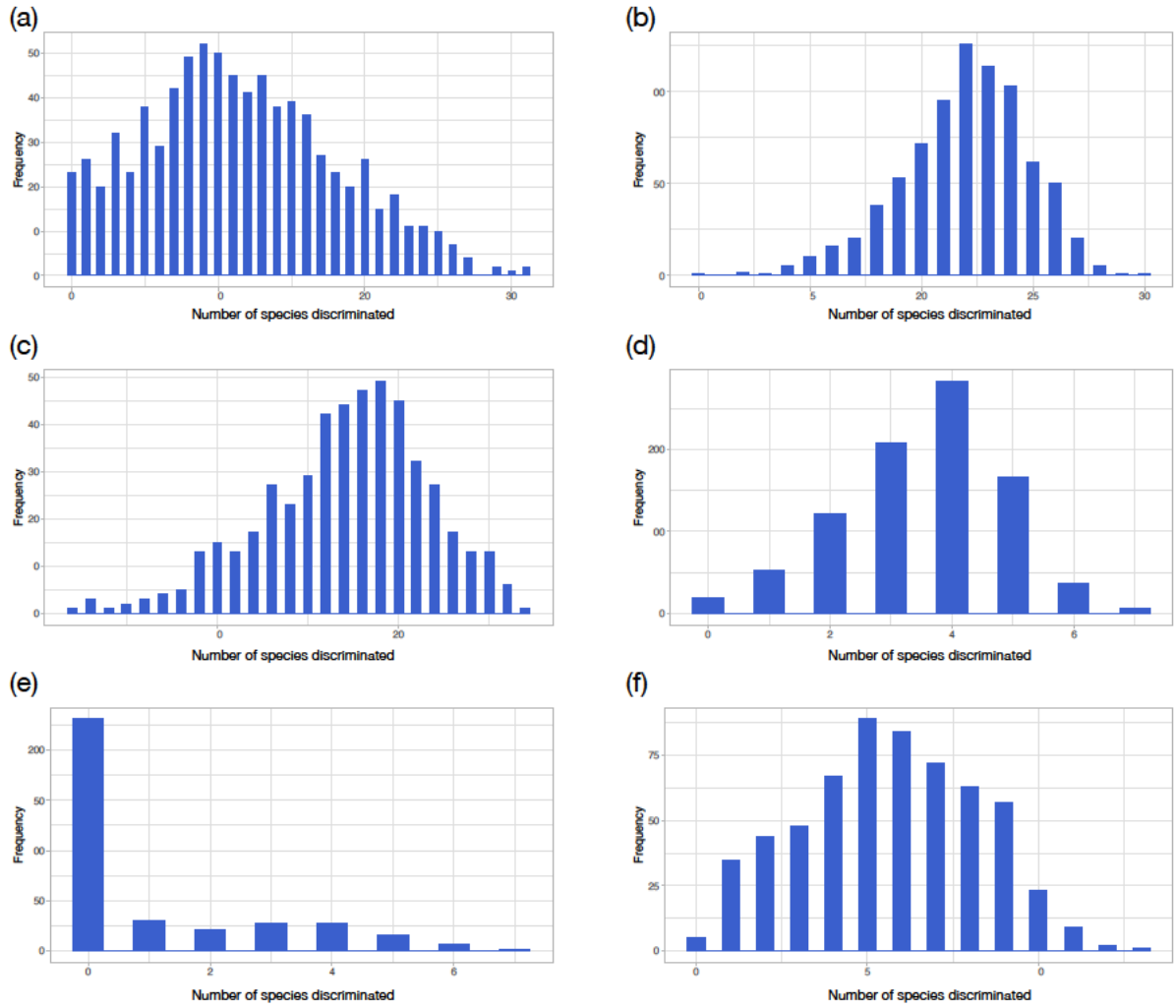


Figure 5.6. Frequency distribution of the number of species discriminated by individual loci. The x-axis is ordered from the genes with the lowest discriminatory power to those with the greatest discriminatory power. (a) *Inga* (b) *Geonoma* (c) *Artocarpus* (d) *Tsuga* (e) *Polemonium* (f) *Capurodendron.*

I then examined the attributes of loci in relation to their discriminatory power. Of the five genera tested (*Capurodendron* was excluded from these tests due variation in the number of individuals per locus), there were some positive correlations between the number of species resolved as monophyletic by a given gene, and its nucleotide diversity across samples in that genus (Table 5.5, Figure 5.5). However, the strength of these correlations varied, and in some cases the relationship was very weak, such as in *Artocarpus* (Table 5.5). The positive correlation between nucleotide diversity and species monophyly for individual genes was statistically significant in only three of the five tested datasets. A similar pattern was detected between the density of SSSNPs and nucleotide diversity (Table 5.5). There was a stronger overall correlation in the relationship between the number of species resolving as monophyletic and the density of SSSNPs,. There is a general pattern of large amounts of variation in these data and imperfect correlations, and hence imperfect predictors of the attributes of the best-performing genes for species discrimination.

Table 5.5. Correlation between nucleotide diversity and density of SSSNPs, and nucleotide diversity and number of species that are monophyletic

Genus	Nuceot de d vers ty and dens ty of SSSNPs (r)	Nuceot de d vers ty and dens ty of SSSNPs (p-va ue)	Dens ty of SSSNPs and number of monophy et c spec es (r)	Dens ty of SSSNPs and number of monophy et c spec es (p-va ue)	Nuceot de d vers ty and number of monophy et c spec es (r)	Nuceot de d vers ty and number of monophy et c spec es (p-va ue)
Artocarpus	0 002	0 940	0 322	<0 001	0 040	0 256
Geonoma	0 154	<0 001	0 206	<0 001	0 159	<0 001
Inga	0 493	<0 001	0 612	<0 001	0 423	<0 001
Polemonium	0 168	0 057	0 395	<0 001	0 159	0 073
Tsuga	0 529	<0 001	0 529	<0 001	0 185	<0 001
Average	0 269		0 413		0 193	

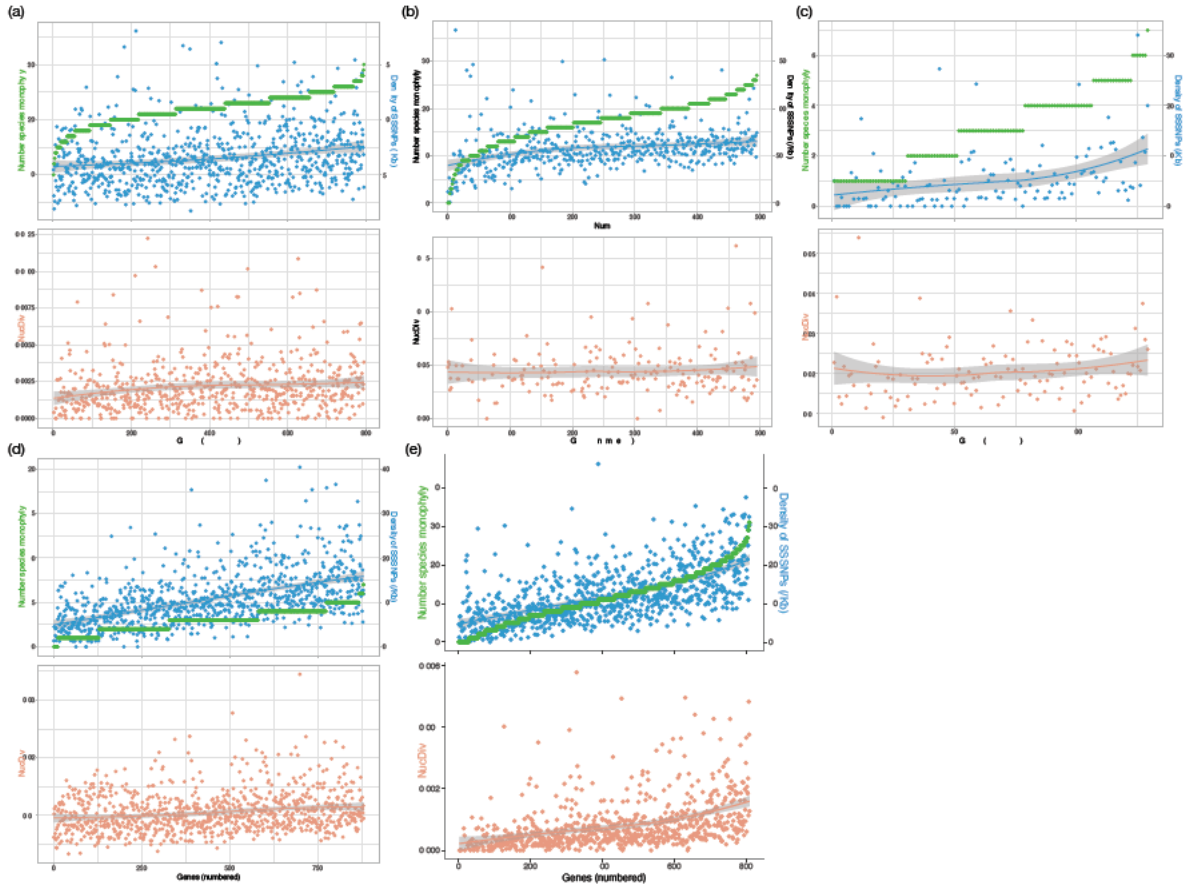


Figure 5.7. The genetic features of individual genes for five genera – (a) *Geonoma* (b) *Artocarpus* (c) *Polemonium* (d) *Tsuga* and (e) *Inga*. The x-axis is ordered by individual loci from the worst to the best-performing genes in terms of the number of species they discriminate. The green dots represent the number of species resolved as monophyletic by each locus; the blue dots represent the density of SSSNPs (per kilobases) for each locus, and the peach dots represent the nucleotide diversity of a given locus across the whole genus.

5.5. Discussion

This study has provided the first synthetic quantification of the species discrimination signal in plant multi-locus nuclear DNA sequence datasets. Although the vast majority of the studies analysed here were not gathered with this purpose in mind, the synthesis is nevertheless informative about general patterns and the nature of differences in the nuclear genome among groups of congeneric plant species.

5.5.1. Monophyly - what proportion of plant species resolve as monophyletic based on multi-locus sequence data?

Of the 1,701 species, evaluated from 149 genera a total of 1,206 species overall resolved as monophyletic (71%). The median percentage of species resolved as monophyletic across all genera was 75%. At the level of individual genera, 37 of the 149 genera (25.8%) had 100% of species resolved as monophyletic, and 75 (50.3%) genera had at least 75% of the species resolving as monophyletic, whereas 14 genera (9.4%) had a monophyly ratio < 25%. There are clearly (and not surprisingly) substantial differences between genera in the proportion of species that resolve as monophyletic. But a key emergent finding from this study is this first quantitative estimate of the proportion of species that do, and do not, resolve as monophyletic based on large amounts of sequence data, over multiple plant genera.

In the future, with increases in sample density compared to this current study, the proportion of species resolving as monophyletic may decrease (e.g. sampling further individuals from further species within these genera may disrupt the monophyly of species compared to these existing sample sets). Likewise, sampling other genera may lead to the monophyly ratio increasing or decreasing depending on the divergence among species in any subsequently sampled genera. However, the current study represents an important starting point, and the sample size of 1701 species from 149 genera is substantial.

Our study did not detect any significant association between growth form, or sequencing technique, with the levels of species monophyly. The results refute the hypotheses that the proportion of species distinguishable is higher in herbaceous than woody plant groups, and that a different level of the proportion of species distinguishable varies among different sequencing methods. However, this observation should be qualified by noting the heterogeneity of study types, and the absence of a balanced sampling design for testing this directly. Direct comparisons of multiple methods, on the same set of samples, across different growth forms would be a much more powerful approach for testing for such associations.

Contextualising the levels of monophyly detected here more widely is difficult given this is the first study to attempt it, and the lack of comparative studies using multi-locus nuclear sequence data to assess plant species monophyly. As noted by Lavin & Pennington (2022), the availability of datasets with multiple individuals per species sampled for large numbers of loci has been a rate-limiting step; an issue which the current chapter has aimed to address. In terms of other data types, based on morphological data, Chrisp & Chandler (1996) noted from a partial survey of two angiosperm families (Fabaceae, Proteaceae) that c 20% of plant species resolved as paraphyletic. Likewise, Rieseberg & Brouillet (1994) (responding to advocates of a

monophyletic species concept), also argued that non-monophyletic species are likely to be common. Several other authors have reviewed the likely circumstances and reasons behind a predicted frequent recovery of non-monophyly among plant species (e.g. Naciri & Linder, 2015; Lavin & Pennington, 2022). Various authors have identified high levels of monophyly from standard barcoding regions (e.g. Fazekas *et al*., 2009; Hollingsworth *et al.,* 2011) but it is difficult to know whether this relates to the attributes of barcoding loci, versus a more general point about the nature of the monophyly of plant species. There are few, if any, direct comparisons between multi-locus nuclear datasets and standard barcoding datasets. In the case of *Inga,* there is a substantial increase in monophyly from 41% standard barcodes (Dexter *et al.,* 2017) to the 65% detected here (Chapters 4,5) using multi-locus nuclear markers (although the sample sets were not the same).

5.5.2. The nature of inter-specific differences among plant species

a) How common are SSSNPs?

In the 29 datasets assessed in detail in this study, SSSNPs were found even in complex groups such as *Salix* (Figure 5.1). 17 genera (58.6% of all) had at least one SSSNP in all studied species. Of the 460 species examined in total, 411 had at least one SSSNP (89.3%). This is an encouragingly high recovery of taxon-specific substitutions, which indicates the potential for further development of molecular diagnostic assays. However, as stressed in Chapter 2, the distribution of these SSSNPs is a relative function of the other samples in each data set, and with subsequent sampling, some SSSNPs may turn out not to be fixed inter-specific differences.

The density of SSSNPs from the meta-analysis conducted here is a noteworthy finding. The overall median density of one SSSNP every 3,098 bp gives a first approximation of how often taxonomically diagnostic SNPs might be encountered (although the range of densities of 0 to 27,262 SSSNPs per Mb is substantial). Understanding the frequency distribution of SSSNPs is interesting in that it provides information on how much random sequence data needs to be generated before encountering a putatively diagnostic SNP (assuming a comprehensive reference database existed). It is also of interest, as a first approximation of what proportion of the genome shows fixed differences between species and how this relates to wider patterns of variation (Malinsky et al., 2015). In the current meta-analysis, the ratio of the median number of SSSNPs per species per genus (222), against the median total number of SNPs per genus (78,719) was 0.3%.

b) How does the presence of SSSNPs relate to species monophyly?

The comparison of SSSNP occurrence mapped against the species that resolved as monophyletic gives the intuitively expected outcome, namely that there is a tendency for monophyletic species to show a high density of SSSNPs (Figure 5.3). This result endorses the hypotheses that species with more fixed differences (species-specific SNP) are prone to be monophyletic. However, what was less intuitive at the outset of this study, was the repeated recovery of SSSNPs in taxa that do not resolve as monophyletic. These occur frequently in the dataset, with a total of 116 species (25.2%) not resolving as monophyletic, but nevertheless containing at least one SSSNP.

These SSSNPs in non-monophyletic species, may be SNPs linked to regions of the genome under selection and thus linked to the cohesiveness of a species (Malinsky et al., 2015). Alternatively, they may just reflect a stochastic process of allele fixation during the history of species divergence (Kimura, 1962), and the SSSNPs consisting of loci that happen to have become fixed at an earlier stage in the history of the species divergence than many other loci in the genome.

5.5.3. How much data are needed to tell plant species apart?

The amount of data required to tell species apart can be evaluated based on either random generation of data, or using targeted knowledge on informative gene regions. Starting from the perspective of randomly gathering multi-locus nuclear sequence data, the current study identified an asymptote in the number of species discriminated at around 2,500 - 3,000 randomly selected SNPs. This is a small fraction of the diversity of the total datasets assessed here (ranging from 0.2 – 49 % of the total number of SNPs in each dataset). Likewise, when the analysis was conducted at the level of individual gene, there was an asymptote at c 50 - 100 loci. This indicates that species can be maximally discriminated with fewer loci than are currently used in many multi-locus plant systematic/evolutionary biology studies. And it is noteworthy that this early asymptote was obtained, as opposed to a slow progressive gain of more species being discriminated when more and more data are added.

The alternative perspective on the 'how much data is needed' question, can be addressed by considering how many 'best-performing' loci are needed to achieve maximal discrimination among species. Here, in four of the six genera evaluated, a single best-performing locus was able to achieve approximately the same level of discriminatory power as the full data (see Table 5.4). In the 'worst' case (e.g. *Inga*) sequencing only 9/810 loci were still able to recover the maximum discriminatory power of the entire dataset (again a very small proportion of the total original data). These results support the hypothesis of increasing the number of markers will increase the species identification success, and there are loci that tell more species apart than others. It also highlights the efficiency benefits of targeting genomic regions of known discriminatory power when the discrimination of species in a particular taxonomic group is of interest.

However, the challenge here is that both the down-sampling of the datasets and the assessment of individual genes' performance in telling species apart are objective to the sequencing techniques. For example, the genomic regions being recovered are usually short for RAD-seq and GBS, and more dispersed and evenly distributed on the genome when compared to full transcriptome sequencing and target capture method. In the scenario of down-sampling by SNPs, the genomic representation of RAD-seq/GBS datasets is effectively reduced while not as effective for full transcriptome sequencing and target capture method, because the selected SNPs still represent similar genomic regions as the full datasets. The sequencing techniques also decide whether the assessment of the 'best-performing' loci could be done or not. With the production of short segments, data from RAD- seq or GBS are naturally not suitable for this analysis because the variation on each segment is too limited to be ranked. There is of course a wider challenge, in that the genomic regions that are maximally efficient in one genus, may not be the same as those in another (Osada et

al., 2005), and until this is further characterised, this has impacts on the development of a universally useful identification system.

5.5.4. Do loci that perform well for telling plant species apart have predictable attributes?

To assess whether there are any attributes of 'the best-performing loci' that are common across different plant groups that might better help understand the selection of nuclear DNA barcoding loci, I examined patterns of diversity across loci. These loci are expected to be or linked to genes that contribute to reproductive isolation (RI) in plants (Rieseberg et al., 2010, Wu, 2001), so they diverge at an early speciation stage when the new sister population started to diverge. The assessment of the genetic diversity of each locus shows that in four out of six tested genera, there is a higher density of SSSNPs/greater recovery of monophyletic species from gene regions with higher genetic diversity.

This trend reflects the correctness of the hypotheses of species that have more fixed differences (species-specific SNP), have a higher nucleotide diversity, and are prone to be monophyletic. However, the relationship is complex, and even in the four datasets with significant correlations, there is a substantial spread in the data (Table 5.5). The most variable genes do not have the most SSSNPs, nor do the most variable genes always tell most of the species apart. Some studies argued that multiple genomic divergence driven factors, such as natural selection, recombination (Pease et al., 2013), and hitchhiking (Fay et al., 2000), are accountable for the imperfect correlation of the genetic diversity and the frequency of fixed variations in a species (Nosil et al., 2012). Further studies are required to evaluate the diversity and the nature of the best-performing gene regions from other plant groups and to increase the sample size for comparison, to better evaluate and understand whether predictions/generalities can be developed about which types of nuclear genes have a predisposition for being useful for telling species apart.

5.6. Conclusions

This meta-analysis provides an initial quantification of the efficacy of mult-locus sequence data from the nuclear genome for plant species discrimination. Of 1,701 species from 149 genera, 71% resolved as monophyletic. This provides the first quantitative assessment of the proportion of plant species that resolve as monophyletic using large scale sequence datasets. Likewise, the study provides the first insights into the frequency distribution of species specific SNPs, and a study of 29 datasets from 21 plant families showed a density of SSSNPs ranging from 0-27,262 per Mb, with a median density of 323 SSSNPs per Mb (a median density of one SSSNP every 3,098 bp). When the data were subsampled to evaluate the minimum amounts of data required for species discrimination, there was an asymptote detected at around 2,500-3,000 randomly selected SNPs from which almost all of the species resolved in the full datasets could be distinguished. Focusing on a selection of six genera, by further selecting just a few best-performing genes from each dataset (<<10), almost all species distinguished using a full complement of genes could be separated, but further work is required to better understand the attributes of loci that are of the maximum value for plant species discrimination.

5.7. Reference

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., . . . Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One,* 3(10), e3376. doi:10.1371/journal.pone.0003376

Bebber, D. P., Carine, M. A., Wood, J. R. I., Wortley, A. H., Harris, D. J., Prance, G. T., . . . Scotland, R. W. (2010). Herbaria are a major frontier for species discovery. *Proceedings of the National Academy of Sciences,* 107(51), 22169. doi:10.1073/pnas.1011841108

Burri, R. Interpreting differentiation landscapes in the light of long-term linked selection. *Evolution Letters*, 2017, 1(3): 118-131.

Cao, M. D., Ganesamoorthy, D., Zhou, C., & Coin, L. J. M. (2018). Simulating the dynamics of targeted capture sequencing with CapSim. *Bioinformatics,* 34(5), 873-874. doi:10.1093/bioinformatics/btx691

Cariou, M., Duret, L., & Charlat, S. (2013). Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. *Ecology and Evolution,* 3(4), 846-852. doi:10.1002/ece3.512

China Plant BOL Group, Li, D. Z., Gao, L. M., Li, H. T., Wang, H., Ge, X. J., . . . Duan, G. W. (2011). Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences,* 108(49), 19641-19646. doi:10.1073/pnas.1104551108

CBOL Plant Working Group (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences,* 106(31), 12794-12797. doi:10.1073/pnas.0905845106

Chase, M. W., Soltis, D. E., Olmstead, R. G., Morgan, D., Les, D. H., Mishler, B. D., . . . Albert, V. A. (1993). Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. *Annals of the Missouri Botanical Garden,* 80(3), 528-580. doi:10.2307/2399846

Crisp, M., & Chandler, G. (1996). Paraphyletic species. Telopea, 6(4), 813-844. doi:10.7751/telopea19963037

Dodsworth, S. (2015). Genome skimming for next-generation biodiversity analysis. *Trends in Plant Science,* 20(9), 525-527. doi:10.1016/j.tplants.2015.06.012

Dong, W. P., Sun, J. H., Liu, Y. L., Xu, C., Wang, Y. H., Suo, Z. L., . . . Wen, J. (2021). Phylogenomic relationships and species identification of the olive genus *Olea* (Oleaceae). *Journal of Systematics and Evolution*. doi:10.1111/jse.12802

Dunning, L. T., & Savolainen, V. (2010). Broad-scale amplification of *matK* for DNA barcoding plants, a technical note. *Botanical Journal of the Linnean Society,* 164(1), 1-9. doi:10.1111/j.1095-8339.2010.01071.x

Eaton, D. A., & Ree, R. H. (2013). Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). *Systematic Biology,* 62(5), 689-706. doi:10.1093/sysbio/syt032

Eaton, D. A. R., Spriggs, E. L., Park, B., & Donoghue, M. J. (2017). Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. *Systematic Biology,* 66(3), 399-412. doi:10.1093/sysbio/syw092

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One,* 6(5), e19379. doi:10.1371/journal.pone.0019379

Fay, J. C. and Wu, C. I. Hitchhiking under positive Darwinian selection. Genetics, 2000, 155(3): 1405-1413.

Fazekas, A. J., Kesanakurti, P. R., Burgess, K. S., Percy, D. M., Graham, S. W., Barrett, S. C., . . . Husband, B. C. (2009). Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? Molecular Ecology Resources, 9 Suppl s1, 130-139. doi:10.1111/j.1755-0998.2009.02652.x

Fitzek, E., Delcamp, A., Guichoux, E., Hahn, M., Lobdell, M., & Hipp, A. L. (2018). A nuclear DNA barcode for eastern North American oaks and application to a study of hybridization in an Arboretum setting. *Ecology and Evolution,* 8(11), 5837-5851. doi:10.1002/ece3.4122

Frazee, A. C., Jaffe, A. E., Langmead, B., & Leek, J. T. (2015). Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics,* 31(17), 2778-2784. doi:10.1093/bioinformatics/btv272

Fu, C. N., Mo, Z. Q., Yang, J. B., Cai, J., Ye, L. J., Zou, J. Y., . . . Gao, L. M. (2022). Testing genome skimming for species discrimination in the large and taxonomically difficult genus *Rhododendron*. *Molecular Ecology Resources,* 22(1), 404-414. doi:10.1111/1755-0998.13479

Gernandt, D. S., Aguirre Dugua, X., Vazquez-Lobo, A., Willyard, A., Moreno Letelier, A., Perez de la Rosa, J. A., . . . Liston, A. (2018). Multi-locus phylogenetics, lineage sorting, and reticulation in *Pinus* subsection *Australes*. *American Journal of Botany,* 105(4), 711-725. doi:10.1002/ajb2.1052

Hale, H., Gardner, E. M., Viruel, J., Pokorny, L., & Johnson, M. G. (2020). Strategies for reducing per-sample costs in target capture sequencing for phylogenomics and population genomics in plants. *Applications in Plant Sciences,* 8(4), e11337. doi:10.1002/aps3.11337

Harvey, M. G., Smith, B. T., Glenn, T. C., Faircloth, B. C., & Brumfield, R. T. (2016). Sequence capture versus restriction site associated DNA sequencing for shallow systematics. *Systematic Biology,* 65(5), 910-924. doi:10.1093/sysbio/syw036

Hebert, P. D., Hollingsworth, P. M., & Hajibabaei, M. (2016). From writing to reading the encyclopedia of life. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 371(1702). doi:10.1098/rstb.2015.0321

Hollingsworth, P. M. (2011). Refining the DNA barcode for land plants. *Proceedings of the National Academy of Sciences, 108*(49), 19451-19452. doi:10.1073/pnas.1116812108

Kates, H. R., Johnson, M. G., Gardner, E. M., Zerega, N. J. C., & Wickett, N. J. (2018). Allele phasing has minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a case study of *Artocarpus*. *American Journal of Botany,* 105(3), 404-416. doi:10.1002/ajb2.1068

Kimura, M. On the probability of fixation of mutant genes in a population. Genetics, 1962, 47: 713-719.

Knapp, H. (2017). Wilcoxon test. United Kingdom: SAGE Publications Ltd.

Kozarewa, I., Armisen, J., Gardner, A. F., Slatko, B. E., & Hendrickson, C. L. (2015). Overview of target enrichment strategies. *Current Protocols in Molecular Biology,* 112, 7 21 21-23. doi:10.1002/0471142727.mb0721s112

Lavin, M., & Pennington, R. (2022). The Implications of Coalescent Conspecific Genetic Samples in Plants. In A. Monro & S. Mayo (Eds.), Cryptic Species: Morphological Stasis, Circumscription, and Hidden Diversity (Systematics Association Special Volume Series, pp. 197-212). Cambridge: Cambridge University Press. doi:10.1017/9781009070553.008

Lepais, O., & Weir, J. T. (2014). SimRAD: an R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches. *Molecular Ecology Resources,* 14(6), 1314-1321. doi:10.1111/1755-0998.12273

Lin, H. Y., Hao, Y. J., Li, J. H., Fu, C. X., Soltis, P. S., Soltis, D. E., & Zhao, Y. P. (2019). Phylogenomic conflict resulting from ancient introgression following species diversification in *Stewartia* s.l. (Theaceae). *Molecular Phylogenetics and Evolution,* 135, 1-11. doi:10.1016/j.ympev.2019.02.018

Linné, C. v. (1753). *Species plantarum* (Vol. 2). Stockholm, Sweden: *Laurentius Salvius*.

Malinsky, M., Challis, R. J., et al. Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science*, 2015, 350(6267): 1493-1498.

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., . . . Turner, D. J. (2010). Target-enrichment strategies for next-generation sequencing. *Nature Methods,* 7(2), 111-118. doi:10.1038/nmeth.1419

Manzanilla, V., Teixidor-Toneu, I., Martin, G. J., Hollingsworth, P. M., de Boer, H. J., & Kool, A. (2022). Using target capture to address conservation challenges: Population-level tracking of a globally-traded herbal medicine. *Molecular Ecology Resources,* 22(1), 212-224. doi:10.1111/1755-0998.13472

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research,* 17(2), 240-248. doi:10.1101/gr.5681207

Mora-Marquez, F., Garcia-Olivares, V., Emerson, B. C., & Lopez de Heredia, U. (2017). ddradseqtools: a software package for in silico simulation and testing of double-digest RADseq experiments. *Molecular Ecology Resources,* 17(2), 230-246. doi:10.1111/1755-0998.12550

Mu, X. Y., Tong, L., Sun, M., Zhu, Y. X., Wen, J., Lin, Q. W., & Liu, B. (2020). Phylogeny and divergence time estimation of the walnut family (Juglandaceae) based on nuclear RAD-Seq and chloroplast genome data. *Molecular Phylogenetics and Evolution,* 147, 106802. doi:10.1016/j.ympev.2020.106802

Naciri, Y., & Linder, H. P. (2015). Species delimitation and relationships: The dance of the seven veils. *Taxon, 64*(1), 3-16. doi:10.12705/641.24

Nicholls, J. A., Pennington, R. T., Koenen, E. J., Hughes, C. E., Hearn, J., Bunnefeld, L., . . . Kidner, C. A. (2015). Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: *Mimosoideae*). *Frontiers in Plant Science,* 6, 710. doi:10.3389/fpls.2015.00710

Nosil, P. and Feder, J. L. Genomic divergence during speciation: causes and consequences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2012, 367(1587): 332-342.

One Thousand Plant Transcriptomes, I. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature,* 574(7780), 679-685. doi:10.1038/s41586-019-1693-2

Osada, N. and Wu, C. I. Inferring the mode of speciation from genomic data: a study of the great apes. *Genetics*, 2005, 169(1): 259-264.

Pease, J. B. and Hahn, M. W. More accurate phylogenies inferred from low-recombination regions in the presence of incomplete lineage sorting. *Evolution*, 2013, 67(8): 2376-2384.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One,* 7(5), e37135. doi:10.1371/journal.pone.0037135

Rieseberg, L. H. and Blackman, B. K. Speciation genes in plants. *Annual Botany*, 2010, 106(3): 439-455.

Rieseberg, L. H., & Brouillet, L. (1994). Are many plant species paraphyletic? Taxon, 43(1), 21-32. doi:10.2307/1223457

Scharmann, M., Wistuba, A., & Widmer, A. (2021). Introgression is widespread in the radiation of carnivorous *Nepenthes* pitcher plants. *Molecular Phylogenetics and Evolution,* 163, 107214. doi:10.1016/j.ympev.2021.107214

Schmickl, R., Liston, A., Zeisek, V., Oberlander, K., Weitemier, K., Straub, S. C., . . . Suda, J. (2016). Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African *Oxalis* (Oxalidaceae). *Molecular Ecology Resources,* 16(5), 1124-1135. doi:10.1111/1755-0998.12487

Simmonds, S. E., Smith, J. F., Davidson, C., & Buerki, S. (2021). Phylogenetics and comparative plastome genomics of two of the largest genera of angiosperms, *Piper* and *Peperomia* (Piperaceae). *Molecular Phylogenetics and Evolution,* 163, 107229. doi:10.1016/j.ympev.2021.107229

So, A. P., Vilborg, A., Bouhlal, Y., Koehler, R. T., Grimes, S. M., Pouliot, Y., . . . Ji, H. P. (2018). A robust targeted sequencing approach for low input and variable quality DNA from clinical samples. *Npj genomic medicine,* 3(1), 2-10. doi:10.1038/s41525-017-0041-4

Stephens, J. D., Rogers, W. L., Heyduk, K., Cruse-Sanders, J. M., Determann, R. O., Glenn, T. C., & Malmberg, R. L. (2015). Resolving phylogenetic relationships of the recently radiated carnivorous plant genus *Sarracenia* using target enrichment. *Molecular Phylogenetics and Evolution,* 85, 76-87. doi:10.1016/j.ympev.2015.01.015

Stephens, J. D., Rogers, W. L., Mason, C. M., Donovan, L. A., & Malmberg, R. L. (2015). Species tree estimation of diploid *Helianthus* (Asteraceae) using target enrichment. *American Journal of Botany,* 102(6), 910-920. doi:10.3732/ajb.1500031

Twyford, A. D. (2014). Testing evolutionary hypotheses for DNA barcoding failure in willows. *Molecular Ecology, 23*(19), 4674-4676. doi:10.1111/mec.12892

Wang, S., Meyer, E., McKay, J. K., & Matz, M. V. (2012). 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nature Methods,* 9(8), 808-810. doi:10.1038/nmeth.2023

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Review Genetics,* 10(1), 57-63. doi:10.1038/nrg2484

Wu, C.-I. The genic view of the process of speciation. *Journal of Evolutionary Biology*, 2001, 14(6): 851-865.

Yao, H., Song, J., Liu, C., Luo, K., Han, J., Li, Y., . . . Chen, S. (2010). Use of ITS2 region as the universal DNA barcode for plants and animals. *PLoS One,* 5(10). doi:10.1371/journal.pone.0013102

Yardeni, G., Viruel, J., Paris, M., Hess, J., Groot Crego, C., de La Harpe, M., . . . Leroy, T. (2021). Taxon-specific or universal? Using target capture to study the evolutionary history of rapid radiations. *Molecular Ecology Resources*. doi:10.1111/1755-0998.13523

Chapter 6 Conclusions and Future Directions

6.1. Summary of findings

In this thesis, I have explored the challenge of improving the resolution of plant DNA barcoding approaches by assessing some of the practical challenges and potential benefits of accessing data from the nuclear genome to support plant species identification.

I first assessed the existing landscape of standard DNA barcoding approaches in plants, and the reasons why standard DNA barcodes often fail to provide species resolution in some groups. I also reviewed why extending DNA barcoding approaches to entire plastomes and complete rDNA sequencing may not lead to materially solving the problem of limited species resolution (Chapters 1 and 2). A key factor limiting progress here is that plastid genomes and rDNA sequence data represent a small number of independent loci, and they themselves have atypical modes of inheritance and dynamics of evolution. Given these limitations, Chapters 1 and 2 highlight the potential and importance of assessing the nuclear genome as a source of multiple independent characters to improve levels of species discrimination in plants

I secondly evaluated some of the key issues to consider in utilizing the nuclear genome for plant species identification (Chapter 2). These include noting how different sampling strategies of taxa can influence measures of species discrimination, and stressing the importance of studies that generate sequence data from multiple nuclear loci, from multiple individuals of multiple congeneric species. Such sample sets are of critical importance in evaluating the resolving power of different approaches. I also noted the expected heterogeneity in species discrimination success depending on the biological attributes of different plant groups, and noted the added complexities when dealing with large recently radiated genera, and also with biologically complex groups in general. Chapter 2 also briefly explored the potential for different genome sampling strategies (e.g. genome skimming, target capture, RAD-sequencing, genotyping by sequencing, transcriptome sequencing) to recover different proportions of coding versus non-coding, and single-copy versus repetitive DNA regions, as well as varying levels of overlap in the loci recovered from different samples. Other technical factors that have the potential to influence recovered patterns of sequence variation within and among species include sequence quality and coverage, and how this impacts the assembly of raw sequencing reads into reliable files for analysis. Another variable that may impact on downstream use of sequence data is how heterozygosity has been treated, as some studies (and associated datasets) mask out potential heterozygosity, and instead call the most common variant in the consensus sequence. Some of these issues are easy to address and accommodate in meta-analyses (for instance by selecting datasets to include or exclude based on the sampling criteria they used). Others, particularly those which involve how other researchers have generated and assembled sequence files are more difficult to standardise, and at least for the analyses involved in this thesis, there is an unavoidable set of variations among studies which is likely to have added some noise to the overall findings.

In Chapter 3, I outlined a new pipeline for handling and analysing nuclear sequence data to tell plant species apart. This pipeline (named NucBarcoder), supports the

searching and selection of datasets for evaluation, the generation of phylogenetic trees and extraction of the proportion of species that resolve as monophyletic, and the distribution of species-specific SNPs and SNPs which show significant frequency differences among species. The pipeline also supports random subsampling of full species datasets, to enable a bootstrap resampling of subsets of loci, at user-specified intervals, followed by phylogenetic tree reconstruction and recording levels of species monophyly at each sub-sampled interval. This enables an assessment of the minimum number of randomly selected loci required to achieve the same levels of resolution as the full datasets. The final stages of the pipeline focus on characterising the best-performing gene regions from each data-set, and evaluating the sequence diversity and characteristics of these loci. All code scripts, workflows, and an example to run the pipeline are available at

https://github.com/Hazelhuangup/Species_specific_alleles_analysis.

In Chapter 4, the scripts developed in Chapter 3, were applied to the legume genus *Inga*. This classic 'difficult test case' for plant DNA barcoding, showed that when using 810 genes and 205,871 SNPs from a target capture dataset, 65% of the 69 species with multiple sampled individuals resolved as monophyletic (a substantial improvement in resolution compared to previous plastid barcoding studies). When these data were subsampled, a random selection of 70 genes or 2500 SNPs, or a combination of nine 'best-performing' genes could achieve levels of species discrimination similar to the full dataset.

The scripts and approaches described in Chapter 3, and applied in Chapter 4, were then utilised in a large-scale meta-analysis in Chapter 5. In this synthesis, a total of 71% of the investigated species from a sample of 149 genera and 1,701 multiple-sampled species resolved as monophyletic. An investigation into the distribution of species-specific SNPs (SSSNPs) from 29 datasets representing 21 plant families revealed a density of SSSNPs from 0 to 27,262 per Mb, with a median density of 323 SSSNPs per Mb. This meta-analysis provides a striking figure of a median density of one SSSNP every 3,098 bp across these 29 different plant datasets. As found for *Inga* in Chapter 4, smaller subsets of the data could be used to resolve similar amounts of species as the full datasets, and this translated to an asymptote in species discrimination at around 2,500-3,000 randomly selected SNPs. Furthermore, in a detailed investigation of six genera, between one and nine pre-selected genes were able to recover equivalent levels of species discrimination compared to several hundred genes from the full datasets.

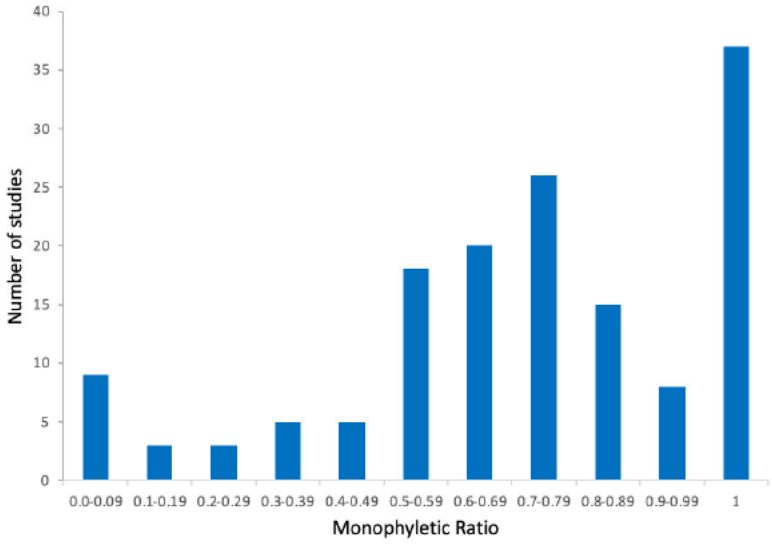


Figure 6.1. Monophyletic Ratio summary histogram. On the X-axis consisted of 11 blocks of the ratio of species in that genus resolving as monophyletic, and on the Y-axis are the numbers of studies falling into each of those categories.

6.2. Caveats and additional desirable work

A strength of the work undertaken in this thesis is the efficient use of existing data to tackle the highly topical question of telling plant species apart with DNA, and this has enabled some broad generalisations and summaries to be made in terms of the number of species that resolve as monophyletic, and the nature of the data that underlie the plant species differences. However, one compromise that this has involved, has been that the analyses have been dependent on the strengths and weaknesses of the original data that were available for analysis.

6.2.1. Weaknesses of available data and limitations of the analyses

A general challenge for the reuse of data from the published literature is uncertainty around some steps taken by researchers in the process leading up to the published datasets. For assessing levels of species discrimination with genetic data, one very practical challenge is how 'misplaced' specimens are dealt with. In this thesis, I checked the datasets for inclusion against circularity (e.g. not including studies for assessing the discriminatory power of datasets for telling species apart, if the data themselves were used to identify the samples in the first place). This is a conceptually straightforward issue, but in practice, the implementation may be more complex. Most studies start with an a priori set of sample identifications, which may then be updated if the sequencing detects major discrepancies which highlight obvious errors. This is sensible routine practice. However, what should be avoided is an a priori assumption that multiple samples per species 'should' group together, and hence the identity of any 'misplaced' samples being wrong, and hence needing to be 'cleaned up' to match their phylogenetic placement. I did not include any datasets where there was evidence of this practice occurring, but given the large-scale nature of the datasets assessed, I was unable check directly with all corresponding authors how any sample reidentifications were handled.

Another general weakness/challenge is the lack of all datasets being available in standardised, easy-to-analysis format, or in some cases, the primary data not being available at all. This results in the discrepancy between the number of datasets (149) that were analysed for species monophyly, and those that were assessed for the distribution of SSSNPs (29). Likewise, even where datasets were available for re-analyses, it was not always clear what steps had been taken in the assembly of datasets, and for instance, how issues such as heterozygosity had been handled. To explore the implications of this particular issue (heterozygosity), I undertook some sensitivity analyses where I compared the levels of species discrimination for datasets in which heterozygosity information was retained, and where I then in turn masked the heterozygosity to the most common variant. The findings of this were the number of monophyletic clades doesn't change much, if any, when including or excluding the heterozygosity information (Table S6).

Additional sensitivity analyses I undertook included the testing impact of different nucleotide substitution models and phylogeny-building software (iqtree2 and RAxML NG) on the number of monophyletic clades. No obvious impact was found (Table S6).

A final, and important limitation of the meta-analysis, is that the scale of the available data remains the major rate-limiting step. As noted here and by others (e.g. Lavin &

Pennington 2022), there is a pressing need to increase the number of datasets that sample multiple individuals from multiple congeneric species for multiple unlinked nuclear markers.

6.2.2. Additional desirable work

At the outset of this thesis, my work plan included a combination of meta-analysis/synthesis of existing data, as well as the generation of new data. In terms of planning new data production, I planned to target sampling of multiple individuals from multiple species from the genus *Salix* in the UK, and to use these data as a direct comparison with the good document evidence for sharing of standard DNA barcodes among species in the genus (Percy et al., 2014; Twyford, 2014). After undertaking desk-based research on the genus in the UK, and preliminary fieldwork, I selected a set of six diploid willow species (*S. lapponum, S. lanata, S. reticulata, S. herbacea, S. myrsinites, S. arbuscula*) that occur at montane sites in Scotland. I selected 10 individuals per species, including samples from populations where the species grows in relative allopatry (spatially separating from other *Salix* species they may hybridise with), as well as populations where the species grow in sympatry with other hybridising willows. This sample set was designed to test the nature of species boundaries among a group of *Salix* species that were subject to different likelihoods of contemporary hybridisation and inter-specific gene flow. The selected populations and species were sampled from existing tissue collections at the Royal Botanic Garden Edinburgh, their DNA extracted by Dr. Michelle Hart, and the samples were sent for whole genome sequencing at the Wellcome Sanger Institute as part of the Darwin Tree of Life Project (https://www.darwintreeoflife.org/). However, a combination of restricted laboratory access due to the COVID-19 pandemic, and then further restrictions to building access due to a major flood in 2021, meant that there were delays in being able to ship and process the samples. By the time the samples were shipped, and then sequenced, it was not possible to analyse the data in time for submission in this thesis. Clearly, the incorporation of sample sets that were explicitly designed to address the questions tackled in this thesis would strengthen the meta-analysis further.

An additional area of desirable additional work includes extending the range of analytical methods that are deployed for assessing the discriminatory power of multi-locus nuclear DNA sequence data for plant species discrimination. In this thesis, I focused on monophyly (a widely used, and well-understood mechanism of distinguishing among species with genetic data), and the distribution of taxonomically informative SNPs as a highly automated unit of data. One obvious area of future work would be to evaluate other methods for telling plant species apart with multi-locus sequence data. Potential areas to explore include model-based approaches such as the multispecies coalescent (MSC) models which are designed to accommodate incongruent information from multiple gene trees and bypass the constraints of a concatenation of multiple alignments, where all genes from a sample set are assumed to follow a single common genealogy (Quatela and Oxelman, 2022). However, the MSC approach remains computationally intensive to apply (Rannala et al., 2020), and as such difficult to scale to a large-scale meta-analysis. An alternative approach that is easier to scale is the use of alignment-free approaches such as k-mer analyses, which can be used for computationally efficient repeated comparisons and matches of

sequences of a given length (k-mer) to the best match in a reference database (Wood et al., 2014, V. Bhange, 2018, Sarmashghi et al., 2019, Bohmann et al., 2020). Of particular interest here is the recent development of tools such as the assembly-free and alignment-free tool, Skmer, which can be used to treat a set of samples as both reference samples and query samples, and to compute genomic distances between samples to identify their closest match and hence taxonomic assignment. Such approaches offer great opportunities for effectively evaluating the species discrimination potential in multi-locus nuclear barcode datasets (Sarmashghi et al., 2019).

6.3. Outstanding priorities for developing nuclear DNA barcoding approaches for plants

To further develop the concept of nuclear DNA barcoding in plants, I finish by identifying a set of future priority actions, and research topics. The need to develop improved methods of telling plant species apart is strong, and hence there is a pressing need for community collaboration to address outstanding data and infrastructure needs.

Priority areas include:

- Targeted studies aiming to generate multi-locus sequence data from multiple individuals from multiple congeneric species, ideally with comprehensive sampling of all known species in a genus.
    - These studies will provide sample sets suitable for robust insights into absolute levels of discriminatory power in individual genera, and also provide a sample set for comparing gains in discriminatory power with standard barcodes and extended plant barcodes (plastomes/rDNA arrays)
- Generation of high quality reference genomes, and whole genome resequencing studies providing comprehensive coverage of inter-specific sequence differences for phylogenetically disparate genera of conservation, ecological or socioeconomic importance.
    - Having comprehensive genome coverage as well as sample coverage in a range of targeted phylogenetically disparate genera allows the nature of inter-specific differences to be directly measured (rather than estimated). Complete genome sequences will also allow *in silico* assessments of how subsamples of the genome would differ from the full sequences, informing the types of an assay that are likely to be most appropriate. This could include *in silico* assessments for instance, of the difference in the performance of universal baits (such as the angiosperm 353 kit) against the use of genome skimming approaches at variable levels of coverage, and/or the use of taxon-specific baits. Reporting of standardised meta-data, and archiving of data and meta-data in a fashion explicitly designed to allow the reanalysis and meta-analysis of the use of multi-locus sequence datasets for plant species discrimination
    - There is a pressing need to enable synergies among individual studies to maximise the efficiency of (re)use of multi-locus sequence data from plants. The curve of data generation is currently at an early stage, and it is likely there will be a rapid growth in the production of relevant datasets in the next few years. Effectively guiding the formatting and meta-data standards for such data could result in significant scientific benefits (in terms of plant species identification) at minimal additional cost, based on highly efficient reuse of data.
- Further optimisation of pipelines and analytical methods for routinely and robustly quantifying the degree to which multi-locus nuclear sequence data tell

plant species apart, and which subsets of the data give rise to the most effective diagnostic assays

    - As the volume of data continue to grow, there is an increased need for highly effective data evaluation and analysis methods. The optimisation of the BOLD database for analysis of CO1 DNA barcodes has been instrumental in understanding the distribution of barcode data among animal taxa. Likewise, the Plant and Fungal Tree of Life Data Explorer (Baker et al., 2022), is an example of a data portal designed to house complex multi-locus data for plants, that is well suited for regular updating of data, and reanalysis of data in light of additional data depositions. This portal focuses on phylogenomic analyses, and a similar informatics framework/portal for plant species discrimination would be extremely useful.

- Development of efficient multi-locus assays using nuclear sequence data, to start trialling how such assays can be used in routine species identification applications.

    - In parallel with efforts to develop a new standardised approach for 'plant DNA barcoding 2.0', it is important to already pursue the development of workable taxon-specific multi-locus barcoding assays, to begin the process now, of identifying and overcoming practical challenges to implementation. This could include developing optimal assays for efficient specimen sample screening (e.g. as was recently done for *Anopheles* species by Makunin *et al.* (Makunin et al., 2021)). Another key priority is to start the process of considering how multi-locus reference datasets could be used to support environmental DNA (eDNA) and mixed sample meta-barcoding studies (Taberlet et al., 2018), which brings the additional challenge of simultaneously handling the resulting mixture of data from multiple loci from multiple species, and as such represents an interface between meta-barcoding and meta-genomic studies.

This list of priorities is not intended to be exhaustive, but rather to identify some of the key high-level priorities to take forward the development and use of multi-locus nuclear DNA barcoding of plants. A key factor that will accelerate the success of such an approach is a collaboration among research groups. This thesis has benefited greatly from data sharing and research collaborations. Major breakthroughs around standardised approaches for using DNA for species identification in DNA barcoding have in the past come from consortia working together to assemble and co-analyses large datasets to understand and refine optimal solutions for species discrimination (China Plant Working Group et al., 2011; Plant Working Group CBOL, 2009; Pawlowski et al., 2012, Schoch et al., 2012). Such an approach would be desirable for developing the next wave of plant DNA barcoding tools and ultimately a standardise version of plant DNA barcode 2.0.

6.4. Reference

Baker, W. J., Bailey, P., Barber, V., Barker, A., Bellot, S., Bishop, D., . . . Forest, F. (2022). A comprehensive phylogenomic platform for exploring the angiosperm tree of life. Systematic Biology, 71(2), 301-319. doi:10.1093/sysbio/syab035

Bohmann, K., Mirarab, S., Bafna, V., & Gilbert, M. T. P. (2020). Beyond DNA barcoding: The unrealized potential of genome skim data in sample identification. *Molecular Ecology*. doi:10.1111/mec.15507

China Plant BOL Group, Li, D. Z., Gao, L. M., Li, H. T., Wang, H., Ge, X. J., . . . Duan, G. W. (2011). Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences,* 108(49), 19641-19646. doi:10.1073/pnas.1104551108

CBOL Plant Working Group (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences,* 106(31), 12794-12797. doi:10.1073/pnas.0905845106

Lavin, M., & Pennington, R. (2022). The Implications of Coalescent Conspecific Genetic Samples in Plants. In A. Monro & S. Mayo (Eds.), Cryptic Species: Morphological Stasis, Circumscription, and Hidden Diversity (Systematics Association Special Volume Series, pp. 197-212). Cambridge: Cambridge University Press. doi:10.1017/9781009070553.008

Makunin, A., Korlevic, P., Park, N., Goodwin, S., Waterhouse, R. M., von Wyschetzki, K., . . . Lawniczak, M. K. N. (2021). A targeted amplicon sequencing panel to simultaneously identify mosquito species and Plasmodium presence across the entire Anopheles genus. *Molecular Ecology Resources*. doi:10.1111/1755-0998.13436

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., . . . de Vargas, C. (2012). CBOL protist working group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biol,* 10(11), e1001419. doi:10.1371/journal.pbio.1001419

Percy, D. M., Argus, G. W., Cronk, Q. C., Fazekas, A. J., Kesanakurti, P. R., Burgess, K. S., . . . Graham, S. W. (2014). Understanding the spectacular failure of DNA barcoding in willows (Salix): does this result from a trans-specific selective sweep? *Molecular Ecology,* 23(19), 4737-4756. doi:10.1111/mec.12837

Quatela A-S, Oxelman B (2022) Chapter 17. Species delimitation. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. doi: 10.3897/avb.e98875

Rannala, B., Edwards, S. V., Leaché, A., & Yang, Z. (2020). The multi-species coalescent model and species tree inference. *Phylogenetics in the Genomic Era*.

Sarmashghi, S., Bohmann, K., MT, P. G., Bafna, V., & Mirarab, S. (2019). Skmer: assembly-free and alignment-free sample identification using genome skims. *Genome Biology,* 20(1), 34. doi:10.1186/s13059-019-1632-4

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., . . . Consortium, F. B. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America,* 109(16), 6241-6246. doi:10.1073/pnas.1117018109

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). *Environmental DNA : for biodiversity research and monitoring* (First edition. ed.). Oxford: Oxford University Press.

Twyford, A. D. (2014). Testing evolutionary hypotheses for DNA barcoding failure in willows. *Molecular Ecology,* 23(19), 4674-4676. doi:10.1111/mec.12892

Snehal, V. B. (2018). K-mer profiling for bacterial identification. *Helix,* 8(5), 4007-4009. doi:10.29042/2018-4007-4009

Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology,* 15(3), R46. doi:10.1186/gb-2014-15-3-r46

Appendices

The **excel file** appended contains supplementary tables, Full description of the data is included in each file.

S1. Dataset_1_for_MR: 149 plant groups included for Monophyletic Ratio calculation (Dataset 1)

S2. Dataset_2_for_SSSNPs: 29 plant groups included for Calculating the abundance and density of species-specific SNPs (SSSNPs) (Dataset 2)

S3. SSSNPs_density_all_spps: All multiple-sampled species list and their density of SSSNPs and Monophyletic status (with full dataset)

S4. Random_label_p-value: p-value of the number of SSSNPs comparison between the orginal label and random label

S5. Genes_diagnosability-Geonom: The number of multiple-sampled species discriminated by individual gene for *Inga* dataset

S6. Sensitivity_test: The impact of nucleotide substitution models, including heterozygosity information, and phylogeny building software on the number of multiple-sampled species being told apart

S7. All_genera_subsampling: The number of multiple-sampled species discriminated by different size of subsamples for 23 genera (Dataset 4)

S8. Results_for_Inga: Detailed results for *Inga* analysis

The **pdf file** appended contains supplementary figures:

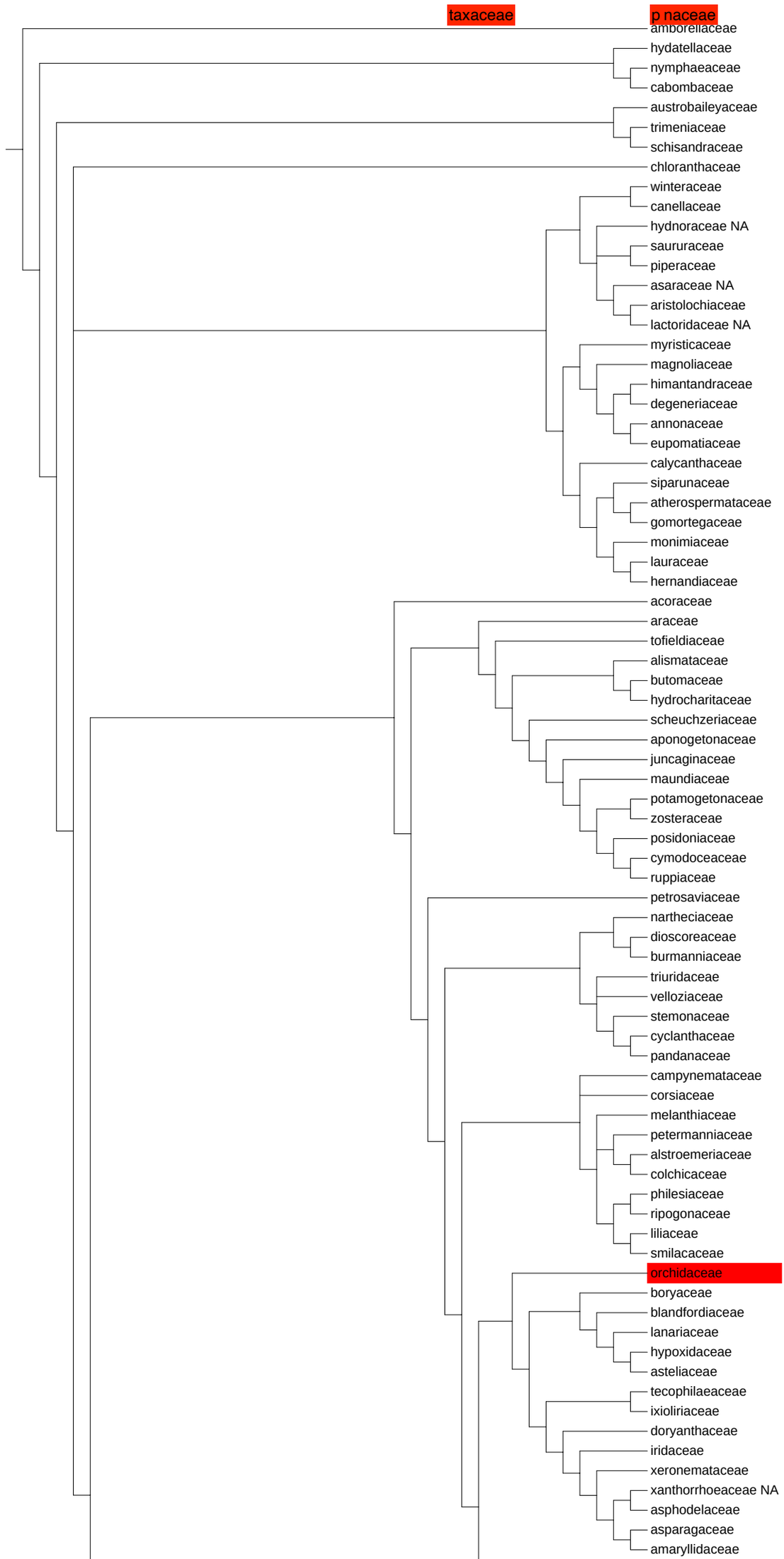
Figure S5.1. Families the 29 genera (Dataset 2) belong to on the Angiosperm tree of life (highlighted in red). One Gymnosperm genera Taxus is drawn on top.

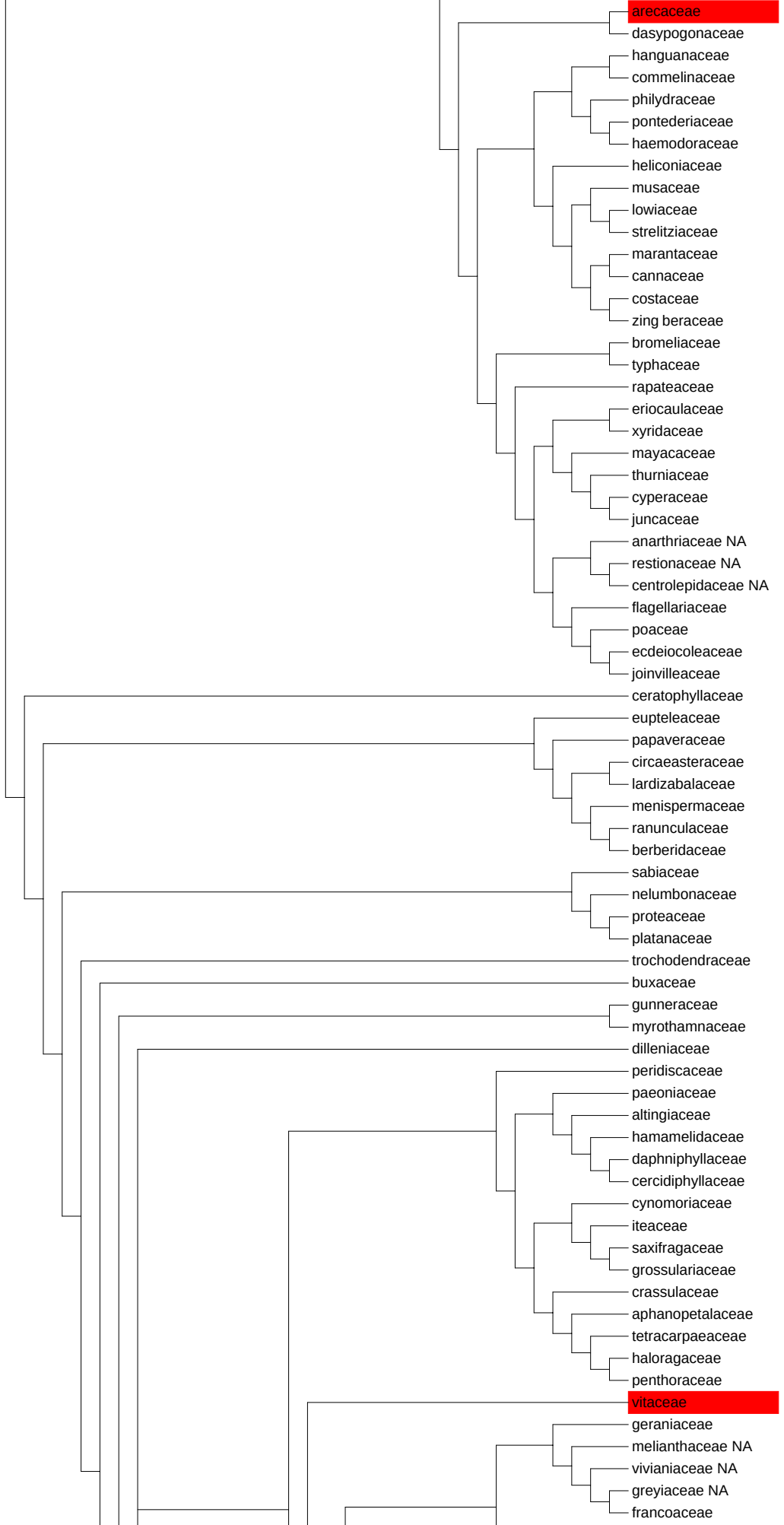
Figure S5.2 – S5.24. The subsampling curve of the individual genera by SNPs. The name of each individual dataset is on the top of each figure.

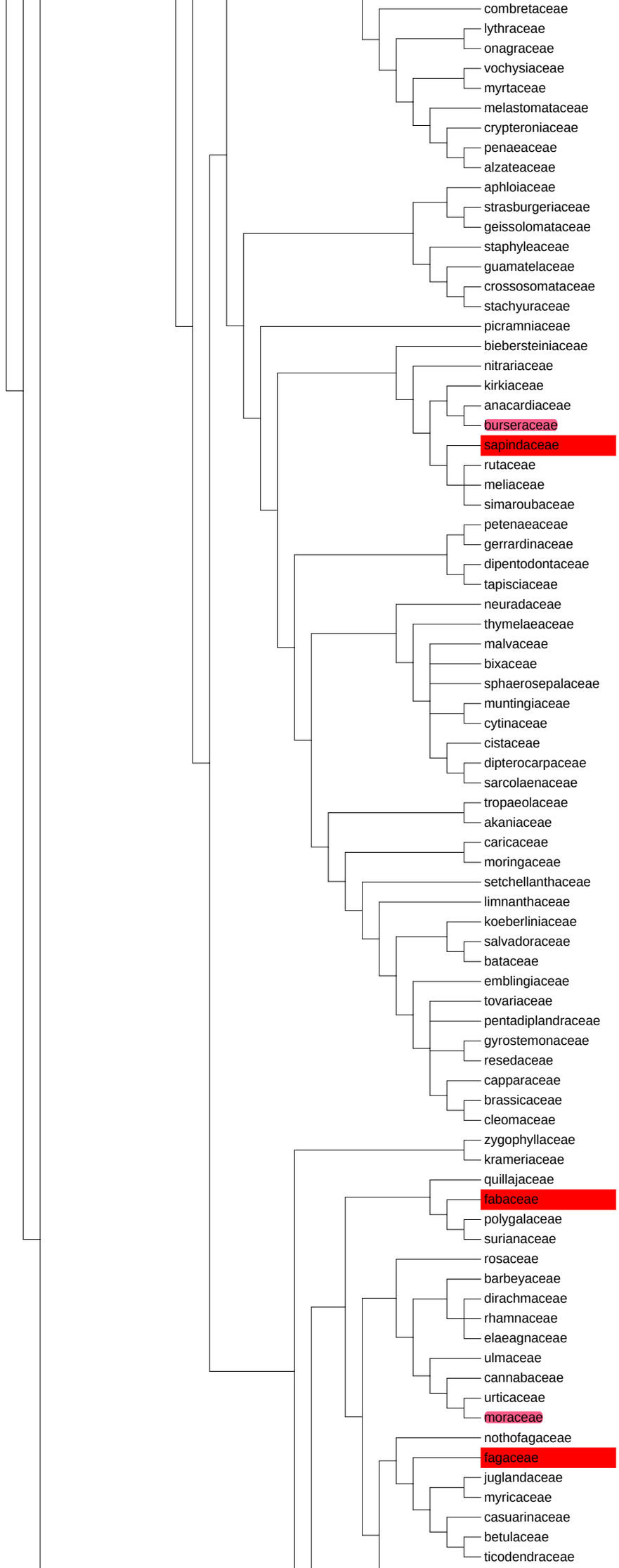
Figure S5.25 – S5.33. The subsampling curves of the individual genera by genes/DNA segments. The name of each individual dataset is on the top of each figure.

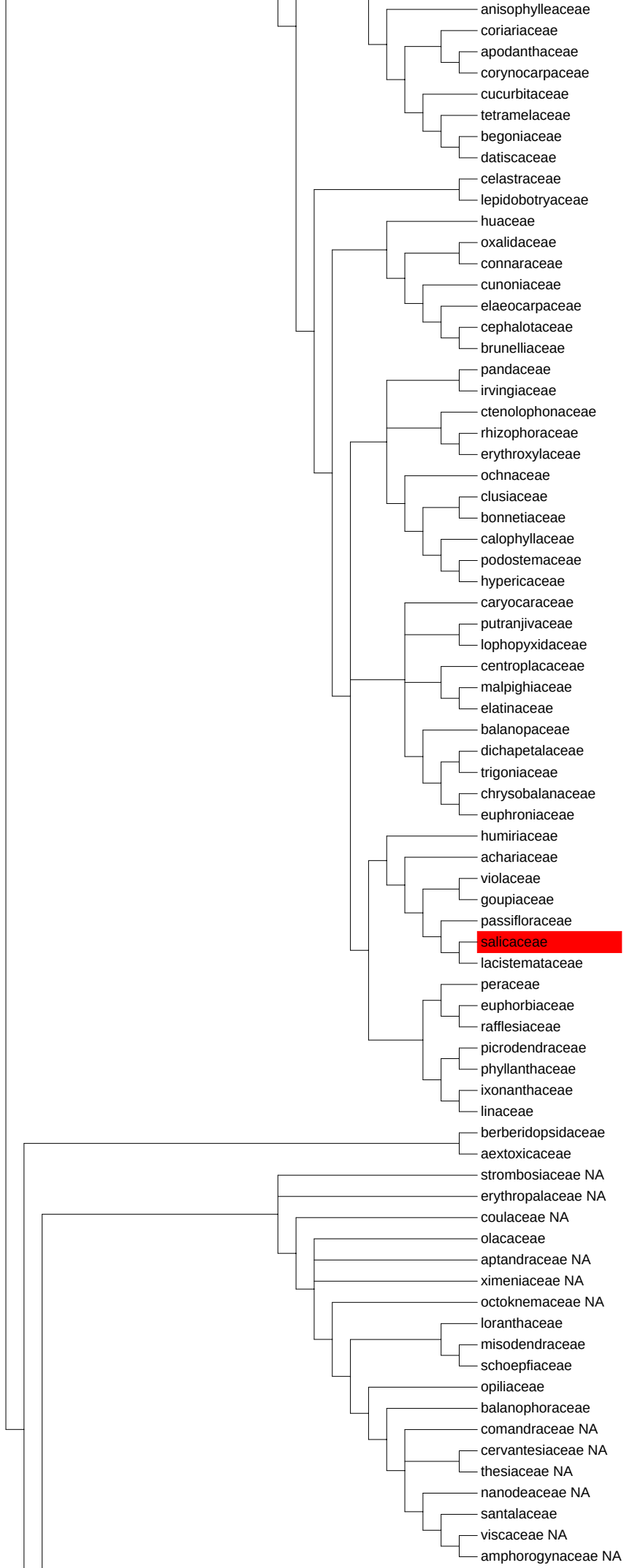
Figure S5.34 – S5.62. Distribution of the number of Species-Specific SNPs for multi-sampled species for all genera analysed (Dataset 2)

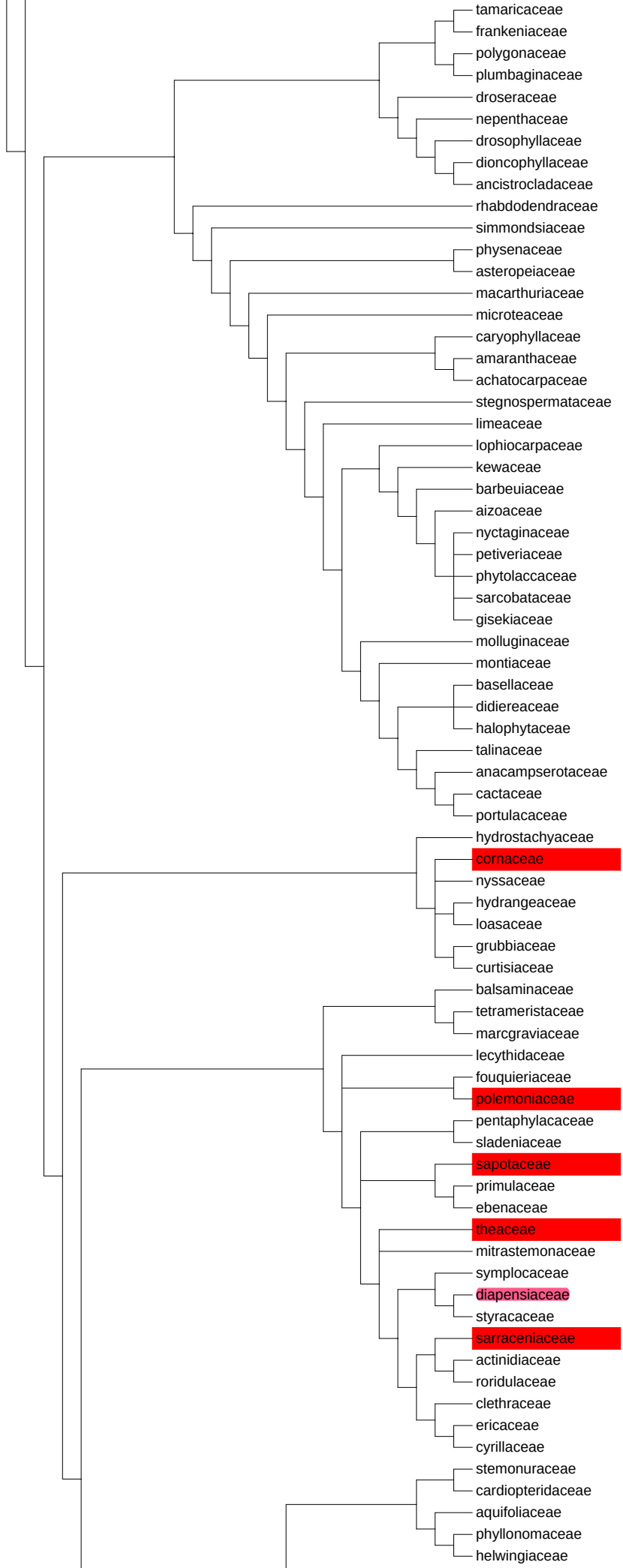
Figure S5.1. Families the 29 genera (Dataset 2) belong to on the Angiosperm tree of life (highlighted in red).
One Gymnosperm genera Taxus is drawn on top.


taxaceae
p naceae
amborellaceae
hydatellaceae
nymphaeaceae
cabombaceae
austrobaileyaceae
trimeniaceae
schisandraceae
chloranthaceae
winteraceae
canellaceae
hydnoraceae NA
saururaceae
piperaceae
asaraceae NA
aristolochiaceae
lactoridaceae NA
myristicaceae
magnoliaceae
himantandraceae
degeneriaceae
annonaceae
eupomatiaceae
calycanthaceae
siparunaceae
atherospermataceae
gomortegaceae
monimiaceae
lauraceae
hernandiaceae
acoraceae
araceae
tofieldiaceae
alismataceae
butomaceae
hydrocharitaceae
scheuchzeriaceae
aponogetonaceae
juncaginaceae
maundiaceae
potamogetonaceae
zosteraceae
posidoniaceae
cymodoceaceae
ruppiaceae
petrosaviaceae
nartheciaceae
dioscoreaceae
burmanniaceae
triuridaceae
velloziaceae
stemonaceae
cyclanthaceae
pandanaceae
campynemataceae
corsiaceae
melanthiaceae
petermanniaceae
alstroemeriaceae
colchicaceae
philesiaceae
ripogonaceae
liliaceae
smilacaceae
orchidaceae
boryaceae
blandfordiaceae
lanariaceae
hypoxidaceae
asteliaceae
tecophilaeaceae
ixioliriaceae
doryanthaceae
iridaceae
xeronemataceae
xanthorrhoeaceae NA
asphodelaceae
asparagaceae
amaryllidaceae


arecaceae
dasypogonaceae
hanguanaceae
commelinaceae
philydraceae
pontederiaceae
haemodoraceae
heliconiaceae
musaceae
lowiaceae
strelitziaceae
marantaceae
cannaceae
costaceae
zing beraceae
bromeliaceae
typhaceae
rapateaceae
eriocaulaceae
xyridaceae
mayacaceae
thurniaceae
cyperaceae
juncaceae
anarthriaceae NA
restionaceae NA
centrolepidaceae NA
flagellariaceae
poaceae
ecdeiocoleaceae
joinvilleaceae
ceratophyllaceae
eupteleaceae
papaveraceae
circaeasteraceae
lardizabalaceae
menispermaceae
ranunculaceae
berberidaceae
sabiaceae
nelumbonaceae
proteaceae
platanaceae
trochodendraceae
buxaceae
gunneraceae
myrothamnaceae
dilleniaceae
peridiscaceae
paeoniaceae
altingiaceae
hamamelidaceae
daphniphyllaceae
cercidiphyllaceae
cynomoriaceae
iteaceae
saxifragaceae
grossulariaceae
crassulaceae
aphanopetalaceae
tetracarpaeaceae
haloragaceae
penthoraceae
vitaceae
geraniaceae
melianthaceae NA
vivianiaceae NA
greyiaceae NA
francoaceae


combretaceae
lythraceae
onagraceae
vochysiaceae
myrtaceae
melastomataceae
crypteroniaceae
penaeaceae
alzateaceae
aphloiaceae
strasburgeriaceae
geissolomataceae
staphyleaceae
guamatelaceae
crossosomataceae
stachyuraceae
picramniaceae
biebersteiniaceae
nitrariaceae
kirkiaceae
anacardiaceae
burseraceae
sapindaceae
rutaceae
meliaceae
simaroubaceae
petenaeaceae
gerrardinaceae
dipentodontaceae
tapisciaceae
neuradaceae
thymelaeaceae
malvaceae
bixaceae
sphaerosepalaceae
muntingiaceae
cytinaceae
cistaceae
dipterocarpaceae
sarcolaenaceae
tropaeolaceae
akaniaceae
caricaceae
moringaceae
setchellanthaceae
limnanthaceae
koeberliniaceae
salvadoraceae
bataceae
emblingiaceae
tovariaceae
pentadiplandraceae
gyrostemonaceae
resedaceae
capparaceae
brassicaceae
cleomaceae
zygophyllaceae
krameriaceae
quillajaceae
fabaceae
polygalaceae
surianaceae
rosaceae
barbeyaceae
dirachmaceae
rhamnaceae
elaeagnaceae
ulmaceae
cannabaceae
urticaceae
moraceae
nothofagaceae
fagaceae
juglandaceae
myricaceae
casuarinaceae
betulaceae
ticodendraceae


anisophylleaceae
coriariaceae
apodanthaceae
corynocarpaceae
cucurbitaceae
tetramelaceae
begoniaceae
datiscaceae
celastraceae
lepidobotryaceae
huaceae
oxalidaceae
connaraceae
cunoniaceae
elaeocarpaceae
cephalotaceae
brunelliaceae
pandaceae
irvingiaceae
ctenolophonaceae
rhizophoraceae
erythroxylaceae
ochnaceae
clusiaceae
bonnetiaceae
calophyllaceae
podostemaceae
hypericaceae
caryocaraceae
putranjivaceae
lophopyxidaceae
centroplacaceae
malpighiaceae
elatinaceae
balanopaceae
dichapetalaceae
trigoniaceae
chrysobalanaceae
euphroniaceae
humiriaceae
achariaceae
violaceae
goupiaceae
passifloraceae
salicaceae
lacistemataceae
peraceae
euphorbiaceae
rafflesiaceae
picrodendraceae
phyllanthaceae
ixonanthaceae
linaceae
berberidopsidaceae
aextoxicaceae
strombosiaceae NA
erythropalaceae NA
coulaceae NA
olacaceae
aptandraceae NA
ximeniaceae NA
octoknemaceae NA
loranthaceae
misodendraceae
schoepfiaceae
opiliaceae
balanophoraceae
comandraceae NA
cervantesiaceae NA
thesiaceae NA
nanodeaceae NA
santalaceae
viscaceae NA
amphorogynaceae NA


tamaricaceae
frankeniaceae
polygonaceae
plumbaginaceae
droseraceae
nepenthaceae
drosophyllaceae
dioncophyllaceae
ancistrocladaceae
rhabdodendraceae
simmondsiaceae
physenaceae
asteropeiaceae
macarthuriaceae
microteaceae
caryophyllaceae
amaranthaceae
achatocarpaceae
stegnospermataceae
limeaceae
lophiocarpaceae
kewaceae
barbeuiaceae
aizoaceae
nyctaginaceae
petiveriaceae
phytolaccaceae
sarcobataceae
gisekiaceae
molluginaceae
montiaceae
basellaceae
didiereaceae
halophytaceae
talinaceae
anacampserotaceae
cactaceae
portulacaceae
hydrostachyaceae
cornaceae
nyssaceae
hydrangeaceae
loasaceae
grubbiaceae
curtisiaceae
balsaminaceae
tetrameristaceae
marcgraviaceae
lecythidaceae
fouquieriaceae
polemoniaceae
pentaphylacaceae
sladeniaceae
sapotaceae
primulaceae
ebenaceae
theaceae
mitrastemonaceae
symplocaceae
diapensiaceae
styracaceae
sarraceniaceae
actinidiaceae
roridulaceae
clethraceae
ericaceae
cyrillaceae
stemonuraceae
cardiopteridaceae
aquifoliaceae
phyllonomaceae
helwingiaceae

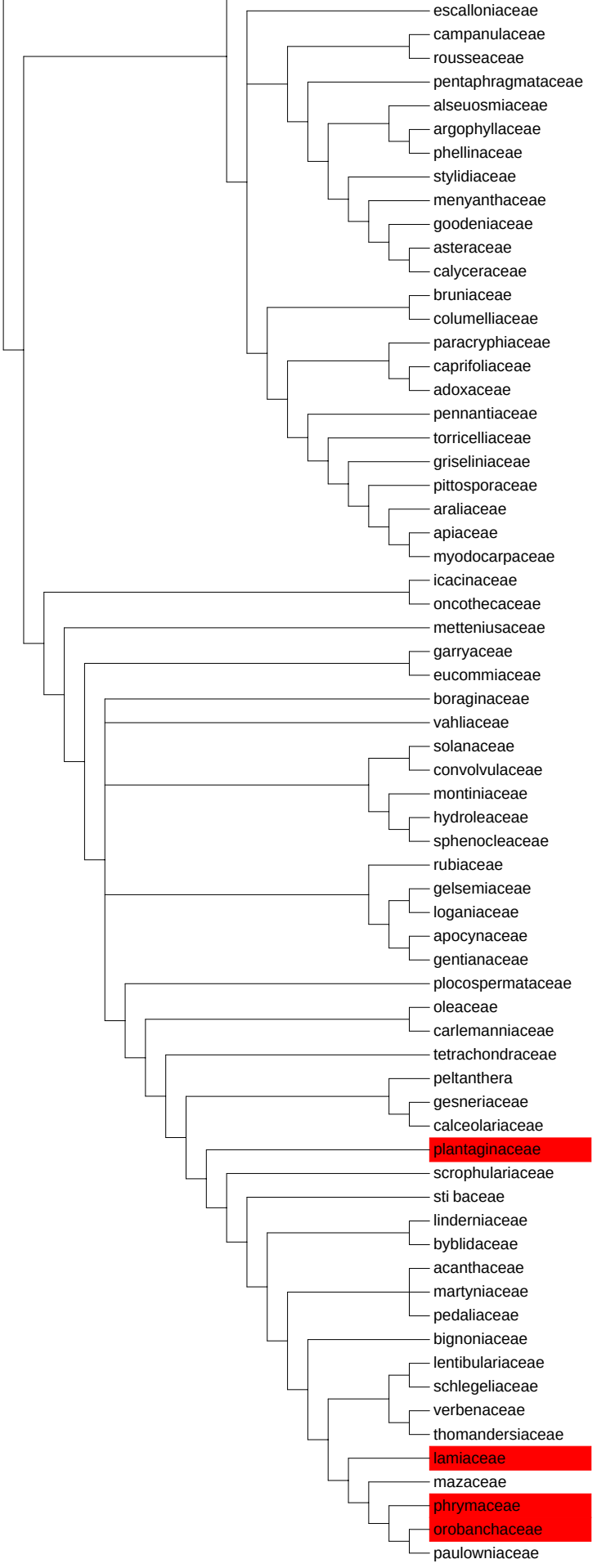
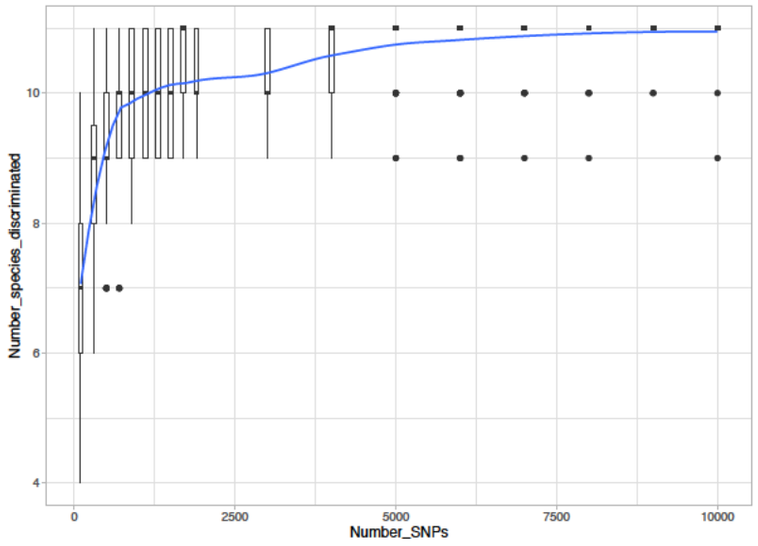
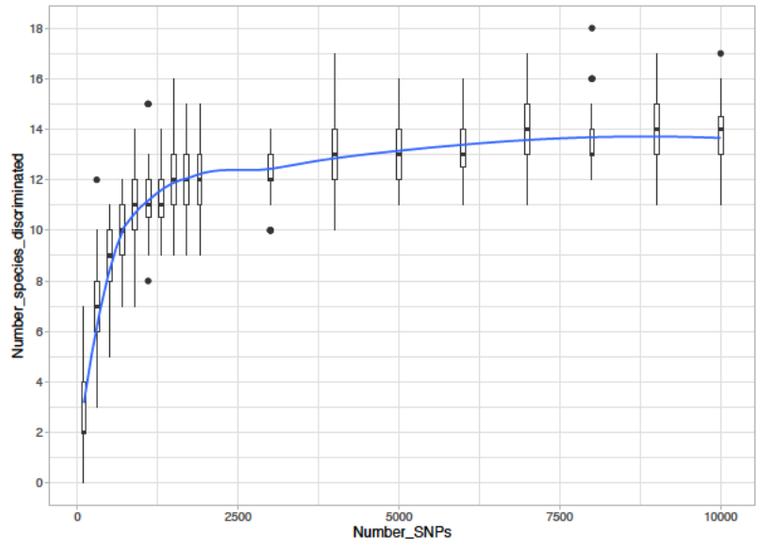

escalloniaceae
campanulaceae
rousseaceae
pentaphragmataceae
alseuosmiaceae
argophyllaceae
phellinaceae
stylidiaceae
menyanthaceae
goodeniaceae
asteraceae
calyceraceae
bruniaceae
columelliaceae
paracryphiaceae
caprifoliaceae
adoxaceae
pennantiaceae
torricelliaceae
griseliniaceae
pittosporaceae
araliaceae
apiaceae
myodocarpaceae
icacinaceae
oncothecaceae
metteniusaceae
garryaceae
eucommiaceae
boraginaceae
vahliaceae
solanaceae
convolvulaceae
montiniaceae
hydroleaceae
sphenocleaceae
rubiaceae
gelsemiaceae
loganiaceae
apocynaceae
gentianaceae
plocospermataceae
oleaceae
carlemanniaceae
tetrachondraceae
peltanthera
gesneriaceae
calceolariaceae
plantaginaceae
scrophulariaceae
sti baceae
linderniaceae
byblidaceae
acanthaceae
martyniaceae
pedaliaceae
bignoniaceae
lentibulariaceae
schlegeliaceae
verbenaceae
thomandersiaceae
lamiaceae
mazaceae
phrymaceae
orobanchaceae
paulowniaceae

Figure S5.2 – S5.24. The subsampling curve of the individual genera by SNPs. The name of each individual dataset is on the top of each figure.

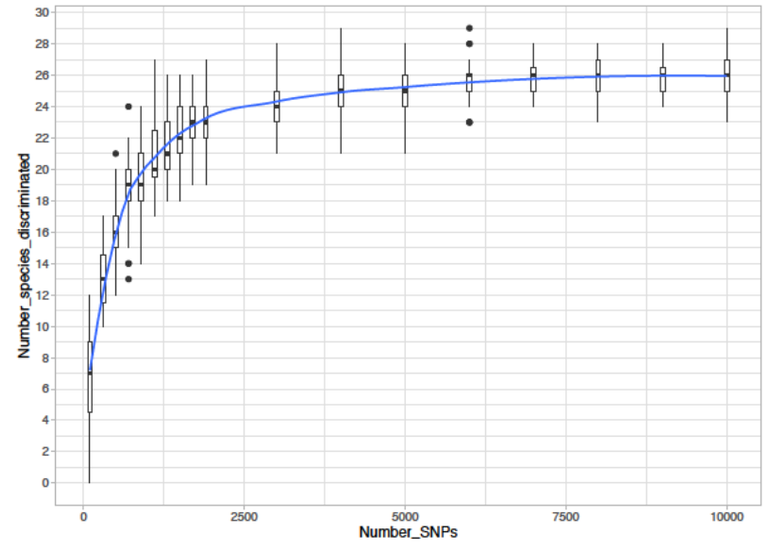
S5.2 Numbers of species discriminated by Monophyly – Aesculus



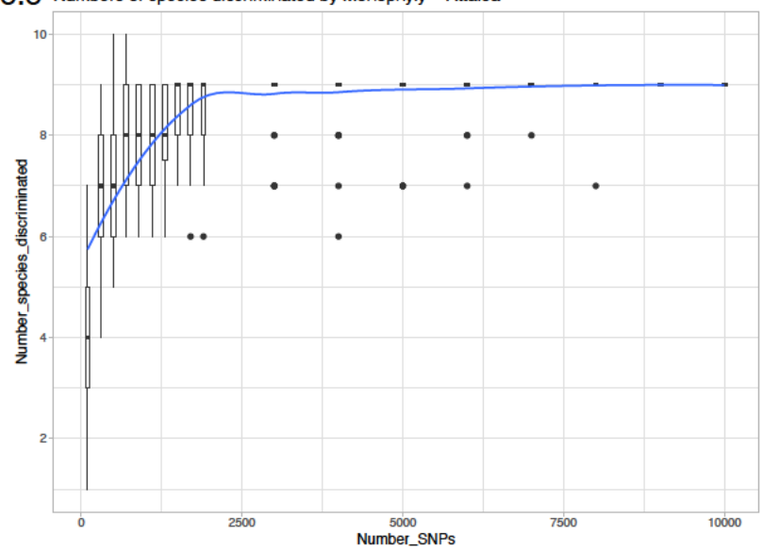
S5.3 Numbers of species discriminated by Monophyly – Antirrhinum



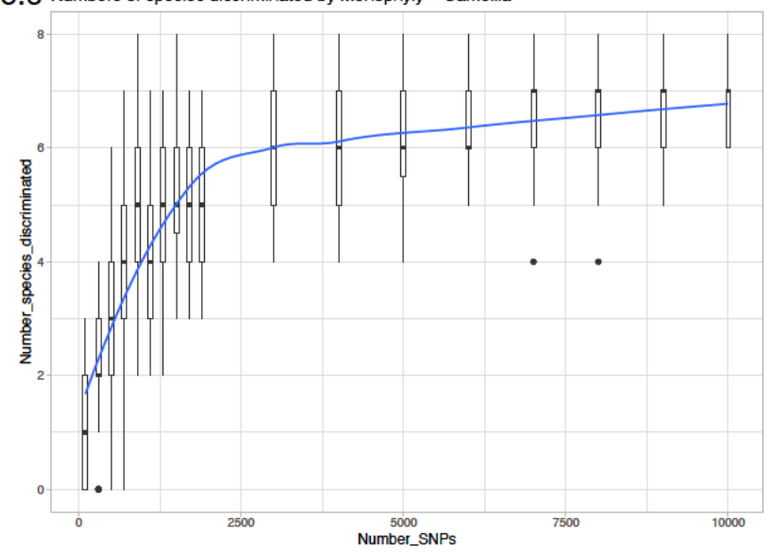
S5.4 Numbers of species discriminated by Monophyly – Artocarpus



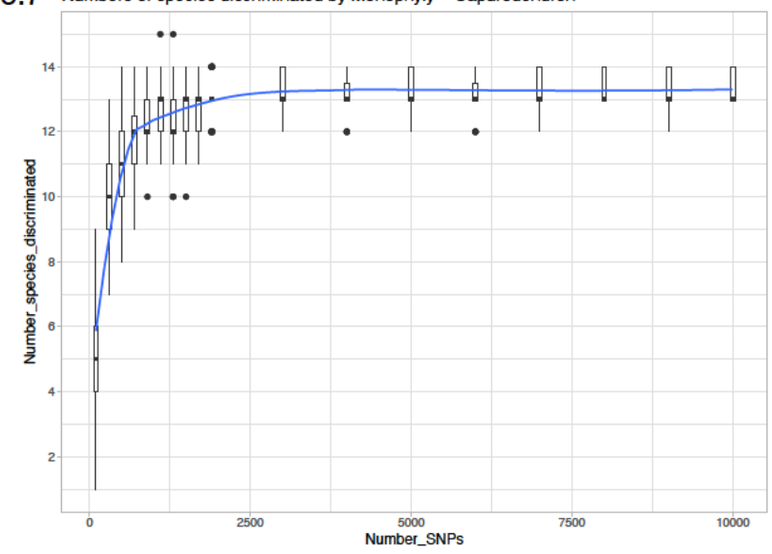
S5.5 Numbers of species discriminated by Monophyly – Attalea



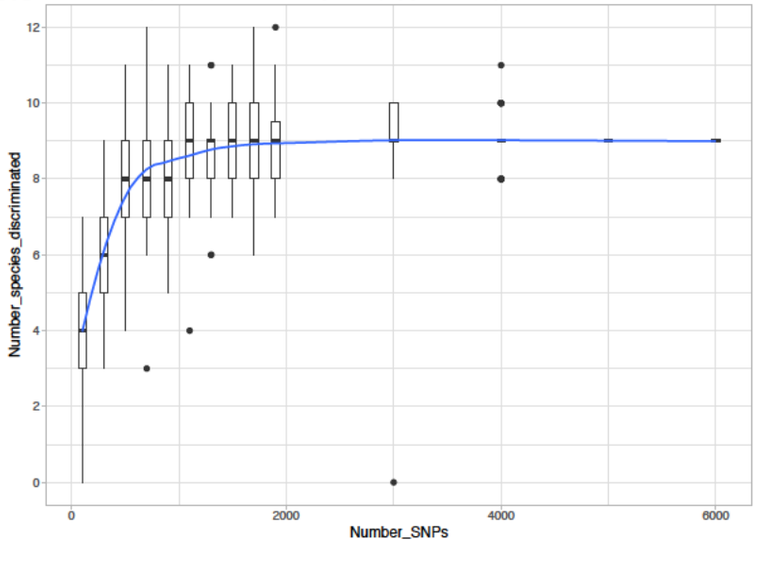
S5.6 Numbers of species discriminated by Monophyly – Camellia



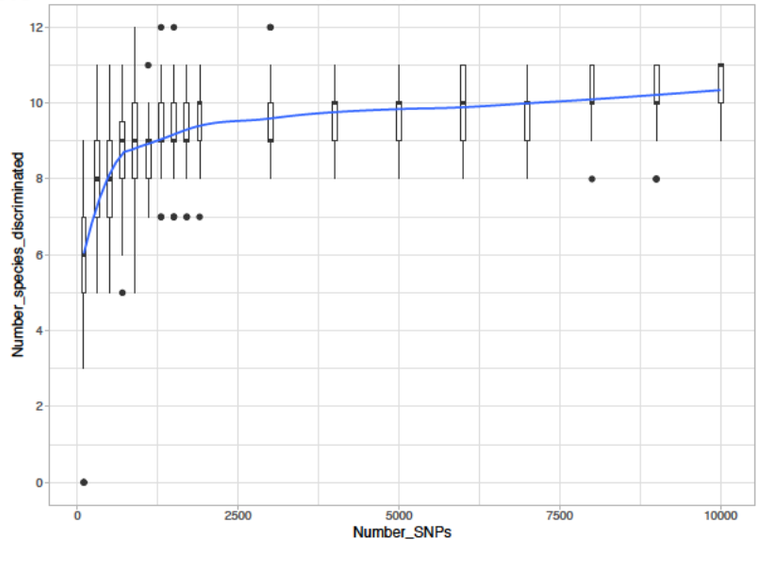
S5.7 Numbers of species discriminated by Monophyly – Capurodendron



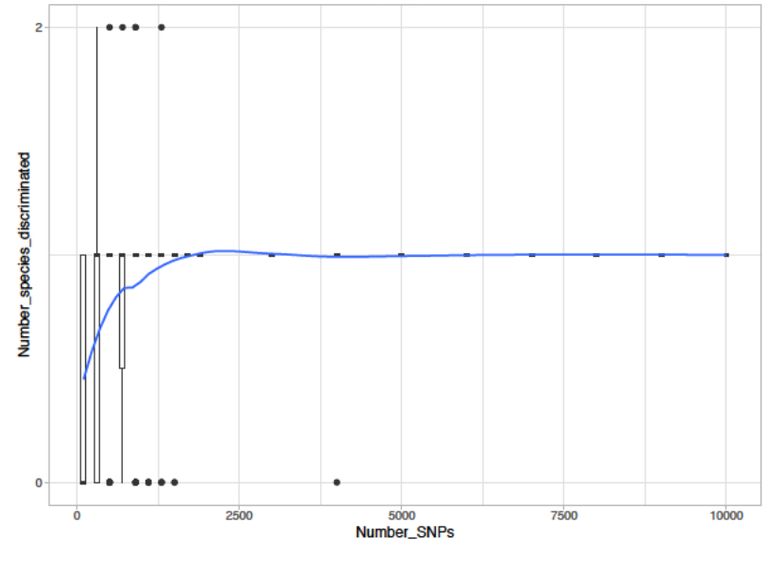
S5.8 Numbers of species discriminated by Monophyly – Commiphora



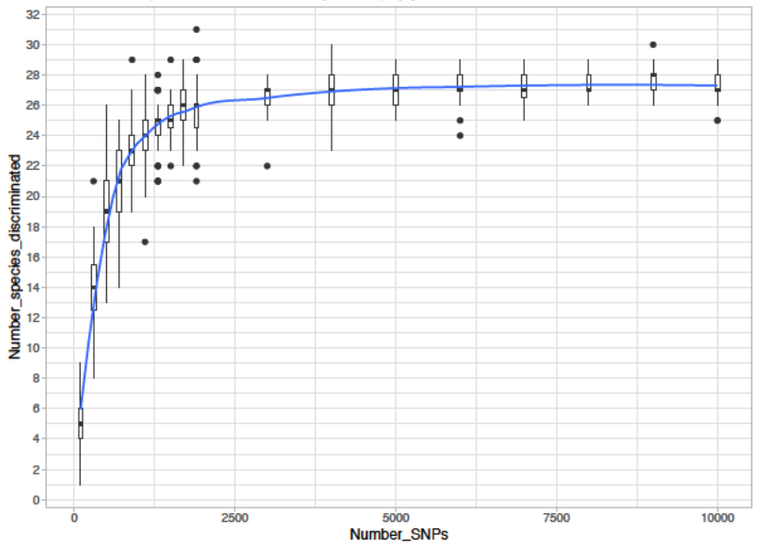
S5.9 Numbers of species discriminated by Monophyly – Cornus



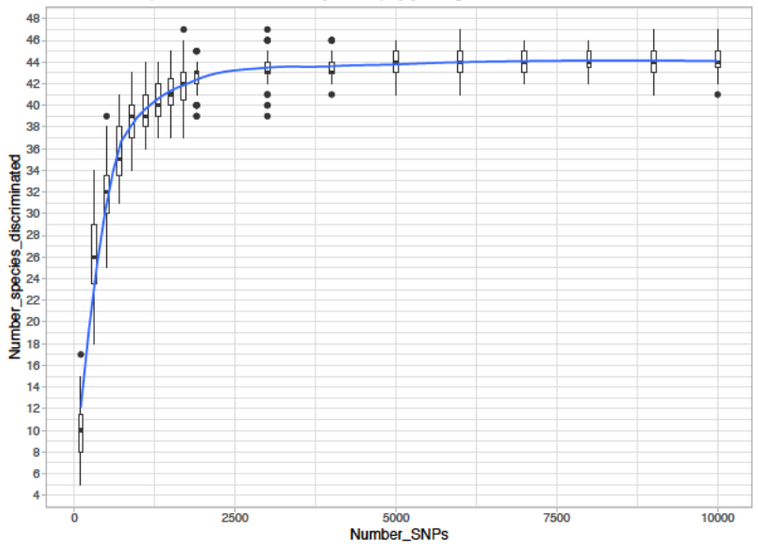
S5.10 Numbers of species discriminated by Monophyly – Dicerandra



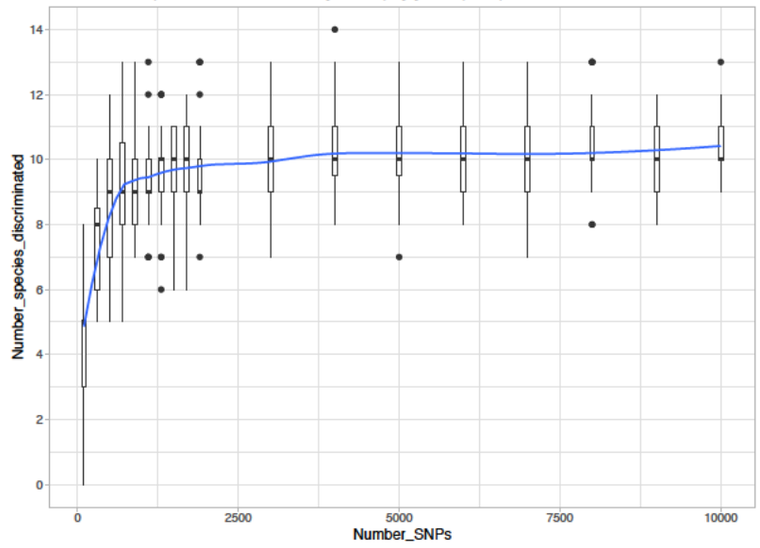
S5.11 Numbers of species discriminated by Monophyly – Geonoma



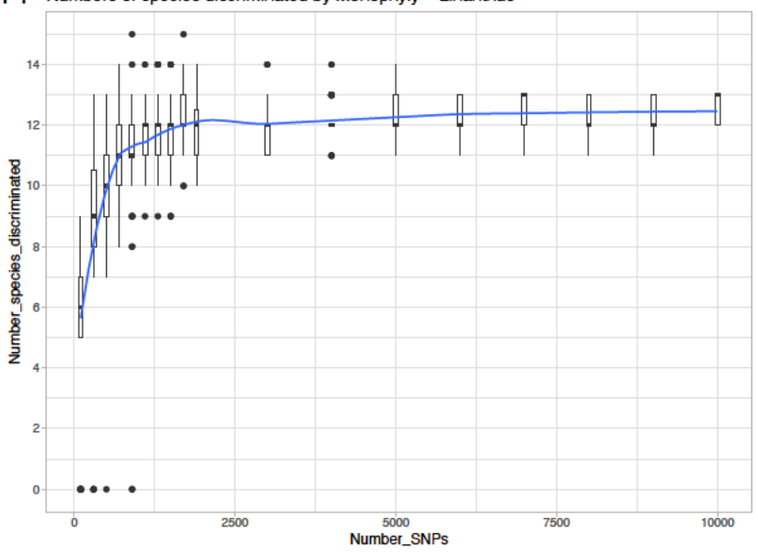
S5.12 Numbers of species discriminated by Monophyly – nga



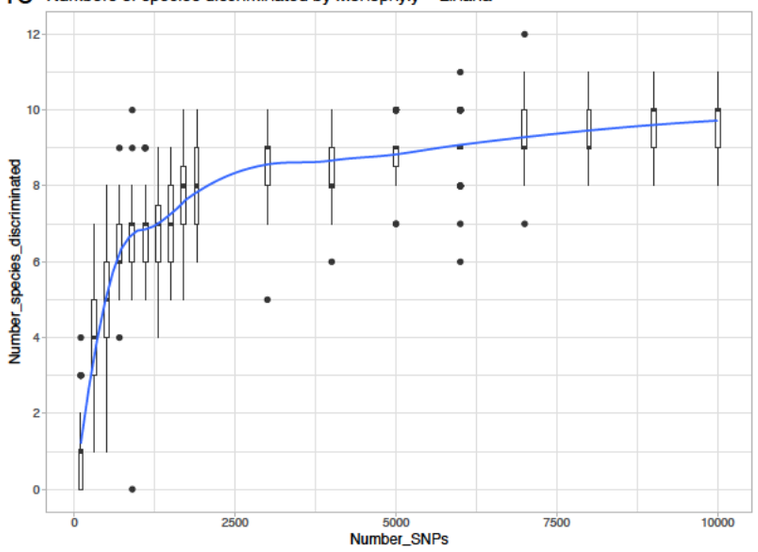
S5.13 Numbers of species discriminated by Monophyly – Leptosiphon



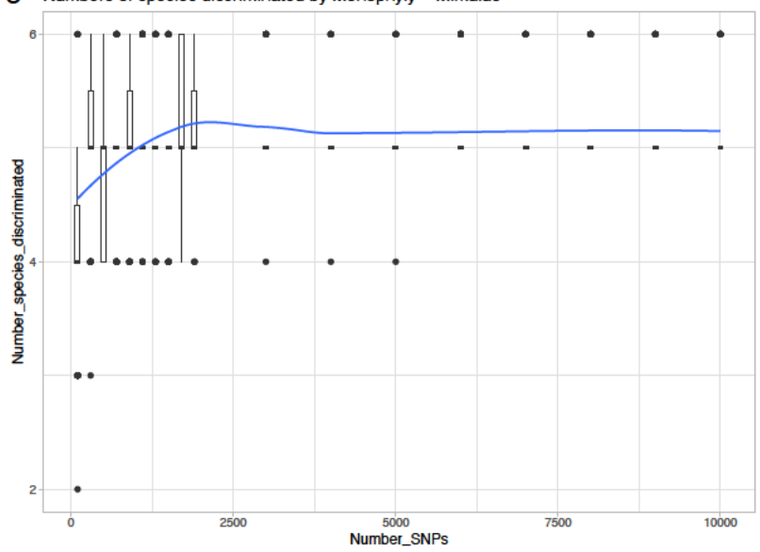
S5.14 Numbers of species discriminated by Monophyly – Linanthus



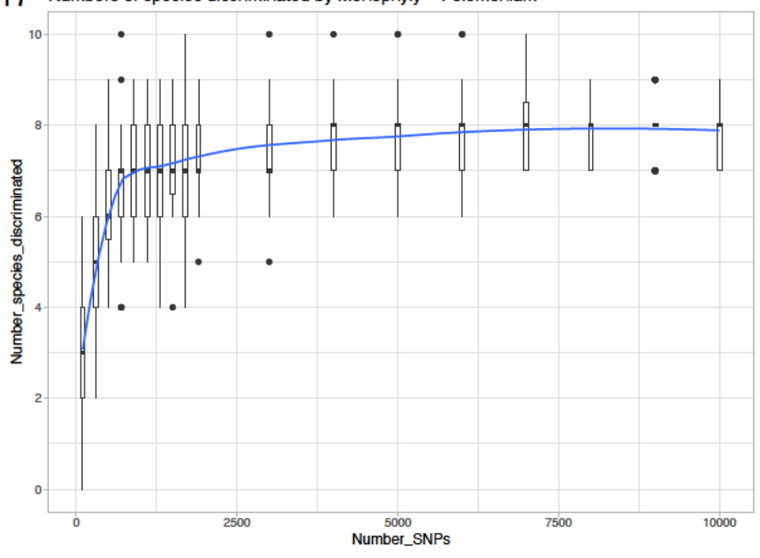
S5.15 Numbers of species discriminated by Monophyly – Linaria



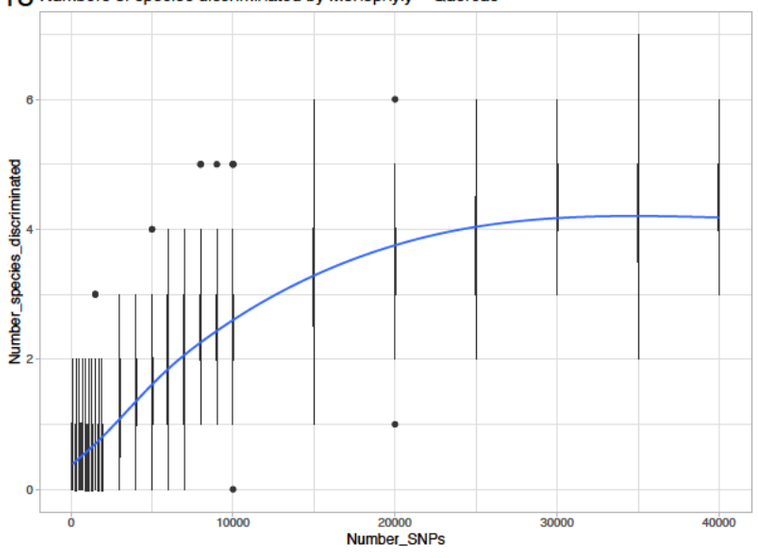
S5.16 Numbers of species discriminated by Monophyly – Mimulus



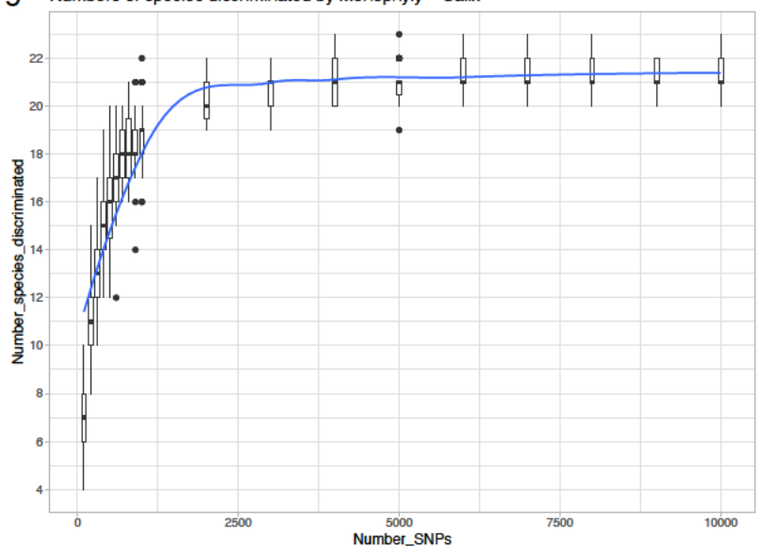
S5.17 Numbers of species discriminated by Monophyly – Polemonium



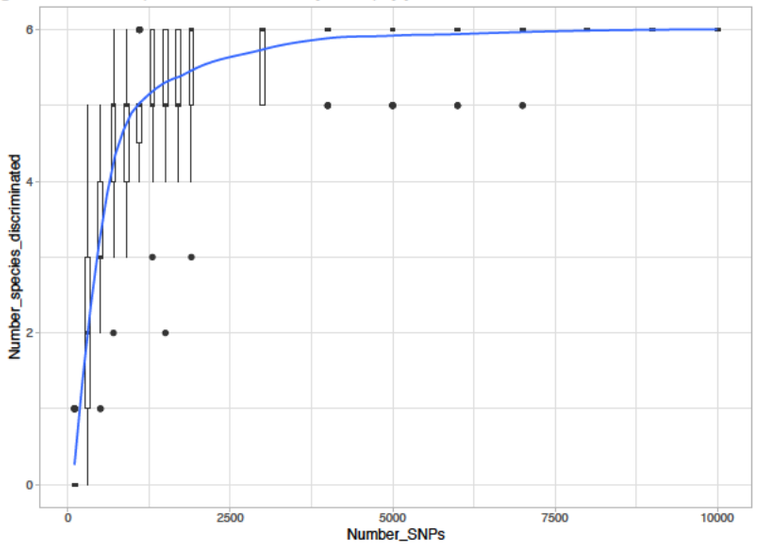
S5.18 Numbers of species discriminated by Monophyly – Quercus



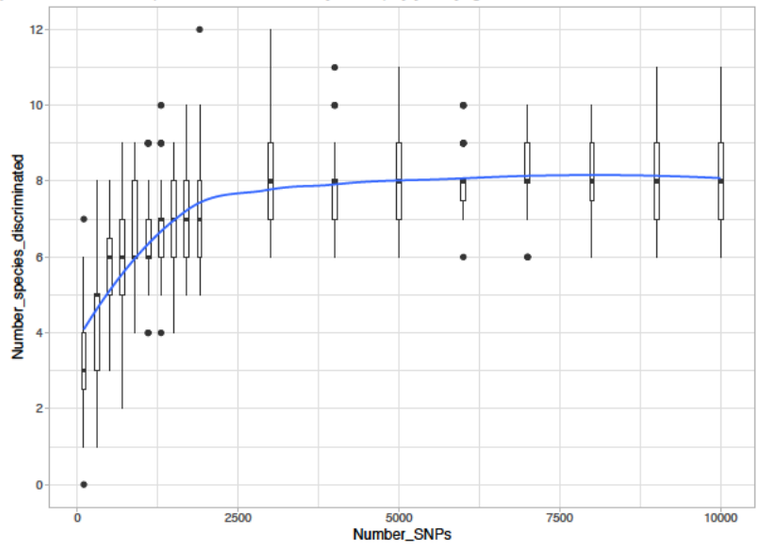
S5.19 Numbers of species discriminated by Monophyly – Salix



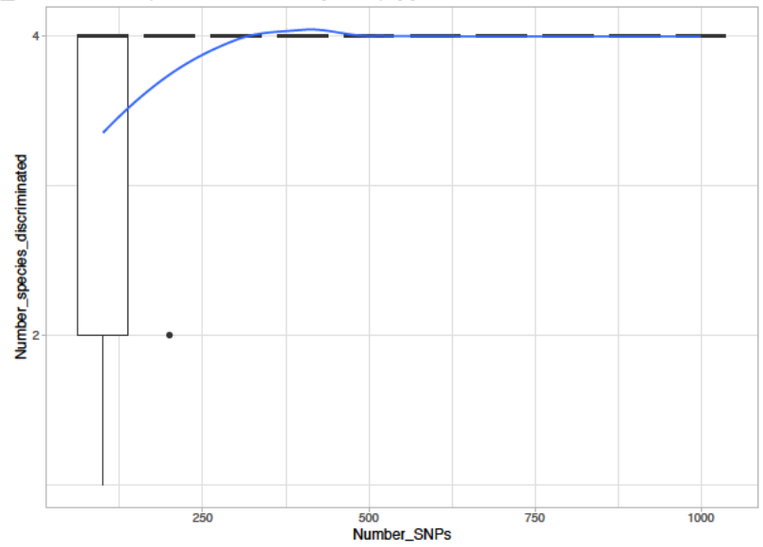
S5.20 Numbers of species discriminated by Monophyly – Sarracenia



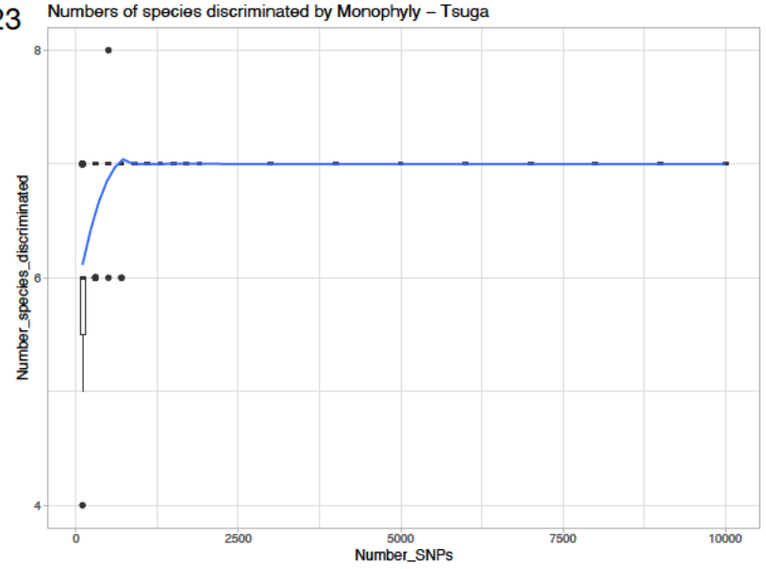
S5.21 Numbers of species discriminated by Monophyly – Syagrus



S5.22 Numbers of species discriminated by Monophyly – Taxus



S5.23 Numbers of species discriminated by Monophyly – Tsuga



S5.24 Numbers of species discriminated by Monophyly – Vitis

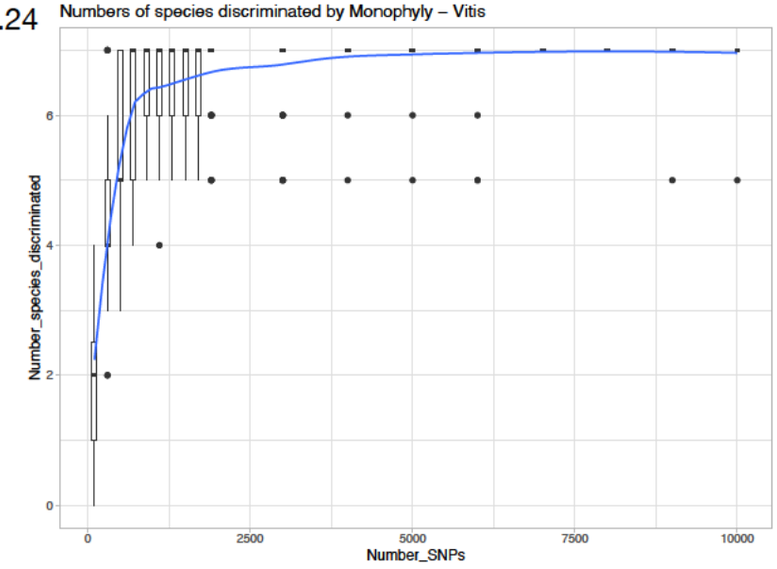
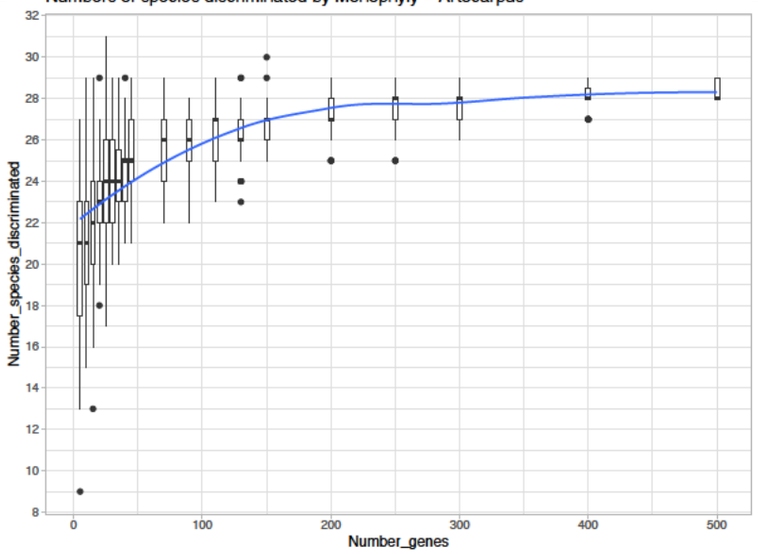
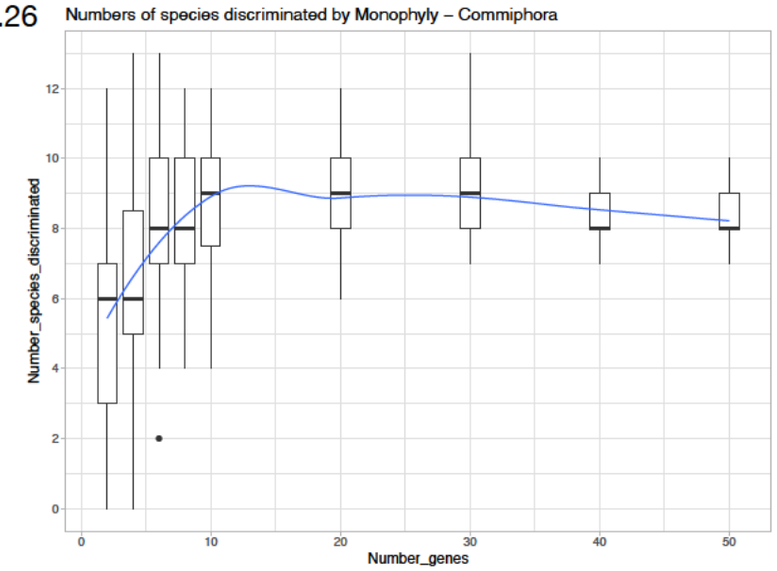


Figure S5.25 – S5.33. The subsampling curves of the individual genera by genes/DNA segments.
The name of each individual dataset is on the top of each figure.

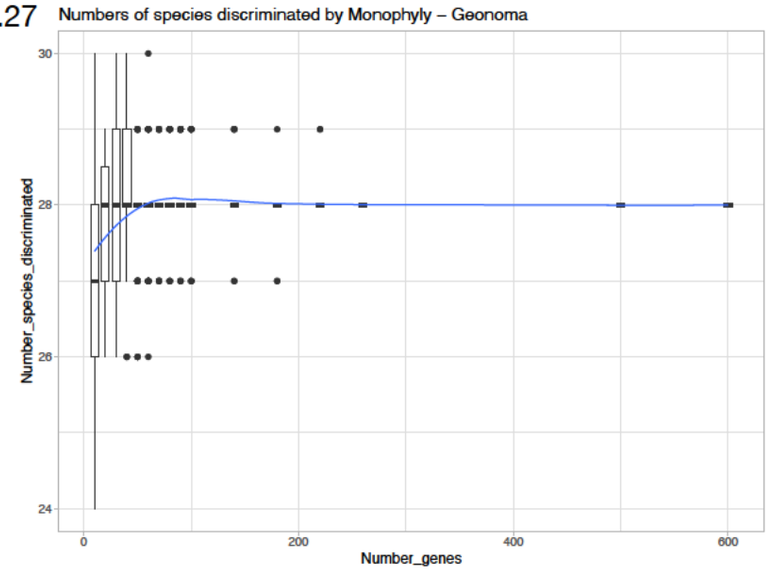
S5.25



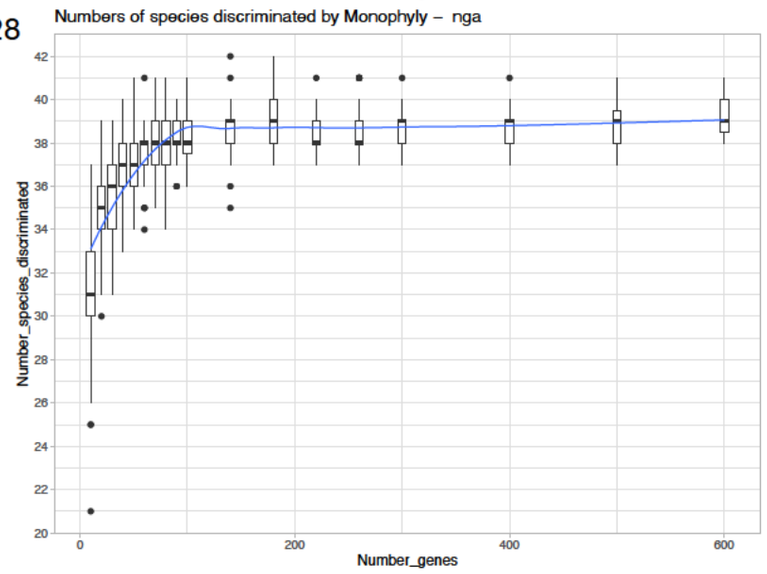
S5.26



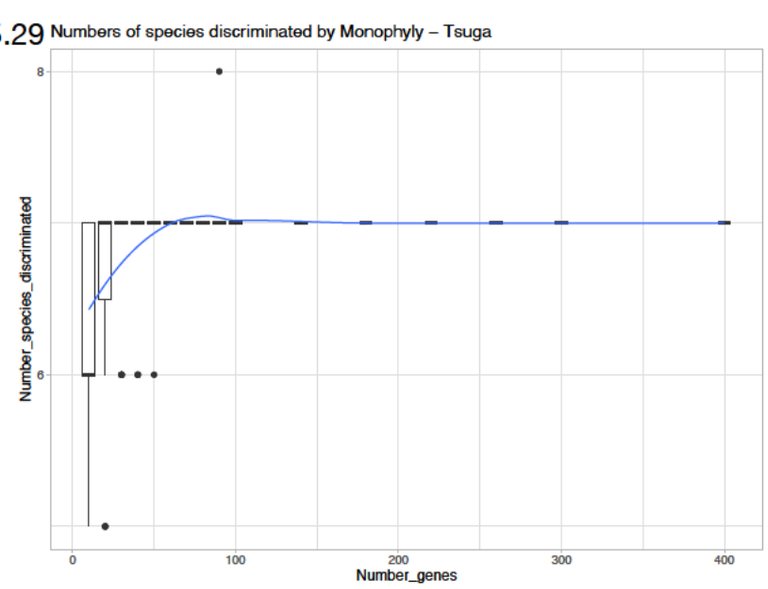
S5.27



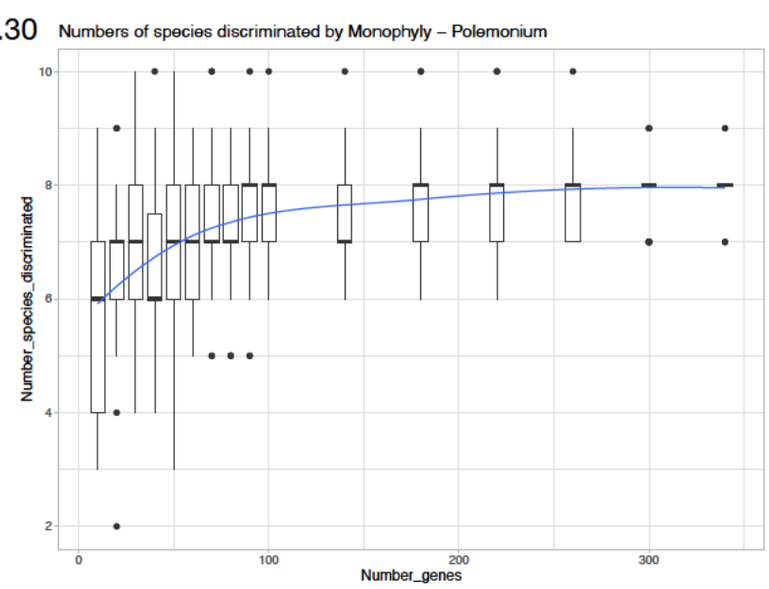
S5.28



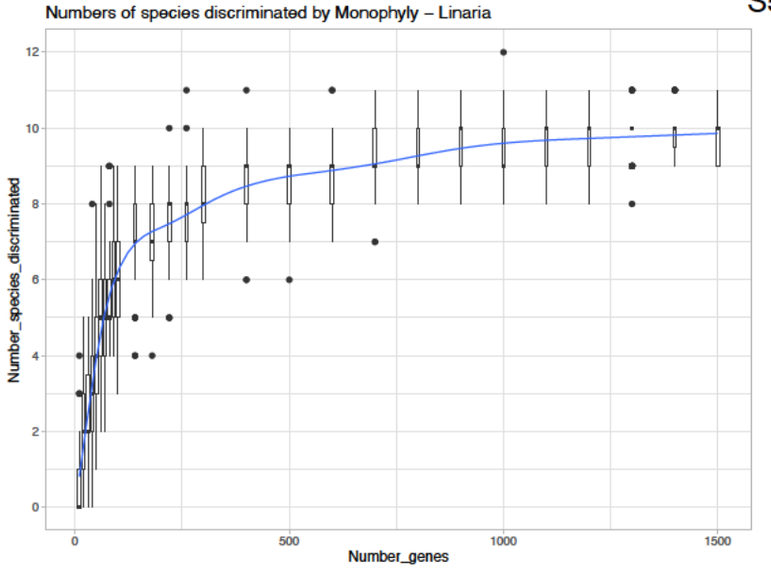
S5.29



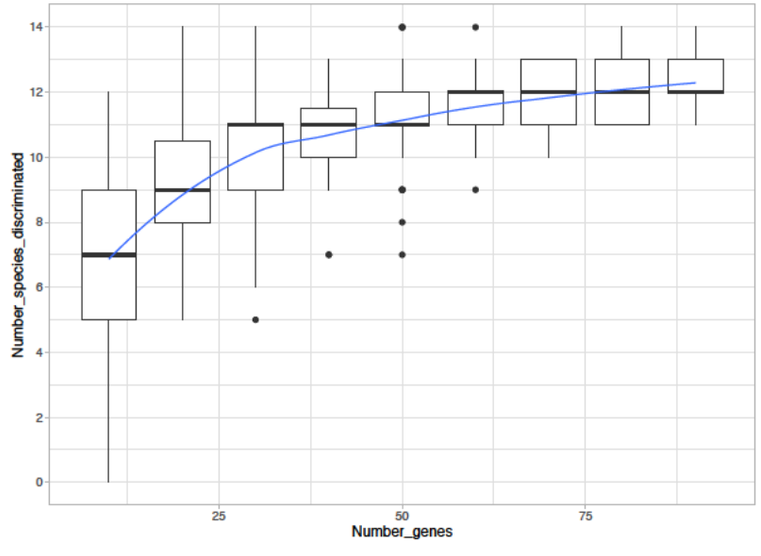
S5.30



S5.31



S5.32



S5.33

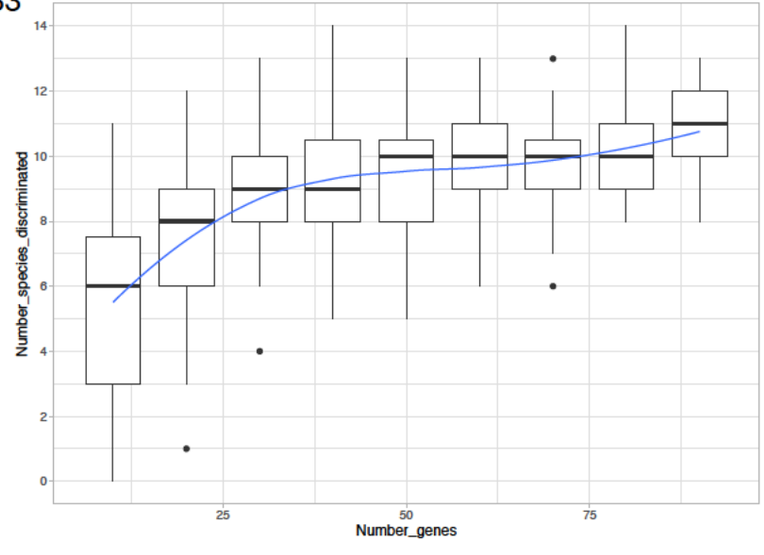
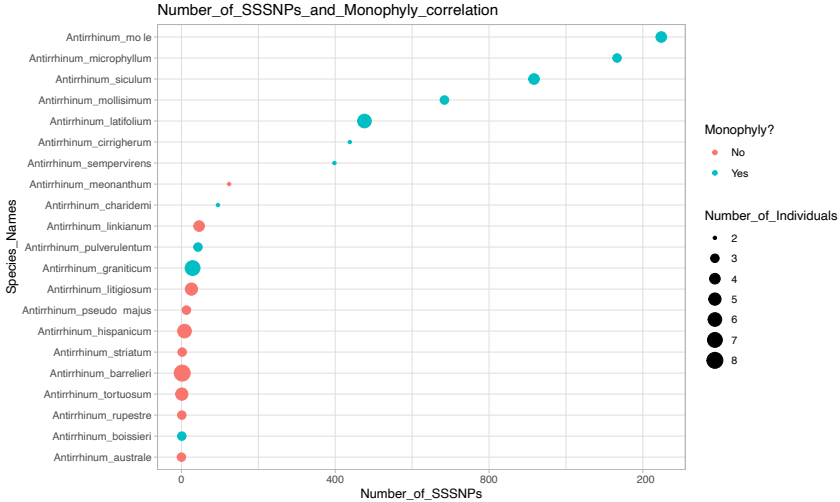
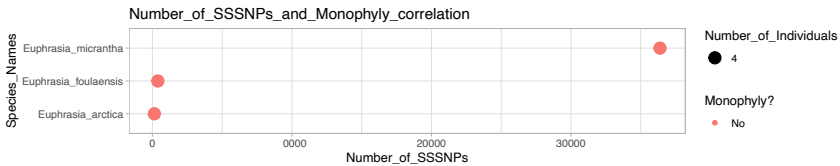


Figure S5.34 – S5.62. Distribution of the number of Species-Specific SNPs for multi-sampled species for all genera analysed (Dataset 2)

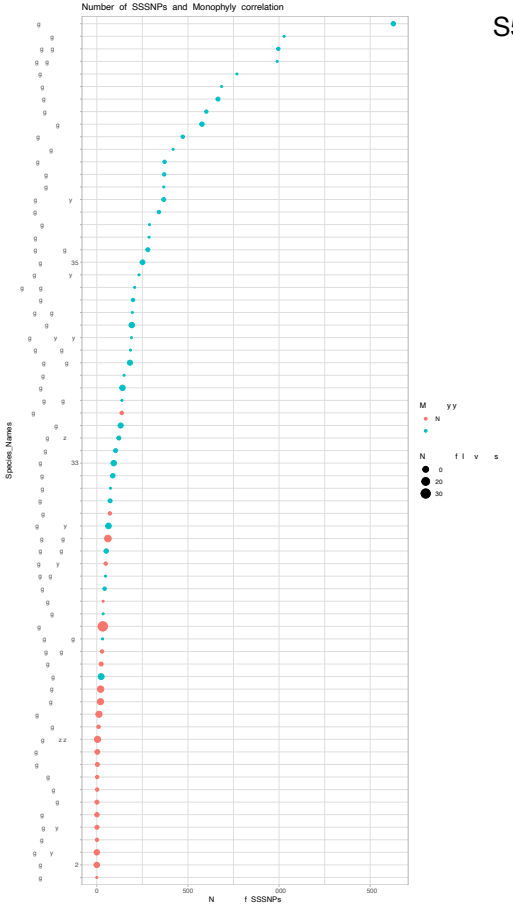
S5.34


Number_of_SSSNPs_and_Monophyly_correlation
Antirrhinum_mo le
Antirrhinum_microphyllum
Antirrhinum_siculum
Antirrhinum_mollisimum
Antirrhinum_latifolium
Antirrhinum_cirrigherum
Antirrhinum_sempervirens
Antirrhinum_meonanthum
Antirrhinum_charidemi
Antirrhinum_linkianum
Antirrhinum_pulverulentum
Antirrhinum_graniticum
Antirrhinum_litigiosum
Antirrhinum_pseudo_majus
Antirrhinum_hispanicum
Antirrhinum_striatum
Antirrhinum_barrelieri
Antirrhinum_tortuosum
Antirrhinum_rupestre
Antirrhinum_boissieri
Antirrhinum_australe
Species_Names
0
400
800
200
Number_of_SSSNPs
Monophyly?
No
Yes
Number_of_Individuals
2
3
4
5
6
7
8

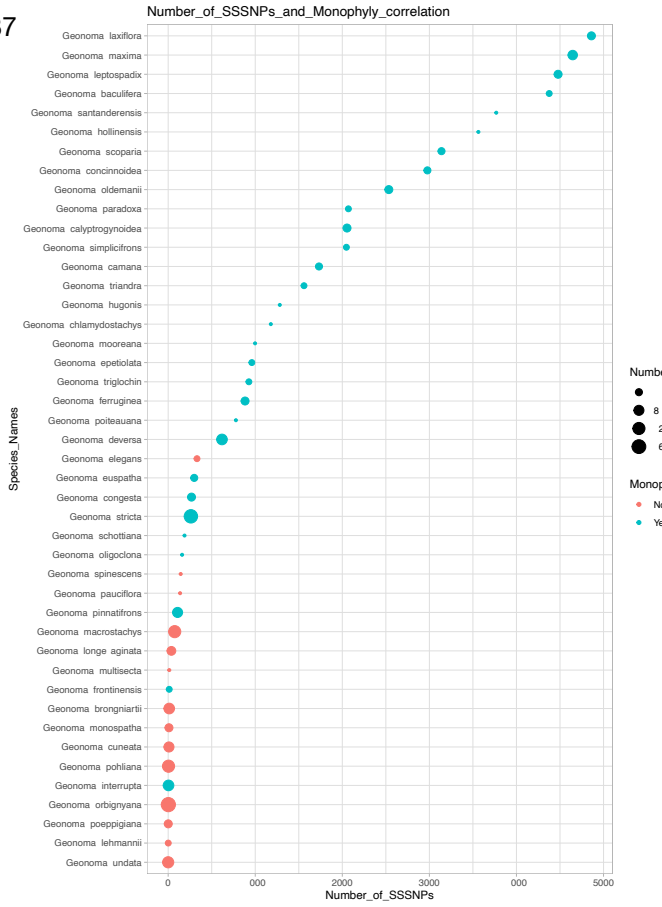
S5.35


Number_of_SSSNPs_and_Monophyly_correlation
Euphrasia_micrantha
Euphrasia_foulaensis
Euphrasia_arctica
Species_Names
0
0000
20000
30000
Number_of_SSSNPs
Number_of_Individuals
4
Monophyly?
No

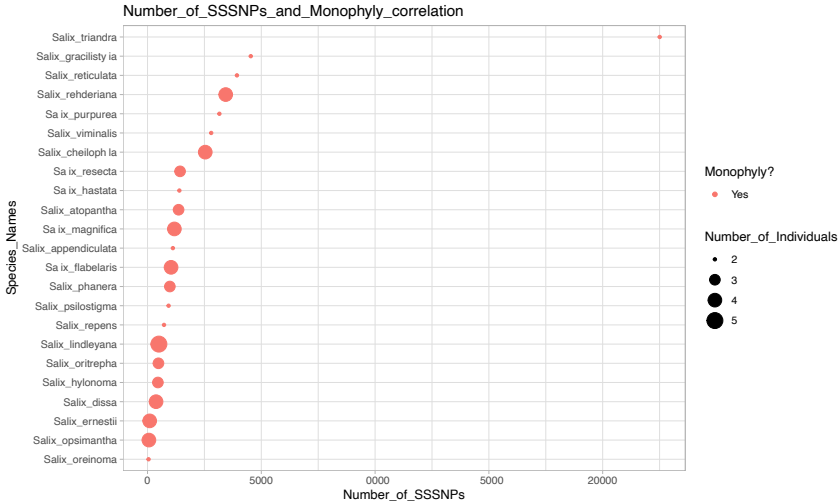
S5.36


Number of SSSNPs and Monophyly correlation
Species_Names
N f SSSNPs
0
500
000
500
35
33
2
M yy
N
N f l v s
0
20
30

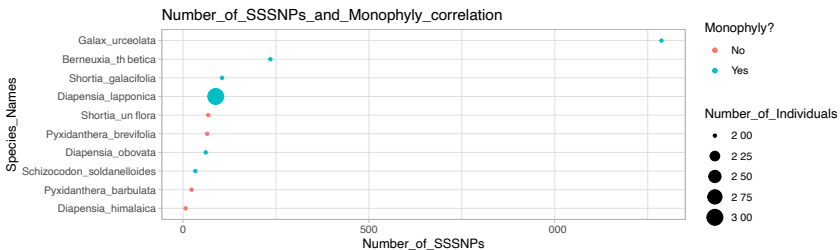
S5.37


Number_of_SSSNPs_and_Monophyly_correlation
Geonoma laxiflora
Geonoma maxima
Geonoma leptospadix
Geonoma baculifera
Geonoma santanderensis
Geonoma hollinensis
Geonoma scoparia
Geonoma concinnoidea
Geonoma oldemanii
Geonoma paradoxa
Geonoma calyptrogynoidea
Geonoma simplicifrons
Geonoma camana
Geonoma triandra
Geonoma hugonis
Geonoma chlamydostachys
Geonoma mooreana
Geonoma epetiolata
Geonoma triglochin
Geonoma ferruginea
Geonoma poiteauana
Geonoma deversa
Geonoma elegans
Geonoma euspatha
Geonoma congesta
Geonoma stricta
Geonoma schottiana
Geonoma oligoclona
Geonoma spinescens
Geonoma pauciflora
Geonoma pinnatifrons
Geonoma macrostachys
Geonoma longe aginata
Geonoma multisecta
Geonoma frontinensis
Geonoma brongniartii
Geonoma monospatha
Geonoma cuneata
Geonoma pohliana
Geonoma interrupta
Geonoma orbignyana
Geonoma poeppigiana
Geonoma lehmannii
Geonoma undata
Species_Names
0
000
2000
3000
000
5000
Number_of_SSSNPs
Numbe
8
2
6
Monop
No
Ye

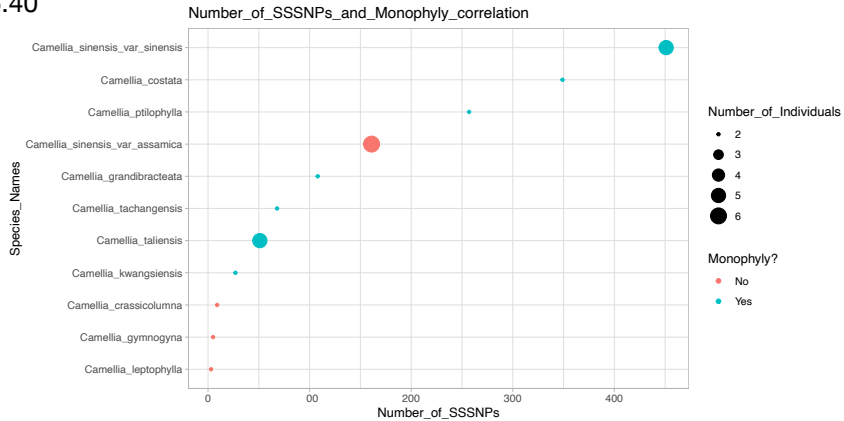
S5.38


Number_of_SSSNPs_and_Monophyly_correlation
Salix_triandra
Salix_gracilisty ia
Salix_reticulata
Salix_rehderiana
Sa ix_purpurea
Salix_viminalis
Salix_cheiloph la
Sa ix_resecta
Sa ix_hastata
Salix_atopantha
Sa ix_magnifica
Salix_appendiculata
Sa ix_flabelaris
Salix_phanera
Salix_psilostigma
Salix_repens
Salix_lindleyana
Salix_oritrepha
Salix_hylonoma
Salix_dissa
Salix_ernestii
Salix_opsimantha
Salix_oreinoma
Species_Names
0
5000
0000
5000
20000
Number_of_SSSNPs
Monophyly?
Yes
Number_of_Individuals
2
3
4
5

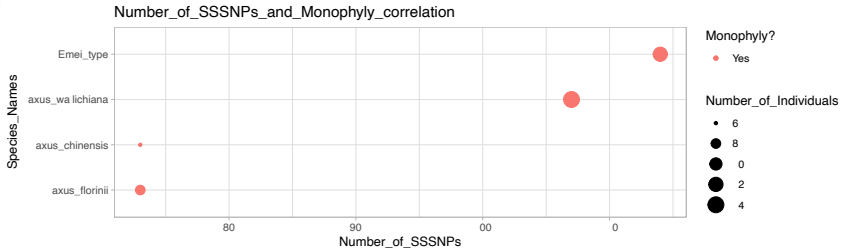
S5.39


Number_of_SSSNPs_and_Monophyly_correlation
Galax_urceolata
Berneuxia_th betica
Shortia_galacifolia
Diapensia_lapponica
Shortia_un flora
Pyxidanthera_brevifolia
Diapensia_obovata
Schizocodon_soldanelloides
Pyxidanthera_barbulata
Diapensia_himalaica
Species_Names
0
500
000
Number_of_SSSNPs
Monophyly?
No
Yes
Number_of_Individuals
2 00
2 25
2 50
2 75
3 00

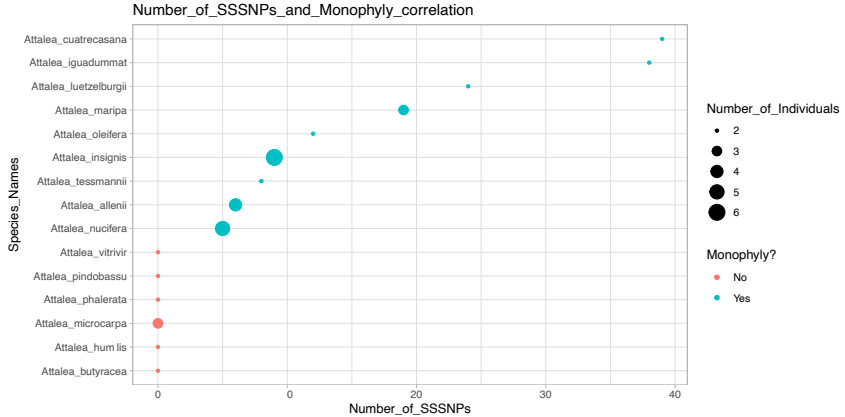
S5.40


Number_of_SSSNPs_and_Monophyly_correlation
Camellia_sinensis_var_sinensis
Camellia_costata
Camellia_ptilophylla
Camellia_sinensis_var_assamica
Camellia_grandibracteata
Camellia_tachangensis
Camellia_taliensis
Camellia_kwangsiensis
Camellia_crassicolumna
Camellia_gymnogyna
Camellia_leptophylla
Species_Names
0
00
200
300
400
Number_of_SSSNPs
Number_of_Individuals
2
3
4
5
6
Monophyly?
No
Yes

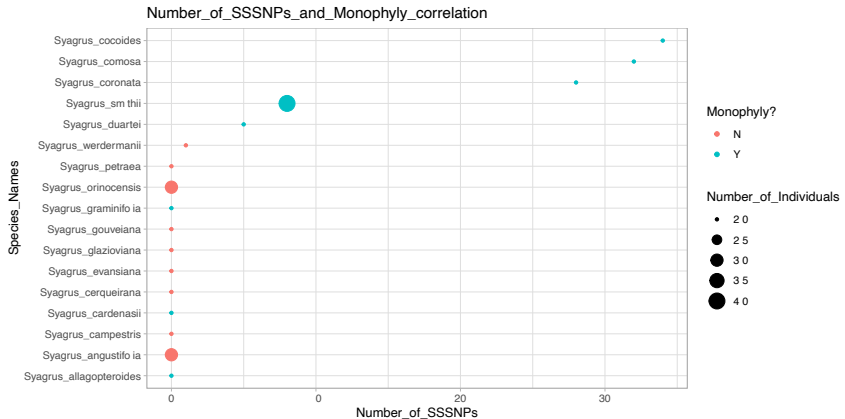
S5.41


Number_of_SSSNPs_and_Monophyly_correlation
Emei_type
axus_wa lichiana
axus_chinensis
axus_florinii
Species_Names
80
90
00
0
Number_of_SSSNPs
Monophyly?
Yes
Number_of_Individuals
6
8
0
2
4

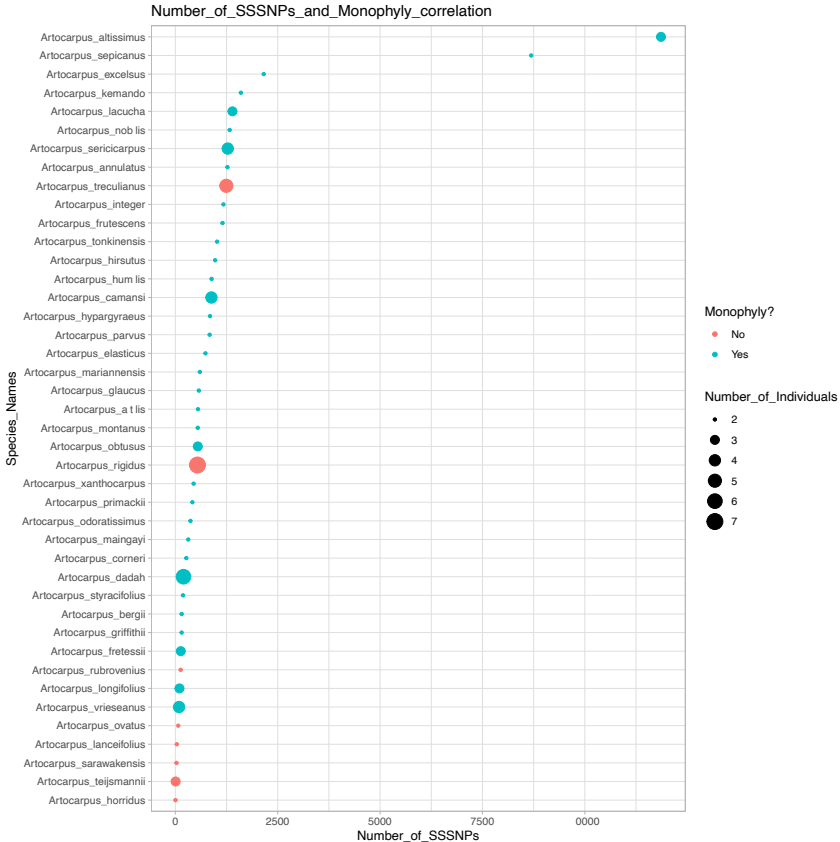
S5.42


Number_of_SSSNPs_and_Monophyly_correlation
Attalea_cuatrecasana
Attalea_iguadummat
Attalea_luetzelburgii
Attalea_maripa
Attalea_oleifera
Attalea_insignis
Attalea_tessmannii
Attalea_allenii
Attalea_nucifera
Attalea_vitrivir
Attalea_pindobassu
Attalea_phalerata
Attalea_microcarpa
Attalea_hum lis
Attalea_butyracea
Species_Names
0
0
20
30
40
Number_of_SSSNPs
Number_of_Individuals
2
3
4
5
6
Monophyly?
No
Yes

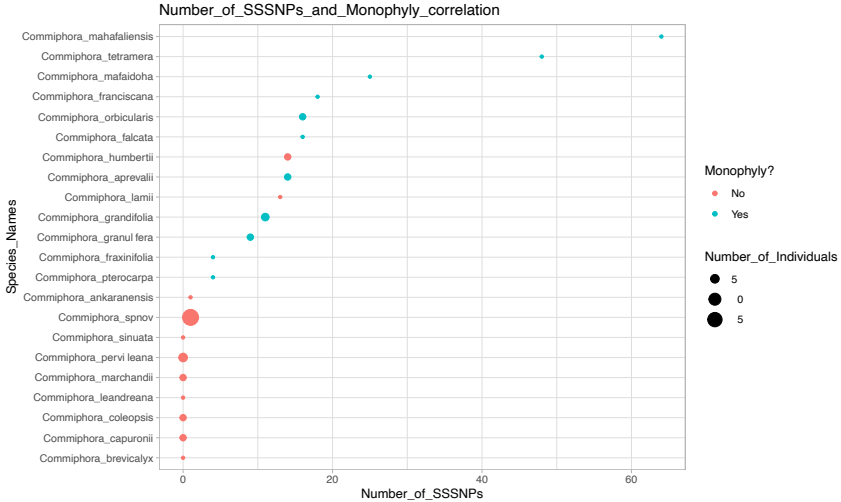
S5.43


Number_of_SSSNPs_and_Monophyly_correlation
Syagrus_cocoides
Syagrus_comosa
Syagrus_coronata
Syagrus_sm thii
Syagrus_duartei
Syagrus_werdermanii
Syagrus_petraea
Syagrus_orinocensis
Syagrus_graminifo ia
Syagrus_gouveiana
Syagrus_glazioviana
Syagrus_evansiana
Syagrus_cerqueirana
Syagrus_cardenasii
Syagrus_campestris
Syagrus_angustifo ia
Syagrus_allagopteroides
Species_Names
0
0
20
30
Number_of_SSSNPs
Monophyly?
N
Y
Number_of_Individuals
2 0
2 5
3 0
3 5
4 0

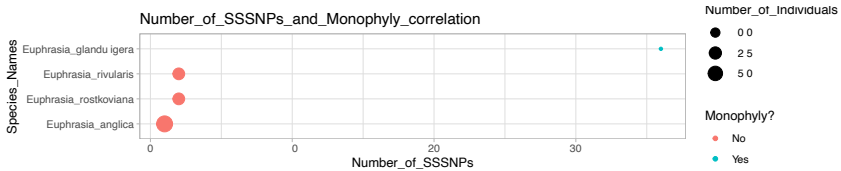
S5.44


Number_of_SSSNPs_and_Monophyly_correlation
Artocarpus_altissimus
Artocarpus_sepicanus
Artocarpus_excelsus
Artocarpus_kemando
Artocarpus_lacucha
Artocarpus_nob lis
Artocarpus_sericicarpus
Artocarpus_annulatus
Artocarpus_treculianus
Artocarpus_integer
Artocarpus_frutescens
Artocarpus_tonkinensis
Artocarpus_hirsutus
Artocarpus_hum lis
Artocarpus_camansi
Artocarpus_hypargyraeus
Artocarpus_parvus
Artocarpus_elasticus
Artocarpus_mariannensis
Artocarpus_glaucus
Artocarpus_a t lis
Artocarpus_montanus
Artocarpus_obtusus
Artocarpus_rigidus
Artocarpus_xanthocarpus
Artocarpus_primackii
Artocarpus_odoratissimus
Artocarpus_maingayi
Artocarpus_corneri
Artocarpus_dadah
Artocarpus_styracifolius
Artocarpus_bergii
Artocarpus_griffithii
Artocarpus_fretessii
Artocarpus_rubrovenius
Artocarpus_longifolius
Artocarpus_vrieseanus
Artocarpus_ovatus
Artocarpus_lanceifolius
Artocarpus_sarawakensis
Artocarpus_teijsmannii
Artocarpus_horridus
Species_Names
0
2500
5000
7500
0000
Number_of_SSSNPs
Monophyly?
No
Yes
Number_of_Individuals
2
3
4
5
6
7

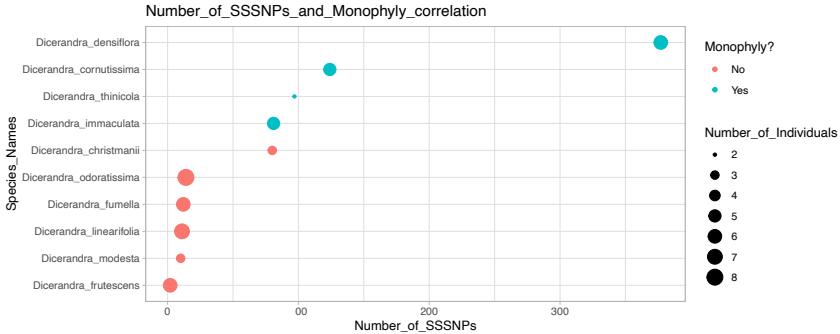
S5.45


Number_of_SSSNPs_and_Monophyly_correlation
Commiphora_mahafaliensis
Commiphora_tetramera
Commiphora_mafaidoha
Commiphora_franciscana
Commiphora_orbicularis
Commiphora_falcata
Commiphora_humbertii
Commiphora_aprevalii
Commiphora_lamii
Commiphora_grandifolia
Commiphora_granul fera
Commiphora_fraxinifolia
Commiphora_pterocarpa
Commiphora_ankaranensis
Commiphora_spnov
Commiphora_sinuata
Commiphora_pervi leana
Commiphora_marchandii
Commiphora_leandreana
Commiphora_coleopsis
Commiphora_capuronii
Commiphora_brevicalyx
Species_Names
0
20
40
60
Number_of_SSSNPs
Monophyly?
No
Yes
Number_of_Individuals
5
0
5

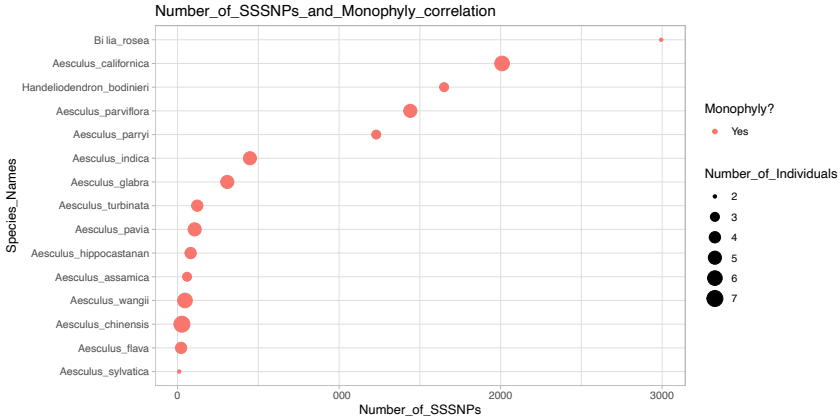
S5.46


Number_of_SSSNPs_and_Monophyly_correlation
Euphrasia_glandu igera
Euphrasia_rivularis
Euphrasia_rostkoviana
Euphrasia_anglica
Species_Names
0
0
20
30
Number_of_SSSNPs
Number_of_Individuals
0 0
2 5
5 0
Monophyly?
No
Yes

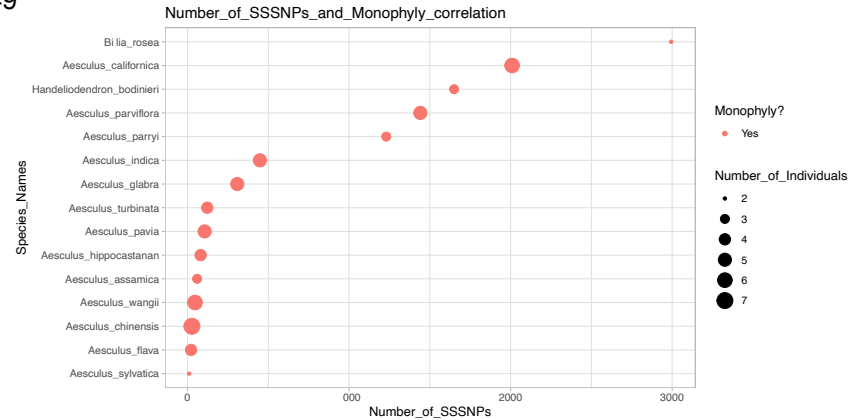
S5.47


Number_of_SSSNPs_and_Monophyly_correlation
Dicerandra_densiflora
Dicerandra_cornutissima
Dicerandra_thinicola
Dicerandra_immaculata
Dicerandra_christmanii
Dicerandra_odoratissima
Dicerandra_fumella
Dicerandra_linearifolia
Dicerandra_modesta
Dicerandra_frutescens
Species_Names
0
00
200
300
Number_of_SSSNPs
Monophyly?
No
Yes
Number_of_Individuals
2
3
4
5
6
7
8

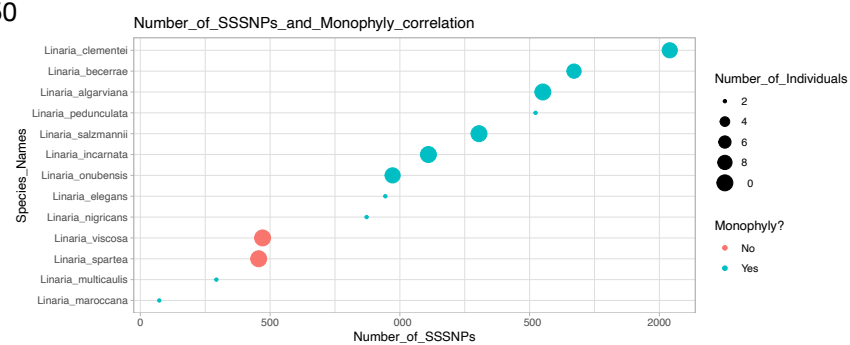
S5.48


Number_of_SSSNPs_and_Monophyly_correlation
Bi lia_rosea
Aesculus_californica
Handeliodendron_bodinieri
Aesculus_parviflora
Aesculus_parryi
Aesculus_indica
Aesculus_glabra
Aesculus_turbinata
Aesculus_pavia
Aesculus_hippocastanan
Aesculus_assamica
Aesculus_wangii
Aesculus_chinensis
Aesculus_flava
Aesculus_sylvatica
Species_Names
0
000
2000
3000
Number_of_SSSNPs
Monophyly?
Yes
Number_of_Individuals
2
3
4
5
6
7

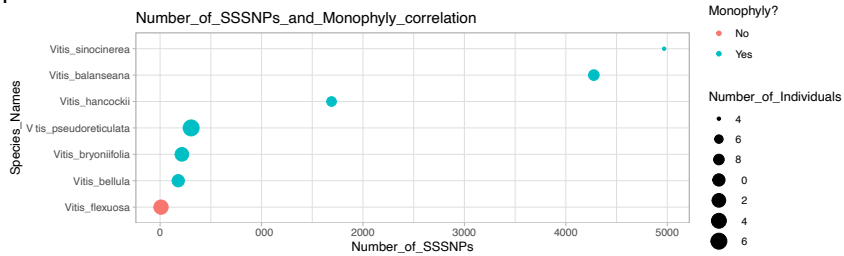
S5.49


Number_of_SSSNPs_and_Monophyly_correlation
Bi lia_rosea
Aesculus_californica
Handeliodendron_bodinieri
Aesculus_parviflora
Aesculus_parryi
Aesculus_indica
Aesculus_glabra
Aesculus_turbinata
Aesculus_pavia
Aesculus_hippocastanan
Aesculus_assamica
Aesculus_wangii
Aesculus_chinensis
Aesculus_flava
Aesculus_sylvatica
Species_Names
0
000
2000
3000
Number_of_SSSNPs
Monophyly?
Yes
Number_of_Individuals
2
3
4
5
6
7

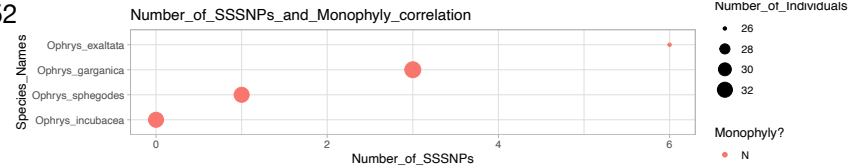
S5.50


Number_of_SSSNPs_and_Monophyly_correlation
Linaria_clementei
Linaria_becerrae
Linaria_algarviana
Linaria_pedunculata
Linaria_salzmannii
Linaria_incarnata
Linaria_onubensis
Linaria_elegans
Linaria_nigricans
Linaria_viscosa
Linaria_spartea
Linaria_multicaulis
Linaria_maroccana
Species_Names
0
500
000
500
2000
Number_of_SSSNPs
Number_of_Individuals
2
4
6
8
0
Monophyly?
No
Yes

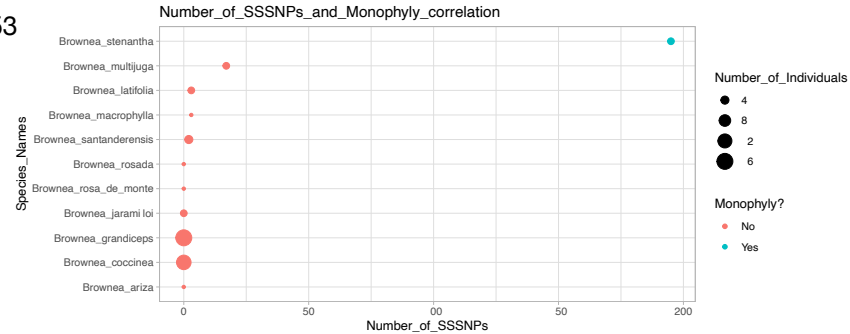
S5.51



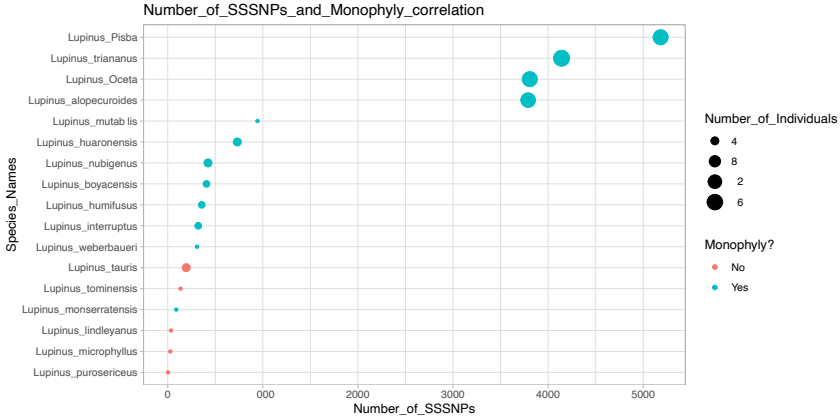
S5.52



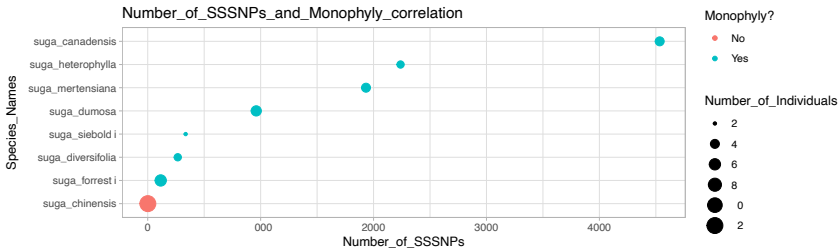
S5.53



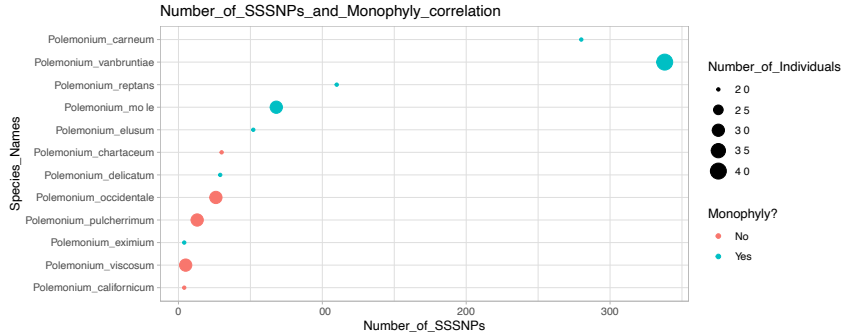
S5.54


Number_of_SSSNPs_and_Monophyly_correlation
Lupinus_Pisba
Lupinus_triananus
Lupinus_Oceta
Lupinus_alopecuroides
Lupinus_mutab lis
Lupinus_huaronensis
Lupinus_nubigenus
Lupinus_boyacensis
Lupinus_humifusus
Lupinus_interruptus
Lupinus_weberbaueri
Lupinus_tauris
Lupinus_tominensis
Lupinus_monserratensis
Lupinus_lindleyanus
Lupinus_microphyllus
Lupinus_purosericeus
Species_Names
0
000
2000
3000
4000
5000
Number_of_SSSNPs
Number_of_Individuals
4
8
2
6
Monophyly?
No
Yes

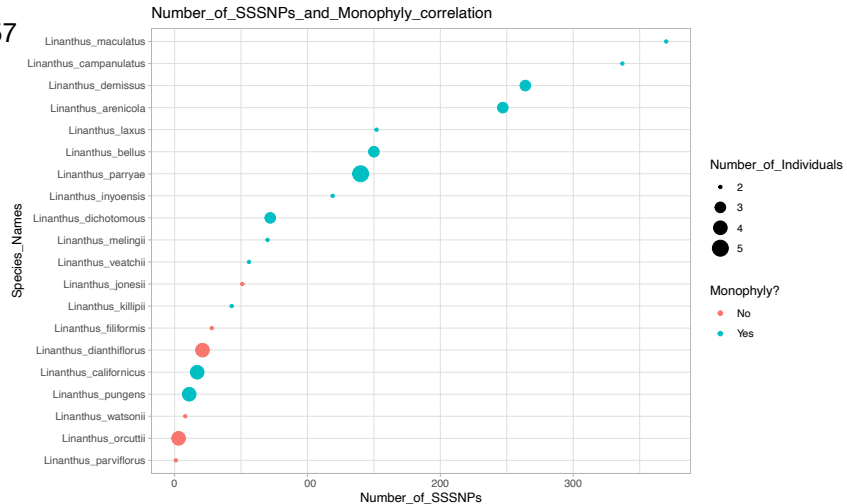
S5.55


Number_of_SSSNPs_and_Monophyly_correlation
suga_canadensis
suga_heterophylla
suga_mertensiana
suga_dumosa
suga_siebold i
suga_diversifolia
suga_forrest i
suga_chinensis
Species_Names
0
000
2000
3000
4000
Number_of_SSSNPs
Monophyly?
No
Yes
Number_of_Individuals
2
4
6
8
0
2

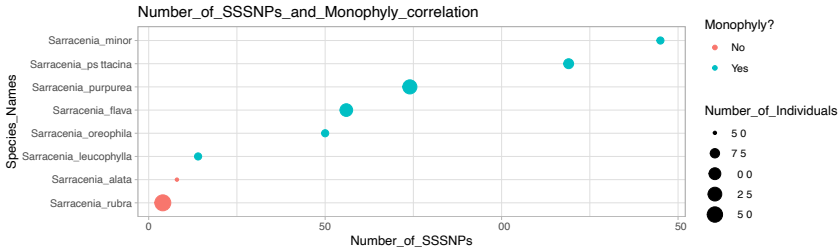
S5.56


Number_of_SSSNPs_and_Monophyly_correlation
Polemonium_carneum
Polemonium_vanbruntiae
Polemonium_reptans
Polemonium_mo le
Polemonium_elusum
Polemonium_chartaceum
Polemonium_delicatum
Polemonium_occidentale
Polemonium_pulcherrimum
Polemonium_eximium
Polemonium_viscosum
Polemonium_californicum
Species_Names
0
00
200
300
Number_of_SSSNPs
Number_of_Individuals
2 0
2 5
3 0
3 5
4 0
Monophyly?
No
Yes

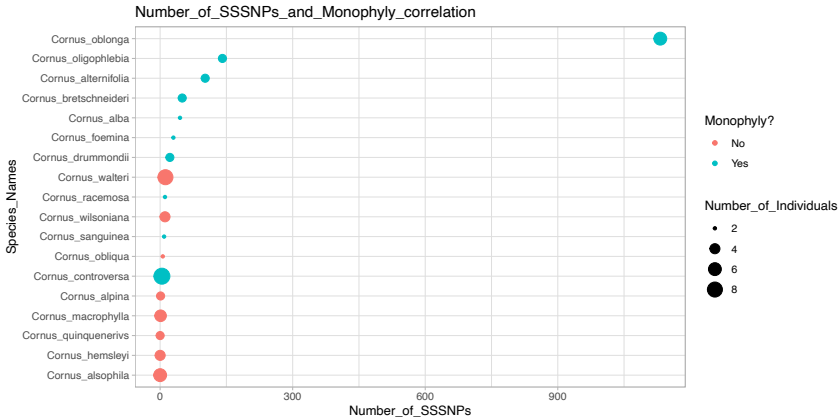
S5.57


Number_of_SSSNPs_and_Monophyly_correlation
Linanthus_maculatus
Linanthus_campanulatus
Linanthus_demissus
Linanthus_arenicola
Linanthus_laxus
Linanthus_bellus
Linanthus_parryae
Linanthus_inyoensis
Linanthus_dichotomous
Linanthus_melingii
Linanthus_veatchii
Linanthus_jonesii
Linanthus_killipii
Linanthus_filiformis
Linanthus_dianthiflorus
Linanthus_californicus
Linanthus_pungens
Linanthus_watsonii
Linanthus_orcuttii
Linanthus_parviflorus
Species_Names
0
00
200
300
Number_of_SSSNPs
Number_of_Individuals
2
3
4
5
Monophyly?
No
Yes

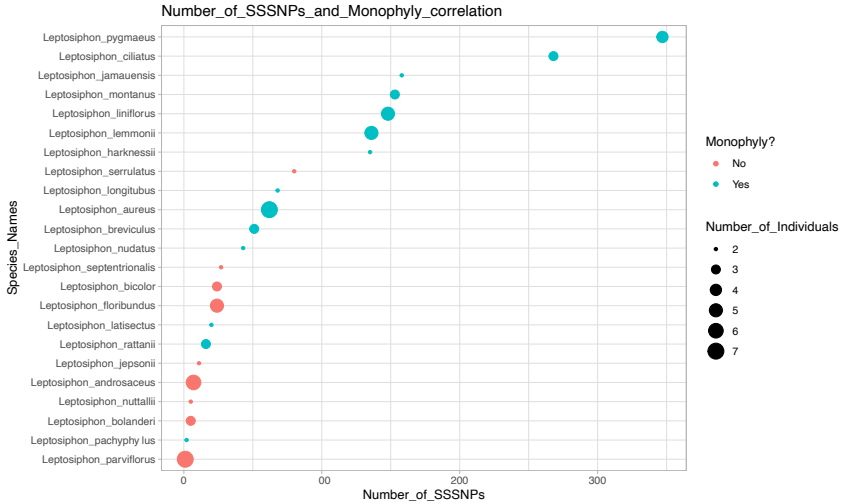
S5.58


Number_of_SSSNPs_and_Monophyly_correlation
Species_Names
Sarracenia_minor
Sarracenia_ps ttacina
Sarracenia_purpurea
Sarracenia_flava
Sarracenia_oreophila
Sarracenia_leucophylla
Sarracenia_alata
Sarracenia_rubra
0
50
00
50
Number_of_SSSNPs
Monophyly?
No
Yes
Number_of_Individuals
5 0
7 5
0 0
2 5
5 0

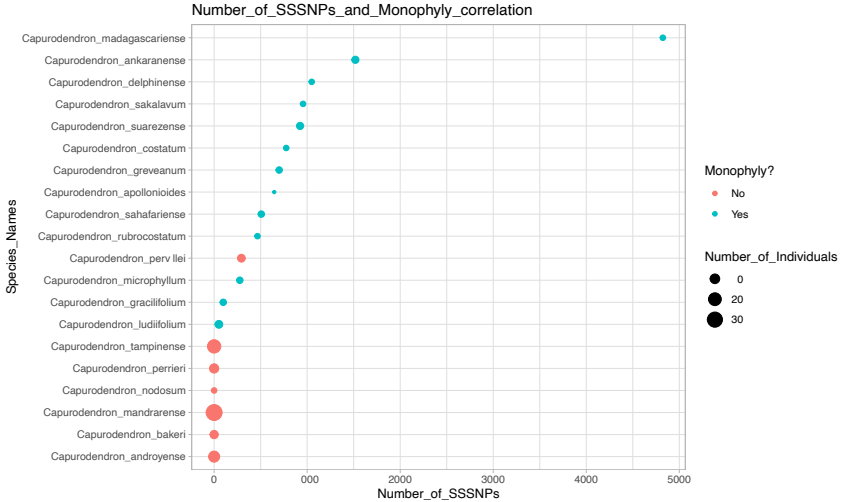
S5.59


Number_of_SSSNPs_and_Monophyly_correlation
Species_Names
Cornus_oblonga
Cornus_oligophlebia
Cornus_alternifolia
Cornus_bretschneideri
Cornus_alba
Cornus_foemina
Cornus_drummondii
Cornus_walteri
Cornus_racemosa
Cornus_wilsoniana
Cornus_sanguinea
Cornus_obliqua
Cornus_controversa
Cornus_alpina
Cornus_macrophylla
Cornus_quinquenerivs
Cornus_hemsleyi
Cornus_alsophila
0
300
600
900
Number_of_SSSNPs
Monophyly?
No
Yes
Number_of_Individuals
2
4
6
8

S5.60


Number_of_SSSNPs_and_Monophyly_correlation
Leptosiphon_pygmaeus
Leptosiphon_ciliatus
Leptosiphon_jamauensis
Leptosiphon_montanus
Leptosiphon_liniflorus
Leptosiphon_lemmonii
Leptosiphon_harknessii
Leptosiphon_serrulatus
Leptosiphon_longitubus
Leptosiphon_aureus
Leptosiphon_breviculus
Leptosiphon_nudatus
Leptosiphon_septentrionalis
Leptosiphon_bicolor
Leptosiphon_floribundus
Leptosiphon_latisectus
Leptosiphon_rattanii
Leptosiphon_jepsonii
Leptosiphon_androsaceus
Leptosiphon_nuttallii
Leptosiphon_bolanderi
Leptosiphon_pachyphy lus
Leptosiphon_parviflorus
Species_Names
0
00
200
300
Number_of_SSSNPs
Monophyly?
No
Yes
Number_of_Individuals
2
3
4
5
6
7

S5.61


Number_of_SSSNPs_and_Monophyly_correlation
Capurodendron_madagascariense
Capurodendron_ankaranense
Capurodendron_delphinense
Capurodendron_sakalavum
Capurodendron_suarezense
Capurodendron_costatum
Capurodendron_greveanum
Capurodendron_apollonioides
Capurodendron_sahafariense
Capurodendron_rubrocostatum
Capurodendron_perv llei
Capurodendron_microphyllum
Capurodendron_gracilifolium
Capurodendron_ludiifolium
Capurodendron_tampinense
Capurodendron_perrieri
Capurodendron_nodosum
Capurodendron_mandrarense
Capurodendron_bakeri
Capurodendron_androyense
Species_Names
0
000
2000
3000
4000
5000
Number_of_SSSNPs
Monophyly?
No
Yes
Number_of_Individuals
0
20
30

S5.62

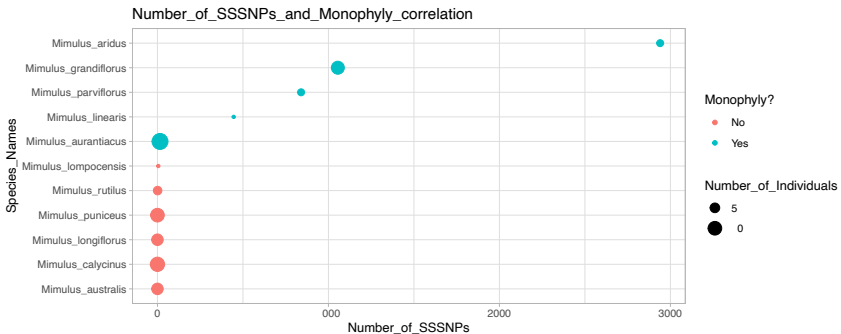

Number_of_SSSNPs_and_Monophyly_correlation
Mimulus_aridus
Mimulus_grandiflorus
Mimulus_parviflorus
Mimulus_linearis
Mimulus_aurantiacus
Mimulus_lompocensis
Mimulus_rutilus
Mimulus_puniceus
Mimulus_longiflorus
Mimulus_calycinus
Mimulus_australis
Species_Names
0
000
2000
3000
Number_of_SSSNPs
Monophyly?
No
Yes
Number_of_Individuals
5
0

Figure S5.63. The proportion of species discriminated using different numbers of sub-sampled SNPs for genera in different sequencing methods.