Phylogenomic Studies in Heathers (*Erica* L.)

DISSERTATION

Seth D. Musker

2022

Phylogenomic Studies in Heathers (*Erica* L.)

DISSERTATION

in fulfilment of the requirements for a

Doctorate in Natural Sciences

(Dr. rer. nat)

Faculty of Biology, Chemistry and Earth Sciences

of the University of Bayreuth

presented by

Seth Daniel Musker

from *Johannesburg, South Africa*

Bayreuth, December 2022

This doctoral thesis was prepared at the Department of Biology, Chair of Plant Systematics in Bayreuth from March 2019 to November 2022 under the supervision of Dr. Nicolai M. Nürk, Prof. Dr. Michael D. Pirie, and Prof. Dr. G. Anthony Verboom, and was funded by the Deutsche Forschungsgemeinschaft (PI 1169/1-2).

This is a full reprint of the thesis submitted to obtain the academic degree of Doctor of Natural Sciences (Dr. rer. nat.) and approved by the Faculty of Biology, Chemistry and Geosciences of the University of Bayreuth.

Date of submission: 01.12.2022

Date of defence: 08.05.2023

Acting dean: Prof. Dr. Benedikt Westermann

Doctoral committee:

Dr. Nicolai Nürk (reviewer)

Prof. Dr. Carl Beierkuhnlein (reviewer)

Prof. Dr. Steven Higgins (chairman)

Prof. Dr. Michael Pirie

I dedicate this thesis to my friends, family, and Gemma.

Thank you all.

Erica praecox, Klein Wellington Sneeukop.

Table of contents

List of figures

List of tables

Nomenclature

Units

bp	Base pair
Gb	Gigabase, billion base pairs
Ma	Mega-annum, million years before present
Mb	Megabase, million base pairs

Acronyms / Abbreviations

CDS	Coding sequence
CFR	Cape Floristic Region
ETS	External Transcribed Spacer
GBS	Genotyping-by-sequencing
GEE	Gene tree estimation error
ILS	Incomplete lineage sorting
Indel	Insertion/deletion
ITS	Internal Transcribed Spacer
LPP	Local posterior probability
MLH	Multilocus heterozygosity
ML	Maximum likelihood
MSA	Multiple sequence alignment
MSC	Multispecies coalescent
PCA	Principal component analysis
PoMo	Polymorphism-aware model
QIRP	Quartet internode resolution probability

SD Standard deviation

SH-alrt Shimodaira–Hasegawa approximate likelihood ratio test

sNMF sparse non-negative matrix factorization

SNP Single nucleotide polymorphism

VCF Variant call format

WGS Whole-genome sequencing

Summary

Systematists study the diversity of life on Earth, aiming to describe its variety of forms and the relationships between them, as well as to understand the processes that influence changes in diversity over time and space. One of the most striking aspects of Earth's biodiversity is that its distribution is highly heterogeneous, varying enormously not just between geographic regions but also between lineages. One place that exemplifies this is the Cape Floristic Region (CFR), a global biodiversity hotspot that hosts roughly 9,000 vascular plant species, of which nearly 70% are found nowhere else.

The CFR flora comprises a taxonomically unusual mixture of lineages whose origins lie in Africa, South America, Australia, and Europe. One of its European-origin components, the heathers (genus *Erica*), stands out as a remarkable example of floristic diversity globally. Out of a global total of around 850 species, almost 700 are found in the CFR, all of which share a single common ancestor that arrived in the region at the earliest around 15 million years ago. Almost immediately after its arrival in the Cape, *Erica* began to rapidly diversify, attaining a large variety of novel forms. The reasons for this exceptional diversity, however, remain unclear.

In this thesis, I aimed to investigate the diversification of Cape *Erica* by applying recently developed genomic methods to infer inter- and intraspecific relationships in much finer detail than has previously been achieved.

I begin by introducing the study of biological diversification in general, in the context of the CFR, and in the context of *Erica*. In the next chapter I develop a suite of resources to better enable genome-scale phylogenetics (i.e., phylogenomics) in *Erica* using a genome sampling approach known as target capture, and show that it provides high quality, informative data. In the third chapter I apply this new resource to an unresolved phylogenetic problem regarding the recent diversification of a charismatic group of Cape *Erica*, the *E. abietina/E. viscaria* clade. This results in the resolution of some long-standing taxonomic questions, uncovers evidence of interspecific hybridisation, but also indicates a high degree of uncertainty regarding phylogenetic relationships at deep and shallow phylogenetic levels alike. However, rather than indicating a lack of statistical power this uncertainty is shown to more likely be a direct consequence of historical biological processes such as incomplete lineage sorting and rapid diversification.

In the fourth chapter I focus in on *E. abietina*, a species complex that shows evidence of recent, rapid phenotypic diversification, aiming to explore the dynamics of diversification in its earliest stages at the interface of micro- and macroevolution. To do so I employ genotyping-by-sequencing, another genome sampling method that is, relative to target capture, better suited to investigating genetic relationships at such a shallow scale. This reveals a highly dynamic system that is a product of the interplay between divergent selection on floral traits, adaptation to different environments, geographic isolation, secondary contact, and both recent and ancient introgression.

Lastly, I conclude with a discussion of what the results of the thesis imply about the modes and drivers of diversification in Cape *Erica*.

Zusammenfassung

Systematiker untersuchen die Vielfalt des Lebens auf der Erde mit dem Ziel, ihre Formenvielfalt und die Beziehungen zwischen ihnen zu beschreiben sowie die Prozesse zu verstehen, die Veränderungen in der Vielfalt über Zeit und Raum beeinflussen. Einer der auffälligsten Aspekte der Biodiversität der Erde ist, dass ihre Verteilung sehr heterogen ist und nicht nur zwischen geografischen Regionen, sondern auch zwischen Abstammungslinien enorm variiert. Ein Beispiel dafür ist die Kapflora (Cape Floristic Region, CFR), ein globaler Biodiversitäts-Hotspot, der rund 9,000 Gefäßpflanzenarten beherbergt, von denen fast 70% nirgendwo sonst zu finden sind.

Die Kapflora umfasst eine taxonomisch ungewöhnliche Mischung von Linien, deren Ursprünge in Afrika, Südamerika, Australien und Europa liegen. Eine ihrer ursprünglich europäischen Komponenten, die Heide (Gattung *Erica*), sticht als bemerkenswertes Beispiel für die weltweite floristische Vielfalt hervor. Von insgesamt rund 850 Arten weltweit kommen fast 700 in der Kapregion vor, die alle einen einzigen gemeinsamen Vorfahren haben, der vor etwa frühestens 15 Millionen Jahren in die Region kam ("Kap *Erica*" Klade). Fast unmittelbar nach seiner Ankunft in der Kapregion, *Erica* begann sich schnell zu diversifizieren und erreichte eine große Vielfalt neuartiger Formen. Die Gründe für diese außergewöhnliche Diversität bleiben jedoch unklar.

In dieser Arbeit wurde die Diversifizierung von *Erica* in der Kapregion untersucht. Moderne genomische Methoden wurden anwendet, um inter- und intraspezifische Beziehungen in viel feineren Detail abzuleiten, als dies bisher möglich war.

Ich beginne damit, das Studium der biologischen Diversifizierung im Allgemeinen im Kontext des Kapflora und im Kontext von *Erica* vorzustellen. Im nächsten Kapitel entwickle ich eine Reihe von Ressourcen, um eine Phylogenetik im Genommaßstab (Phylogenomik) in *Erica* zu ermöglichen, indem ich einen als "target capture" bekannten Ansatz zur Genomprobenahme verwende, und zeige, dass er qualitativ hochwertige, informative Daten liefert. Im dritten Kapitel wende ich diese neue Ressource auf ein ungelöstes phylogenetisches Problem bezüglich der jüngsten Diversifizierung einer charismatischen Gruppe von Kap *Erica*, der *E. abietina/E. viscaria*-Klade. Dies führt zur Lösung einiger seit langem bestehender taxonomischer Fragen, deckt Hinweise auf interspezifische Hybridisierung auf, weist aber auch auf ein hohes Maß an Uneindeutlichkeit in Bezug auf phylo-

genetische Beziehungen auf tiefer und flacher phylogenetischer Ebene hin. Anstatt auf einen Mangel an statistischer Aussagekraft hinzuweisen, zeigt sich jedoch, dass diese Unsicherheit eher eine direkte Folge historischer biologischer Prozesse ist, wie unvollständige Allelsortierung (incomplete lineage sorting) und explosieve Diversifizierung.

Im vierten Kapitel konzentriere ich mich auf *E. abietina*, einen Artenkomplex, der Hinweise auf eine kürzliche, schnelle phänotypische Diversifizierung zeigt, mit dem Ziel, die Dynamik der Diversifizierung in ihren frühesten Stadien an der Schnittstelle von Mikro- und Makroevolution zu untersuchen. Dazu verwende ich Genotyping-by-Sequencing, eine weitere genomische Methode, die im Vergleich zur "target capture" besser geeignet ist, um genetische Beziehungen in einem so flachen Maßstab zu untersuchen. Dies offenbart ein hochdynamisches System, das ein Produkt des Zusammenspiels zwischen unterschiedlicher Selektion auf florale Merkmale, Anpassung an unterschiedliche Umgebungen, geografische Isolation, sekundären Kontakt und sowohl rezenter als auch alter Introgression ist.

Ich schließe mit einer Diskussion darüber, was die Ergebnisse der Dissertation über Tempo und Modus der Diversifizierung in Kap *Erica* implizieren.

Chapter 1

Introduction

1.1 Biodiversity — patterns and processes

The distribution of species diversity across the tree of life is uneven relative to time, such that a clade's age does not predict its size (Rabosky et al., 2012). What, then, enables some lineages to diversify more rapidly and more prolifically than others? This fundamental question in evolutionary biology branches into more specific problems. How do the links between more or less freely interbreeding individuals in a population become severed, creating divergent lineages (Rieseberg and Willis, 2007; Templeton, 1981), and why does this appear to happen more often in some clades than in others (The Marie Curie SPECIATION Network, 2012)? To what extent do macroevolutionary patterns reflect the microevolutionary forces that generated them and that are currently at play (Aristide and Morlon, 2019; Overcast et al., 2021; Weber et al., 2017)? What are the dynamics of diversification in its earliest stages (Gottlieb, 2004)?

1.2 The Cape Floristic Region

Just as none of these questions can be considered in isolation, we also cannot ignore their geographic context — the arenas in which processes that influence life's diversity take place. At the south-western tip of Africa lies the Cape Floristic Region (CFR), a global hotspot of botanical diversity with over 9000 vascular plant species (Manning and Goldblatt, 2012). In the context of the world's five

Mediterranean-type ecosystems, the CFR's unusually small size makes it by far the most diverse per unit land area (Manning and Goldblatt, 2012). Apart from this, the CFR has several other unusual features that make it particularly intriguing to systematists, whose primary goal is to understand the patterns and processes underlying the diversity of life on Earth.

One of the most prominent features of the CFR is the Cape Fold Belt, an extensive mountain range that dominates the landscape. Its rocks mostly belong to the Cape Supergroup, a group of sandstone Formations whose 500-million year history has, in some sense, culminated in the spectacular floristic diversity of the CFR. During the early stages of the Supergroup's formation in the early Ordovician (Shone and Booth, 2005), plants had only just begun to colonise the Earth's land surface (Morris et al., 2018). Without abundant vascular plants to stabilise the soil, rivers' banks were highly mobile and the hard, sandy sediments they deposited became spread over large areas as they shifted and changed course over millions of years (Shone and Booth, 2005). Once buried, those sand deposits would go on to form kilometres-thick columns of extremely durable rock. Later, when plants eventually began to constrain the movement of rivers, sand deposition became much more localised and much finer particles came to dominate sedimentary deposits, which went on to form relatively soft rocks. Long after these rock layers had formed, around 250 Ma, the formation of Pangaea coincided with a mountain-building event of grand scale that subjected the deeply buried sandstone beds to immense pressure, under which they buckled, folded and deformed (Hansma et al., 2016). Over time the overlying, softer rocks were eroded away and the underlying sandstones were exposed, and because of their extensive deformation, the sandstones stood out from the surrounding landscape as rugged mountains. This exposure is thought to have happened by at least 145 Ma in the Late Jurassic to Early Cretaceous (Muir et al., 2017), and since then the landscape is believed to have changed very little due to extremely slow erosion rates (Scharf et al., 2013).

The irony of the history of the Cape landscape is that, had the absence of land plants half a billion years ago not allowed for the formation of its rocks, its present-day floristic hyper-diversity may never have developed — at least not in its present form. This is because the Cape's rocks give its landscape many of the features that are thought to have fostered its diversity. Firstly, the very slow pace of erosion has provided the Cape flora with millions of years of a relatively stable landscape. Secondly, the mountains have acted as a buffer against past climate change, causing rain, lowering

temperatures, and thus shielding plants from drought (Bradshaw and Cowling, 2014). Thirdly, the extremely low mineral complexity of the sandstones has given the CFR some of the most nutrient-poor soils on Earth (Stock and Verboom, 2012). Consequently, its extremely specialised flora (Verboom et al., 2017), known collectively as fynbos, has evolved in relative isolation as very few plant groups can tolerate such extreme edaphic conditions (Lu et al., 2022; van Santen and Linder, 2020). The Cape flora's low extinction rates have been attributed to these three factors (Cowling et al., 2015; Verboom et al., 2009). Lastly, the Cape's rugged topography is thought to have increased the pace of diversification. By introducing physical barriers, the Cape's mountains and valleys have acted to inhibit gene flow between populations of mountain-adapted plants (Verboom et al., 2015). At the same time, the ruggedness has given rise to a heterogeneous landscape of sharp climatic and hydrological gradients associated with slope, aspect, and elevation, inducing fine-scale niche partitioning and high rates of adaptive divergence (Araya et al., 2010; Goldblatt and Manning, 2002).

Of course, many other factors have undoubtedly played important roles in the Cape flora's diversification. Other important abiotic factors include edaphic variation (Schnitzler et al., 2011), fire (Cowling, 1987), and fine-scale geomorphic evolution (Cowling et al., 2009; Hoffmann et al., 2015). Biotic factors have certainly also been at play. Ecological interactions such as pollination and competition are thought to have had important roles in both promoting and constraining diversification (Johnson, 1996; Slingsby et al., 2014; van Der Niet et al., 2014). Adding to all of this complexity is the fact that not all plant groups respond to such external factors in the same way — each has its own set of traits that determine its sensitivity and means of responding to change and variation (Donoghue, 2008). When imagining diversification as a function of time, the abiotic environment, biotic interactions, and traits (Nürk et al., 2020), it becomes clear that no single parameterisation of this function could possibly explain the diversity of a region such as the CFR. Instead, moving towards a deeper understanding of the region's diversity will continue to require detailed investigation into the factors underpinning the diversification of its individual lineages.

1.3 The genus *Erica*

The CFR is dominated by a relatively small number of lineages, one of which is the heathers, *Erica* L. (Ericaceae). Just under 700 of the CFR's 9000-odd plant species belong to *Erica*, making it well over twice the size of the next-largest genus in the CFR, *Aspalathus* L. (Fabaceae), which holds about 273 species (Manning and Goldblatt, 2012). There is strong evidence to suggest that all Cape *Erica* belong to a single clade (Pirie et al., 2016) whose ancestors are thought to have slowly dispersed southwards from Europe (around 40 Ma) via the African Highlands, eventually reaching the CFR by around 10 Ma, after which they underwent a remarkable surge of diversification (Pirie et al., 2019). This diversity of species is matched by a similarly impressive variety of forms. Cape heathers range from miniscule, creeping herbs (e.g., *E. oxycoccifolia* Salisb.) to tall, almost tree-like shrubs (e.g., *E. brachialis* Salisb.), and their flowers come in a variety of colours, shapes, and sizes, reflecting an array of pollination syndromes (Rebelo et al., 1985). Within *Erica* this degree of floral diversity is unique to the Cape species, far exceeding that of the rest of the genus. The unevenness of *Erica* diversity is also striking in a geographic context: outside of the CFR its range spans Europe and includes Madagascar and the African highlands, and yet this vast area hosts fewer than 200 species. Even within the CFR, the south-western region is a mini-hotspot of *Erica* diversity (Fig. 1.1). Clearly, something about the CFR has catalysed speciation in *Erica* in a way that no other region appears to have done, and investigating what that was may well help to reveal the dynamics involved in the evolution of the CFR's flora more broadly (Linder, 2003).

It has been suggested that frequent pollinator shifts have contributed to plant diversification in the CFR in general (e.g., Johnson, 1996), and in Cape *Erica* this idea is supported by the apparent lability of floral traits throughout the phylogeny (Pirie et al., 2011). Insect pollination predominates (*ca.* 80% of species; Rebelo et al., 1985), though the variety of floral colours, shapes, sizes, and scents implies a similar variety of insect pollinators (Fig. 1.2; Newman and Johnson, 2021; Rebelo et al., 1985; van Der Niet et al., 2014). Wind-pollination is also fairly common, and some of the most abundant and widespread species are anemophilous (e.g., *E. hispidula*). As in many other Cape lineages, pollination by birds is also fairly common (Rebelo et al., 1984). Bird-pollinated *Erica* species occur almost exclusively in the CFR, and most are pollinated by the CFR-endemic Orange-breasted Sunbird

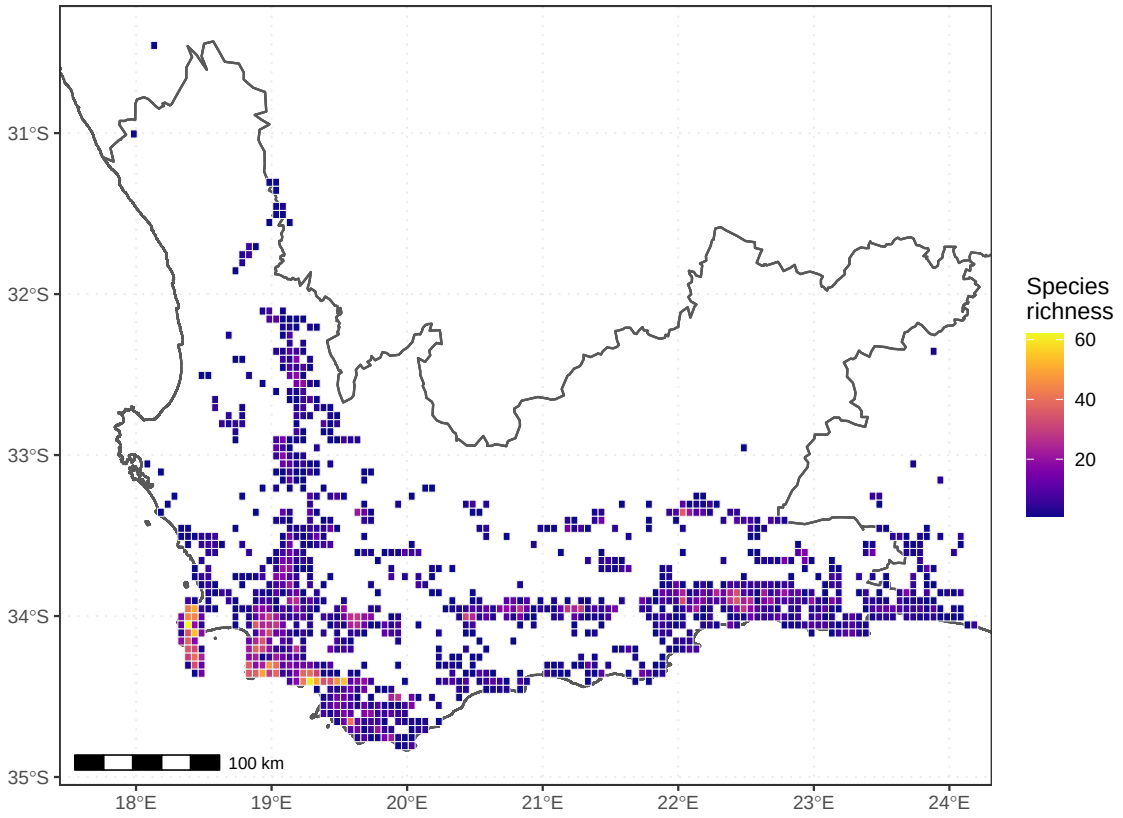


Fig. 1.1 Map showing the distribution of species richness in *Erica* in the CFR. The centre of diversity lies in the south-western Cape, much like the rest of the Cape flora. Research-grade observations were downloaded from `iNaturalist.org` on 20.11.2022. The border of the Western Cape province is shown for reference.

(*Anthobaphes violacea*; Coetzee, 2016; Coetzee et al., 2020; Rebelo et al., 1984). These nectarivores derive much of their nectar from *Erica* species and have long (18-26 mm), narrow, downward-curving bills (Roberts et al., 2005). As such, sunbird-pollinated *Erica* tend to have corolla tubes that match the dimensions of the birds' bills (Barnes et al., 1995; Rebelo et al., 1984, 1985), whereas insect-pollinated species usually have short corollas (Rebelo et al., 1984). Insect-pollinated species typically have pink flowers, whereas bird-pollinated species often have red flowers (Oliver and Oliver, 2002, 2005; Rebelo et al., 1984, 1985) presumably to match the visual systems of birds (Shrestha et al., 2013).

In Cape *Erica* there appear to have been many independent shifts between the various modes of pollination. For example, although a few major clades hold most of the bird-pollinated species, they also hold many taxa with short (< 10 mm), open corolla tubes indicative of insect pollination. As such, morphological and phylogenetic evidence suggests that shifts between these floral types have

Fig. 1.2 A small selection of animal visitors to flowers of Cape *Erica* species. *Left*: *E. abietina* subsp. *abietina* is probed by a female Orange-breasted Sunbird (*Anthobaphes violacea*). *Centre*: A long-proboscid fly (family Nemestrinidae) prepares to probe a flower of *E. daphniflora*. *Right*: A Cape Honey Bee (*Apis mellifera* subsp. *capensis*) enters a flower of *E. abietina* subsp. *constantiana* to drink nectar and collect pollen (personal observation). Photographs taken from `iNaturalist.org` (observation IDs are shown).

been frequent (Oliver and Oliver, 2002, 2005; Pirie et al., 2017, 2011). At the same time they have not been only one-directional. In one example, within the *E. plukenetii* L. species complex whose forms are predominantly sunbird-pollinated, a range-restricted population (*E. plukenetii* subsp. *breviflora*) has been shown to have shifted to moth pollination via a reduction in petal length, a shift from red to white flowers, and the synthesis of a range of scent compounds (Le Maitre et al., 2019a; van Der Niet et al., 2014). Although roughly 80% of Cape *Erica* have insect-pollination traits, compared to around 15% with bird-pollination traits (Rebelo et al., 1985), the frequency of bird-insect transitions suggests that diversifying selection has been at play. Though less effort has been put into studying other types of pollinator shifts in *Erica*, such as between insect and wind pollination or between different types of insect pollination (e.g., generalist pollinators *versus* long-proboscid flies; Newman and Johnson, 2021), these are likely to have been similarly influential.

What remains unclear is how and to what extent such shifts have stimulated speciation and contributed to Cape *Erica* diversity, and whether additional factors, such as geographic isolation and other forms of niche divergence, have played supporting or superior roles. This lack of certainty may stem from a lack of data. While the usual difficulties of inferring historical processes of diversification from present-day patterns still apply (Via, 2009), in recent years it has become clear that genomic tools can nevertheless provide significant insights even in non-model groups (McCormack et al., 2013). To date, very little genomic research has focused on *Erica* (Kadlec et al., 2017; Le Maitre et al.,

2019b). However, the genus, in particular the Cape clade, presents as an excellent candidate in which a genomic approach may bear fruit. There are two main avenues of research in which this seems likely to be the case. Firstly, without the ability to generate robust phylogenetic hypotheses it is difficult to make inferences about the macroevolutionary processes that influence the dynamics of diversification. Despite considerable effort involving extensive taxon sampling and employing several molecular phylogenetic markers, the Cape *Erica* clade's phylogenetic relationships have proved exceptionally difficult to resolve (Pirie et al., 2017, 2016). Secondly, studying the processes that operate to drive divergence at the microevolutionary scale – and which may ultimately manifest as macroevolutionary diversity patterns – can benefit substantially from an understanding of the genetic relationships between the populations involved (Avise et al., 1987, 2000). Population genetic studies in Cape *Erica* are few in number (Ojeda et al., 2016; van Der Niet et al., 2014), while population genomics has never been applied. The goal of this thesis was to contribute to resolving these shortcomings by taking phylogenomic and population genomic approaches to the study of diversification in Cape *Erica*.

Chapter 2

Improving phylogenomic resources for *Erica*

2.1 Background

The advent and ever-decreasing cost of next-generation sequencing (NGS) technologies has paved the way for genome-scale phylogenetics (i.e., "phylogenomics") in non-model organisms (McCormack et al., 2013). Angiosperm genomes vary considerably in size, structure and composition but are generally large and complex (Murat et al., 2012). These factors make whole-genome sequencing (WGS) not only prohibitively expensive but also impractical for most plant phylogeneticists, who typically rely on multiple sequence alignments (MSAs) of orthologous genes or loci for phylogenetic inference (e.g., Minh et al. 2020, Stamatakis 2014; but see e.g., Springer et al. 2019; Zhao et al. 2021). One particularly effective and increasingly popular method that enables phylogenomics is target capture (also termed hybrid selection; Gnirke et al., 2009). In basic terms, this technique subsamples the genome by capturing genomic regions of interest, thereby excluding unwanted regions prior to sequencing. This is accomplished with the use of "baits": short, typically 120 bp long, synthetic biotinylated RNA fragments. During library preparation, thousands of these baits with various sequences bind to matching sequences in the genomic DNA in a process called "hybridisation", forming complexes that are then isolated from the rest of the genomic DNA using streptavidin-coated magnetic beads. These "enriched" libraries can then be sequenced with a much greater degree of

multiplexing (i.e. combining multiple, indexed samples in one sequencing run) than would otherwise be possible, making the method highly cost-effective.

In principle, baits can be designed to capture any genomic region provided that the region's sequence is known or can be approximated to within a certain lower threshold of similarity (typically 80% sequence identity). In practice, however, deciding which regions to target and being able to successfully capture them are each undertakings that come with significant challenges. In this chapter I begin by introducing these challenges and discussing them in the context of the genus *Erica*. I then develop and implement a design approach that aims to strike a balance between the various conflicting requirements of a multi-purpose target set. This involved (1) refining a pre-existing target set using data derived from it, followed by (2) adding more targets derived from several recently published high-quality *Rhododendron* genomes, and (3) using new WGS data from three *Erica* species to quality check the new targets and produce *Erica*-specific versions of many of them. Lastly, I assess the new target set's completeness, quality and informativeness for *Erica* phylogenetics, and evaluate how different target design choices influence these variables.

2.1.1 Considerations when designing a target set

Orthology

One of the basic assumptions of all tree reconstruction methods that use MSAs is that all sequences trace back *only* via successive speciation events to a single common ancestor, that being the root of the tree. This property, a form of homology termed orthology (Avise and Robinson, 2008; Fitch, 1970), is not always straightforward to verify because the signal of orthology tends to erode over time. This can happen abruptly following events such as genome duplication, chromosomal rearrangements (e.g., inversions) and horizontal gene transfer, but it is also an inevitable result of smaller mutations – insertions, deletions and substitutions – accumulating over very long time periods. Considerable effort has been put into identifying orthologous genes that have been retained across deeply divergent groups and can therefore be fairly reliably recovered using target capture. For angiosperms these most notably include the "Angiosperms353" bait set (Johnson et al., 2019) and the "mostly single-copy" gene set identified by De Smet et al. (2013). The Angiosperms353 set was designed for ease-of-use and

universality, and comes in the form of just over 75,000 bait sequences targeting 353 genes identified by Johnson et al. (2019) as being well-conserved and mostly single copy across over 600 angiosperm genomes. A bait "kit" containing the synthesised bait sequences ready for use in target capture library preparation is commercially available. In contrast, the De Smet et al. set is typically used in conjunction with the software MARKERMINER (Chamala et al., 2015), which takes one or more reference transcriptomes from members or close relatives of the angiosperm group being studied and identifies within them the copy (or copies) of each of the De Smet et al. genes. The output is a custom set of targets from which bait sequences can be designed and synthesised prior to target capture.

Paralogy

Whole-genome duplications and other types of polyploidisation have been common and highly influential throughout the evolutionary history of plants (Soltis and Soltis, 2020; Tank et al., 2015). Typically the period following polyploidisation involves the gradual loss of redundant gene copies and the "diploidisation" of the genome, but gene copies are also often retained and adapted to perform slightly different functions (Soltis et al., 2015). Individual genes can also be duplicated in isolation without the occurrence of polyploidisation. Duplications that are retained to the present day result in paralogs (Fitch, 1970): gene copies whose sequences are similar and once shared a common ancestor, and are therefore homologs, but which have evolved separately in the genome since the duplication event, and are therefore not orthologs. Paralogs present both opportunities and challenges for phylogenomics. The sequence similarity of the gene copies means that a bait designed to target one copy is likely to also effectively capture the other(s), providing the researcher with two (or more) genes for the price of one. This can significantly improve species tree inference power if each copy can be treated as an independent locus (Gardner et al., 2021; Ufimov et al., 2022).

The problem of paralogy for phylogeneticists is that in order to be used as independent loci, each gene copy first needs to be distinguished *within* each species and then correctly grouped *across* species; the latter especially is a task that can pose considerable challenges. Of particular importance is the time between the duplication event and the next speciation event: if duplication happens shortly before speciation, relatively few mutations can accumulate in each copy before they go on to evolve independently in each daughter lineage. In distantly related species that share an ancient "duplication-

speciation" event, the paralogs may be more different from each other within species than they are between species. This problem can be compounded by further duplications or independent gene losses (Li et al., 2020). Nevertheless, methods that take on this task have recently been proposed (Ufimov et al., 2022; Zhou et al., 2022). These typically involve computationally demanding steps such as genotype calling and multiple sequence alignment, and have yet to be extensively tested to determine their accuracy across a range of scenarios. An alternative to separating paralogs is to model or account for gene loss and duplication within the analysis (Smith and Hahn, 2021). One recent species tree reconstruction method, ASTRAL-PRO (Zhang et al., 2020), has shown promise in this regard. However, systematists nowadays use target capture data for a variety of analyses apart from species tree reconstruction, most of which cannot (yet) account for paralogy (e.g., phylogenetic network inference, Solís-Lemus et al. 2017; demographic history modelling, Gronau et al. 2011). Therefore, to ease computational burden and reduce the risk of false inferences due to model violations, a versatile target set should ideally have low rates of paralogy.

Informativeness

The ultimate test of a target set's utility is its power to answer the questions posed by the researcher. For phylogenomics this boils down to absolute sequence variation, which typically needs to be sufficient to resolve relationships at multiple levels in the phylogeny. In this context the multispecies coalescent (MSC) model (Avise et al., 1987; Degnan and Rosenberg, 2009; Maddison, 1997) and the process of incomplete lineage sorting (ILS), is particularly relevant. ILS happens when more than one variant (i.e., allele) of a locus is retained in a species following speciation; in other words, the two alleles are not completely "sorted" into the two daughter lineages. If the daughter that inherited both alleles undergoes another speciation event, one of its two daughter lineages can end up with the same allele that was retained in the ancestral lineage. Even though the true species tree might have the topology (A,(B,C)), this sequence of events results in the phylogeny of the locus – the "gene tree" – having a different topology, either (B,(A,C)) or (C,(A,B)). The time between the first and second speciation events is crucial in determining what proportion of the genome is subject to ILS: the shorter it is, the less time there will be for the (eventual) ancestral allele to be lost or for mutations to generate new alleles, making ILS more likely for any given locus (Degnan and Rosenberg, 2009; Maddison,

1997; Townsend et al., 2012). Time periods during which many speciation events happen in quick succession exacerbate the problem of ILS, and in extreme cases can even cause most loci to have phylogenies that misrepresent the "true" species tree (Degnan and Rosenberg, 2006).

The fact that gene trees can misrepresent the species tree is a major problem in phylogenetics, as failing to account for it can potentially cause errors in species tree inference (Jiang et al., 2020). Because of this, several methods have been developed to estimate species trees while accounting for ILS (e.g., Chifman and Kubatko, 2014; Douglas et al., 2022; Zhang et al., 2018). Although generally very powerful and robust to high degrees of ILS, an important weakness of such methods is that they can be confounded when the number of loci is small and/or loci have relatively few variable sites, making them uninformative (Huang et al., 2020; Molloy and Warnow, 2017). To improve the informativeness of their data, researchers can choose to sequence more, longer and/or more variable loci. Variable loci may appeal to those with a limited sequencing budget wanting to maximise the number of loci, but two key factors make them potentially problematic. Firstly, the regions being targeted may fail to be captured if they are too divergent from the bait sequences (Gnirke et al., 2009), which is naturally more likely for loci with high mutation rates and when taxa are more distantly related. Secondly, loci with high rates of sequence evolution are more likely to have had mutations that are informative at deeper phylogenetic levels "written over" by subsequent mutations, and are also more susceptible to homoplasy (i.e., convergence falsely interpreted as common ancestry) – both of which compromise the accuracy of phylogenetic inference (Graybeal, 1994; Yang, 1998).

Most target capture approaches to phylogenomics try to balance the trade-off between informativeness and reliability by basing each target locus on the coding sequence (CDS) of a gene (e.g., Johnson et al., 2019) and ultimately relying on the small proportion of "flanking sequence" captured at the ends of the exons to (usually partially) assemble the more variable intronic regions (Johnson et al., 2016). Most tests have however shown that the phylogenetic informativeness of individual loci is a major limiting factor for "summary coalescent" species tree reconstruction methods (Gatesy and Springer, 2014; Meiklejohn et al., 2016; Roch and Warnow, 2015). This is because, with the exception of the very recently introduced wASTRAL method (Zhang and Mirarab, 2022), summary coalescent methods implement the MSC model by summarising sets of gene (i.e., locus) trees under the assumption of no gene tree estimation error (GEE). The less informative the loci are, the more

likely this assumption is to be violated. Summary coalescent methods are nevertheless presently the dominant species tree inference method owing partly to the computational limitations of even the most advanced sequence-based approaches (Douglas et al., 2022), but also because they have been shown to be highly accurate as long as there are enough loci and GEE is relatively low (Molloy and Warnow, 2017). For concatenation-based phylogenetic inference, preferring more informative loci has also been shown to improve the chance of recovering the correct tree topology especially in parts of the tree that are difficult to resolve due to limited sequence variation (e.g., Salichos and Rokas, 2013). Therefore, despite the potential risks inherent to more variable loci, they are generally preferable especially for researchers working on recalcitrant groups with a high degree of ILS (Meiklejohn et al., 2016).

Because loci recovered from highly conserved genomic regions, such as the Angiosperms353 genes (Johnson et al., 2019) and ultra-conserved elements (UCEs; Faircloth et al., 2012), are by their nature relatively invariant and thus often suffer from reduced phylogenetic informativeness, recent workers have proposed alternatives to the existing status quo which specifically aim to capture more variable loci. Karin et al. (2019) proposed a method to identify long, rapidly evolving exons and applied it successfully to squamate reptiles. Zhang et al. (2019) identified highly variable orthologs in Lepidoptera and other insect genomes, designed PCR primers from their conserved regions, and used the primers to amplify the loci in a pooled DNA sample of five distantly related taxa to generate custom baits. This approach, though laborious, resulted in excellent capture efficiency across the entire lepidopteran order. An approach that specifically aims to recover intronic sequences has been developed in which, rather than using target capture, PCR amplification using exon-derived primers serves to effectively isolate the targeted genes (Li et al., 2010, 2017). Though effective across broad phylogenetic scales, this approach is labour-intensive and is ill-suited to degraded DNA (Li et al., 2017), making it generally unsuitable for "museomics" (Raxworthy and Smith, 2021) and "herbariomics" (Brewer et al., 2019), for which target capture is highly effective. Target capture-based studies in which intronic sequences have explicitly been used for bait design are uncommon, presumably due to concerns about target capture efficiency. Folk et al. (2015) and de Sousa et al. (2014) took such an approach and each reported excellent target capture efficiency resulting in highly informative data; however, both studies involved closely related taxa (with maximum divergence

times of < 10 Ma) which makes it hard to assess how well the approach might perform across deeper phylogenetic levels.

Locus length

Another important aspect of target set design that relates to variability is sequence length. All else being equal, the number of phylogenetically informative sites will increase roughly linearly as more DNA is sequenced. However, the number of distinct genealogies underpinning the variation along a sequence will also be greater for longer sequences. A phylogeny reconstructed from any DNA sequence will represent the (weighted) average phylogeny of the sequence's underlying recombination blocks, whose number and size will depend primarily on the rate of recombination and the size (number of tips) and length (in years) of the tree being inferred; this has been called the "recombination ratchet" (Gatesy and Springer, 2014; Springer and Gatesy, 2016) because successive recombination events increasingly subvert the genealogical consistency of the sequence neighbourhood. When a locus contains more than one recombination block its tree will not recapitulate a single independent coalescent history, and summary coalescent methods will violate this assumption of the MSC model (Springer and Gatesy, 2016). One implication of this is that if a gene's introns are extremely long (i.e., tens of kilobases), the strategy of only targeting exons will result in the recovery of "supercontigs" – separately assembled sequences stitched together based on their order of mapping to the target reference sequence (Johnson et al., 2016) – that are essentially chimeric in that they consist of multiple "coalescence genes" (c-genes) each with an independent genealogy (Springer and Gatesy, 2018, 2016). Knowing whether a gene is likely to have large introns can therefore be valuable when deciding whether to include it in a target set.

Practical limitations

Custom bait designs are priced based on the target "footprint", the total size of the bait set required to capture the full set of targets. This will depend on several factors, including the total length of all targets combined, sequence complexity and uniqueness, and tiling (the degree to which neighbouring baits overlap). In general, researchers with a limited budget who wish to develop a custom target set face a trade-off of more loci *versus* longer loci. Kadlec et al. (2017), facing such a trade-off, aimed

to design a target set capable of resolving relationships among a group of extremely closely related *Erica* species as well as recover loci across Ericaceae. Choosing to maximise variability, they filtered an initial set of 4,649 potential putatively single-copy genes based on their predicted length, ending up with a target set consisting of 132 *Rhododendron* transcripts with a median predicted length of > 2 kb. In comparison, the Angiosperms353 targets have an average length of 738 bp (Johnson et al., 2019). Applying their targets to a set of *Erica* samples, Kadlec et al. (2017) obtained aligned sequence matrices with a mean length of 1,810 bp with 2.6–26.1% variable sites.

2.1.2 The challenge of *Erica* phylogenomics

The genus *Erica* comprises over 800 species and is distributed in Europe and Africa. However, most species (*ca.* 690) are confined to the Cape Floristic Region (CFR) of South Africa, all of which appear to have a single common ancestor (Pirie et al., 2016). This "Cape" clade shows clear indications of recent and rapid diversification which accelerated upon its arrival in the region, with a crown age of 6.0-15.0 Ma and net diversification rates of 0.28-0.97 species.Ma^{-1} – notably higher than in other CFR radiations (Pirie et al., 2016). This surge is responsible for the genus's status as by far the largest in the CFR (Manning and Goldblatt, 2012) and its potential to shed light on the causes of the region's extreme floristic diversity (Linder, 2003). At the same time, it makes it extremely difficult to recover robust phylogenetic hypotheses at the species level, a fact that is well illustrated by the low nodal support values throughout the Cape clade in the most recently published *Erica*-wide phylogeny (Pirie et al., 2016), which was based on a relatively small number of "traditional" plant phylogenetic markers, such as ITS and various chloroplastic regions, which can be affordably sequenced and have been available to botanical systematists since long before the advent of next-generation sequencing.

The democratisation of phylogenomics precipitated much enthusiasm among systematists, in particular those working on difficult phylogenetic problems, who envisioned a new era in which long-standing problematic relationships could finally be conclusively resolved (Delsuc et al., 2005). However, due to their size and complexity the reality is that phylogenomic data sets require considerable care when being designed, generated, curated, and analysed, and failure to do so can in the worst cases produce misleading results and spurious inferences (Gatesy et al., 2019; Hahn and Nakhleh, 2016; McKain et al., 2018; Reddy et al., 2017). With this in mind, I set out to design a novel target

set whose primary purpose would be to produce sequence data that are appropriate and effective for phylogenomic analysis of relationships among closely related *Erica* species, but which would be flexible enough to also be useful for studying higher-level relationships (e.g., between African and European *Erica*; between genera within Ericaceae) and lower-level relationships (e.g., between closely-related taxa in species complexes; between populations within species; between individuals within populations). At the same time, in order to inform future work I aimed to investigate the impacts of alternative target set design choices on downstream analyses. After developing this new target set, I aimed to address the following questions:

1. Can genomic resources be used to predict the presence and paralogy of potential targets?
2. Do different target identification methods provide data with different qualities?
3. What are the costs and benefits of targeting intronic regions?
    (a) Does it reduce or increase target capture success and efficiency?
    (b) Does it result in more phylogenetically informative data?

2.2 Methodological overview

The Kadlec et al. (2017) target set was derived from *Rhododendron* (*R. scopulorum* Hutch.; Matasci et al., 2014), and since those authors had tested the target set by conducting a target capture and sequencing experiment on *Erica* samples (see Section 2.1.1), I used those data to produce *Erica*-derived versions of their targets. The present work's project funding also allowed for a larger target footprint than that of the Kadlec et al. set. Subsequent to that study several highly complete and well-annotated *Rhododendron* genomes were published, bringing their number from zero in 2017 to three by the end of 2020 (Soza et al., 2019; Yang et al., 2020; Zhang et al., 2017). I therefore used these genomes to develop two additional sets of candidate targets. I then used high-depth shotgun WGS data from three species of *Erica* to refine all three target sets and to build draft genome assemblies. Next, I used those assemblies to, where possible, generate *Erica*-derived versions of the targets, including introns and other non-coding sequences. Finally, I assessed each target set's ability to produce useful data for *Erica* phylogenomics and compared *Rhododendron*- and *Erica*-derived targets in this regard.

I distilled the product of much of the programming effort required to develop and assess the target set into a user-friendly suite of open-source command-line tools, TARGETVET, with the aim of contributing to the ever-growing phylogenomic community. The source code and a detailed account of the tool's functionality and usage (with example code) are available at `github.com/SethMusker/TargetVet`. A diagram illustrating TARGETVET's functionality is provided in Fig. 2.1, while pertinent details are provided in the following sections.

2.3 Whole-genome shotgun sequencing and assembly

I developed a custom protocol for DNA extraction from *Erica* leaf material, which is known to be highly recalcitrant (Bellstedt et al., 2010), by adapting and making some important modifications to the protocol outlined by Inglis et al. (2018). The details of these modifications, along with the full protocol itself, are presented in Appendix A.

Genomic DNA was extracted from fresh leaf material of three *Erica* species growing in the University of Bergen (UiB; Norway) arboretum following the custom protocol. These species were (1) *E. cinerea* L. which is widespread across western Europe; (2) *E. trimera* (Engl.) Beentje which is widespread in the East African highlands; and (3) *E. cerinthoides* L. which is widespread in the CFR and further east in South Africa. Library preparation and sequencing was conducted by the Genomics Core Facility at UiB. Sequencing was done using a single Illumina NovaSeq 6000 SP flowcell to generate 2 x 150 bp paired-end reads.

Raw reads were trimmed using FASTP (parameters: *–trim_poly_g –poly_g_min_len 8 –trim_tail1 3 –trim_tail2 3 –length_required 50 –overrepresentation_analysis –qualified_quality_phred 20 –unqualified_percent_limit 30 –average_qual 20*; Chen et al., 2018), after which duplicate removal was performed using *clumpify.sh* from BBTOOLS v. 38.90 (BBMap - Bushnell B. - `sourceforge.net/projects/bbmap/`) (parameters: *dedupe optical adjacent reorder=p dupedist=12000*). Overlapping read pairs were merged using *bbmerge-auto.sh* from BBTOOLS (parameters: *adapter=default rem k=60*), keeping un-merged pairs. Read quality was checked with FASTQC (`www.bioinformatics.bbsrc.ac.uk/projects/fastqc`) and MULTIQC (Ewels et al., 2016).

Draft genomes were assembled using ABYSS v.2.2.5 (Jackman et al., 2017; Simpson et al., 2009)

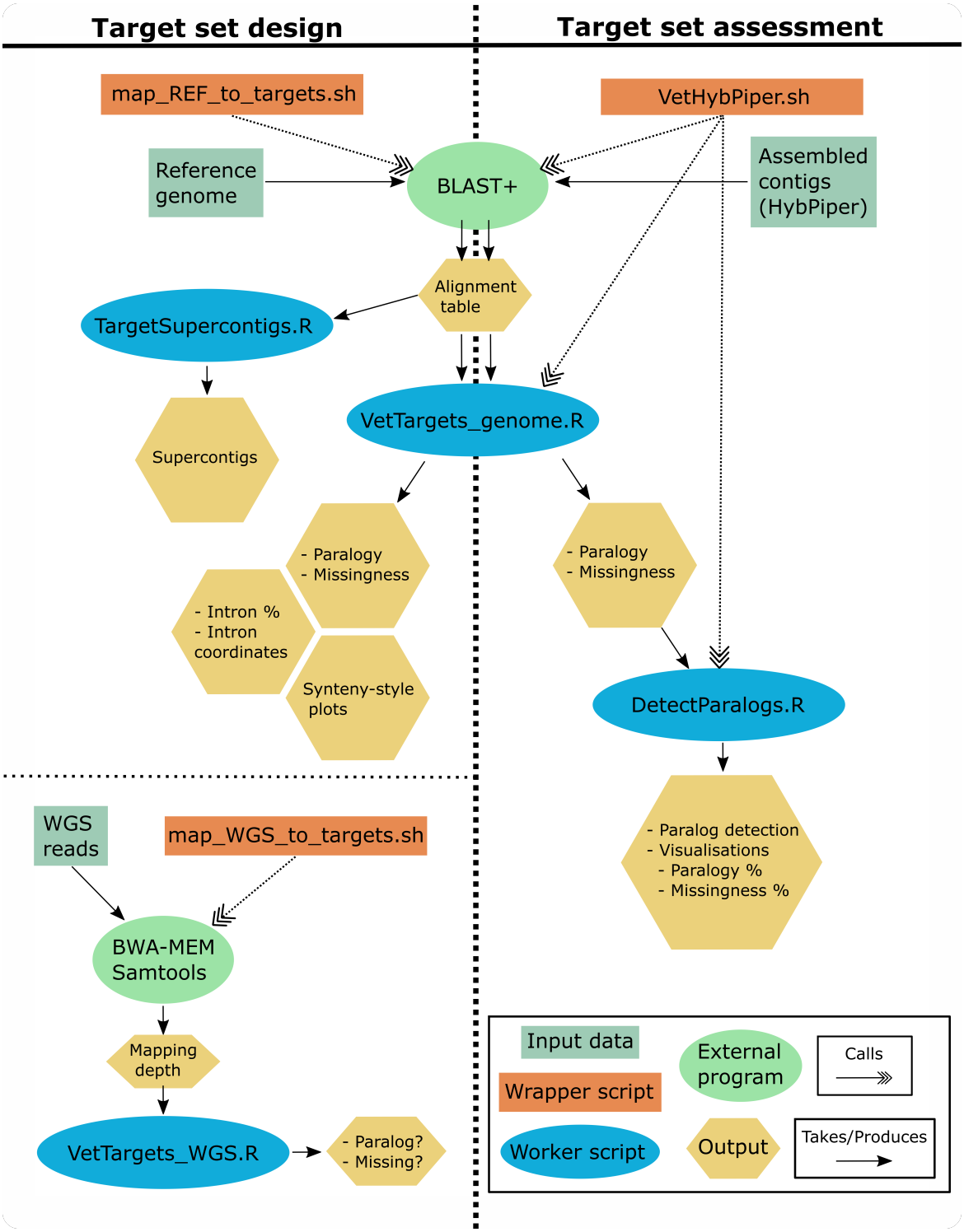


Fig. 2.1 Graphical illustration of the functionality of TARGETVET.

using both merged and un-merged reads with parameters *k=96 l=40 s=1000*. Assembly statistics such as N50 and L50 were calculated by ABYSS and BBTOOLS *stats.sh*. To further assess genome completeness on the basis of gene recovery, I used Benchmarking Universal Single-Copy Orthologs (BUSCO) v.5.0.0 (Simão et al., 2015). BUSCO searches the assembly for genes that are confidently thought to be single-copy and reports completeness- and duplication-related statistics. I ran BUSCO separately for each assembly with the same parameters: Reference universal single-copy orthologs were from the "eudicots_odb10" lineage dataset version 2020-09-10, which consists of 2326 genes from 31 species, and METAEUK v.4 (Karin et al., 2020) was used as the gene predictor. I summarised the BUSCO results using the bundled script *generate_plot.py* which uses GGPLOT2 (Wickham, 2016).

2.3.1 Genome assembly results

The quality of the draft genome assemblies of *Erica cinerea*, *E. trimera* and *E. cerinthoides* varied considerably (Table 2.1; Fig. 2.2). The much greater contiguity of the *E. cinerea* assembly compared to that of the other species was most notable. This was most likely a result its much smaller genome size as approximated by the total sequence length of the assemblies (Table 2.1), combined with the sample having *ca.* 20% more reads. The *E. cinerea* assembly also had much better completeness based on the BUSCO results, likely due to its greater contiguity. The low proportions of duplicated BUSCOs suggest that the three species are all diploid. Overall, the assemblies are of reasonable quality and should prove useful for genomic studies in *Erica* beyond the present work.

2.4 Designing a target set for *Erica* phylogenomics

2.4.1 Refining the Kadlec et al. target set

Refinement method. Kadlec et al. (2017) conducted their target capture experiment using several species of Cape *Erica*. Because a major objective of my project was to resolve relationships in the *E. abietina/E. viscaria* clade, I retrieved the reads from the single sample of *E. grandiflora* – the only member of that clade in the sample set – and used HybPiper v.1.3.1 (Johnson et al., 2016) to assemble the 134 targets (132 that Kadlec et al. identified, plus the two "universal" loci [rpb2 and topoisomerase B] that they added for comparative purposes). Additional programs made use of by

Table 2.1 Assembly statistics of the three newly assembled *Erica* draft genomes.

	E. cinerea	*E. trimera*	*E. cerinthoides*
Read statistics			
Number of read pairs	340,904,000	282,465,000	284,039,000
% reads merged	50.69%	43.84%	43.97%
Mean insert size	306.8 bp	299.1 bp	303.4 bp
Assembly statistics			
Scaffold sequence total	353.050 Mb	708.005 Mb	679.014 Mb
Number of scaffolds	286,992	1,852,782	1,463,182
Number of scaffolds > 50 kb	670	51	1
% genome in scaffolds > 50 kb	13.11%	0.43%	0.01%
Scaffold N50	5,597	124,874	73,631
Scaffold L50	15,727 bp	616 bp	1,028 bp
Max scaffold length	192,106 bp	121,715 bp	54,438 bp
Mean (SD) GC content	39.5% (0.92%)	44.9% (1.08%)	40.3% (0.89%)

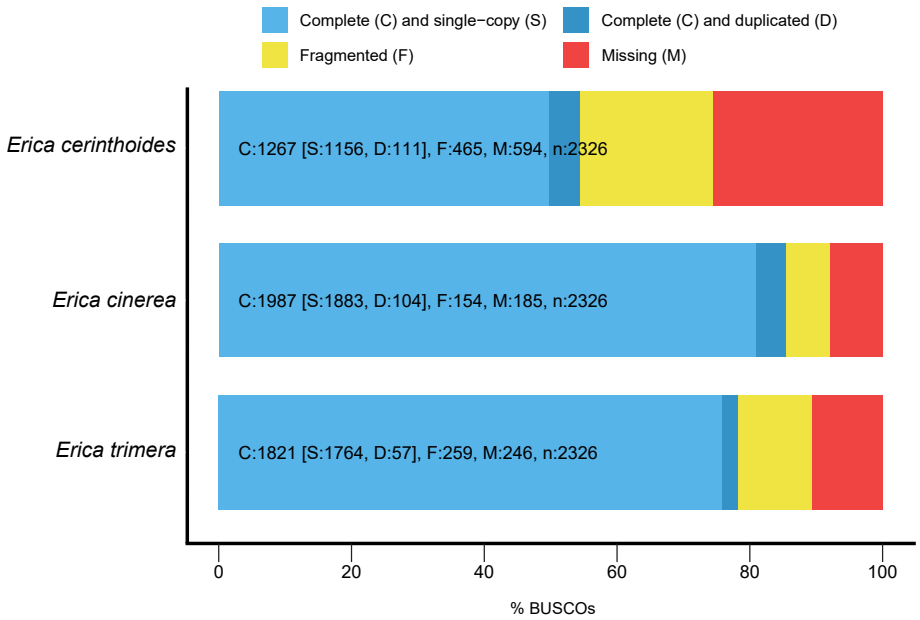


Fig. 2.2 Graphical summary of the BUSCO results for the three assembled *Erica* draft genomes. Despite their fragmented nature, the genomes have reasonably good gene recovery rates.

HybPiper were BWA-MEM v.0.7.17 (Li, 2013) for mapping reads to the targets; SPADES v.3.13.0 (Bankevich et al., 2012) for contig assembly; and EXONERATE v.2.2.0 (Slater and Birney, 2005) for identifying exon-intron boundaries to allow HybPiper to generate supercontigs. The final supercontigs were taken as potential representatives of their targets prior to further refinement (see Section *2.4.3*). I refer to this approach to target design as the "Refinement" method. It is important to note that not all supercontigs contained the full set of exons present in their respective *Rhododendron* transcript-based target.

2.4.2 Identifying new targets

MarkerMiner method. I used MarkerMiner v.1.2 with the *Vitis vinifera* single-copy reference genes, setting the minimum transcript length to 900 bp. Three *Rhododendron* CDS files were used to find matches: (1) *R. simsii* Planch. (`ftp.ncbi.nlm.nih.gov/genomes/all/GCA/014/282/245/GCA_014282245.1_ASM1428224v1/GCA_014282245.1_ASM1428224v1_cds_from_genomic.fna.gz`, accessed 02.11.2020; Yang et al., 2020), (2) *R. williamsianum* Rehder & E.H.Wilson (`ftp.ncbi.nlm.nih.gov/genomes/all/GCA/009/746/105/GCA_009746105.1_ASM974610v1/GCA_009746105.1_ASM974610v1_cds_from_genomic.fna.gz`, accessed 02.11.2020; Soza et al., 2019), and (3) *R. delavayii* Franch. var. *delavayi* (`ftp.cngb.org/pub/gigadb/pub/10.5524/100001_101000/100331/Gene/Rhododendron_delavayi.cds.fa`, accessed 02.11.2020; Zhang et al., 2017). As considerably fewer genes were identified from *R. williamsianum*, I discarded genes not found in both *R. simsii* and *R. delavayii*. I kept only the longest sequence out of the three potential *Rhododendron* targets. Lastly, I used *BLASTn* (e-value: $1e^{-5}$, BLAST v.2.10.1+; Altschul et al., 1997) to identify targets already present in the Kadlec et al. target set and removed them from the MarkerMiner set if there was at least one match.

NewTargets method. I adapted the work of McLay et al. (2021) to the task of designing taxon-specific targets for genes belonging to the Angiosperms353 target set (`github.com/chrisjackson-pellicle/NewTargets`). I used the script *BYO_transcriptome.py* (parameters: *-no_n -discard_short -length_percentage 0.7*) to search for *Rhododendron* versions of the Angiosperms353 genes, using the "Mega353" gene set – which is an expanded Angiosperms353 set with many

additional taxa representing each sequence (McLay et al., 2021) – as the reference. The same three *Rhododendron* CDS files that were used with MarkerMiner were used as the input transcriptomes. *BYO_transcriptome.py* uses HMMER3 (Mistry et al., 2013) to build hidden Markov model profiles of the reference genes and identify homologous sequences in the transcriptomes from which new targets are sought. The chosen settings disabled the formation of chimeric sequences by grafting and discarded transcripts whose length was < 70% that of the mean of the reference sequence homolog. I extracted the longest of the three potential *Rhododendron* targets and discarded those shorter than 1,000 bp. I used *BLASTn* as before to identify and remove any targets already present in the MarkerMiner or Refinement sets.

2.4.3 Filtering the target sets using WGS read depth

Because shotgun sequencing represents a largely unbiased method of deriving sequences from a genome, I reasoned that read mapping depth information could be used to infer presence/absence and paralogy of the candidate targets in *Erica*. Specifically, missing targets should have a depth of zero, while duplicated regions should have a depth roughly twice that of the mean across all targets (assuming most targets are single-copy). *Erica cinerea* has a considerably smaller genome than most *Erica* species with genome size data (Mugrabi De Kuppler, 2013), including *E. trimera* (based on the assembly size) and *E. cerinthoides* (Mugrabi De Kuppler, 2013), which implies a lower rate of paralogy and/or more missing genes. I therefore mapped the WGS reads from the latter two species separately to the potential targets using BWA-MEM v.0.7.17 with default parameters, used SAMTOOLS v.1.11 (Danecek et al., 2021) to keep only hits with mapping quality > 20, and then calculated read depth at each position using BAMTOOLS v.2.1.1 (Barnett et al., 2011). I removed any target whose median depth deviated by more than one standard deviation from the mean of the medians across all targets for either of the two *Erica* species. This process was repeated for each target set separately (Refinement, MarkerMiner and NewTargets).

Additionally, for the Refinement set I applied the above process separately to the *E. grandiflora*-derived targets and the original transcript-derived targets, and added transcript-derived targets to the target set if they passed the filters but their *E. grandiflora*-derived counterpart failed. I wrote a pair of command-line scripts (*map_WGS_to_targets.sh* and *VetTargets_WGS.R*) which I added to

TARGETVET and which automate this process and can be applied to any data when provided with one or more WGS read files and a set of target sequences (Fig. 2.1).

Because MarkerMiner identified many more genes than could be added to the target set given the total footprint available to the project (Fig. 2.3), I implemented a pre-filtering step for the MarkerMiner genes prior to using the WGS reads as above for further filtering. As off-target reads from target capture experiments are essentially equivalent to shotgun reads (Costa et al., 2021), I used the off-target reads from the Kadlec et al. experiment to identify the MarkerMiner genes that were most likely to be present in *Erica*. Reads were pooled across all *Erica* samples (n = 25) in the Kadlec et al. data and mapped to the MarkerMiner genes with NEXTGENMAP v.0.5.5 (Sedlazeck et al., 2013). I chose to use NEXTGENMAP because it tolerates greater levels of sequence divergence than BWA-MEM (Sedlazeck et al., 2013), which was useful given that the number of off-target reads was relatively small. Depth per position was determined using BAMTOOLS, and depth of each gene was calculated as total depth divided by gene length. I first kept genes with > 80% of their length having depth ≥ 1, then kept genes with depth between the mean and two standard deviations above the mean across all genes. Finally, I discarded genes that were shorter than 1,500 bp.

2.4.4 Extracting *Erica*-derived targets

I next aimed to produce *Erica*-derived versions of the new MarkerMiner and NewTargets sets. I chose to use only the *E. cinerea* assembly as it was by far the most contiguous and complete of the three. I removed any scaffolds in the assembly <500 bp long. The targets were translated to protein sequences using EMBOSS (Madeira et al., 2022) and these were then mapped to the *E. cinerea* draft genome assembly using *tBLASTn* (adding the option *-max_target_seqs 50000* to ensure that all matches were returned; Shah et al., 2019). I kept matches with sequence identity $\geq$ 70% and E-value < $1e^{-6}$, and only kept targets if >70% of their length mapped to a single *E. cinerea* scaffold (i.e., discarding any that mapped to more than one scaffold). I calculated the length of the mapped region in the *E. cinerea* genome as the difference between the largest end position and the smallest start position of the blast matches, giving an estimate of the total gene length including exons and introns. I extracted these genomic sequences using RSAMTOOLS v.2.10.0 (Morgan et al., 2021).

The read depth-based filtering procedure described above was repeated for the genomic sequences

to help ensure that they were present and single-copy across their full length in other *Erica* species. Genomic sequences that failed read depth filtering were reverted to their *Rhododendron* transcript version (which had already passed the filters), while those that passed were substituted in for their corresponding *Rhododendron* transcripts.

2.4.5 Target set design results

Refinement method

Of the 134 Kadlec et al. targets, two were found to be almost identical (sequence similarity = 99.8%, identical length), so one of them was arbitrarily discarded. *Erica grandiflora* supercontigs were assembled for all targets, of which 92 passed the WGS depth-based filtering. Of the remaining targets, the transcript sequence of a further 13 passed the filtering, bringing the total number of targets in the Refinement set to 105.

MarkerMiner method

A total of 1,572 mostly single-copy genes were identified by MarkerMiner as being present in at least one of the three *Rhododendron* transcriptomes (Fig. 2.3). Of these, 1,293, 1,217 and 999 were present in *R. simsii*, *R. delavayi*, and *R. williamsianum*, respectively. Of the 1,021 genes present in both *R. simsii* and *R. delavayi*, 16 were discarded as they had significant BLAST hits to Kadlec et al. targets. The pre-filtering step based on off-target read depth and sequence length ($\geq$1,500 bp) reduced the number of genes from 1,005 to 129, while the WGS depth-based filtering further reduced the set to 114 genes. A total of 71 of these genes had good matches in the *E. cinerea* genome, all of which passed depth-based filtering. This left 43 genes represented by their transcript sequence in the final MarkerMiner set.

NewTargets method

Of the 353 genes in the Mega353 reference set, 348 were found in at least one of the three *Rhododendron* transcriptomes and 101 of these were longer than 1,000 bp. Of these, 87 passed WGS depth-based filtering, 59 of which had good matches in the *E. cinerea* genome. Seven of these failed

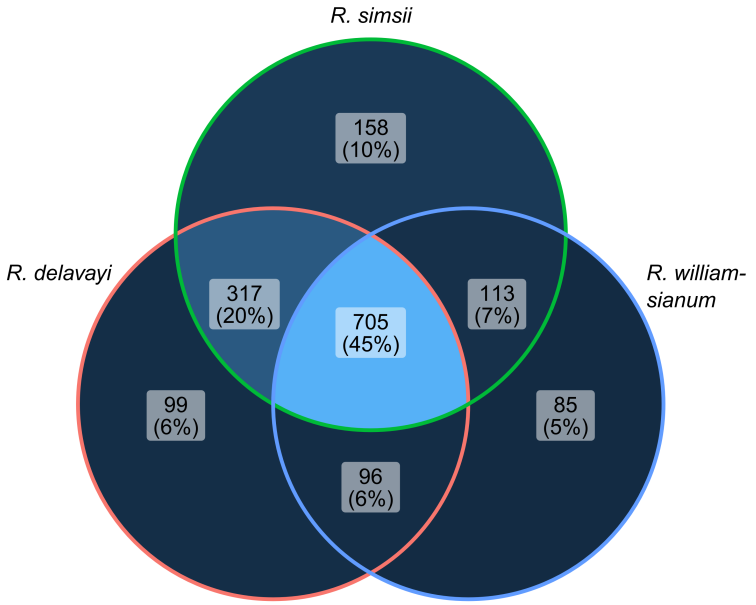


Fig. 2.3 Venn diagram showing the number of genes initially identified by MarkerMiner for each of the three *Rhododendron* transcriptomes.

depth-based filtering and were reverted to their transcript form, leaving 52 genomic sequences and 35 transcript sequences in the final NewTargets set.

Combined target superset

After all of the above steps the final combined target "superset" consisted of 303 targets with a combined length of 1,161,538 bp, and is herein referred to as the "Erica303" set.

2.5 Evaluating the target set's quality

2.5.1 DNA extraction and sequencing

The final target set was used in a target capture experiment including 295 samples, mostly of Cape *Erica* species. DNA was extracted using a custom protocol (see Appendix A). Bait design (3X tiling), bait synthesis, library preparation and sequencing were carried out by Daicel Arbor BioSciences (Ann Arbor, MI 48103, United States). Samples were paired-end sequenced using an Illumina NovaSeq 600 instrument to 2 x 150 bp. To quality-filter, trim and deduplicate the raw reads I used FASTP

v.0.23.2 (parameters: *–detect_adapter_for_pe –dedup –overrepresentation_analysis –trim_poly_g –qualified_quality_phred 20 –unqualified_percent_limit 30 –average_qual 20 –length_required 100*).

2.5.2 Target assembly

To investigate the effects of target source (i.e., *Rhododendron* CDS *versus Erica* genome) and marker identification method (i.e., Refinement, MarkerMiner and NewTargets) on aspects of target recovery and assembly, I assembled the targets from all 295 samples using HybPiper v.2.0.1. I ran HybPiper's *assemble* module using BWA-MEM v.0.7.17 for read mapping, SPADES v.3.15.3 for assembly (with kmer values of 33 and 77), EXONERATE v.2.4.0, and BBTOOLS v.38.92.

Prior to assembly with HybPiper, in order to ease computational burden I used *reformat.sh* from BBTOOLS to randomly subsample each sample's reads to one million read pairs. Given a total target footprint of 1,161,538 bp and assuming a mean read pair length of *ca.* 290 bp (to account for trimming and pair overlaps), this gives an expected mean coverage of

$$\frac{read\,length \times no.\,reads}{footprint} = \frac{290 \times 1,000,000}{1,161,538} \approx 250X.$$

2.5.3 Quantifying paralogy and capture efficiency

Assessing paralogy and missingness

To investigate paralogy I first used HybPiper's length-based criterion which, on a per-sample basis, flags a target as a potential paralog if its second-longest contig is above a certain proportion (which I set to 0.75, the default) of the length of the longest contig. Secondly, I developed a custom coverage-based approach which characterises paralogy and identifies paralogs across the full sample set. I incorporated the approach into a command-line utility in the form of a bash script (*VetHybPiper.sh*), which acts largely as a wrapper around BLAST and several custom R scripts that are part of TARGETVET (Fig. 2.1). A graphical illustration of the method is provided in Fig. 2.4, and it proceeds as follows:

1. For each sample,

    i. map all assembled contigs to the target sequences using BLAST;

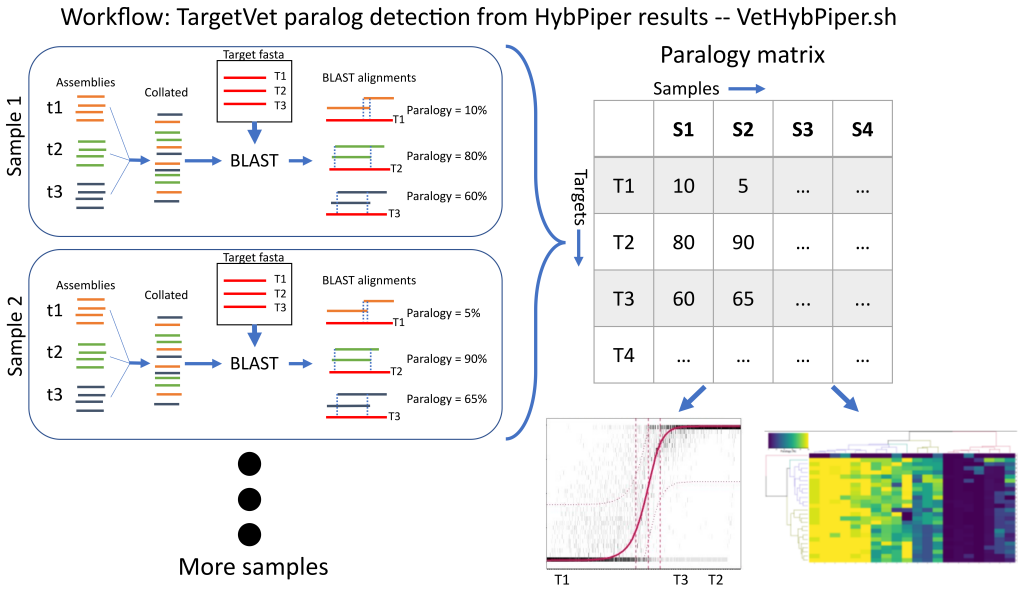


Fig. 2.4 Graphical illustration of paralogy estimation using TARGETVET's *VetHybPiper.sh* script.

ii. remove matches below given thresholds of length (by default, 150 bp) and sequence similarity (by default, 70%);

iii. for each target, calculate each site's coverage (c) by counting how many BLAST matches from different contigs map to it;

iv. define L as the total length of the target in base pairs (i.e., number of sites) and l_c as the number of sites with coverage = c;

v. estimate each target's paralogy (P) as the fraction of its length with $c \geq 2$, ignoring missing regions, i.e.,

$$P = \frac{l_{c\geq 2}}{L - l_0}.$$

2. Flag targets as potential paralogs if P is unusually high across all samples.

Using the above definitions, missingness (M) can be estimated as the fraction of the target's length with $c = 0$, i.e.,

$$M = \frac{l_0}{L},$$

and copy number (C) can be estimated as the mean coverage across s sites ignoring sites with $c = 0$, i.e.,

$$C = \frac{1}{n}\sum_{s=1}^{n} c_s \quad ; \quad c_s > 0.$$

Estimates of *P*, *M* and *C* were derived from two separate *BLASTn* mapping results: one in which the actual target sequences were used as the reference, and one in which the CDS versions of the targets were used as the reference. To remove putative paralogs, I discarded targets with mean *P* (across 295 samples) > 40% according to either of the two BLAST results (n = 13). To remove targets that were poorly recovered, I discarded those with mean *M* > 40% according to the BLAST result based on the target sequences (n = 5). This resulted in a "clean" superset comprising 285 targets which I refer to as "Erica285". Unless otherwise stated, all further analyses used the Erica285 superset.

Assessing target and intron capture efficiency

I used HybPiper's *stats* module to collect transcript and supercontig lengths for all samples. To test whether *Erica*-derived targets had greater capture efficiency, I used separate fixed effect models for each marker identification method to model supercontig length as a function of target source, including sample as a fixed effect to account for random variance while also allowing the sample effect to vary by transcript length to account for the tendency for longer transcripts to have longer supercontigs.

Exon-derived baits are only able to capture intronic sequences flanking the exons, meaning that sequence coverage drops off considerably with increasing distance from the nearest exon (Gnirke et al., 2009). I therefore hypothesised that, because they included intronic sequences, targets derived from *Erica* genomic sequences would be better at recovering introns than targets from *Rhododendron* CDS sequences, and that this difference would be most pronounced when the gaps between exons were larger. This logic predicts that as gene length increases there should be a decline in relative intron length for CDS-derived targets but no such decline (or a less pronounced decline) for genome-derived targets. To test this prediction, I determined the intron sequence length of each gene for each sample using the gene models inferred by the *protein2genome* model of EXONERATE, part of the HybPiper pipeline, and set the intron length to zero if no intronic region was identified. I used separate fixed effects models for each target identification method to model intron length as a function of gene length and target source, including sample as a fixed effect. I included the source by gene length interaction term to test whether the slope of the relationship between gene length and intron length was significantly lower for CDS-targeted genes. As a proxy for the gene's true length I used the maximum

gene length inferred by EXONERATE out of all samples. This was likely to be an underestimate for many CDS-targeted genes, especially longer genes whose full intronic sequence may not have been recovered in any sample, meaning that estimated differences in slope were likely to be underestimates of the true difference. Models and significance tests were run using FIXEST (Bergé, 2018).

2.5.4 Target capture experiment results

Paralogy

Overall paralogy was low across the target superset according to both length- and coverage-based analyses (Figs. 2.5 and 2.6, respectively), although the length-based method was apparently less sensitive. These results suggest that the WGS depth-based filtering method was largely successful in identifying paralogs. *P* was largely unaffected by whether it was estimated using the actual targets or their CDS versions (Fig. 2.7), with the exception of two Refinement targets that had high CDS-based *P* but low target-based *P*.

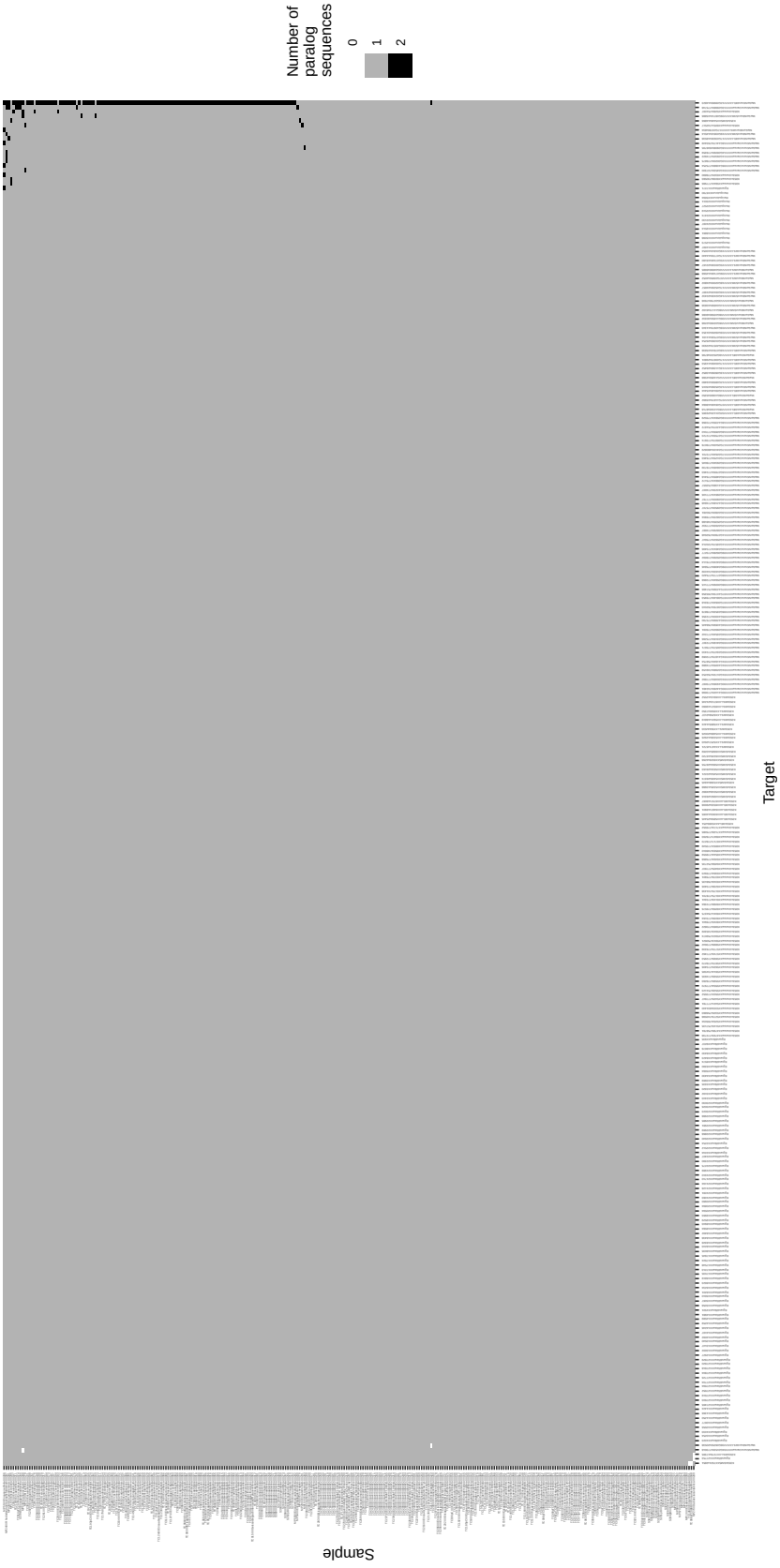


Fig. 2.5 Heatmap showing the number of paralogs (i.e., number of gene copies) identified by HybPiper's length-based method, in which a targets is flagged for a given sample if its second-longest assembled contig is more than 70% the length of its longest assembled contig. Targets and samples are arranged by mean number of copies.

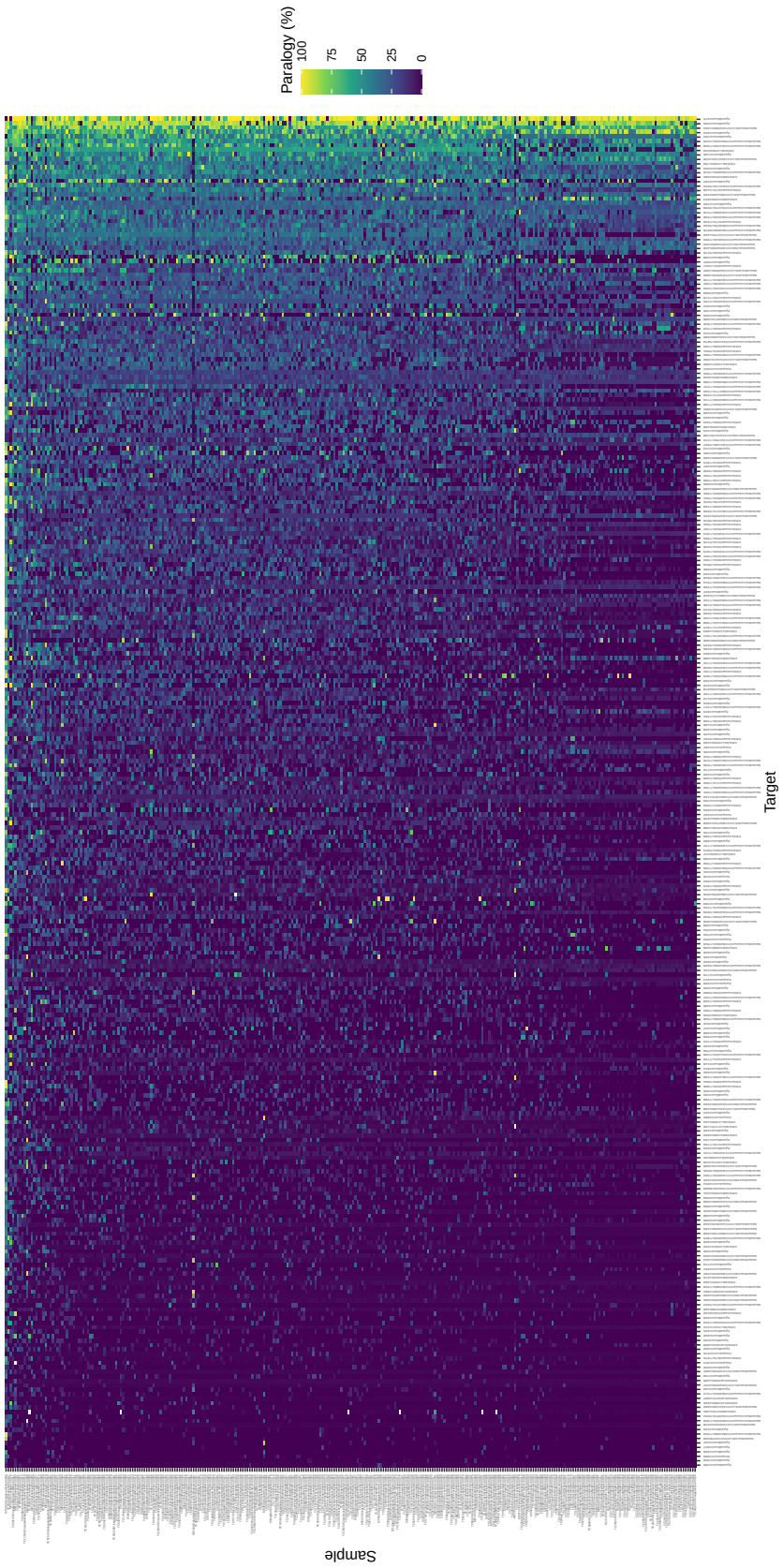


Fig. 2.6 Heatmap showing paralogy (*P*), the estimated proportion of a target's length covered by more than one assembled contig, for all samples and all loci in the Erica303 superset. Values of *P* were calculated from BLAST results using the actual target sequences. Targets and samples are arranged by mean *P*. This plot is a direct product of the TARGETVET script *VetHybPiper.sh*.

Table 2.2 Results of the fixed effects models of supercontig length as a function of target source showing that longer supercontigs were recovered by *Erica* genome-derived targets identified using NewTargets and MarkerMiner, whereas longer supercontigs were recovered by *Rhododendron* CDS-derived targets identified using the Refinement method. R^2 indicates the fit of the full model, while Within R^2 indicates the fit when fixed effects are ignored. Numbers in brackets are standard errors.

	MarkerMiner	NewTargets	Refinement
Source = *Rhododendron* CDS: intercept	-1,162.2 bp*** (20.6)	-1,647.2 bp*** (20.7)	1,075.0 bp*** (18.7)
Observations	32,155	23,010	28,910
R^2	0.264	0.176	0.099
Within R^2	0.037	0.077	0.035
Fixed effects			
Sample	✓	✓	✓
Transcript length × Sample	✓	✓	✓

Signif. codes: *** = 0.01, ** = 0.05, * = 0.10

Most samples showed similar paralogy patterns (Fig. 2.8), with the notable exception of the single *Erica spiculifolia* sample, which had a mean P of 47.0% (27.3% SD), 142 targets with $P > 50\%$, and a mean copy number (C) of 1.65 (0.491 SD). *Erica spiculifolia* has a 1.5-fold higher chromosome number (2n = 36) than most *Erica*, which typically have 2n = 24 (Nelson and Oliver, 2005), making ploidy the most likely explanation for this finding.

Target recovery

Genome-derived targets produced significantly longer supercontigs than CDS-derived targets for the MarkerMiner (1,162 bp longer) and NewTargets (1,647 bp longer) sets, but significantly shorter supercontigs for the Refinement set (1,075 bp shorter; Table 2.2). Nevertheless, R^2 values were generally low even when accounting for variance explained by CDS length and sample identity (highest $R^2 = 0.264$, highest within-$R^2 = 0.077$), suggesting that variation in supercontig length was not well-predicted. This was most likely because supercontig length was not primarily determined by CDS length but rather by true target length (i.e., including introns), which could not be modelled because true target lengths were unknown for the CDS-derived targets. Nevertheless, the significantly shorter CDS-derived supercontigs in the MarkerMiner and NewTargets sets illustrate the benefits of using genome-derived targets.

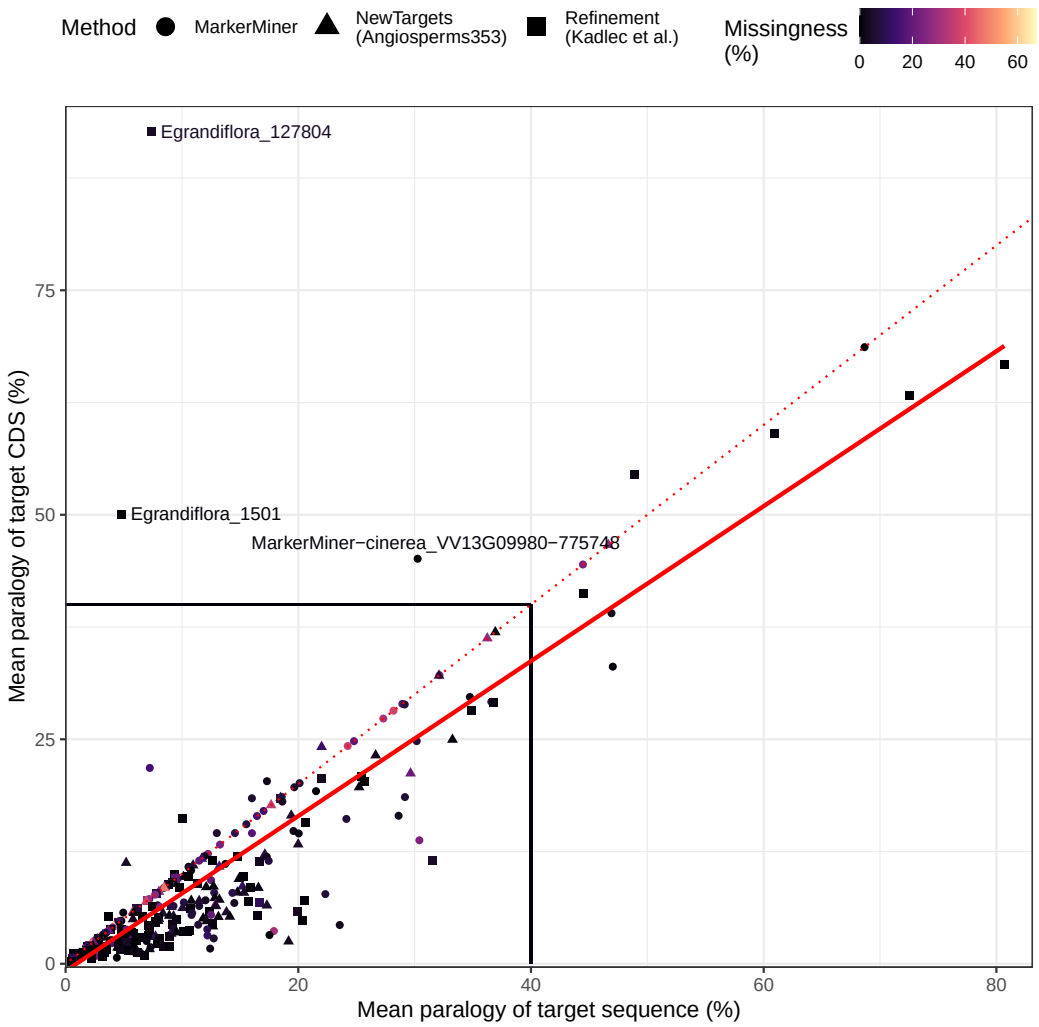


Fig. 2.7 Paralogy (*P*) estimated using the actual target sequences *versus* using their CDS versions. The solid line shows the linear regression line while the dashed line shows the 1:1 line. Points colours indicate missingness (*M*).

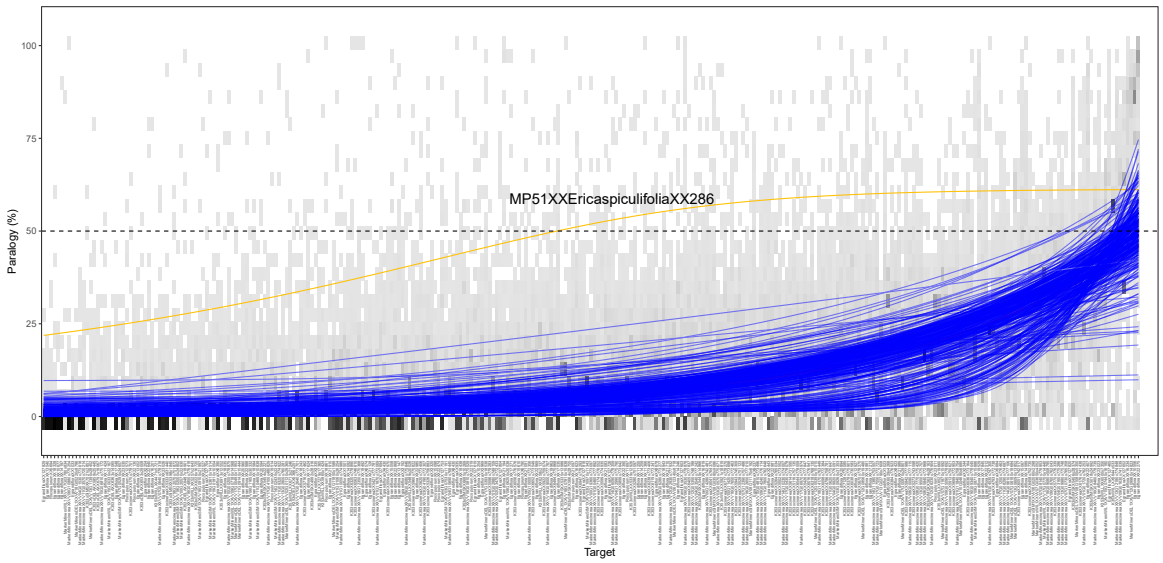


Fig. 2.8 Patterns of paralogy (*P*) per sample. Targets (x-axis) are arranged in ascending order by mean *P* across all samples. Curves show the predicted *P* for each sample obtained from n-parameter logistic regressions. The single sample that deviated from the mean *P* by more than 20% on average across all targets is highlighted (yellow line) and labelled. This plot is a direct product of the TARGETVET script *VetHybPiper.sh*.

Intron recovery

The analysis of intron length in relation to gene length suggested that *Erica*-derived targets captured relatively more intronic sequence (Table 2.3, Fig. 2.9). Specifically, for the MarkerMiner and NewTargets sets intron length increased with gene length more steeply for the genome-derived target sets (MarkerMiner: slope = 0.721, NewTargets: slope = 0.781) than for the CDS-derived sets (MarkerMiner: slope = 0.650, NewTargets: slope = 0.598). For the Refinement set the slope difference was reversed (CDS-derived: slope = 0.826, genome-derived: slope = 0.648), however, the intercept difference estimate showed that the CDS-derived supercontigs had, on average, less intronic sequence than the genome-derived supercontigs (Fig. 2.9). While it is possible that sequence similarity could explain these results (i.e., *Erica*-derived baits capture *Erica* DNA more effectively than *Rhododendron*-derived baits), the high capture efficiency of the CDS-derived baits (Table 2.2) suggests that target capture was not hampered by sequence divergence. Rather, the results supported the hypothesis that explicitly targeting introns results in improved intron recovery by mitigating the decline in capture efficiency further from exons.

Table 2.3 Results of the fixed effects models of intron length as a function of target source and gene length, showing that longer introns were recovered by *Erica* genome-derived targets identified using NewTargets and MarkerMiner, whereas longer introns were recovered by *Rhododendron* CDS-derived targets identified using the Refinement method. The relationship was unaffected by sample identity ($R^2 \approx$ Within R^2). Numbers in brackets are standard errors.

	MarkerMiner	NewTargets	Refinement
Gene length × Source = *Erica* genome: slope	0.721***	0.781***	0.648***
	(0.003)	(0.005)	(0.002)
Gene length × Source = *Rhododendron* CDS: slope	0.650***	0.598***	0.826***
	(0.005)	(0.003)	(0.005)
Source = *Rhododendron* CDS: intercept	18.0	607.2***	-1,428.7***
	(21.1)	(27.2)	(18.4)
Observations	33,599	24,691	30,957
R^2	0.900	0.904	0.795
Within R^2	0.900	0.904	0.795
Fixed effects			
Sample	✓	✓	✓

Signif. codes: *** = 0.01, ** = 0.05, * = 0.10

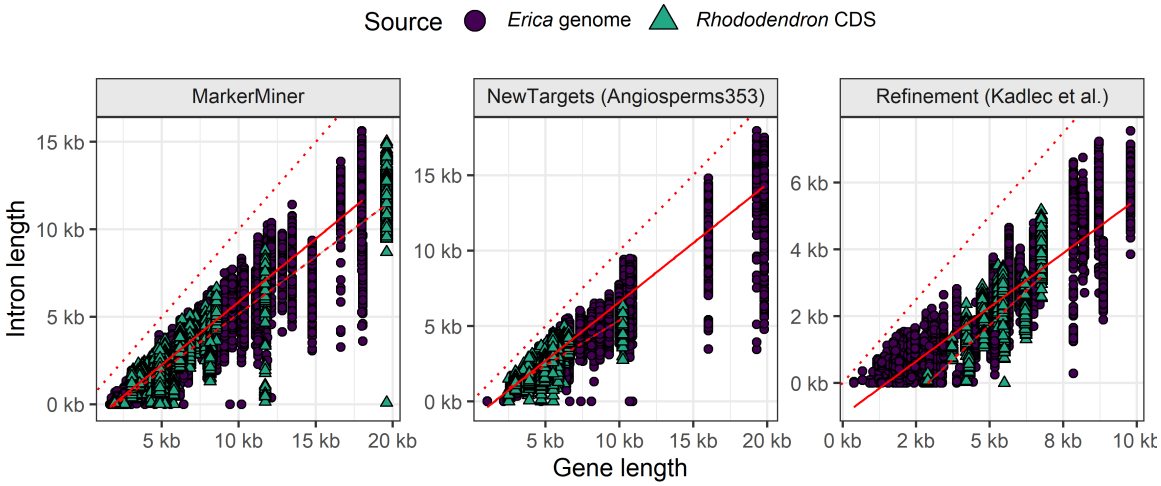


Fig. 2.9 The relationship between gene length and intron length depends on the source of the target and the method of target set design. For MarkerMiner and NewTargets targets, the slope is steeper for genome-derived targets (solid lines) than for CDS-derived targets (dashed lines). For Refinement targets, the slope is steeper for CDS-derived targets, though these also have relatively less intronic sequence on average. The dotted lines indicate the 1:1 line. Results of the statistical tests to compare the slopes are given in Table 2.3.

2.6 Evaluating the target set's phylogenetic utility

To assess the usefulness of the targets for phylogenomics, I selected a subset of 32 samples including three outgroup samples (*Daboecia*, *Rhododendron*, and *Calluna*) and eight European, one Madagascan, one East African, and 19 Cape *Erica* (details in Table B.1). I aimed to characterise the ability of the target sets to (1) recover well-established relationships based on previous work, and (2) resolve relationships between Cape *Erica* clades that have shown evidence of recent and rapid diversification (Pirie et al., 2011, 2016). I investigated how these properties were affected by the presence or absence of paralogs or largely missing targets (Erica303 *versus* Erica285), as well as target source (*Rhododendron* CDS *versus Erica* genome) and marker identification method (Refinement, MarkerMiner and NewTargets). I restricted the analyses to supercontig sequences in order to maximise sequence length and thus variation (Bagley et al., 2020).

Multiple sequence alignment

Supercontig MSAs were generated using the L-INS-i algorithm of MAFFT (Katoh and Standley, 2013), after which poorly aligned ends of individual sequences were recoded as missing using a custom modification of HERBCHOMPER (Gardner, 2021), which is an open-source fork of the HERBCHOMPER repository and is available at `github.com/SethMusker/HerbChomper_MSA`. The original HERBCHOMPER algorithm takes a user-specified sequence in an MSA (the "reference") and calculates sequence identity between the reference and another user-specified sequence (the "target") along a sliding window of a given number of nucleotides, with two rounds (forward and reverse) each of which starts from one end of the alignment and works inwards. Each round recodes as gaps ("-") any target nucleotides that fall within a window whose sequence identity (relative to the reference sequence in that window) falls below a given threshold, and stops when the sequence identity of a window reaches the threshold. My modified implementation, "*herbchomper_consensus_allseqs.R*", calculates the majority-rule consensus of the alignment using SEQINR (Charif and Lobry, 2007) and uses that as the reference sequence to recode each individual sequence in the alignment separately. I used a sliding window of 50 bp and a sequence identity threshold of 0.8 for all MSAs. Finally, gappy regions of the MSAs were removed using CLIPKIT *smart-gap* (Steenwyk et al., 2020), which aims to

remove gappy regions without introducing potential errors caused by excessive trimming (Tan et al., 2015).

2.6.1 Species tree concordance

Species tree inference

Species trees were estimated using a concatenation method and a summary coalescent method. For the concatenation method, IQ-TREE v.2.2.0 (Minh et al., 2020) was used with an edge-linked proportional partition scheme, setting each target as a separate initial partition. MODELFINDER (Kalyaanamoorthy et al., 2017) was used for substitution model estimation and partition merging (to reduce over-fitting) while only examining the top 25% of partitioning schemes (Lanfear et al., 2014) to reduce computational burden. Branch support values were estimated using ultrafast bootstrap (UFBoot; Hoang et al., 2018) and SH-alrt (Guindon et al., 2010) with 1,000 replicates each.

For the summary coalescent method I used a modification of the ASTRAL method (Zhang et al., 2018), Weighted ASTRAL - Hybrid (wASTRAL-h) v.1.8.2.3 (Zhang and Mirarab, 2022), which weights quartets by both branch length and local support values to provide more accurate species tree inferences than the unweighted ASTRAL algorithm. Herein I refer to wASTRAL-h simply as ASTRAL. As input for ASTRAL, gene trees were estimated by maximum-likelihood (ML) using IQ-TREE with two independent runs to improve the tree search after automated substitution model selection using MODELFINDER, with UFBoot (1,000 replicates) used to estimate branch support. I ran wASTRAL-h with the flag *"–moreround"* to increase the number of placement and subsampling rounds from four to 16 for a more thorough search of the tree space and to specify support values as bootstrap (range 0–100).

As a means of assessing the impact of paralogs and poorly recovered loci on phylogenetic inference, I ran both IQ-TREE and ASTRAL analyses separately on the Erica303 and Erica285 target sets.

Topological concordance

I compared trees inferred using different marker sets and different methods using *cophylo* from PHYTOOLS (Revell, 2012). To assess the results in the context of previous work, I also compared the newly inferred trees to the most recent *Erica*-wide phylogeny (Pirie et al., in prep.), which is based on several "traditional" loci, including ITS, ETS, and several chloroplastic markers and was inferred by Pirie et al. (in prep.) using RAxML v.8.0.0 (Stamatakis, 2014) with standard non-parametric bootstrapping (100 replicates) and originally included 752 tips. I trimmed the tree to include only the species or subspecies shared between the sample sets (n = 30) using the APE function *drop.tips*.

2.6.2 Phylogenetic informativeness

Lastly, I aimed to investigate the effects of marker identification method and target source on phylogenetic informativeness. AMAS (Borowiec, 2016) was used to determine the number of parsimony-informative sites in each alignment. PHYINFORMR (Dornburg et al., 2016) was used to estimate Quartet Internode Resolution Probability (QIRP), which is a measure of phylogenetic informativeness that accounts for sequence substitution rate variation, tree depth, and internode length. I estimated QIRP for the crown of the clade consisting of the *E. abietina/E. viscaria* clade, the *E. massonii* clade, and the *E. corifolia* clade. All of these clades were recovered with good support by Pirie et al. (2016). I refer to this as the "VMC clade", and chose to focus on it due to (1) its young crown age (*ca.* 5 Ma; Pirie et al., 2016) and (2) the very short internodal branches separating the three crowns of the constituent sub-clades (all < *ca.* 1 million years; Pirie et al., 2016). I estimated an ultrametric tree (as required by PHYINFORMR) based on the concatenation phylogeny using *chronos* in APE (Paradis, 2013; Paradis and Schliep, 2019). I estimated site substitution rates using IQ-TREE v.2.2.0 (Minh et al., 2020), using the empirical Bayesian method and the best model and partition-merging scheme as estimated for the concatenation-based phylogenetic analysis.

2.6.3 Species tree concordance results

The presence of paralogs and poorly recovered genes had no effect on species tree topology and little effect on branch support (Figs. 2.10, 2.11). In contrast, the effect of phylogenetic reconstruction

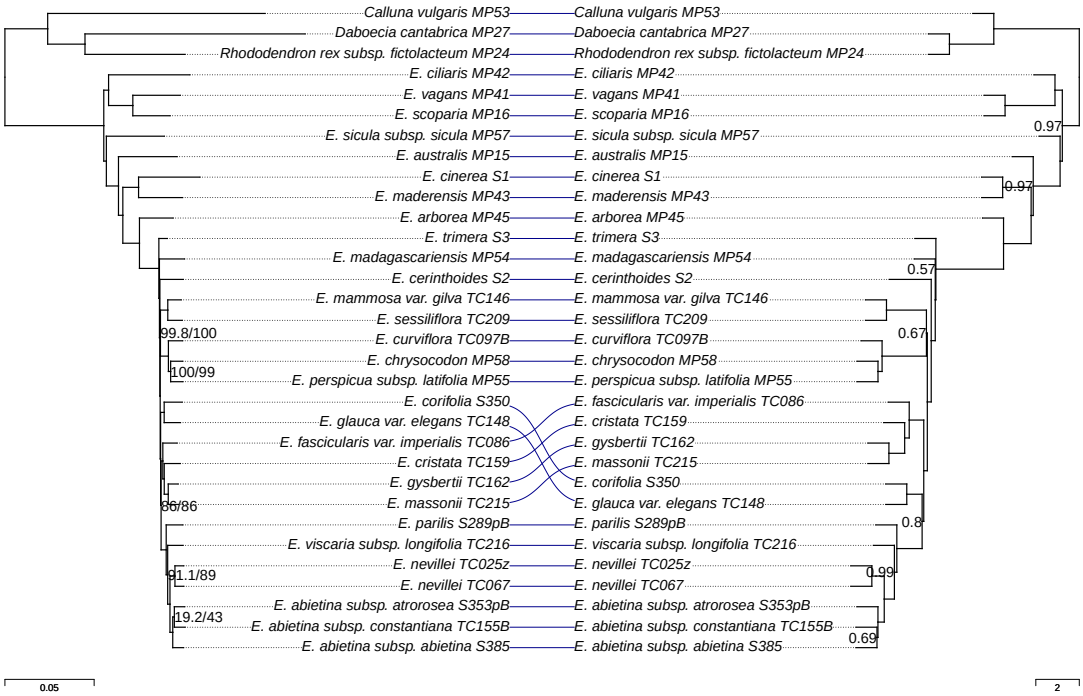


Fig. 2.10 Tanglegram comparing the phylogenies inferred by concatenation (IQ-TREE; *Left*) and by ASTRAL (*Right*) using the full Erica303 target superset. For the concatenation tree, branch lengths are in substitutions per site and node labels are SH-alrt/UFBoot percentages. For the ASTRAL tree, branch lengths represent coalescent units (except for terminal branches which are arbitrarily set to 1 as they are not estimated by ASTRAL) and node labels show posterior probability support. Nodes with full support are unlabelled. The trees are fully bifurcating and are rooted along the branch between the *Erica* and non-*Erica* samples arbitrarily for display purposes.

method was notable. In general, branch support values were higher in the concatenation trees than in the ASTRAL trees. Trees inferred using the two methods differed in the topology of the "VMC clade": concatenation recovered the *E. corifolia* clade as sister to the *E. abietina/E. viscaria* and *E. massonii* clades, i.e., the topology (C,(M,V)), whereas ASTRAL recovered the topology (M,(C,V)). However, this resolution had relatively low local posterior probability (PP = 0.8) in the ASTRAL trees (Figs. 2.10, 2.11) and low support (SH-alrt/UFBoot = 86/86) in the concatenation tree based on the Erica303 set (Fig. 2.10), and therefore the conflict was not strongly supported.

There were also some discrepancies between the "traditional" marker-based phylogeny of Pirie et al. (in prep.; hereafter "Pirie tree") and the phylogenies inferred here (Figs. 2.12, 2.13). Regarding the "VMC clade", the Pirie tree agreed with the ASTRAL tree topology (M,(C,V)). On the other hand, both concatenation and ASTRAL inferred a different placement of *E. australis* than the Pirie tree, a

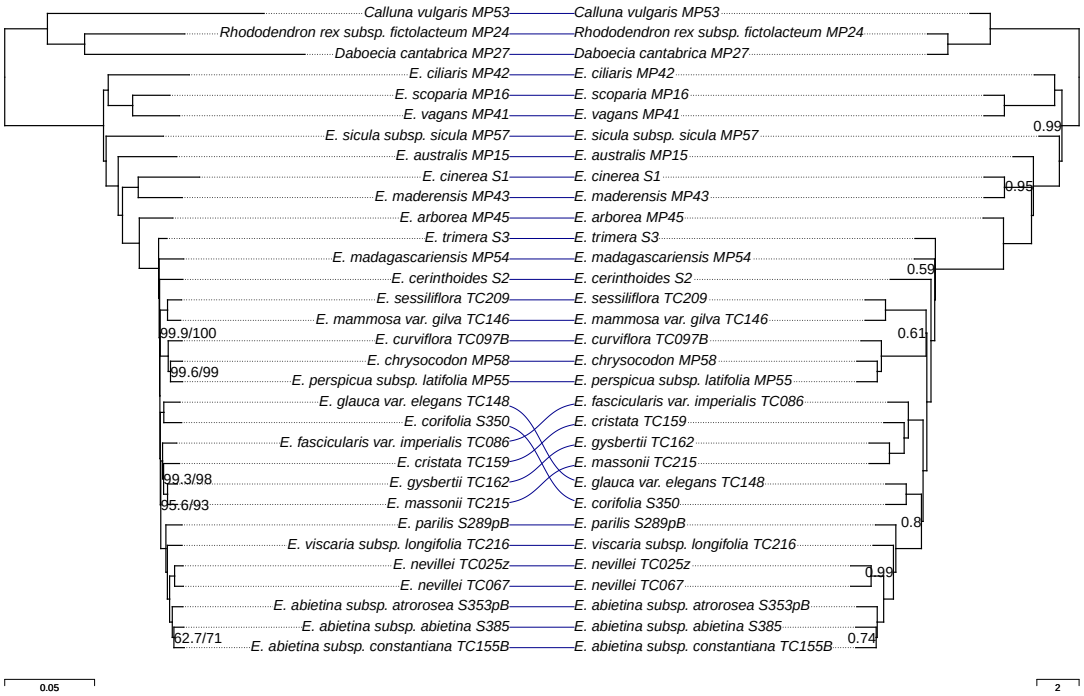


Fig. 2.11 Tanglegram comparing the phylogenies inferred by concatenation (IQ-TREE; *Left*) and by ASTRAL (*Right*) using the Erica285 target superset, which excludes putative paralogs and genes with excessive missing data. Further details follow Fig. 2.10.

conflict that was strongly supported according to branch support values. There were also some much weaker conflicts. For example, the Pirie tree grouped *E. trimera* with *E. arborea* with low support (bootstrap = 50%), whereas the phylogenies inferred here confidently placed *E. arborea* outside the clade of African and Madagascan species.

In summary, there were some topological conflicts between the Pirie tree and the newly inferred trees, as well as between the trees inferred by different methods using the new targets, but only one of the conflicting relationships was strongly supported (the placement of *E. australis* in the Pirie *versus* the newly inferred trees). Overall, the relationships inferred using the new targets were mostly concordant with prior expectations based on previous work and also produced much more strongly supported phylogenies, with limited conflict within the "VMC clade" localised at a single node surrounded by very short branches.

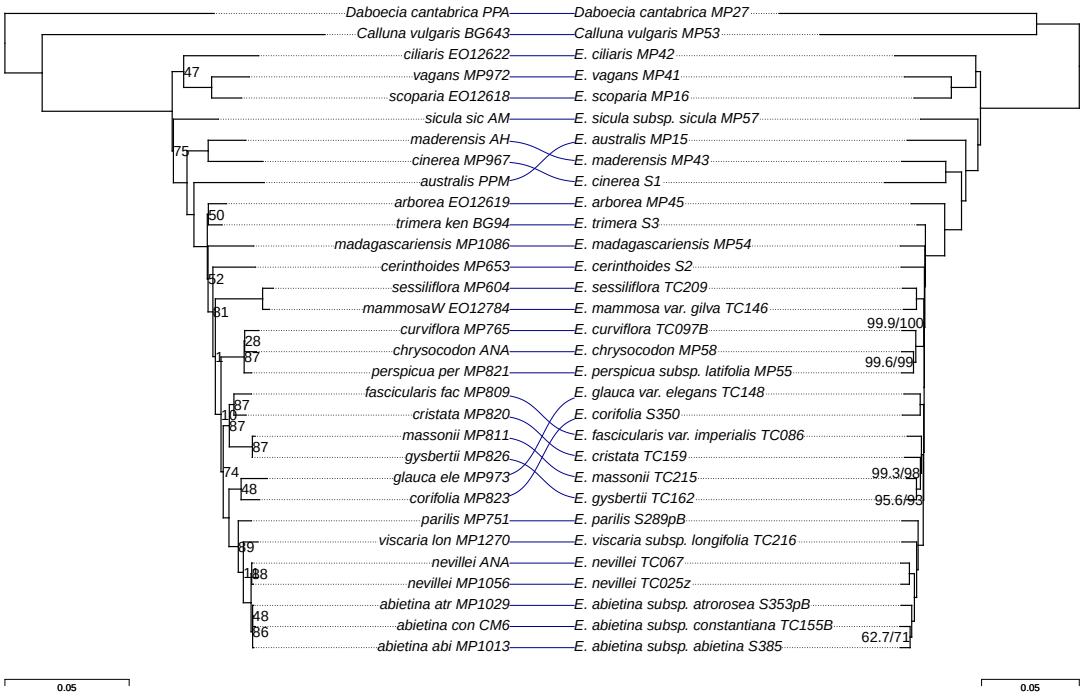


Fig. 2.12 Tanglegram comparing the phylogenies inferred by Pirie et al. using traditional markers (in prep.; *Left*) and by concatenation using the Erica285 superset (*Right*). For the Pirie tree, branch lengths are in substitutions per site and node labels show bootstrap percentage.

2.6.4 Phylogenetic informativeness results

Parsimony informative sites

Table 2.4 shows that the supercontig alignments from CDS-derived targets had a significantly smaller number of PI sites than did the genome-derived alignments for the MarkerMiner and NewTargets sets, but significantly more for the Refinement sets (MarkerMiner, mean difference = -130 sites; NewTargets, mean difference = -223 sites; Refinement, mean difference = 180 sites). In contrast, the proportion of PI sites was slightly greater in CDS-derived alignments for all methods, though the mean difference never exceeded 1%. However, R^2 values were low for all models, indicating that overall PI did not depend strongly on target source.

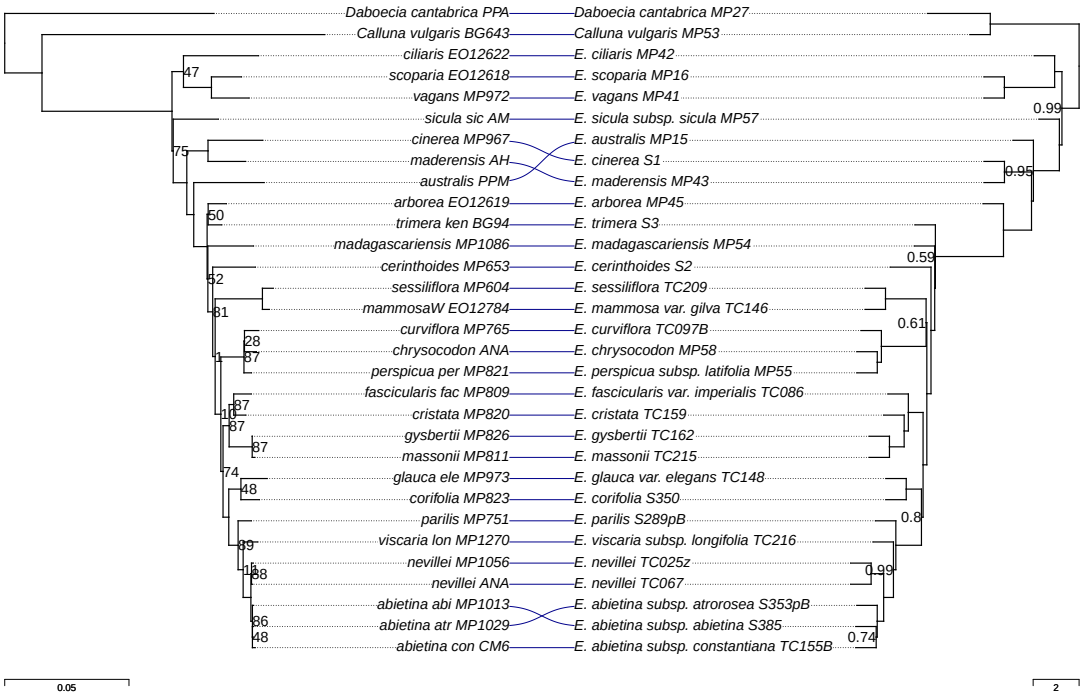


Fig. 2.13 Tanglegram comparing the phylogenies inferred by Pirie et al. using traditional markers (in prep.; *Left*) and by ASTRAL using the Erica285 superset (*Right*). For the Pirie tree, branch lengths are in substitutions per site and node labels show bootstrap percentage. The trees are fully bifurcating and are rooted along the branch between the *Erica* and non-*Erica* samples arbitrarily for display purposes.

QIRP and PI

Overall, QIRP was relatively high (mean = 0.80 $\pm$ 0.15 SD), indicating that the target set was informative for young, short internodes. The proportion of PI sites showed no relationship with QIRP. In contrast, QIRP generally had a clear positive relationship with the number of PI sites, but although the shape of the relationship was the same for all methods for the genome-derived alignments, it differed between methods for the CDS-derived alignments (Fig. 2.14). Genome-derived alignments showed an asymptotic trend for all three methods, with QIRP increasing until *ca.* 1,000 PI sites, at which point most alignments had QIRP > 0.9. CDS-derived alignments showed a mixture of trends. The MarkerMiner alignments fell into two distinct groups, one with higher QIRP regardless of PI, though both groups showed a positive trend. The NewTargets alignments had lower QIRP than their genome-derived counterparts, matching the low-QIRP group of MarkerMiner alignments in trend and absolute values. The Refinement alignments showed no clear trend, though they generally had much

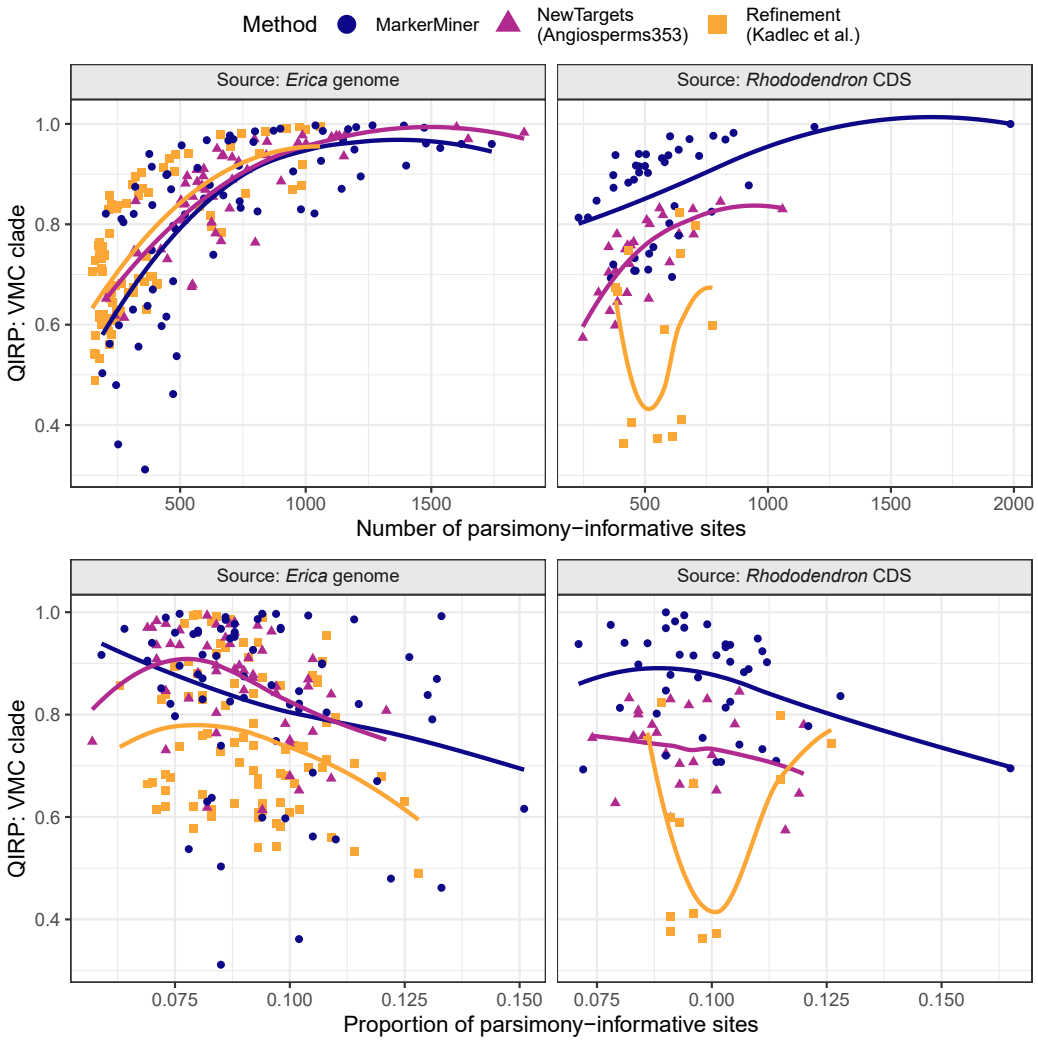


Fig. 2.14 Quartet Internode Resolution Probability (QIRP) at the crown of the "VMC clade" in relation to proportion (*top*) and number (*bottom*) of parsimony-informative sites, and target source and method. Lines show loess model fits with span = 1.

Table 2.4 Results of the fixed effects models of parsimony-informative (PI) sites (number and proportion) as a function of target source for supercontig alignments, using the Erica285 set. More PI sites were recovered by *Erica* genome-derived targets identified using NewTargets and MarkerMiner, whereas fewer were recovered using the Refinement method. In contrast, the proportion of PI sites was slightly greater in *Rhododendron* CDS-derived targets for all methods, though the mean difference never exceeded 1%. Numbers in brackets are standard errors.

	MarkerMiner		NewTargets		Refinement	
	Number	Prop. (%)	Number	Prop. (%)	Number	Prop. (%)
(Intercept)	717.3***	9.44***	720.8***	8.80***	374.2***	9.11***
	(44.3)	(0.217)	(41.2)	(0.172)	(25.6)	(0.143)
Source = *Rhododendron* CDS	-130.2*	0.491	-223.5***	0.598**	180.5**	0.802**
	(72.3)	(0.354)	(71.4)	(0.299)	(70.2)	(0.393)
Observations	109	109	78	78	98	98
R^2	0.029	0.018	0.114	0.050	0.064	0.042
Adjusted R^2	0.020	0.008	0.103	0.038	0.055	0.032

Signif. codes: *** = 0.01, ** = 0.05, * = 0.10

lower QIRP than the other methods. The smaller range of PI sites for the CDS-derived alignments is important to note, as most had fewer than 1,000 PI sites, the point at which genome-derived alignments reached consistent QIRP highs.

For a given number of PI sites, QIRP values of genome-derived alignments were much higher than those of CDS-derived alignments for the Refinement set (linear model: $F(1,96) = 27.0$, $R^2 = 0.21$, $p < 0.001$), but not for the other sets (NewTargets: $F(1,76) = 2.82$, $R^2 = 0.023$, $p = 0.097$; MarkerMiner, $F(1,107) = 2.93$, $R^2 = 0.018$, $p = 0.090$; Fig. 2.15). This revealed that, despite their shorter lengths, the Refinement targets produced relatively more informative alignments per nucleotide base pair.

QIRP and introns

Regardless of target source, the proportion of intron sequence had a strong and significant positive relationship to QIRP (Fig. 2.16) for the NewTargets alignments (best-fit linear model = QIRP ~intron prop. + source, $F(2,75) = 65.2$, $R^2 = 0.63$, $p < 0.001$) and a weaker but still significant relationship for the Refinement alignments (best-fit linear model = QIRP ~intron prop. + source, $F(2,95) = 11.4$, $R^2 = 0.18$, $p < 0.001$). The same positive relationship applied to the MarkerMiner alignments except that its slope varied with source (best-fit linear model = QIRP ~intron prop. * source, $F(3,105) = 24.9$, $R^2 = 0.40$, $p < 0.001$), though the slope difference was only near-significant (difference = -0.17 $\pm$ 0.097 SD, $t = -1.77$, $p = 0.079$).

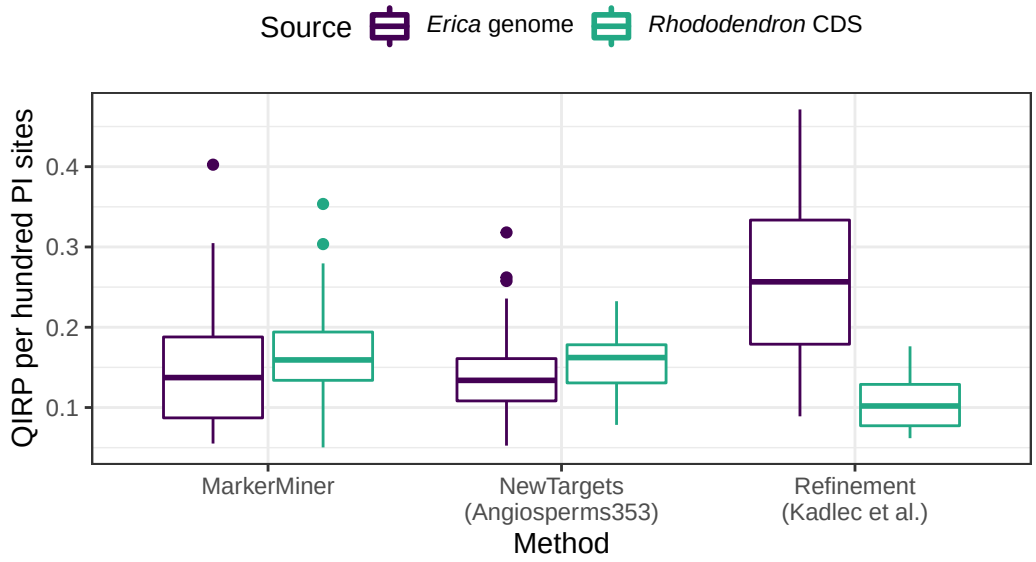


Fig. 2.15 QIRP per hundred PI sites in relation to target source and method.

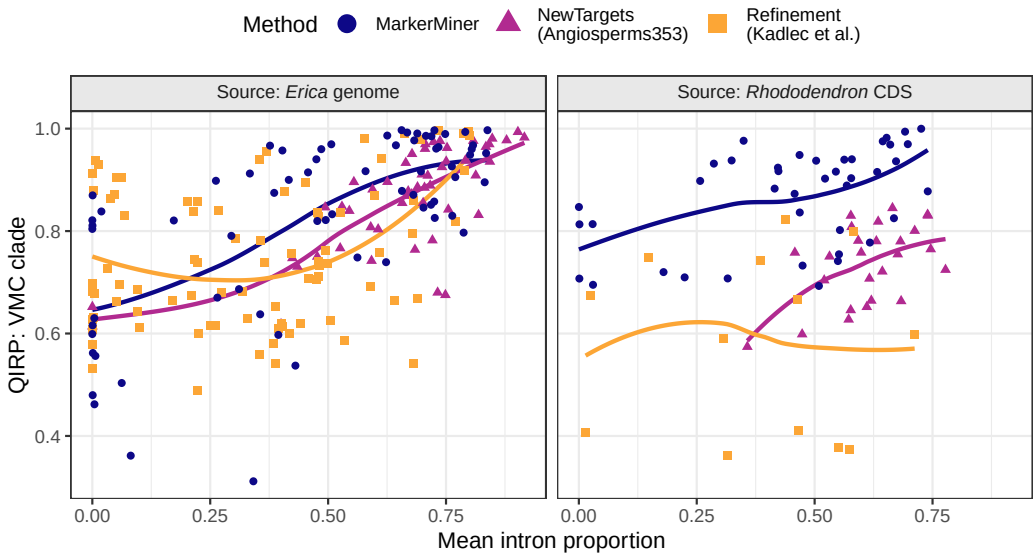


Fig. 2.16 QIRP at the crown of the "VMC clade" in relation to the proportion of intronic sequence, target source and method. Lines show loess model fits with span = 1.

2.7 Conclusions

In this chapter, I developed and tested a new target set for *Erica* phylogenomics using a variety of methods. Overall, I was able to implement effective measures that kept the rate of paralogy and missingness in the resulting target capture data to very low levels. Post-assembly refinement of the target set only reduced the number of targets from 303 to 285, suggesting that the target design approaches effectively identified most undesirable loci. As such, these targets can be expected to be applicable to future phylogenomic studies in *Erica*. Furthermore, good target recovery in the three non-*Erica* samples tested (*Rhododendron rex*, *Calluna vulgaris*, and *Daboecia cantabrica*) suggests that the targets could also be applied to these genera, and perhaps even to more distant relatives (i.e., in Ericaceae beyond the Ericoideae).

Looking beyond the specific target set, I expect that the various target design methods presented here will be generally applicable to any plant group. These include using the NewTargets method of McLay et al. (2021) for target discovery, using assembled targets from a closer relative to iteratively refine an earlier target set (Kadlec et al., 2017), and using WGS and off-target reads from a previous target capture experiment to predict paralogy and presence of candidate targets in the study species. To aid others in implementing several of these approaches, I developed and made freely available an open-source toolkit, TARGETVET.

The results of this chapter demonstrate that the new target set has excellent phylogenetic informativeness, and one of the major reasons for this was the inclusion of intronic sequences in the targets used for bait design. Although this approach has rarely been attempted (de Sousa et al., 2014; Folk et al., 2015), the results indicated high capture efficiency of introns even for Cape *Erica* species, despite the target source being a European *Erica* more than 40 million years diverged (Pirie et al., 2016, Fig. 2.7). Targeting introns appeared to improve their downstream assembly and contiguity, as targets including introns recovered a larger proportion of intronic sequence relative to target length (Fig. 2.9, Table 2.3). Finally, the proportion of intronic sequence correlated well with phylogenetic informativeness (Fig. 2.16). These results should encourage researchers working in phylogenomics to include introns in their targets, where possible, in order to improve the phylogenetic informativeness of their data.

Chapter 3

Phylogenomics of the *Erica abietina/E. viscaria* clade

3.1 Background

3.1.1 Diversity and distribution

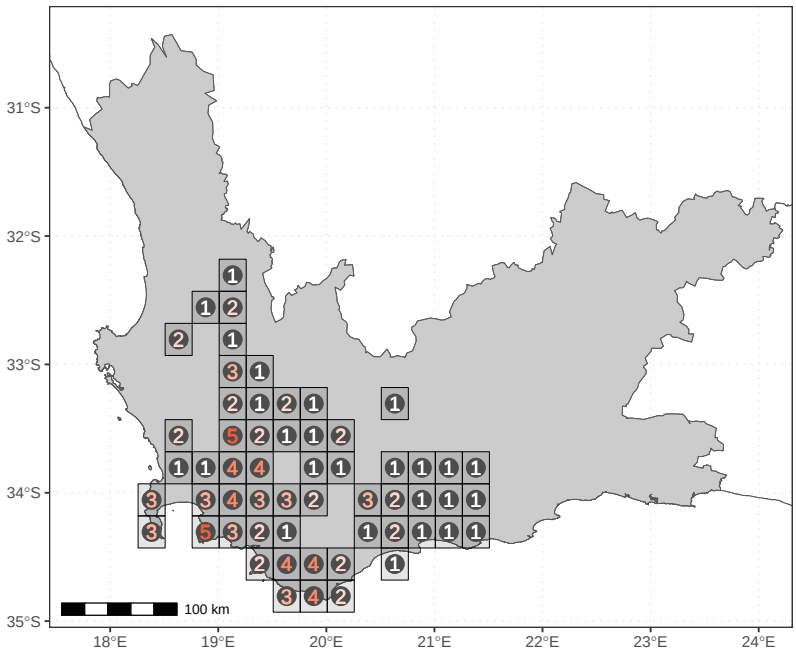
While the Cape *Erica* clade is, as a whole, remarkable for its high species richness, considerable trait variation, and rapid diversification (Manning and Goldblatt, 2012; Pirie et al., 2016), the *Erica abietina/E. viscaria* clade stands out as an exemplary microcosm of all of these factors. It has been estimated to have a crown age of 2-3 Ma (Pirie et al., 2016) and is currently known to hold 19 species (Pirie et al., 2017) and a total of at least 29 taxa, including subspecies and varieties (Table 3.1). Geographically, its diversity follows much the same pattern as the rest of Cape *Erica*: although found throughout much of the CFR, the mountains of the south-western Cape are its centre of diversity and endemism (Fig. 3.1; Pirie et al., 2022, 2019). Additionally, its species range from being very widespread (e.g., *E. grandiflora*, *E. vestita*, *E. parilis*), to highly range-restricted (e.g., *E. filamentosa*, *E. hibbertia*, *E. petrusiana*, *E. situshiemalis*), to effectively limited to a very small area by being confined to high elevations near mountain peaks (e.g., *E. doliiformis*, *E. phillipsii*). Its species also occupy a range of soil types. Most are found on the sandstone-derived sands which dominate the CFR, a few are largely confined to soils derived from shale (e.g., *E. latiflora*, *E. petrusiana*, *E. regia*

subsp. *regia*) or limestone (e.g., *E. regia* subsp. *mariae*), and some occupy a variety of soil types (e.g., *E. grandiflora*, *E. abietina*).

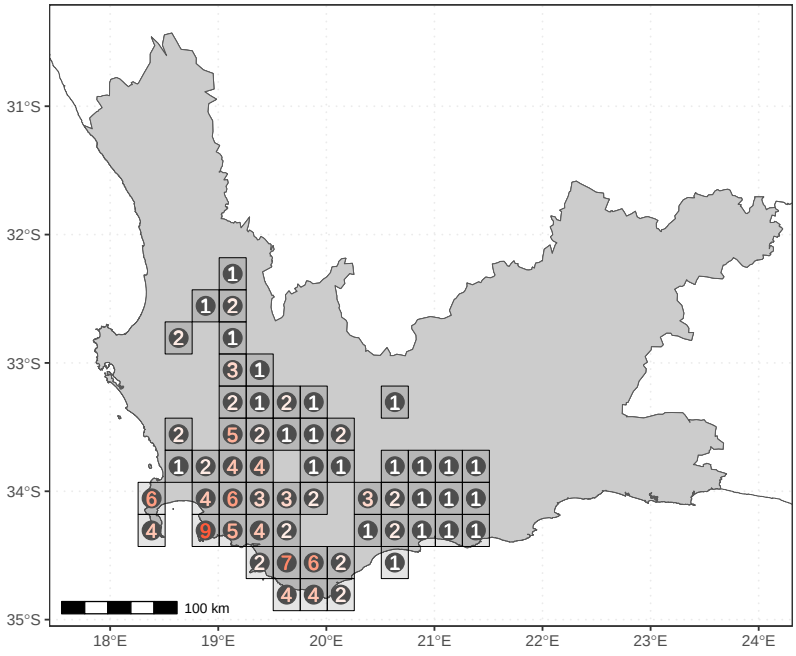
3.1.2 Phenotypic variation

All of the *E. abietina/E. viscaria* clade's species appear to share a distinctive combination of vegetative and floral traits: (1) habit upright, never sprawling or straggling; (2) leaves narrow with margins rolled under, arranged in whorls of six; (3) flowers borne singly at the ends of very short lateral branchlets (giving the appearance of being axillary on the main stem) and arranged into a spike-like synflorescence comprising one to several whorls at or near the tips of the main stems; (4) bracts placed medially on the pedicel or proximate to the calyx; and (5) anthers small and either entirely lacking or occasionally having very reduced appendages below the thecae. Within these parameters, however, traits vary considerably, both between and within species (Table 3.1):

- **Vegetative traits.**
  - Habit: plants range from small shrublets seldom exceeding ca. 30 cm in height (e.g., *E. doliiformis*) to tall shrubs > 150 cm tall (e.g., *E. abietina* subsp. *atrorosea*).
  - Fire survival strategy: reseeding (most species); resprouting (*E. viscaria* subsp. *macrosepala*); avoiding (in rock crevices, e.g., *E. doliiformis*, *E. hibbertia*, *E. nevillei*, *E. quadrisulcata*, *E. situshiemalis*).
- **Floral traits.**
  - Size and shape: short, *ca.* 5-8 mm, bell- or cup-shaped, (e.g., *E. axilliflora*, *E. viscaria* subsp. *viscaria*); short to medium, *ca.* 5-16 mm, urn-shaped (e.g., *E. phillipsii*, *E. doliiformis*); medium to long, *ca.* 15-30 mm, curved or straight tubes (e.g., *E. viscaria* subsp. *longifolia*, *E. grandiflora*).
  - Colour: white, pink, red, yellow, orange, green, and various combinations thereof (see Fig. 3.8).
  - Scent: none (most species); sweet (*E. viscaria* subsp. *viscaria*); lemony (*E. abietina* subsp. *constantiana*).



(a)



(b)

Fig. 3.1 Species richness (a) and species + infraspecies richness (b) in the *E. abietina/E. viscaria* clade, based on "research grade" observations from `iNaturalist.org` (accessed 20.11.2022). Grid cells are quarter degrees, roughly 24 km by 28 km.

3.1.3 Taxonomy and phylogeny

Most of the *E. abietina/E. viscaria* clade's infraspecific taxa are held by just two species: *E. abietina sensu lato* (four subspecies; Pirie et al., 2017) and *E. viscaria* (six subspecies; Oliver and Oliver, 2002). Both of these species show considerable variation in the colour, shape and – occasionally – scent of the flowers. Most of this variation is delineated by the various subspecies, but Oliver and Oliver (2002) suggested that future work, including molecular analyses, might uncover additional variation warranting taxonomic recognition especially within the extremely variable *E. viscaria* subsp. *longifolia*. Unfortunately, Pirie et al. (2017) found that traditional plastid and nuclear phylogenetic markers were entirely unable to resolve the phylogenetic relationships between the subspecies within either species. More broadly, they were also unable to confidently resolve many aspects of the clade's phylogeny owing largely to a lack of phylogenetic signal but also as a result of considerable discordance between their nuclear and plastid phylogenies. Notably, their analyses placed several species within a reasonably well-supported clade that they called the "*viscaria*-clade", within which most branches were unresolved.

Another major concern arising from the Pirie et al. (2017) study was that the monophyly of several species (e.g., *E. abietina sensu* Oliver and Oliver (2002), *E. vestita*, *E. viscaria*) could not be confirmed. Despite a general lack of phylogenetic resolution, they were able to resolve the paraphyly of *E. abietina* by refining its taxonomy. Firstly, they placed *E. abietina* subsp. *aurantiaca* E.G.H.Oliv. & I.M.Oliv. along with *E. abietina* subsp. *perfoliosa* E.G.H.Oliv. & I.M.Oliv. within a resurrected *E. grandiflora*, as *E. grandiflora* subsp. *grandiflora* and *E. grandiflora* subsp. *perfoliosa*, respectively. Despite being unable to confirm the monophyly of *E. grandiflora* as a whole, they based this decision on the morphological similarity of the two new subspecies and their finding that all of their samples fell within the "*viscaria*-clade" (rather than the "*abietina*-clade"). Secondly, they showed conclusively that *E. abietina* subsp. *petraea* E.G.H.OIiv. & I.M.Oliv. was not closely related to the rest of *E. abietina*, and raised it to species level as *E. situshiemalis*.

The combination of substantial trait variation alongside considerable phylogenetic and taxonomic uncertainty make the *E. abietina/E. viscaria* clade an excellent study system for investigating patterns and processes of diversification in Cape *Erica*, and a perfect target for phylogenomic analysis. The

aim of this chapter, therefore, was to reconstruct the clade's phylogeny using the target set designed and refined in Chapter 2, with the goal of furthering our understanding of its diversification.

3.2 Methods and results

3.2.1 Taxon sampling

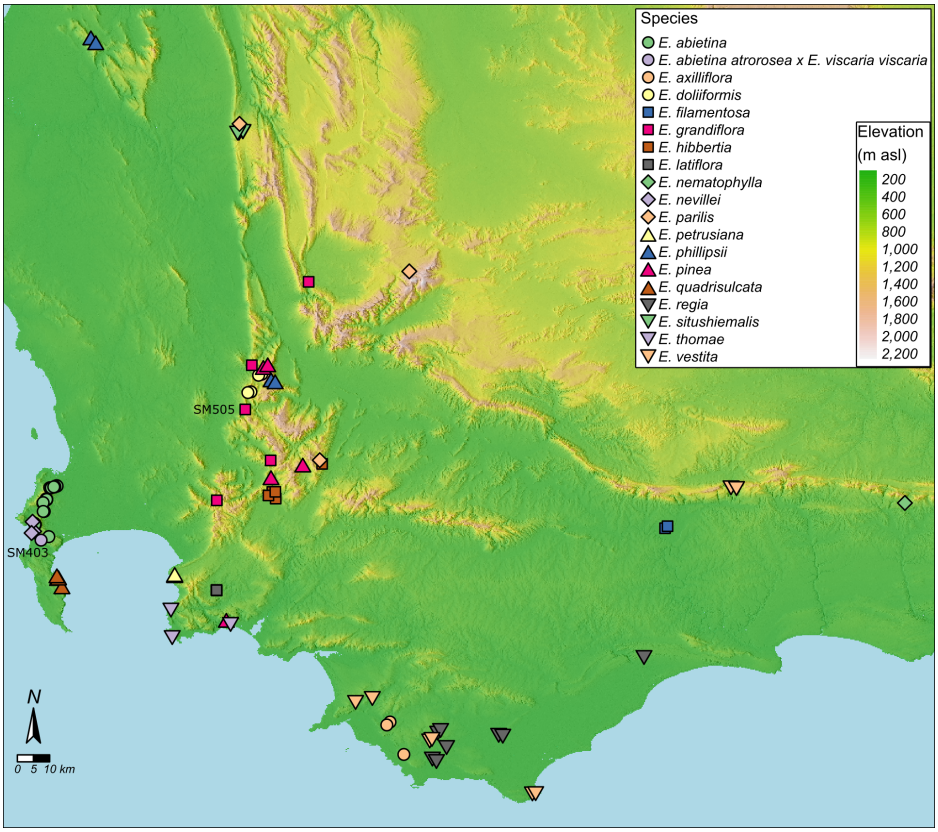
At least one specimen of each of the nineteen species belonging to the *E. abietina/E. viscaria* clade (*sensu* Pirie et al., 2017, 2016) was sequenced, along with all but two subspecies and varieties therein (Oliver and Oliver, 2002), totalling 133 samples (Tables 3.1, B.1; Fig. 3.2). Extra sampling effort was placed on the species *E. abietina* and *E. viscaria* in order to investigate the relationships between their many subspecies.

The sample set included specimens of some taxa that have never been sequenced for phylogenetic analysis. For the first time, *E. petrusiana* was sequenced. This highly localised and poorly-known species was described by Oliver and Oliver (2002), who noted its close affinity to *E. viscaria* subsp. *viscaria* but justified a species-level description owing largely to its unusual (in *Erica*) combination of floral features, being both yellow and short-tubed. A specimen of the extremely localised *E. viscaria* subsp. *gallorum* was also sequenced for the first time, along with three similar specimens of uncertain taxonomic status. These belong to a recently-discovered population (see `inaturalist.org/observations/11312498`) that appears to be restricted to a single hillside some 60 km south-east of the known range of *E. viscaria* subsp. *gallorum*, with several major biogeographic barriers in between, and have short corollas like *E. viscaria* subsp. *gallorum* but a much more lax habit. *Erica thomae* var. *tenax* was also sequenced for the first time, as was *E. casta* (*sensu lato*), which Oliver and Oliver (2002) considered to be a gracile form of *E. regia* subsp. *regia*, and which I here refer to as *E. regia* var. *casta* for the sake of clarity. Lastly, one specimen with several features suggesting that it is a hybrid between *E. abietina* subsp. *abietina* and *E. viscaria* subsp. *viscaria* (see Fig. 3.3) was sequenced in order to test the hypothesis of its parentage.

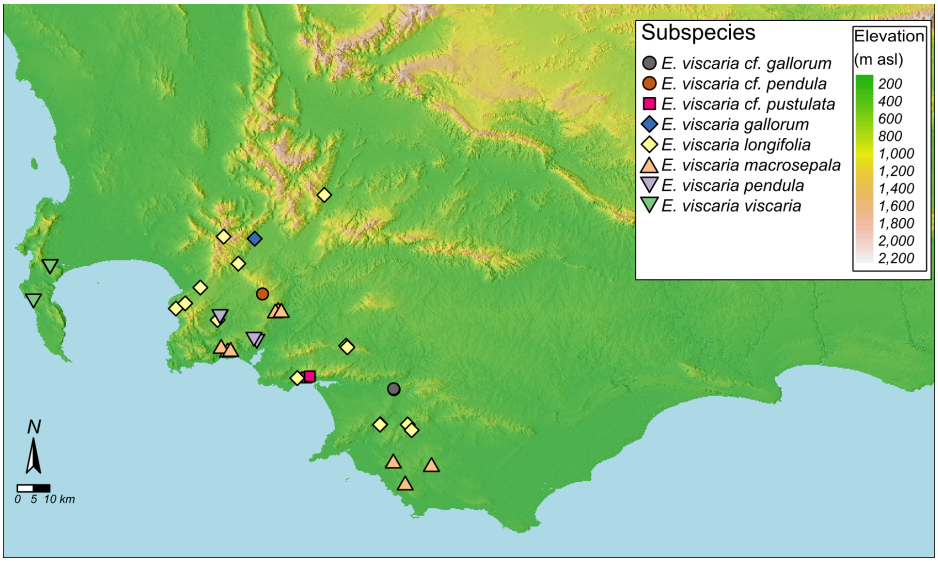
Taxa that were not sampled were (1) *E. viscaria* subsp. *pustulata* (H.A.Baker) E.G.H.Oliv. & I.M.Oliv which could not be located at its type locality, although plants with the characteristic pustulated corolla but with corolla length > ca. 15 mm (as opposed to ca. 7 mm as per the type) were

Table 3.1 Details of the taxa from the *E. abietina/E. viscaria* clade that were analysed in this chapter. The number of samples (n) and details of flower corolla length (L, in mm), colour (C), and scent (S) are given. Colour codes are: R = red; P = pink; O = orange; Y = yellow; W = white; G = green; lower case = colour towards corolla tip.

Taxon	Author(s)	n	L/C/S
E. abietina subsp. *abietina*	L.	7	18-26/R/-
E. abietina subsp. *atrorosea*	E.G.H.OIiv. & I.M.Oliv.	2	18-22/P/-
E. abietina subsp. *constantiana*	E.G.H.OIiv. & I.M.Oliv.	5	8-11/P/✓
E. abietina subsp. *diabolis*	E.G.H.OIiv. & I.M.Oliv.	8	11-14/P/-
E. abietina subsp. *atrorosea* x *constantiana*	N/A	1	15/P/-
E. axilliflora	L.Bolus	3	6-8/P/?
E. doliiformis	Salisb.	4	11-16/P/-
E. filamentosa	Andrews	2	8-9/P/-
E. grandiflora subsp. *grandiflora*	L.f.	4	25-30/O,R,Ry/-
E. grandiflora subsp. *perfoliosa*	(E.G.H.Oliv. & I.M.Oliv.) E.G.H.Oliv. & Pirie	1	20-30/Y/-
E. hibbertia	Andrews	5	27-34/Ry/-
E. latiflora	L.Bolus	3	5-10/P/?
E. nematophylla	Guthrie & Bolus	1	10-12/W,P/?
E. nevillei	L.Bolus	3	25-30/R/-
E. parilis	Salisb.	3	5-9/Y/-
E. petrusiana	E.G.H.OIiv. & I.M.Oliv.	4	5-9/Y/-
E. phillipsii	L.Bolus	4	5-8/P/-
E. pinea	Thunb.	5	23-27/W,P,Yw/-
E. quadrisulcata	L.Bolus	5	26-30/O/-
E. regia subsp. *regia*	Bartl.	3	14-20/R,Wr/-
E. regia subsp. *mariae*	(Guthrie & Bolus) E.G.H.OIiv. & I.M.Oliv.	3	18-22/R/-
E. regia var. *casta*	Guthrie & Bolus	2	12-14/W/-
E. situshiemalis	E.G.H.Oliv. & Pirie	4	18-20/Y/-
E. thomae "pink" (cf. var. *porteri*)	N/A	1	20-25/Pw/-
E. thomae var. *tenax* (Variant B)	L.Bolus	1	22-30/G/-
E. thomae var. *thomae* (Variant A)	L.Bolus	1	22-30/W/-
E. vestita	Thunb.	8	16-24/W,P,R/-
E. viscaria subsp. *viscaria*	L.	4	5-9/P/✓
E. viscaria subsp. *gallorum*	(L.Bolus) E.G.H.OIiv. & I.M.Oliv.	1	5-10/P/-
E. viscaria subsp. *longifolia*	(Bauer) E.G.H.OIiv. & I.M.Oliv.	14	12-20/W,P,R,G,Yw,Ry,Pw/-
E. viscaria subsp. *macrosepala*	E.G.H.OIiv. & I.M.Oliv.	9	15-20/G/-
E. viscaria subsp. *pendula*	E.G.H.OIiv. & I.M.Oliv.	4	12-18/W/-
E. viscaria cf. subsp. *gallorum*	(L.Bolus) E.G.H.OIiv. & I.M.Oliv.	3	7/P/-
E. viscaria cf. subsp. *pendula*	N/A	1	15/Pw/-
E. viscaria cf. subsp. *pustulata*	(H.A.Baker) E.G.H.OIiv. & I.M.Oliv.	3	15/G/-
E. abietina atrorosea x *E. viscaria viscaria*		1	14/P/-



(a)



(b)

Fig. 3.2 Maps showing sampling localities for (a) species level and (b) for *E. viscaria*, subspecies level. Samples suspected of being recent hybrids are labelled in (a). The base maps show elevation and hillshade (270°).

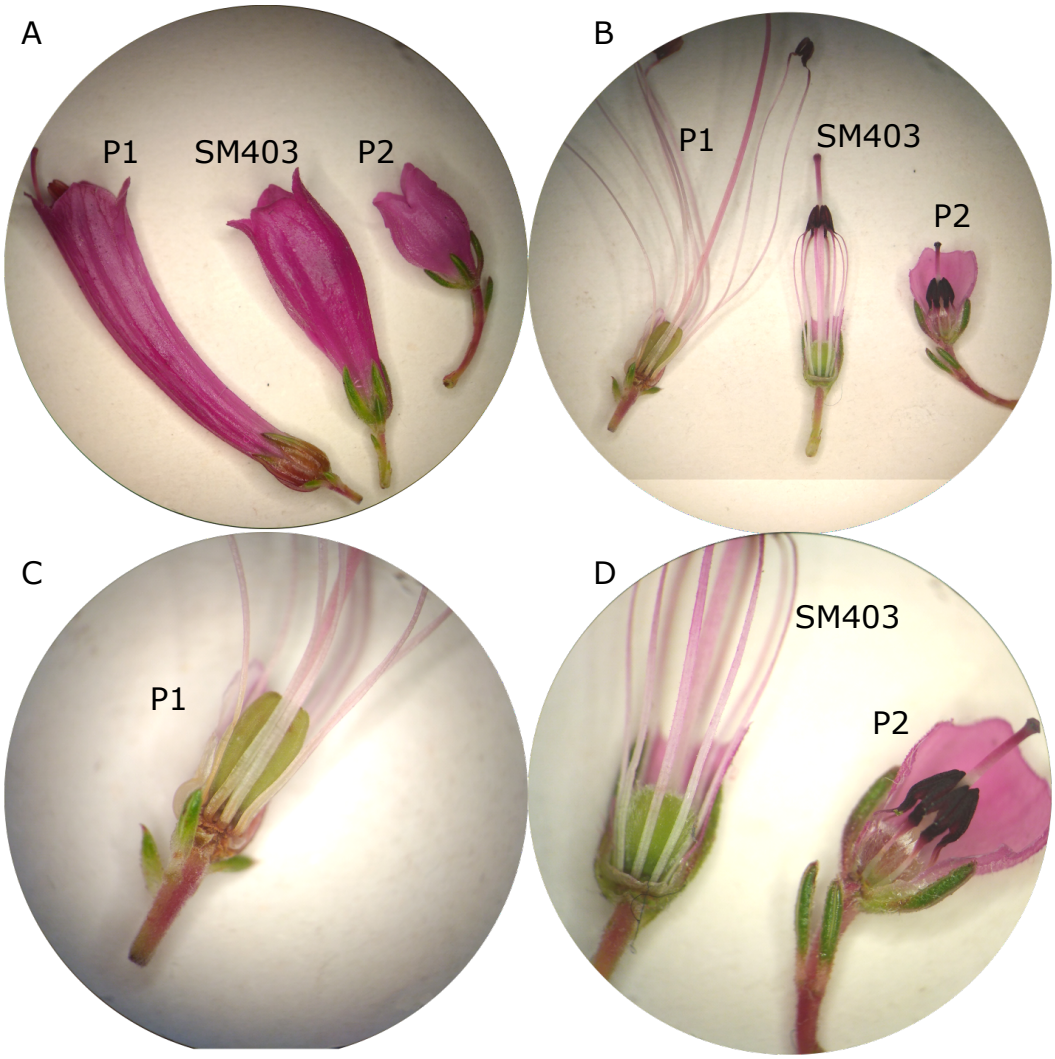


Fig. 3.3 Flower images taken under a dissecting microscope of the sample SM403, suspected of being a hybrid between *E. abietina* subsp. *atrorosea* (P1) and *E. viscaria* subsp. *viscaria* (P2), alongside representatives of its putative parent species. The flowers were all collected on the same day at the same locality, and the same three flowers were used in all images. SM403 possesses characters broadly intermediate between P1 and P2: in the shape and length of the corolla (A); width and shape of the sepals (A); length of the style and filaments (B); and shape, size and hairiness of the ovary (C,D). Magnification: A,B = 10x; C,D = 20x.

located and collected nearby, of which three were sequenced; and (2) *E. thomae* var. *porteri*, which is highly localised and could not be located due to a recent wildfire, although a specimen that matched the type in having a pink-white corolla but differed due to the corolla not being exceptionally slender was collected from nearby and was labelled as *E. thomae* "pink".

All specimens were collected in the wild as part of this project with permission from the relevant authorities (see Acknowledgements), except for the single sample of *E. nematophylla* Guthrie &

Bolus, which was used with permission from the collector, E.G.H. Oliver. This sample was collected in 2011 and had been silica-dried and kept in cold storage. DNA extraction and sequencing are detailed in Chapter 2, Section 2.5.1.

3.2.2 Phylogenetic inference

I aimed to infer the phylogeny of the *E. abietina/E. viscaria* clade using both concatenation- and coalescent-based methods. In an effort to minimise systematic error, I chose to only use the low-paralogy, low-missingness Erica285 subset of genes (see Chapter 2, Section 2.5.3). First, to verify the monophyly of the clade I chose a subset of *Erica* samples (plus *Calluna vulgaris*; n = 178) belonging to species spread throughout the "Cape *Erica*" clade (Pirie et. al, in prep.; Pirie et al., 2016). The HybPiper-assembled supercontigs of genes in the Erica285 targets set were retrieved for these samples, following which multiple sequence alignment (including "chomping" and trimming) and phylogeny inference using IQ-TREE and wASTRAL-h was conducted. Methods followed those detailed in Chapter 2, Section 2.6, with two exceptions: for the concatenation (IQ-TREE) analysis the approximate Bayes (aBayes; Anisimova et al., 2011) method of estimating branch support was used in addition to UFBoot and SH-alrt, and partition merging was not done because of computational limitations.

These analyses confirmed the monophyly of the *E. abietina/E. viscaria* clade, and identified *E. filiformis* and *E. stokoei* as the clade's closest relatives among the sample set, which is in line with recent results using traditional phylogenetic markers (Pirie et al., in prep.). The two methods, however, disagreed on the branching order of these samples, with ASTRAL suggesting *E. filiformis* and concatenation suggesting *E. stokoei* to be sister to the *E. abietina/E. viscaria* clade (Fig. 3.4).

To infer the *E. abietina/E. viscaria* clade phylogeny I included the samples of the *E. abietina/E. viscaria* clade (n = 133) along with the samples of *E. stokoei*, *E. filiformis* and *E. massonii* to be used for rooting. Rather than simply filtering the supercontig MSAs generated for the previous analysis, I reran the alignment and subsequent trimming steps from scratch in order to improve alignment accuracy, which may have been compromised by the larger number of samples and greater phylogenetic distance among samples in the previous analysis. To determine the quality of the alignments, alignment statistics were calculated using AMAS (Borowiec, 2016). The alignments

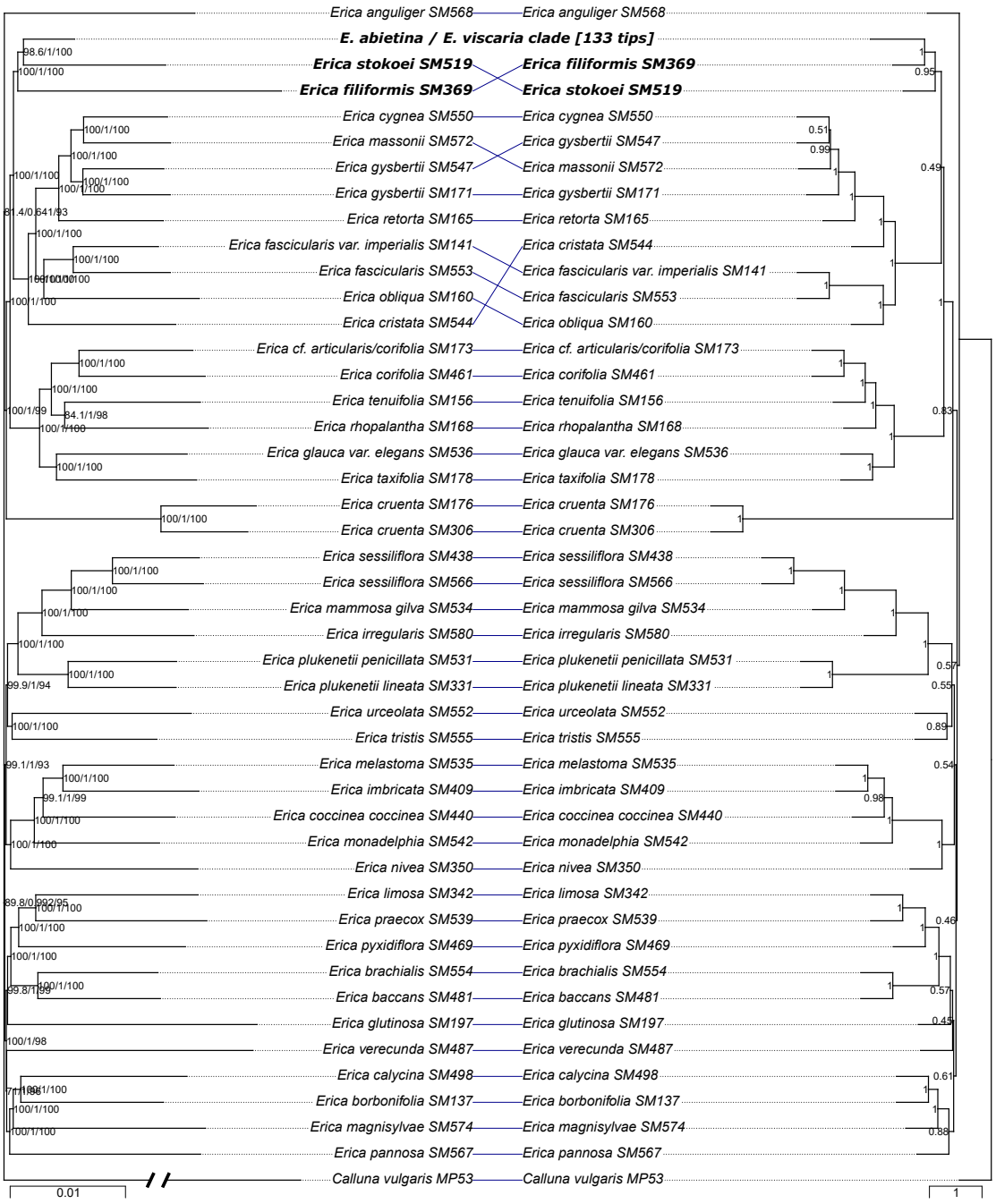


Fig. 3.4 Tanglegram comparing the phylogenies inferred using concatenation (*Left*) and ASTRAL (*Right*). For the concatenation tree, branch lengths are in units of expected substitutions per site (note that the branch leading to *C. vulgaris* has been shortened) and node labels show SH-alrt/aBayes/UFBoot support. For the ASTRAL tree, branch lengths are in coalescent units (except for terminal branches arbitrarily set to 0.5) and node labels show local posterior probability support. In bold are the *E. abietina / E. viscaria* clade and its closest outgroups among the sample set.

were highly complete: >99% of the 285 genes had all 136 samples; the proportion of missing bases per gene ranged from 1.5% to 44% (mean = 23%) between genes; and missingness was low on a per-sample basis, with mean missingness ranging from 20% to 33% (mean = 23%) across all alignments. GC content per gene ranged from 36.3% to 50.4% (mean = 40.6% $\pm$ 2.5% SD). The concatenated matrix had 1,774,435 sites, of which 127,235 (7.2%) were parsimony informative and 294,164 (16.6%) were variable. The proportion of parsimony informative sites per gene ranged from 2.7% to 14.7% (mean = 7.3%).

Tree inference methodological details were the same as for the expanded sample set (above) except that partition merging was done for the concatenation-based method as the reduced sample size made it computationally feasible.

3.2.3 Comparison between phylogenetic inference methods

The concatenation and ASTRAL trees for the *E. abietina/E. viscaria* clade were largely concordant (Fig. 3.5). The three well-supported conflicts were (1) the placement of SM403, a suspected hybrid *E. abietina* subsp. *atrorosea* x *E. viscaria* subsp. *viscaria* (Fig. 3.3); (2) the placement of one sample identified as *E. grandiflora* subsp. *grandiflora* (SM505); and (3) the placement of the species *E. quadrisulcata*, *E. nevillei*, *E. doliiformis* and *E. phillipsii*. Conflicts (1) and (2) are addressed below in the context of hybridisation (Subsection 3.2.4). In conflict number (3), the concatenation tree recovered the four aforementioned species as a fully supported clade sister to the clade consisting of *E. abietina* and *E. grandiflora* (clade "AG"), while the ASTRAL tree recovered *E. quadrisulcata* and *E. nevillei* as a clade which was sister to clade AG (albeit with somewhat low support, LPP = 0.83), with *E. doliiformis* and *E. phillipsii* branching off earlier.

Perhaps the most notable overall difference between the trees was that all measures of branch support were universally greater in the concatenation tree than the ASTRAL tree (Fig. 3.5). Some authors have argued that gene tree estimation error (GEE) significantly compromises methods such as ASTRAL by inducing a false signal of ILS in which gene tree conflict is a product of GEE rather than ILS (Richards et al., 2018; Springer and Gatesy, 2016). However, branch support values of the gene trees in the Erica285 set used to infer the ASTRAL tree of the *E. abietina/E. viscaria* clade were generally high (Fig. 3.6), and wASTRAL-h takes branch support into account when estimating the

species tree (Zhang and Mirarab, 2022). It therefore seems likely that, rather than being caused by GEE, the low LPP values of several internal branches in the ASTRAL tree (highlighted in Fig. 3.8) reflect a genuine signal of incongruence between gene trees caused by ILS and/or ancient introgression (Giarla and Esselstyn, 2015; Sayyari and Mirarab, 2016). This implies, firstly, that the support values of the concatenation-based tree are inflated, which is a typical result in phylogenomic studies (e.g., Arcila et al., 2021; Rodríguez et al., 2017; Roycroft et al., 2019). Secondly, and more importantly, it implies that the concatenation-based phylogeny may be less accurate than the ASTRAL phylogeny, as has consistently been demonstrated to be the case when ILS is anything other than negligible, using both simulated and empirical data (Bagley et al., 2020; Jiang et al., 2020; Zhang and Mirarab, 2022).

3.2.4 Evidence of recent hybrids

Both phylogenetic inference methods recovered SM403 in intermediate positions between its putative parents, though concatenation placed it relatively closer to the *E. abietina* clade. The concatenation tree also placed SM403 on the longest terminal branch in the tree (Fig. 3.9; note that ASTRAL does not infer terminal branch lengths). Phylogenomics in empirical systems has shown that species tree reconstruction methods tend to place hybrid individuals on relatively long terminal branches in positions roughly intermediate between their parent species (Chan et al., 2020; Dolinay et al., 2021; Pyron et al., 2022). The placement and branch length of this sample, combined with its intermediate morphological features (Fig. 3.3; McDade, 1990), therefore provide strong evidence that SM403 is indeed a hybrid between these two relatively distantly related taxa.

The anomalous placement of SM505 is more difficult to explain than that of SM403 because the specimen presents morphologically as *E. grandiflora* subsp. *grandiflora*. The discordance between the concatenation and ASTRAL trees may be noteworthy: in the ASTRAL tree, SM505 occupied the earliest branching position in a clade comprising *E. abietina*, *E. nevillei*, *E. quadrisulcata* and the rest of the *E. grandiflora* samples, whereas in the concatenation tree it was placed at the earliest branching position in a monophyletic *E. grandiflora* clade. It is well-known that concatenation-based species tree inference is less sensitive than coalescence-based methods to conflicting topological signals between different regions of the genome, such as might be caused by ILS or introgression, especially when those signals are relatively weak (Giarla and Esselstyn, 2015; Jiang et al., 2020). The

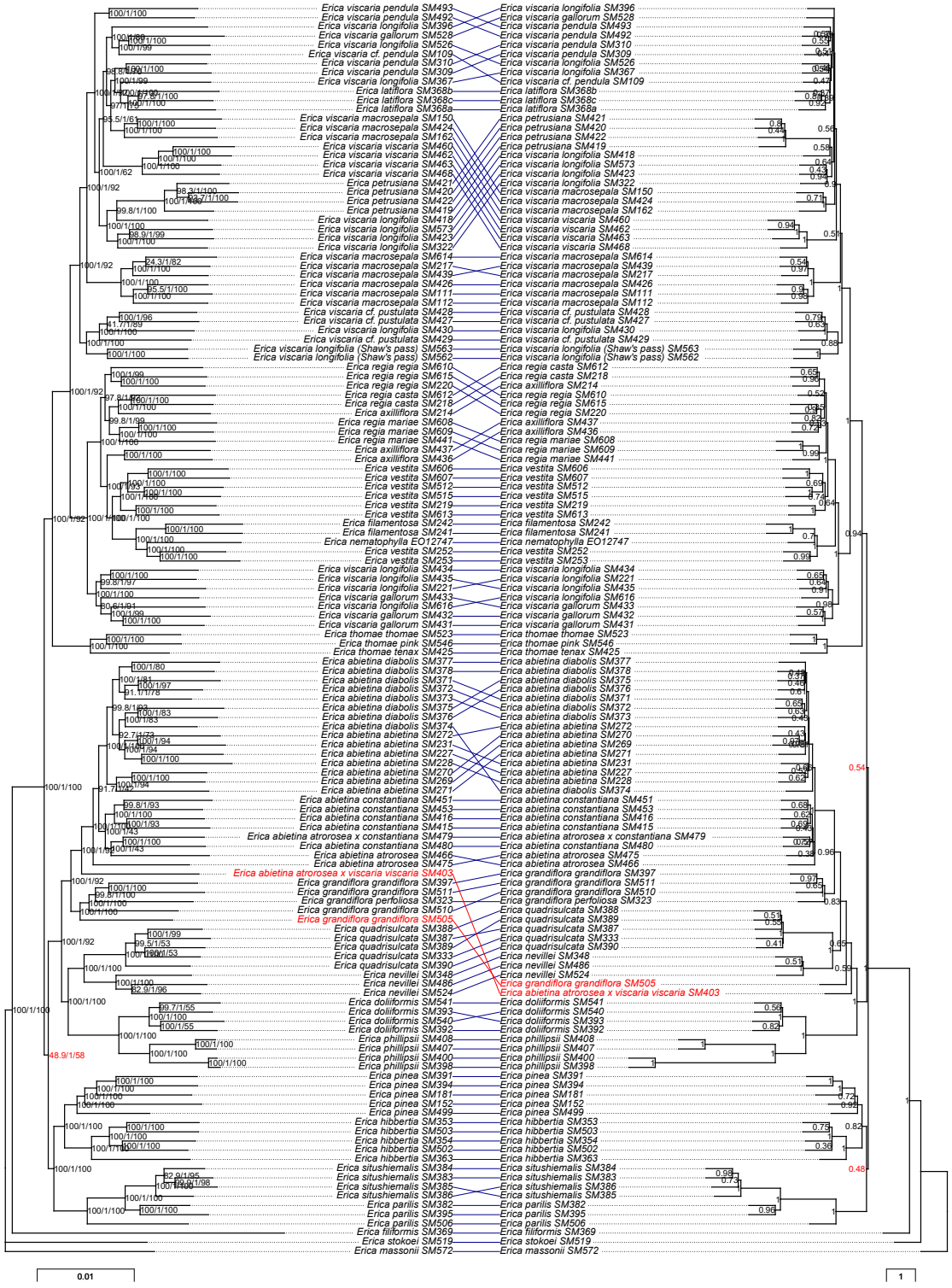


Fig. 3.5 Tanglegram comparing the *E. abietina/E. viscaria* clade trees inferred using concatenation (*Left*) and ASTRAL (*Right*). Highlighted with red text are the two putative hybrids and their relative positions as well as the support values of the branches that indicate a basal polytomy. Other details as in Fig. 3.4.

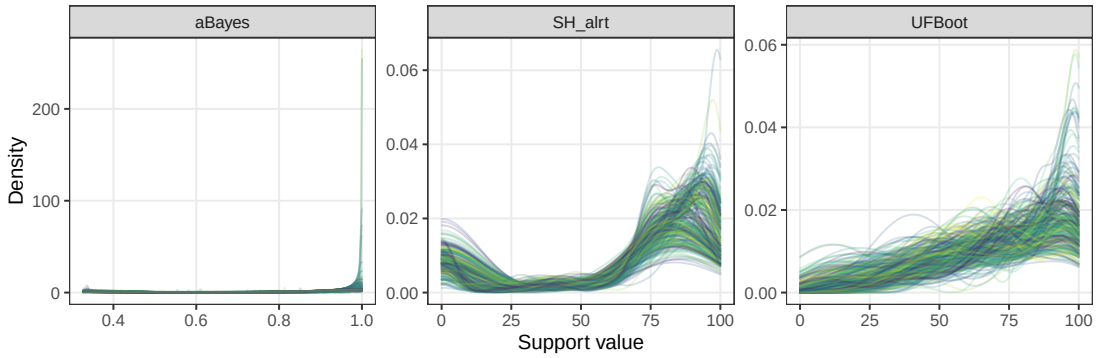


Fig. 3.6 Density plots showing the distribution of branch support values across the 285 ML gene trees inferred using IQ-TREE for the *E. abietina/E. viscaria* clade samples. Colours are used arbitrarily to help distinguish individual gene trees.

observed incongruence between methods might therefore suggest that SM505 is a late-generation hybrid deriving most of its ancestry from *E. grandiflora*, which would have obscured the signal of mixed ancestry from the concatenation approach. There is some circumstantial evidence that could support this possibility. The only other species in the *E. abietina/E. viscaria* clade that occurs at the same locality as SM505 (Du Toitskloof pass; latitude,longitude: -33.699780,19.068118) is *E. pinea*, and this appears to be one of relatively few localities where these two species occur side by side. Geographic proximity, and the fact that the species share the sunbird pollination syndrome (Rebelo et al., 1985), suggests that pollen exchange between the two species is likely. Furthermore, there are several observations from Du Toitskloof of plants with a combination of *E. grandiflora*- and *E. pinea*-like traits (Fig. 3.7). When further considering the strong evidence for hybridisation between *E. abietina* subsp. *abietina* and *E. viscaria* subsp. *viscaria* (see above), which have different pollination syndromes and are even more distantly related than are *E. grandiflora* and *E. pinea* (Fig. 3.5), it seems likely that viable hybrids between these species do occur and at least possible that they have in the past back-crossed with *E. grandiflora* to produce individuals such as SM505.

The effect of hybrids on phylogenetic inference

To determine whether the two putative hybrids, SM403 and SM505, could have negatively affected the accuracy of phylogenetic inference and/or caused some of the low LPP values at surrounding clades, I excluded these samples from the MSAs before repeating the gene tree inference and wASTRAL-h

analyses. I then compared, using *comparePhylo* from APE, the new "no hybrids" ASTRAL tree to the original ASTRAL tree with the two hybrids pruned out. This showed that the inclusion of hybrids had a negligible effect: of a total of 133 internal nodes, 114 were present in both trees; these common nodes did not have notably different LPP values on average (paired t-test: mean difference = 0.007232, t = 1.013, df = 111, p-value = 0.3133) nor at nodes surrounding the hybrids (Fig. 3.8); and the nodes that were unique to each tree were shallow (node depth: no hybrids tree median, maximum = 4, 22; original tree median, maximum = 3, 20) and had low LPP values (no hybrids tree mean = 0.50 $\pm$ 0.13 SD; original tree mean = 0.49 $\pm$ 0.09 SD). I therefore concluded that the ASTRAL topology was robust to the presence of putative hybrids.

3.2.5 The *E. abietina/E. viscaria* clade phylogeny

Conflict and resolution

Although both phylogenetic analyses indicated non-trivial uncertainty in the topology of the tree (Fig. 3.5), they nevertheless provided much better resolution than Pirie et al. (2017) were able to achieve using traditional markers (several plastid genes and nuclear ITS and ETS), and many relationships that could not be resolved by those authors were confidently resolved here. These included several groupings at the interspecific level. *Erica parilis* and *E. situshiemalis* were found by both phylogenetic methods to form a well-supported clade, as were *E. hibbertia* and *E. pinea* (albeit with relatively low support from ASTRAL; Fig. 3.5). Pirie et al. (2017) found a well-supported clade, which they named the "*abietina*-clade", consisting of three species endemic to the mountains of the Cape Peninsula: *E. abietina*, *E. nevillei*, and *E. quadrisulcata*. Although this grouping was also recovered here, both concatenation and ASTRAL trees additionally included all of the *E. grandiflora* samples in a "new-*abietina*-clade" (Fig. 3.5), whereas the Pirie et al. (2017) analysis placed their *E. grandiflora* samples at various positions within their "*viscaria*-clade". Strong support was also found by both methods for a clade that contained *E. doliiformis* and *E. phillipsii* along with the "new-*abietina*-clade", which was in agreement with the results of Pirie et al. (2017), notwithstanding the inclusion of *E. grandiflora*.

Regarding the Pirie et al. (2017) *viscaria*-clade, the concatenation and ASTRAL analyses both re-

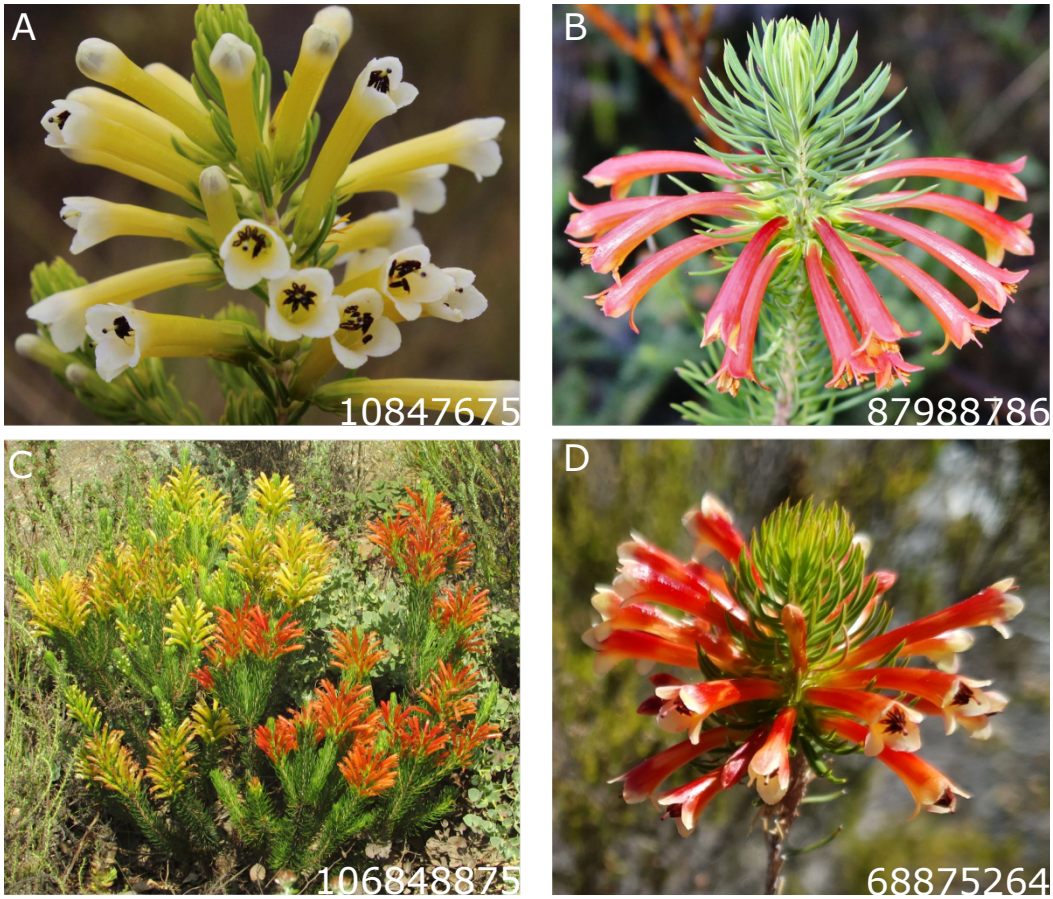


Fig. 3.7 Images of plants observed at Du Toitskloof pass, illustrating typical *E. pinea* (A), typical *E. grandiflora* subsp. *grandiflora* (B), the two taxa occurring and flowering side by side (C), and a putative hybrid between these two species (D). Images are from `iNaturalist.org` and are shown with their associated unique observation identifier. Another putative hybrid individual can be seen at `inaturalist.org/observations/107460733`. Characters typical of *E. pinea* from this locality that are evident in (A) are (1) corolla yellow at the base and white at the tip; (2) corolla widening from the base, first gradually then abruptly, then narrowing slightly just before the tip to create a "bulbous" appearance; and (3) tips of the corolla lobes relatively rounded. Typical *E. grandiflora* characters evident in (B) are (1) corolla uniform orange; (2) corolla widening just before the tip; and (3) corolla lobes reflexed, pointed.

covered a well-supported clade, which I term the "new-*viscaria*-clade", which comprised *E. axilliflora*, *E. filamentosa*, *E. latiflora*, *E. nematophylla*, *E. petrusiana*, *E. regia*, *E. vestita*, and *E. viscaria*. While the Pirie et al. (2017) *viscaria*-clade also contained these species, it additionally held *E. grandiflora*, *E. hibbertia* and *E. pinea*, which in the present work were placed in other clades (see above). Finally, *E. thomae*, another taxon that could not be placed confidently by Pirie et al. (2017), was placed with good support (UFBoot = 92, LPP = 0.94) as sister to the "new-*viscaria*-clade".

Deep unresolved relationships. Both phylogenetic methods identified the base of the *E. abietina/E. viscaria* clade as having very short branches with correspondingly low support values (Fig. 3.5). Although rampant ILS can produce low LPP values (Sayyari and Mirarab, 2016), low bootstrap support values in concatenation-based analyses generally result from a lack of phylogenetic signal (Salichos and Rokas, 2013), which suggests that the results are best interpreted as indicating a basal polytomy, i.e., a multifurcation. Thus, although several broad clades could be confidently identified (see above), the present data set and analyses could not resolve the relationships between them.

Shallow unresolved relationships. According to the ASTRAL analysis, three clades stood out as having shallow unresolved relationships between taxa: the terminal grade of the "core-*viscaria*-clade"; the "RAV clade" comprising *E. regia*, *E. axilliflora*, and *E. vestita* from the Agulhas plains (i.e., the coastal region east of the Hottentots Holland mountains and south-west of the Langeberg mountains); and the *E. abietina* complex (Fig. 3.5). In the case of the latter, it is clear that the four subspecies of *E. abietina* are closely related and either do not represent genetically distinct entities or the relationships between the subspecies cannot be resolved by the target capture data. This complex is the subject of Chapter 4 and I will therefore not discuss it further here. The relatively limited sampling of the "RAV clade", combined with topological conflict between tree reconstruction methods (Fig. 3.5), makes it difficult to interrogate the lack of resolution of its internal nodes. Similarly, the topology of the terminal grade of the "core-*viscaria*-clade" was discordant between the concatenation and ASTRAL trees, and its internal branches were generally very short and were assigned low support values by ASTRAL and occasionally also by UFBoot (Fig. 3.5).

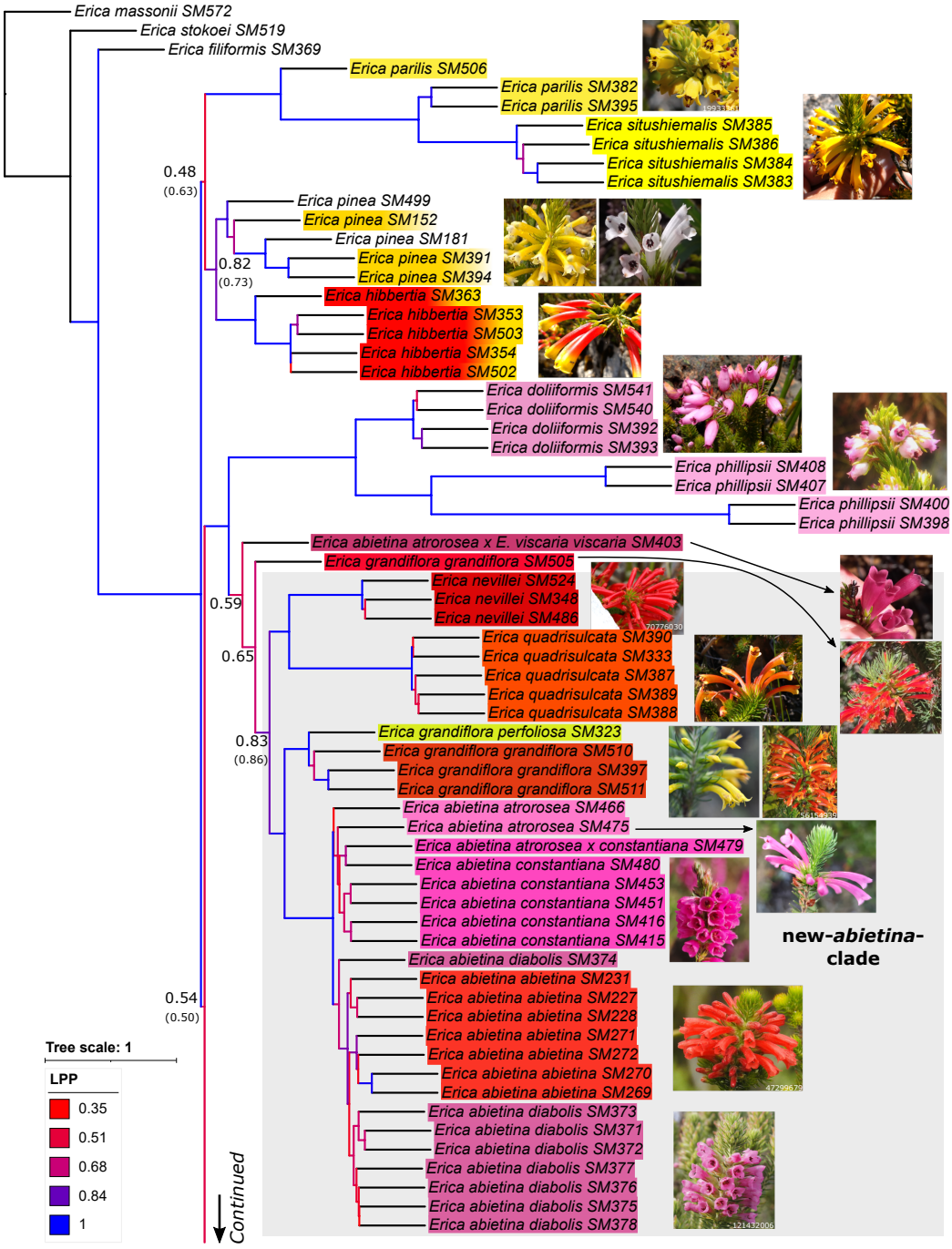


Fig. 3.8 Phylogeny of the *E. abietina/E. viscaria* clade inferred by wASTRAL-h. Non-terminal branch lengths are in coalescent units. Branch colour indicates local posterior probability support; support is indicated for branches above the species level with LPP < 0.9 (in brackets are LPP values recovered when hybrids were excluded prior to gene tree inference). Coloured boxes around tip labels illustrate typical flower colouration (no box = white flower). Images depict representative specimens (labels = unique iNaturalist ID; star = image from Oliver and Forshaw (2012).)

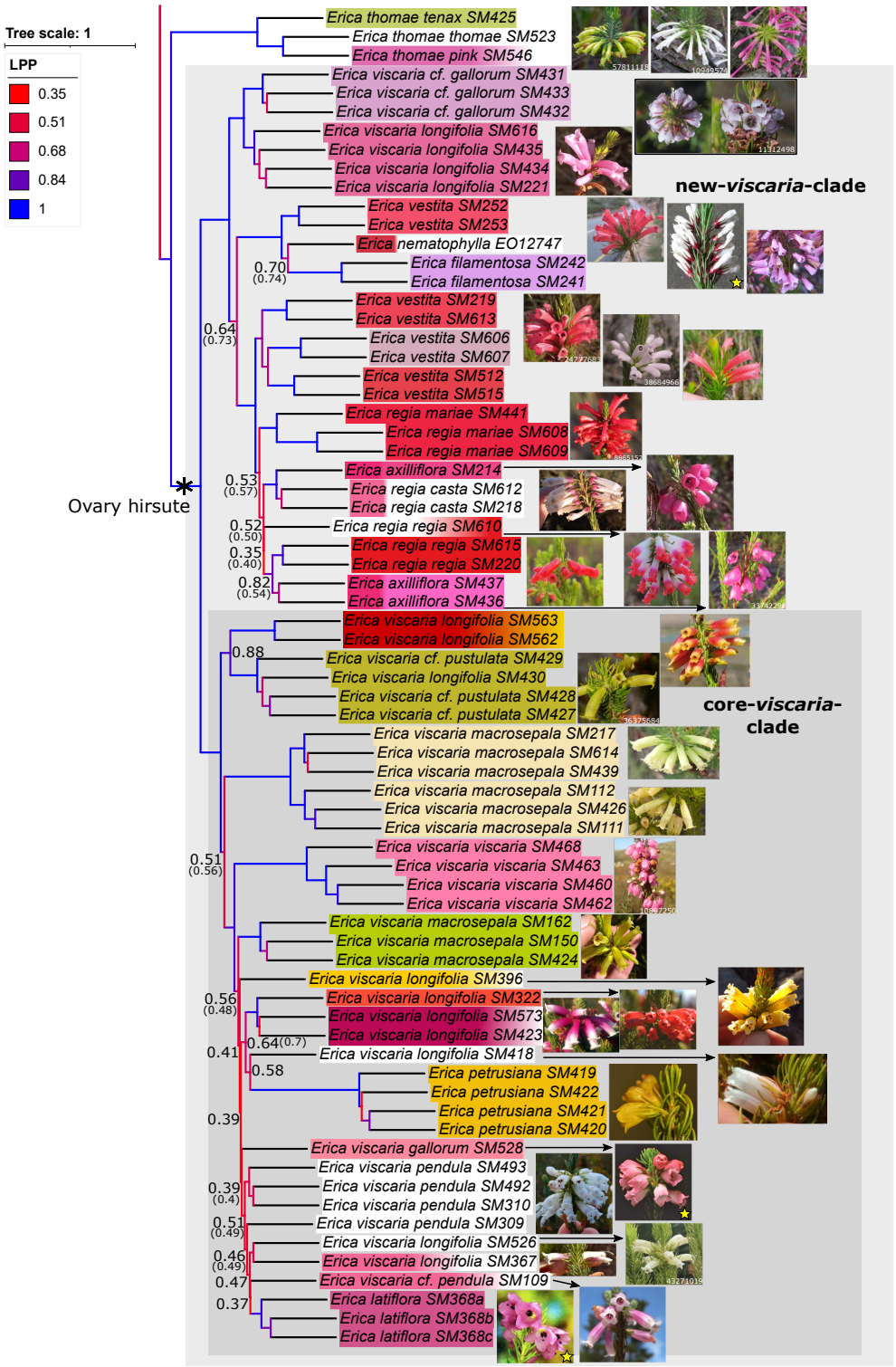


Fig. 3.8 (*Continued.*)

Paraphyly

Paraphyly is said to be present when phylogenetic tree inference suggests that a taxon is not monophyletic, thus conflicting with the hypothesis that the taxon is an independent evolutionary lineage. Several taxa in the *E. abietina/E. viscaria* clade were confidently resolved as paraphyletic, including at the species level. By far the most notable instance of paraphyly concerned *E. viscaria*, for which most samples were placed in a well-supported clade (which I term the "core-*viscaria*-clade"; Fig. 3.8) that also included *E. petrusiana* and *E. latiflora*, while seven samples were placed as sister to the rest of the samples in the "new-*viscaria*-clade". The latter seven samples all originated from the Agulhas plains region, and comprised two distinct morphotypes: one comprising three specimens that have certain features in common with *E. viscaria* subsp. *gallorum*, notably the short, pink corolla tube (see Section 3.2.1), and the other morphotype consisting of four specimens that all have long, entirely pink corollas. The latter specimens appear to belong to a morphotype that Oliver and Oliver (2002, p.56) placed in *E. viscaria* subsp. *longifolia*, describing this form as being confined to the Bredasdorp area (i.e., the Agulhas plains). These authors noted that the form was difficult to distinguish morphologically from specimens of *E. vestita* from the same region, and suggested that the two might hybridise.

Out of all of the subspecies of *E. viscaria* described by Oliver and Oliver (2002), only one – the nominate *E. viscaria* subsp. *viscaria* – was found to be monophyletic (Fig. 3.8). At the same time, however, the many short and poorly-resolved branches within the "core-*viscaria*-clade" meant that the only subspecies that was confidently found to be paraphyletic was *E. viscaria* subsp. *longifolia*. Nevertheless, paraphyly was considerable within this subspecies: apart from the aforementioned Bredasdorp form, the extremely distinct form with red-and-yellow flowers from Shaw's pass and the sample with greenish flowers from Hermanus were confidently recovered (concatenation: full support; ASTRAL: LPP = 0.88; Fig. 3.5) as belonging to the earliest branching clade within the "core-*viscaria*-clade" that also contained the forms resembling *E. viscaria* subsp. *pustulata* (also from Hermanus). *Erica viscaria* subsp. *macrosepala* showed somewhat weaker evidence of paraphyly, while the rest of the specimens identified as *E. viscaria* (several subspecies) were placed within a poorly-resolved terminal grade (see above 3.2.5).

Another instance of paraphyly concerned *E. vestita*. According to the ASTRAL tree, the two *E.*

vestita samples from the Langeberg mountains were confidently grouped with *E. filamentosa* and *E. nematophylla*, which also occur in or near the Langeberg range, while the six samples from the Agulhas plains formed a reasonably well supported clade (LPP = 0.74) that was closer to *E. axilliflora* and *E. regia* – taxa which are endemic to the Agulhas plains. On the other hand, the concatenation tree found *E. vestita* to form a fully-supported clade that also contained *E. filamentosa* and *E. nematophylla* (Fig. 3.5).

Yet another instance of paraphyly concerned *E. axilliflora*. According to the ASTRAL tree, one sample grouped with *E. regia* var. *casta* and the other two samples grouped with *E. regia* subsp. *regia* (Fig. 3.8). The concatenation tree also inferred paraphyly in *E. axilliflora*, but showed disagreement with the ASTRAL tree in the samples' precise placement (Fig. 3.5). The two samples that consistently grouped together were collected at the same locality (Murasie) while the other was collected at Carruthers Hill, *ca.* 10 km to the south-east. Phenotypically these samples appear similar, and although the plants at Carruthers Hill were noted as being relatively tall (*ca.* 100 cm) compared to those from Murasie (*ca.* 30-40 cm), this may have been due to differences in age or phenotypic plasticity (e.g., in response to local edaphic variation). These caveats, combined with the short branches and low branch support values within the *E. axilliflora/E. regia* clade, mean that such paraphyly could reflect either homoplasy (i.e., convergent evolution) or hemiplasy (i.e., phenotypic inheritance via ILS; Avise and Robinson, 2008).

Nestedness

Paraphyly and nestedness are, arguably, two ways of describing the same pattern in which taxa are non-monophyletic, and it can sometimes be difficult to decide how a topology is best characterised. For example, strictly speaking both tree reconstruction methods found – with good support (LPP = 1; UFBoot = 92) – that *E. axilliflora*, *E. regia*, *E. filamentosa*, *E. nematophylla*, and *E. vestita* were nested within *E. viscaria* (Fig. 3.5), but it is arguably more parsimonious to describe *E. viscaria* as paraphyletic (see below, Section 3.2.5). As another example, according to the concatenation tree *E. filamentosa* and *E. nematophylla* were nested within an otherwise monophyletic *E. vestita* (full support; Fig. 3.5), though ASTRAL could only place these taxa with certainty within a broader clade that also contained *E. axilliflora* and *E. regia* (Fig. 3.8). In this case, the uncertainty and conflict

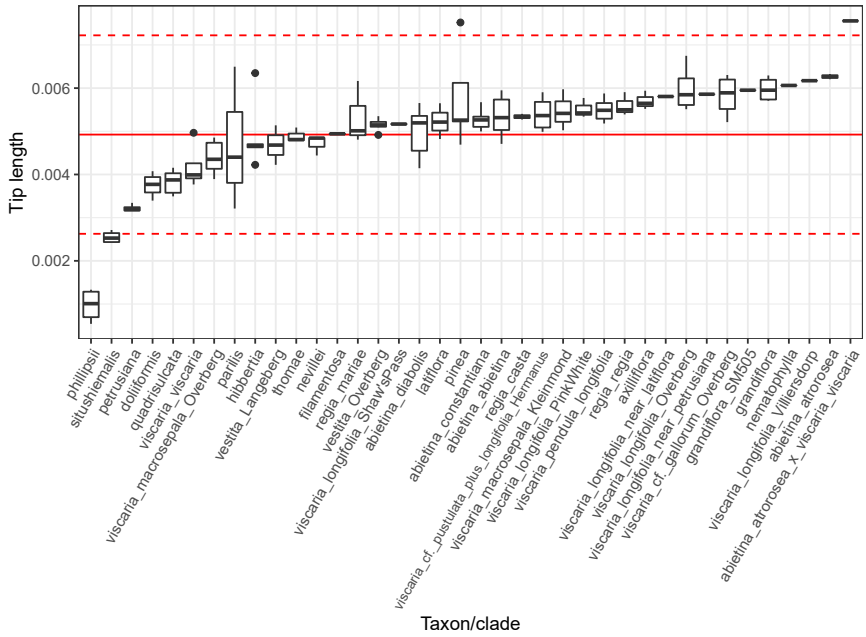


Fig. 3.9 Box plots showing the distribution of terminal branch lengths (in units of expected substitutions per site), taken from the concatenation tree displayed in Fig. 3.5, within clades supported by both concatenation and ASTRAL trees. The solid red line indicates the global mean, and the dashed lines the 95% confidence intervals assuming a normal distribution. Note the long terminal branch of the sample identified as a putative F1 hybrid between *E. abietina* subsp. *atrorosea* and *E. viscaria* subsp. *viscaria*; and the short terminal branches of the samples of *E. petrusiana*, *E. situshiemalis*, and especially *E. phillipsii*.

between methods makes the distinction between paraphyly and nestedness unclear. Lastly, it could either be said that *E. parilis* is paraphyletic, or that *E. situshiemalis* is nested within *E. parilis* (Fig. 3.8). In this case, the limited sampling of *E. parilis*, which is relatively widespread (Oliver and Oliver, 2002), makes the distinction uncertain.

In some instances, however, nestedness was clear-cut. One instance concerned *E. petrusiana*, which both methods found to be nested well within the "core-*viscaria*-clade", although ASTRAL showed uncertainty in its exact placement. Similarly, *E. latiflora* was also deeply nested within the "core-*viscaria*-clade". *Erica viscaria* subsp. *viscaria* was the only subspecies of *E. viscaria* that was clearly monophyletic, which implies that it too was nested within the "core-*viscaria*-clade". An interesting feature shared by *E. petrusiana* and *E. viscaria* subsp. *viscaria* was that both sets of samples were subtended by unusually long branches, both in terms of coalescent units (ASTRAL) and expected substitutions per site (concatenation; Fig. 3.5). Overall, these three taxa emerge as distinct units from what is effectively a polytomy.

Terminal branch lengths

In the concatenation tree (Fig. 3.5) the lengths of the terminal branches (i.e., "tip lengths") represent an estimate of the amount of sequence variation that is unique to each sample. Tip lengths were relatively uniform across most taxa and clades, but there were several notable outliers (Fig. 3.9). There were two samples with exceptionally long branches: one belonged to the putative hybrid SM403 (see Section 3.2.4) while the other belonged to the only specimen of *E. pinea* that originated from the highly localised and geographically isolated population of this species from near the town of Kleinmond (SM152). *Erica phillipsii* stood out as having extremely short terminal branches, while at the same time also having long branches subtending each pair of samples (each of which comprised samples from a single locality; Fig. 3.5). This same pattern was present, though less strongly, in *E. situshiemalis* and *E. petrusiana* (Fig. 3.5). These three species all have relatively small geographic distributions (Esterhuysen, 1963; Oliver and Oliver, 2002; Pirie et al., 2017), although so do several other species with typical tip lengths (e.g., *E. filamentosa*, *E. latiflora*; Oliver and Oliver, 2002).

Phenotypic variation

The most obvious aspect of phenotypic variation across the phylogeny was that there was no apparent phylogenetic signal in the colour and shape of the corolla, suggesting that these traits are highly labile. On the other hand, characteristics of the ovary appear to have phylogenetic significance: species in the "new-*viscaria*-clade" all have hirsute ovaries, whereas the rest of the species in the *E. abietina/E. viscaria* clade have glabrous or, occasionally, sparsely pubescent ovaries (Oliver and Forshaw, 2012; Oliver and Oliver, 2002, Fig. 3.8). Notably, within the "new-*viscaria*-clade" the distribution and orientation of the hairs is variable and, furthermore, may be phylogenetically structured (Fig. 3.10). Of particular interest is that in *E. regia*, *E. vestita*, and their close relatives, the hairs are exclusive to the upper portion of the ovary (see detailed drawings by I. M. Oliver in Oliver and Forshaw, 2012), whereas Oliver and Oliver (2002, p. 56) described *E. viscaria* as a whole as having "ovary...covered with erect dense fairly long, white hairs." The centre image in Figure 3.10 shows the ovary of a plant belonging to the Bredasdorp form of *E. viscaria* subsp. *longifolia* (Fig. 3.8). This specimen (along with the other similar specimens from the Agulhas plains) was identified as this subspecies based on

Fig. 3.10 Images taken under a dissecting microscope of flowers of three specimens with sequence data, dissected to show the ovary. From left to right: *E. viscaria* subsp. *pendula* ("core-*viscaria*-clade") collected from the Hottentots Holland mountains; *E. viscaria* subsp. *longifolia* (not "core-*viscaria*-clade") collected from the Bredasdorp district; *E. vestita* also collected from the Bredasdorp district. Note the differences in ovary shape and the distribution and orientation of the hairs. Magnification ≈ 20x; images not to scale.

its long (*ca.* 20 mm) corolla bearing longitudinal ridges and sparse, short bristle-like hairs (Oliver and Oliver, 2002); however, its ovary is not entirely hairy – instead, the sides of the ovary are nearly glabrous while the top has long, dense hairs much like *E. vestita* and *E. regia*.

Geographic structure

In order to investigate whether phylogenetic relationships were correlated with geography, I projected the IQ-TREE topology of the "core-*viscaria*-clade", along with the seven samples identified as *E. viscaria* that fall outside this clade, onto geographic space using the PHYTOOLS *phylo.to.map* function. This revealed a high degree of phylogeographic signal (Fig. 3.11). Furthermore, geography generally corresponded closely to phylogenetic relatedness, while taxonomy and, by extension, morphology did not (see Fig. 3.8). For example, the closest relative of SM528 (*E. viscaria* subsp. *gallorum*; short pink flowers) was its geographic neighbour SM526, a sample identified as *E. viscaria* subsp. *longifolia* with long white flowers. *Erica petrusiana* (short yellow flowers) showed a similar pattern: its closest geographic neighbour SM418 (*E. viscaria* subsp. *longifolia*, long pink-white flowers) was also its closest relative. Exactly the same pattern applied to *E. latiflora* (short pink flowers). The sister relationship between the specimens from near Hermanus (SM427-430; long greenish flowers with or without pustules) and those from Shaw's Pass (SM562-563; long red-yellow flowers) also

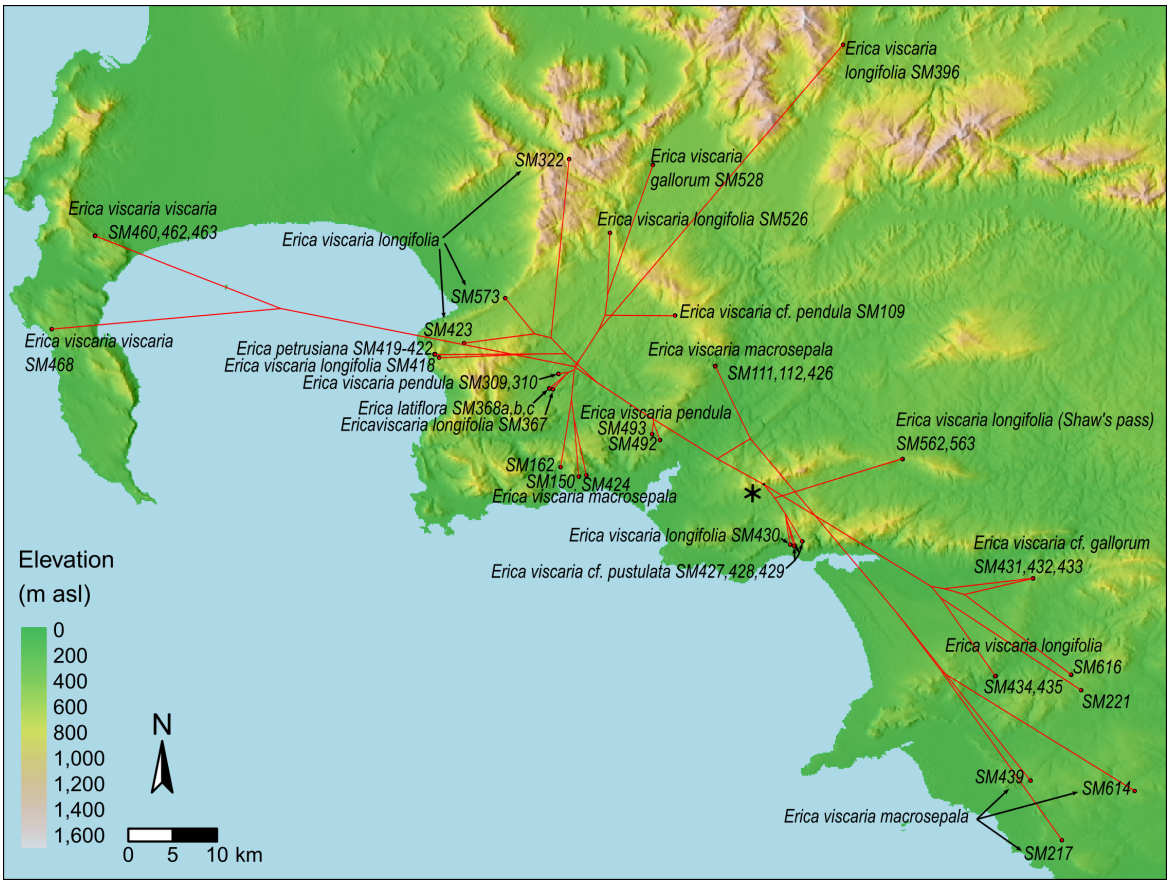


Fig. 3.11 The subtree of the concatenation-based phylogeny including all samples belonging to the "core-*viscaria*-clade" plus the seven samples identified as *E. viscaria* that fall outside this clade, projected onto a topographic map of the south-western CFR. An asterisk is shown at the "root" node which separates the two clades. Note that branch lengths and node coordinates are not meaningful.

corresponded with geographic proximity.

In some instances, however, close relatives were geographically distant from each other. The six specimens of *E. viscaria* subsp. *macrosepala* that formed a well-supported clade comprised two sample sets, each forming its own clade, that were separated by *ca.* 60 km and several mountain ranges. *Erica viscaria* subsp. *viscaria* (which is endemic to the Cape Peninsula) was also not most closely related to its closest geographic neighbours, although it should be noted that its position within the clade was highly uncertain (the two branches directly subtending it had UFBoot = 61 and 62, respectively; Fig. 3.5).

3.3 Taxonomic implications

Several taxa whose monophyly could not be confirmed by Pirie et al. (2017) were here confidently found to be monophyletic. These included *E. hibbertia* and *E. pinea*, as well as *E. grandiflora* (Figs. 3.5, 3.8). The latter case was especially interesting. Notwithstanding the possible admixed sample SM505, *E. grandiflora* was monophyletic and placed confidently as sister to *E. abietina*. Pirie et al. (2017) resolved the paraphyly of *E. abietina sensu* Oliver and Oliver (2002) by resurrecting *E. grandiflora* and describing two subspecies, *E. grandiflora* subsp. *grandiflora* and *E. grandiflora* subsp. *perfoliosa* (see Section 3.1.3). Although Pirie et al. (2017) could not confirm the monophyly of *E. grandiflora* itself, they expected that its resurrection would be robust to phylogenomic analyses because their analysis suggested a closer relationship to *E. viscaria* than to *E. abietina*, within which it was lumped by Oliver and Oliver (2002). Instead, the phylogenomic analyses presented here suggest that the robustness of the taxonomic change made by Pirie et al. (2017) comes from the genetic and morphological distinctness of *E. grandiflora*, and support the opinion of Oliver and Oliver (2002) of a close relationship between *E. grandiflora* and *E. abietina* (Fig. 3.8). Results such as this highlight the power of phylogenomic analysis to overcome the limitations of traditional molecular phylogenetics and resolve taxonomic uncertainties.

On the other hand, several taxa were found to be paraphyletic. The most complex example of this concerned *E. viscaria*. Because of the morphological distinctiveness of the samples that resemble *E. viscaria* but were found to be more closely related to the "RAV clade" than the "core-*viscaria*-clade" (see Section 3.2.5 and Figs. 3.5, 3.10), rather than casting doubt on the reliability of the phylogenetic results or claiming that the "RAV clade" is nested within *E. viscaria*, it is arguably most appropriate to characterise *E. viscaria* as paraphyletic. This would imply that *E. viscaria* comprises at least two species (in the sense of being separately evolving lineages; *sensu* de Queiroz, 2007). Reconciling these phylogenetic results with the taxonomy of the "core-*viscaria*-clade" is equally difficult, firstly because of the general lack of phylogenetic resolution, secondly because of the non-monophyly of the subspecies of *E. viscaria* within it, and thirdly because two taxa recognised at the species level are nested within it (*E. petrusiana* and *E. latiflora*, Fig. 3.5). Similar though less extreme examples of paraphyly/nestedness came from within the "RAV clade", which along with the "core-*viscaria*-clade"

formed the "new-*viscaria*-clade". While it may be that even more data are required to resolve these uncertainties, it could also be that the standard phylogenetic model in which species arise by sequential bifurcations is simply not appropriate to describe the evolutionary dynamics within the clade (Crouch et al., 2021, see Chapter 5).

The general lack of support for the monophyly of many of the subspecies of *E. viscaria* (Fig. 3.5) suggests that the phenotypic characters which Oliver and Oliver (2002) considered to be useful in distinguishing taxa are not good indicators of common ancestry. The authors themselves in some sense acknowledged that *E. viscaria* subsp. *longifolia* was effectively a "bin" to which they assigned a large variety of forms. In some cases, the lack of correspondence between phylogeny and phenotype is, with hindsight, not unexpected. For example, *E. viscaria* subsp. *macrosepala* was defined based on its broad, robust sepals. In Cape *Erica*, nectar robbing by sunbirds and large carpenter bees (genus *Xylocopa* Latreille) in which the base of the flower is pierced in order to access its nectar, is common (Rebelo et al., 1985, pers. obs.). Broad, robust sepals may have evolved in many species as a means of counteracting such theft, which not only robs the plant of costly nectar but also does not result in pollination. Considering this, convergent evolution of this trait seems a strong possibility and could explain the apparent paraphyly of this subspecies. Of course, the same pattern can emerge via hemiplasy, in which deep coalescence (i.e., ILS) of the gene(s) underlying a trait gives the appearance of trait convergence (i.e., homoplasy; Avise and Robinson, 2008). Given that ILS is evidently common throughout the clade (Fig. 3.8), hemiplasy could underlie this and many of the of the clade's other apparent trait convergences (e.g., *E. axilliflora*). Regardless of their evolutionary origin traits do not, however, define lineages, and the present results indicate a concerning lack of correspondence between monophyly and taxonomic circumscription. This should encourage a critical re-examination of the criteria used to delimit species in this clade, and in *Erica* as a whole.

3.4 Hybridisation and introgression

Various lines of evidence suggest that interspecific gene flow may have played an important role in the history of the *E. abietina/E. viscaria* clade. While ancient introgression is nowadays frequently inferred in phylogenomic studies (e.g., Burbrink and Gehara, 2018; Chan et al., 2020; Lin et al., 2019),

its genomic signature is notoriously difficult to distinguish from ILS especially when diversification occurred recently and rapidly (Folk et al., 2018; Knowles et al., 2018; Li et al., 2019). In contrast, present-day hybridisation is much easier to infer confidently because its genomic signatures are relatively unambiguous (Dolinay et al., 2021) and admixed individuals can be investigated directly.

The first instance of evidence for present-day hybridisation was expected, despite the relatively distant relationship between the putative parents: the sample SM403 showed phenotypic characteristics intermediate between its two putative parents (*E. abietina* subsp. *atrorosea* and *E. viscaria* subsp. *viscaria*) which suggested it was a first-generation hybrid (Fig. 3.3; McDade, 1990), and both phylogenetic analyses supported this prediction (Fig. 3.5). The second instance was, however, unexpected: SM505, which phenotypically resembles *E. grandiflora* subsp. *grandiflora*, showed anomalous phylogenetic placement in the coalescent-based analysis but not in the concatenation-based analysis (Fig. 3.5). I interpret these results as indicating that SM505 is a late-generation hybrid between *E. grandiflora* subsp. *grandiflora* and *E. pinea* whose ancestry primarily derives from the former, and base this interpretation on (1) the expected difference in the sensitivity of the two phylogenetic methods to admixture, (2) local co-occurrence of the two species, and (3) evidence of individuals with intermediate phenotypes found at the same locality (see Section 3.2.4 and Fig. 3.7). Based on this case, I tentatively hypothesise that back-crossing in hybrids of these two species is biased towards *E. grandiflora* subsp. *grandiflora*. At least a dozen naturally-occurring first-generation hybrids in Cape *Erica* have been reported based on phenotypic characteristics (Adamson and Salter, 1950; Oliver, 1977, 1986; Oliver and Oliver, 2005) and many more have been artificially produced in cultivation (Nelson and Oliver, 2004; Oliver and Oliver, 2002, 2005). However, to my knowledge the cases of SM403 and SM505 are the first in which molecular evidence has indicated interspecific hybridisation in the wild in Cape *Erica*.

Arguably, evidence of present-day hybridisation also suggests that introgression has occurred in the past, especially – as in the present study – if it is shown to occur between distant relatives and if late-generation hybrids are detected. At least in plants, recurrent back-crossing in which hybrids are biased towards reproducing with one of the parent species is thought to be the primary means by which the genome of one species is first infiltrated by that of another, allowing for a point to eventually be reached when portions of the genome of the second species have become fixed in the genome of the

first (Baack and Rieseberg, 2007; Rieseberg and Wendel, 1993). In other words, it is one of the first steps in the process of introgression. This makes the singular case of SM505 particularly important, in that it highlights a need for further efforts to investigate the frequency, nature and consequences of hybridisation in Cape *Erica* in order to evaluate its role in the (ongoing) evolution of the clade.

The example of SM505 indicates that the details of discordance between coalescent- and concatenation-based phylogenetic analyses can reveal evidence regarding introgression. In the case of *E. vestita*, although both coalescent- and concatenation-based analyses suggested that the species was paraphyletic, they disagreed on the exact nature of the paraphyly (Fig. 3.5). While ASTRAL placed the specimens from the Agulhas plains with *E. regia* and *E. axilliflora*, concatenation placed them as sister to the specimens from the Langeberg (alongside *E. filamentosa* and *E. nematophylla*). Interestingly, Pirie et al. (2017) also uncovered some evidence of paraphyly in *E. vestita*: while most of their samples of this species grouped with *E. filamentosa* and *E. nematophylla*, one grouped with *E. axilliflora* and *E. regia*. Unfortunately, the latter sample ("vestita_ANA; SANBI,176/05") was prepared by Mugrabi De Kuppler (2013) from a plant in cultivation and its provenance was not reported. Nevertheless, an interesting aspect of the Pirie et al. (2017) data was that support for the grouping of "vestita_ANA" with *E. axilliflora* and *E. regia* came mainly from chloroplast markers, while support for the placement of the rest of their samples with *E. filamentosa* and *E. nematophylla* came mainly from nuclear markers. This may hint at cytonuclear discordance, which is relatively common in angiosperms (e.g., Nge et al., 2021; Soltis and Kuzoff, 1995), has been inferred in European *Erica* (Mugrabi De Kuppler et al., 2015), and is thought to result from ancient hybridisation followed by chloroplast capture (Soltis and Kuzoff, 1995). Unfortunately, the present data set did not allow for the assembly of sufficient chloroplast sequence data to test this hypothesis directly, as mapping rates to the chloroplast genome of *E. versicolor* (GenBank accession MW282955.1) were generally very low: after mapping all reads to the genome (excluding one of the inverted repeat regions; total 139,229 bp) using *NextGenMap* v.0.5.5 (Sedlazeck et al., 2013, parameters: *-Q 13 –affine*), the median depth per sample ranged from 0X to 306X but the median across all samples was only 2X. Nevertheless, if "vestita_ANA" did indeed originate from the Agulhas region, this result might have been a sign, supporting speculation by Oliver and Oliver (2002), of gene flow between *E. vestita* and one (or more) of the other taxa with which it co-occurs on the Agulhas plains. This would also lend credibility to an

interpretation of the discordant signals of paraphyly recovered by the two phylogenetic methods as evidence of historical gene flow between these species.

3.5 Evidence for an early burst of speciation

One of the most striking features of the *E. abietina/E. viscaria* clade phylogeny is the evidence of a basal polytomy subtending the major sub-clades (Fig. 3.8). A similar feature has plagued avian phylogeneticists for decades: the base of the Neoaves clade has frequently been left as a large, unresolved polytomy, and intense efforts to resolve it, including various large phylogenomic data sets and analysis methods, have met mixed results (Jarvis et al., 2014; Kuhl et al., 2020; Prum et al., 2015; Reddy et al., 2017). However, most avian systematists agree that this uncertainty is attributable to a burst of speciation early in the evolutionary history of modern birds (Berv et al., 2022; Brusatte et al., 2014). One interpretation, therefore, of the basal polytomy in the *E. abietina/E. viscaria* clade is that it indicates an early burst of speciation. Elevated rates of diversification are apparent in the Cape *Erica* clade as a whole, with significant upward shifts appearing to be associated with the arrival of the genus in the CFR some 6-15 Ma, and another occurring some time later (Pirie et al., 2016). The origin of the *E. abietina/E. viscaria* clade has been dated to just 2-3 Ma, implying that its origin might have coincided with the second shift. While an analysis of the rate and timing of diversification of the clade is beyond the scope of the present work, I expect that the data generated here could be useful in that regard, especially now that molecular dating methods are becoming feasible for larger numbers of loci (Douglas et al., 2022).

3.6 Conclusions

The aim of this chapter was to apply a phylogenomic approach to the challenging phylogenetic problem posed by the *E. abietina/E. viscaria* clade, with the goals of resolving previous uncertainties and shedding light on the evolutionary dynamics at play. Various uncertainties, such as the phylogenetic affinities of certain enigmatic taxa, have indeed been confidently resolved, but many other uncertainties remain. Firstly, relationships within the "core-*viscaria*-clade" could not be confidently resolved. In this case, target capture may not be the most suitable approach, as low levels of sequence divergence

suggest that population genomic methods such as genotyping-by-sequencing (Elshire et al., 2011), combined with more thorough sampling, might provide better resolution (McCormack et al., 2013). Another persistent uncertainty concerns the deeper relationships within the clade, which could not be confidently resolved, perhaps indicating a hard polytomy. However, while these uncertainties leave us without clear answers to questions of taxonomy and phylogeny, they reveal much about evolutionary dynamics, signifying both recent and ancient bursts of speciation and highlighting *E. viscaria* as an especially interesting and potentially fruitful study system. Finally, there are indications that a reticulate model may best explain the evolutionary history of the clade, warranting future efforts to disentangle the signatures of introgression and incomplete lineage sorting.

Chapter 4

Recent and ongoing diversification in the *Erica abietina* species complex

4.1 Background

There is increasing recognition that to understand what drives and limits diversification we need to investigate systems in which species limits are uncertain and factors such as hybridisation, introgression, genetic drift and selection interactively influence the speciation process (Donoghue and Sanderson, 2015; Sobel et al., 2010; Via, 2009). Due to the advent of next-generation sequencing, researchers are now able to investigate the genomics of diversification in great detail (McCormack et al., 2013). These advances have shown that "textbook" cases of speciation are relatively rare, with the small aperture provided by previous genetic techniques capturing only a tiny fraction of what is visible through the genomic lens. For example, well-established species are often shown to exhibit extremely low mean genome-wide divergence, but very high divergence at a few large-effect loci that resist gene flow due to divergent selection and/or genomic location (e.g., Kautt et al., 2020; Mořkovský et al., 2018; Porter et al., 2021; Puntambekar et al., 2020). On the other hand, phenotypically indistinguishable populations are often found to be deeply divergent (e.g., Blair et al., 2019), highlighting the prevalence of cryptic diversity across the tree of life (Balkenhol et al., 2009). Introgression, far from being the traditionally-viewed homogenising force (Templeton, 1981), appears rampant in rapidly diversifying groups (Nosil, 2008) and may even accelerate evolutionary change by spreading advantageous alleles

(e.g., in Lake Malawi cichlid fishes; Svardal et al., 2019) or whole chromosome segments (e.g., in *Heliconius* butterflies; Jay et al., 2018) to generate new trait combinations and drive speciation across a range of phylogenetic scales (e.g., Bougie et al., 2021; Schley et al., 2020).

Erica abietina is a species complex with several phenotypically distinct forms that vary in flower length, colour, shape and scent as well as growth form and distribution, and which have been classified into four subspecies based largely on this variation (Oliver and Oliver, 2002; Pirie et al., 2017, Table 4.1, Figs. 4.1, 4.2). The subspecies of *E. abietina* are clearly extremely closely related despite their phenotypic and geographic range differences (Pirie et al., 2017, see also Chapter 3, Fig. 3.8), suggesting that much of their diversity has emerged in the recent past and/or that they are not strongly reproductively isolated. The complex is confined to the Cape Peninsula, a hotspot of floristic diversity even in the context of the CFR as a whole (Cowling et al., 1996; Simmons and Cowling, 1996), with over 2200 plant species (Trinder-Smith et al., 1996) of which 158 (including $\geq$ 39 *Erica* species) are endemic (Helme and Trinder-Smith, 2006). The Cape Peninsula is a *ca.* 50 km x 15 km mountain range of rugged topography at the south-western tip of South Africa that is largely surrounded by ocean and isolated from the rest of the Cape Fold Mountains by a large (> 40 km wide) sandy plain whose low elevation placed it below sea level during Pleistocene interglacials (Adamson, 1959). All of this suggests that *E. abietina* might be experiencing diversifying forces in a "continental island" system (Hughes and Eastwood, 2006).

Genomic methods allow for detailed investigation into the relationships between closely related lineages, and among the most popular are a family of methods that reduce the complexity of a genomic sample prior to sequencing using restriction enzymes (reviewed in Puritz et al., 2014b). A huge diversity of these enzymes evolved in bacteria as a means of fighting (i.e., "restricting") viruses, and they operate by recognising a short nucleotide sequence and cleaving the virus's DNA strand at that "cut site" (Felice et al., 2019). By adding one or two of these enzymes to a sample of isolated genomic DNA, the long DNA strands are "digested" in a predictable manner to produce a "library" of short fragments all associated with the enzyme cut site that can be isolated from the rest of the genome and, once sequenced, relatively easily assembled and aligned (Puritz et al., 2014a; Rochette et al., 2019). Several features of the restriction-enzyme associated digest (RAD) family of methods make it well-suited to the problems that the *E. abietina* complex presents (Puritz et al., 2014b). Firstly, it

Table 4.1 Characteristics and geographic ranges of the subspecies of *Erica abietina*.

Subspecies	Corolla length (mm)	Corolla colour	Flower scent	Sepals	Ovary	Range
E. a. abietina	18–26	Crimson to dark red	None	Lanceolate-attenuate/ acuminate; pilose	Elongate obovoid; pubescent	Northern CP: TM high plateau, N and W slopes
E. a. diabolis	11–14	Rose pink	None	Lanceolate-attenuate/ acuminate; pilose	Obovoid; pubescent	Northern CP: Devil's Peak
E. a. constantiana	8–11	Pale to deep rose pink	Sweet, lemony	Lanceolate-attenuate/ acuminate; glabrous to sparsely puberulous	Squat obovoid; puberulent	Central CP: S slopes of TM to Chapman's Peak
E. a. atrorosea	18–22	Rose to deep rose pink	None	Lanceolate-acute; glabrous to sparsely puberulous	Ellipsoid; glabrous	Widespread on CP: E slopes of TM S to Cape Point

CP: Cape Peninsula; TM: Table Mountain.

provides a relatively unbiased sample of the genome, especially if a restriction enzyme with a short, abundant cut site is used (e.g., Elshire et al., 2011). Secondly, depending on the library preparation method and the genome itself, it can provide a large amount of data suitable for detecting genetic differentiation and diversity at the level of populations and even individuals (Szarmach et al., 2021). Thirdly, it is highly cost-effective because the reduced library complexity allows for relatively little sequencing effort to sufficiently sequence many samples simultaneously (Sonah et al., 2013).

In this chapter I take a RAD-type sequencing approach (genotyping-by-sequencing, GBS; Elshire et al., 2011) to addressing the following questions regarding the evolution of the *E. abietina* species complex:

- To what extent does the current taxonomy reflect genetic patterns?
- How prevalent is gene flow between the taxa, and what role has it played in the group's evolution?
- What is the history of floral trait evolution, and what role have floral trait shifts played in the group's evolution?

4.2 Methods and results

4.2.1 Sample collection and sequencing

I collected fresh leaf material from at least six individuals of each formally recognised taxon in addition to several individuals of putative hybrid origin (Tables 4.1,B.2) and attempted to sample from

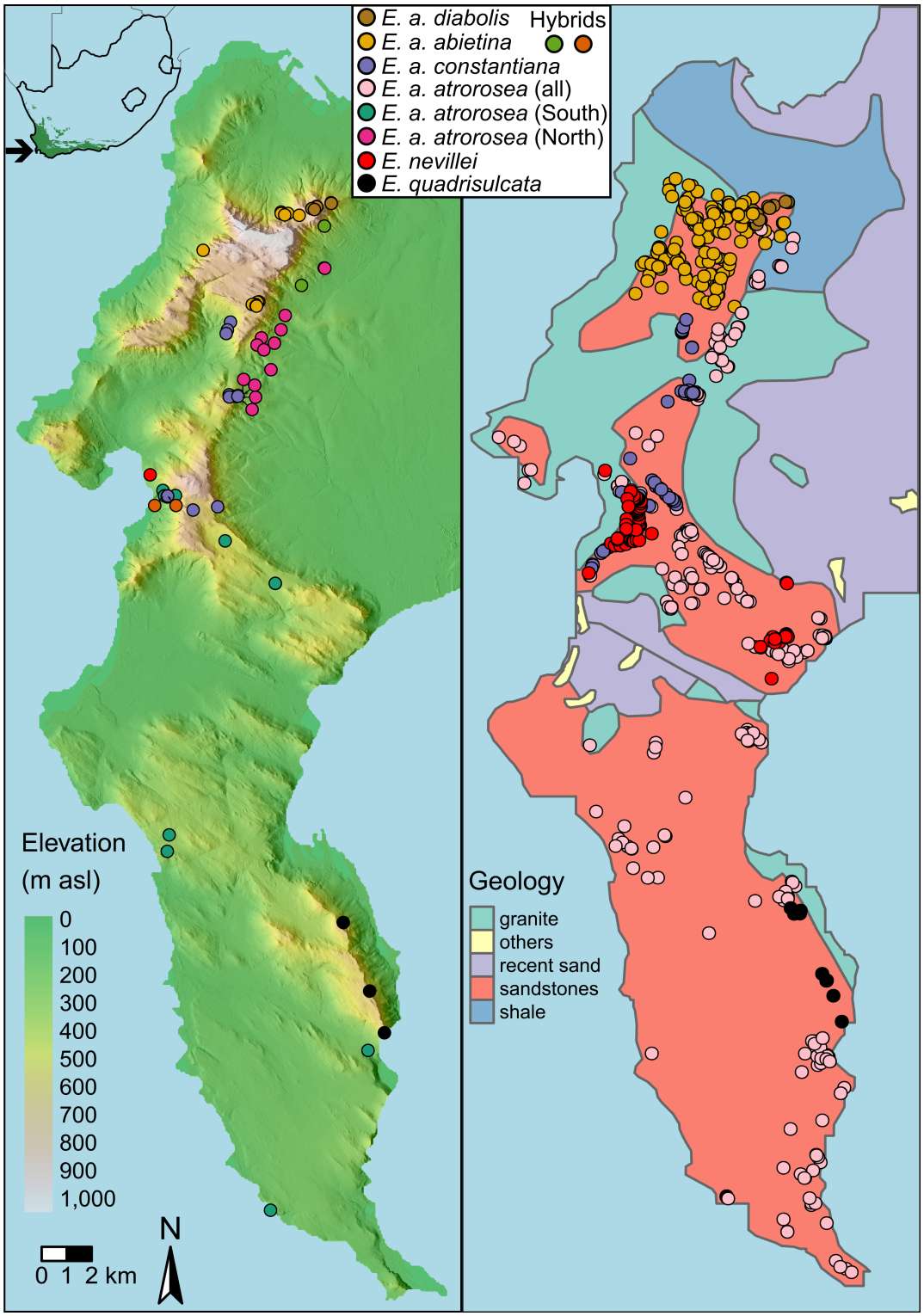


Fig. 4.1 Maps of sampling and subspecies' ranges. *Left:* Elevation and hillshade with localities of collected samples. *Right:* Underlying geology with localities of research grade observations from `iNaturalist.org` (Date accessed: 31.08.2021). *Inset:* Map of South Africa showing the extent of the Fynbos biome and the location of the Cape Peninsula.

Fig. 4.2 Photographs of the four subspecies of *Erica abietina*. From *left* to *right*: *E. a. abietina*, *E. a. diabolis*, *E. a. constantiana*, *E. a. atrorosea*. Text inset indicates `iNaturalist.org` observation ID.

across the full geographic range of each taxon (Fig. 4.1). I followed the taxon concepts (Meier, 2017) outlined by (Oliver and Oliver, 2002) and refined by Pirie et al. (2017) when identifying specimens. In addition, I sampled from the two putatively closest outgroups of *E. abietina*, *E. nevillei* (n = 1) and *E. quadrisulcata* (n = 3), both of which are endemic to the Cape Peninsula. For convenience I refer to subspecies in abbreviated form (e.g., *E. a. abietina* refers to *E. abietina* subsp. *abietina*).

DNA extraction followed the protocol in Appendix A. Library preparation and sequencing was done by Novogene Genome Sequencing Company Ltd. (Beijing, China), following the original Elshire et al. (2011) GBS protocol but with MseI as the restriction enzyme (cut site T/TAA). Libraries were paired-end sequenced in two separate batches to 144 bp (after barcode removal) using an Illumina NovaSeq 6000 instrument. The two batches were sequenced to different depths, which was done partly to estimate the effect of sequencing effort on genotyping quality. Using a lower sequencing effort is more cost effective but could compromise data quality by reducing the accuracy of genotype calls and increasing the rate of missing data. The first batch consisted of 12 samples sequenced to an estimated 120 Mb each, while the second batch consisted of 54 samples sequenced to an estimated 240 Mb each (i.e. batch 2 had much greater sequencing effort). One sample was included in both batches to test the effect of sequencing effort on genotype calling accuracy, which was found to be negligible (see below). This was done by calling SNPs for each sample separately using FREEBAYES (see below) and investigating the rate of discordance in genotype calls.

To evaluate whether the libraries were sequenced sufficiently to cover most of their complexity, I estimated read redundancy for each sample using *bbcountunique.sh* from BBTOOLS v. 38.90 (BBMap - Bushnell B. - `sourceforge.net/projects/bbmap/`) with default parameter values. In

this analysis read pairs are inspected in random chunks of 25 000, and for each chunk the proportion of unique read pairs (including all previous chunks) is estimated. A well-sequenced library should show a rapid decline in the proportion of unique reads, tending towards zero as more reads are inspected (i.e., high redundancy).

4.2.2 Data processing and variant calling

Raw reads were quality-checked with FASTQC (`http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc`) and MULTIQC (Ewels et al., 2016). Adapter removal, quality trimming and read filtering was done with FASTP (Chen et al., 2018, parameters: *–overrepresentation_analysis –trim_poly_g –qualified_quality_phred 20 –unqualified_percent_limit 30 –average_qual 20 –length_required 100*). The remaining reads were aligned to the draft reference genome of *Erica cerinthoides* (excluding contigs < 400 bp; see Chapter 2) using BWA-MEM (Li, 2013) with default parameters, followed by alignment sorting and indexing using SAMTOOLS (Danecek et al., 2021). Alignments were merged using BAMTOOLS (Barnett et al., 2011).

Variants were called for all samples simultaneously using FREEBAYES v. 1.3.4 (Garrison and Marth, 2012, parameters: *–min-base-quality 3 –min-mapping-quality 10 –skip-coverage 10000 –use-best-n-alleles 4*). Genotypes with read depth < 3X were recoded as missing using VCFTOOLS (Danecek et al., 2011). BCFTOOLS (Danecek et al., 2021) was used to for further variant filtering (Table 4.2). Next, I used *vcfallelicprimitives* from VCFLIB (Garrison, 2012) to decompose complex variants into SNPs where possible, and then removed non-SNP variants and non-biallelic SNPs to obtain a dataset of purely biallelic SNPs. I used the BCFTOOLS *fill-tags* plugin to test for excess heterozygosity. I applied further filters as appropriate and used a naming scheme based on these filters (see Table 4.2). For the analyses that assume loci are unlinked, I filtered out SNPs potentially in linkage disequilibrium (LD) using PLINK (Purcell et al., 2007, parameters: *–indep-pairwise 50 5 0.2*). Table 4.2 outlines the variant calling results and the effects of quality filtering.

For estimating sequence diversity and divergence as well as for individual-level phylogeny inference, I generated an "all sites" VCF file by supplementing the SNP_m10 set (SNPs-only, missingness < 10%; Table 4.2) with invariant sites. This was done by re-running FREEBAYES as above but adding the "*–report-monomorphic*" switch, followed by removing sites with missingness >

Table 4.2 Table summarising the variant filtering applied and resulting features of the genotype matrices.

Set name	Filtering criteria (sites kept)	Total sites	Contigs	SNPs	Invariant sites	Genotyping rate
Variant sites only						
VAR_RAW	All variant types; Read mapping quality $\geq$ 10.	7 800 988	186 414	6 373 677	-	0.43
SNP_HQ	Decompose complex variants; Biallelic SNPs only; GT DP $\geq$ 3; Missingness $\leq$ 50%; MAF $\geq$ 0.01; AB = 0 OR 0.25 < AB < 0.75; 0.9 < MQM/MQMR < 1.1.	771 049	45 804	771 049	-	0.81
SNP_HQ_ExcessHet	Excess heterozygosity: p-value > 0.2.	738 178	45 609	738 178	-	0.81
SNP_m10	Missingness $\leq$ 10%.	292 849	20 697	292 849	-	0.96
SNP_m10_LD	Linkage disequilibrium.	179 927	20 695	179 927	-	0.96
SNP_m10_LD_maf04	MAF $\geq$ 0.04.	94 883	18 969	94 883	-	0.96
Variant plus invariant sites						
ALL_m10	Missingness $\leq$ 10%.	4 993 526	22 920	292 849	4 700 677	0.97

GT: genotype; DP: read depth; MAF: minor allele frequency; AB: allele balance;
MQM(R): mean mapping quality of reads supporting alternate (reference) allele

0.1, decomposing complex variants as above, keeping only invariant sites, and finally combining the invariant sites with the SNP_m10 set using BCFTOOLS. .

4.2.3 Sequencing and bioinformatics results

Sequencing. The first and second batches of sequencing resulted, respectively, in a mean of 0.942 ($\pm$0.168 SD) and 1.96 ($\pm$0.395 SD) million read pairs per sample. For all samples, > 99% of read pairs passed filtering with FASTP. Mean GC content across all samples was 39.0% ($\pm$0.349% SD) and did not differ between batches (linear model, F(1,64) = 1.564, p = 0.216). The read redundancy analysis showed that batch 2 had more unique reads than batch 1, but also that batch 2 showed somewhat diminishing returns as the number of unique reads encountered began to plateau beyond *ca.* 1.5 million read pairs (Fig. 4.3). The shape of the curves, with the proportion of unique reads encountered dropping rapidly as the number of reads inspected grew, suggested that for both batches sequencing effort was sufficient to cover most of the library complexity.

Read mapping. After read mapping, a mean of 67.4% ($\pm$2.94% SD) of read pairs mapped properly to the *E. cerinthoides* draft genome. There was a small but statistically significant difference in mapping rate between the two batches (batch 1: 65.1%, batch 2: 67.9%; linear model, F(1,64) = 9.84,

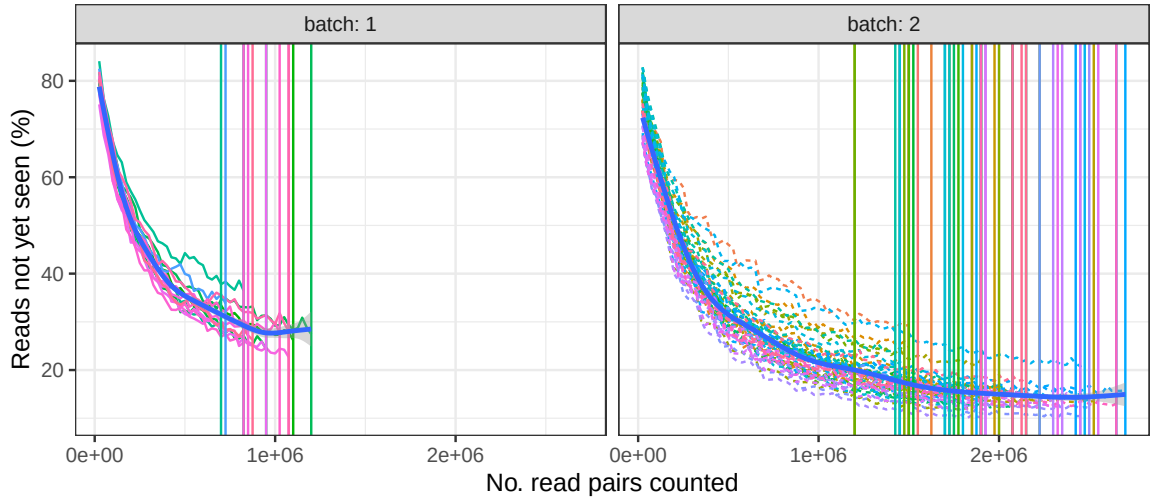


Fig. 4.3 Read redundancy analysis results from *bbcountunique.sh*. Read pairs are inspected in chunks of 25 000, and for each chunk the proportion of read pairs (including all previous chunks) that are completely unique is estimated. Colours are to aid in distinguishing estimates from different individuals. Vertical lines show read pair totals.

p = 0.00258). Mean mapping depth was 5.62X (±0.50 SD) and 9.10X (±1.63 SD) for batches 1 and 2, respectively.

Concordance between sequencing batches. For the individual that was sequenced in both batches, mean genotype depth was 16.9X in batch 1 and 31.4X in batch 2 when genotypes were called separately for each sample. Of the 29,998 sites (including all variant types) called with confidence as heterozygous based on batch 2 reads, 235 were called as homozygous and 185 were called as other genotypes based on batch 1 reads; therefore, assuming the batch 2 calls to be correct gives an error rate of 1.4% induced by lower sequencing effort. When including only SNPs, this putative error rate dropped to just 0.82% (of the 24,753 heterozygous SNPs, 205 were called as homozygous). Given this relatively negligible error rate I assumed that genotype calls for batch 1 individuals were accurate despite their lower depth.

Variant calling. Variant calling followed by stringent quality filtering resulted in a large and highly complete SNP data set (Table 4.2). Individuals sequenced in batch 1 had more missing genotypes in the SNP_m10 set (mean = 8.81% ± 3.09% SD) than batch 2 individuals (mean = 2.50% ± 1.86% SD), while the highest missingness in any individual was 13.4%. Genotype calls had lower read counts for

batch 1 individuals (mean = 10.3, range = 8.1-12.7) than batch 2 individuals (mean = 22.6, range = 12.5-32.5). Missingness was generally higher in the outgroups (*E. nevillei*: 9.14%; *E. quadrisulcata*: 7.01-11.2%), though this was not appreciable, suggesting that allele dropout did not affect library preparation or variant detection.

4.2.4 Analysis of population structure

To assess population structure without any prior assumptions about group membership, I employed three complementary analyses. Firstly, I ran a Principal Component Analysis (PCA) on the SNP_m10_-LD_maf04 set using the ADEGENET v.2.1.5 (Jombart and Ahmed, 2011) function *glPca*. Secondly, I employed the admixture model (Pritchard et al., 2000) for values of K (the number of ancestral populations) ranging from 2 to 8 using the sparse nonnegative matrix factorization (sNMF) algorithm (Frichot et al., 2014) implemented in LEA (Frichot and François, 2015), with default parameter values. For this I used the SNP_m10_LD_maf04 set and excluded the outgroups *E. nevillei* and *E. quadrisulcata*. For each K I ran 100 independent repetitions of the algorithm and summarised the outputs using the CLUMPAK method (Kopelman et al., 2015) implemented in STARMIE (`https://github.com/sa-lee/starmie`). I generated bar plots of ancestry proportions using POPHELPER (Francis, 2017). Lastly, to better visualise the connections between individuals as well as explore the hierarchical structure of genetic variation, I employed network analysis using NETVIEW (Steinig et al., 2016, `https://github.com/esteinig/netview`). With the ALL_m10 set, I used PIXY (Korunes and Samuk, 2021) to calculate the harmonic mean of absolute sequence divergence (d_{XY}) between individuals to generate the genetic distance matrix required by NETVIEW. I ran the network inference algorithm with values of k (the maximum number of mutual nearest neighbours) of 5, 10, 15 and 20, each time also estimating the minimum spanning tree to ensure a connected network was returned. I visualised each network with IGRAPH (Csardi and Nepusz, 2006) using the Kamada-Kawai spring-based layout algorithm (Kamada et al., 1989) to make the lengths of the connecting edges proportional to their associated genetic distance, and coloured the edges based on whether they were unique to the minimum spanning tree.

Based on these analyses I identified two genetically and geographically distinct clusters within *E. a. atrorosea*, which I grouped separately for further analyses that required group assignments (see

below). I refer to the northern cluster as *E. a. atrorosea* (North) and to the southern cluster as *E. a. atrorosea* (South).

4.2.5 Population structure analysis results

All three analyses of population structure revealed the existence of considerable genetic variation distinguishing various groups of individuals. The PCA eigenvalues (Fig. 4.4, *inset*) exhibited a steep decline in explained variance from axes 1 to 4 followed by a plateau from axes 5 to 7, after which they declined gradually. This pattern suggested the existence of either five or eight clusters in the data, as $n - 1$ axes are required to distinguish n clusters. In contrast, the sNMF-based cross-entropy criterion suggested $K = 1$ or $K = 2$ to be optimal given a 5% genotype masking rate, although higher values of K generally recovered sensible groupings of individuals in concordance with the PCA and NETVIEW analyses (Figs. 4.5,4.6).

The first principal component axis (PC1) primarily distinguished the two outgroups from *E. abietina* and showed *E. nevillei* to be the closer of the two outgroups. Within *E. abietina*, PC2 and, to a lesser extent, PC1 distinguished *E. a. abietina* plus *E. a. diabolis* from the rest of the subspecies, and the sNMF results also recovered these two groups as the most important ancestral clusters at $K = 2$. Notably, there was consistent support for two distinct genetic clusters within *E. a. atrorosea*: One group (*E. a. atrorosea* [South]) consisted of individuals from the southern parts of the Cape Peninsula ranging from Cape Point to Silvermine, while the other (*E. a. atrorosea* [North]) comprised northern individuals collected along the lower eastern slopes of Table Mountain. *E. a. atrorosea* (North) fell between the two major groups while still being closer to *E. a. atrorosea* (South) and *E. a. constantiana*. The NETVIEW analysis (Fig. 4.6) revealed more fine-scale patterns of genetic structure and relatedness, especially at lower values of k (the maximum number of connections allowed between individuals). Within *E. a. abietina*, individuals sampled from different parts of Table Mountain were clearly recovered as belonging to distinct network clusters at $k = 5$. Within *E. a. atrorosea* (North), the four southernmost individuals were consistently recovered as distinct from and largely unconnected to other populations, unlike the rest of *E. a. atrorosea* (North). NETVIEW also more consistently recovered *E. a. abietina* and *E. a. diabolis* as distinct from each other than the PCA or sNMF analyses did.

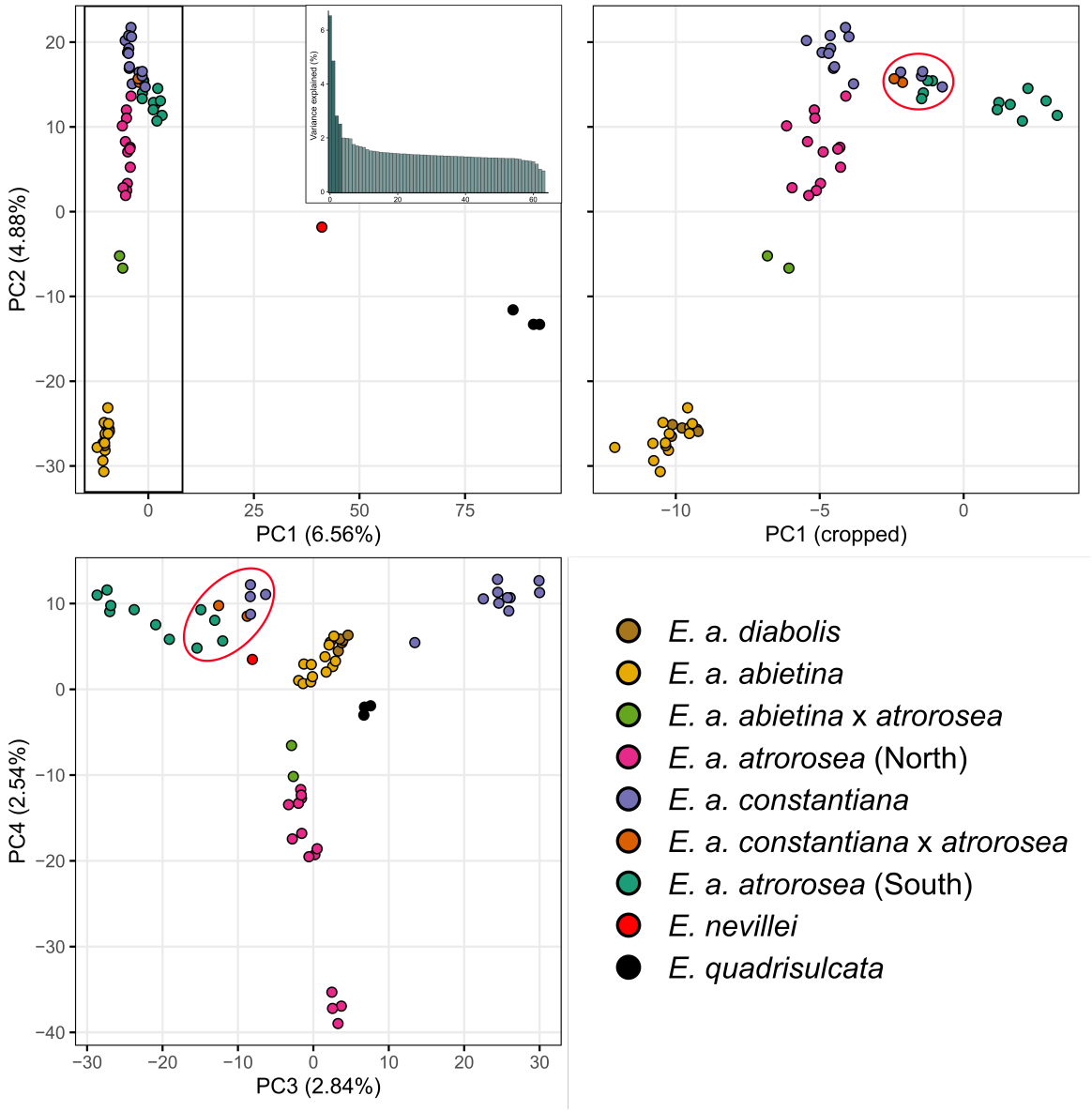


Fig. 4.4 PCA results. The box in the first plot indicates the area covered by the second plot. Samples from Blackburn Ravine are indicated by red ellipses. The *inset* shows a plot of the variance explained by each PCA axis (also shown in brackets in axis titles). Note especially the close relationship between *E. a. abietina* and *E. a. diabolis* and their distinctness from the other subspecies; the positioning of the two samples identified as *E. a. abietina* x *E. a. atrorosea* hybrids based on morphology; and the positioning of samples from Blackburn Ravine.

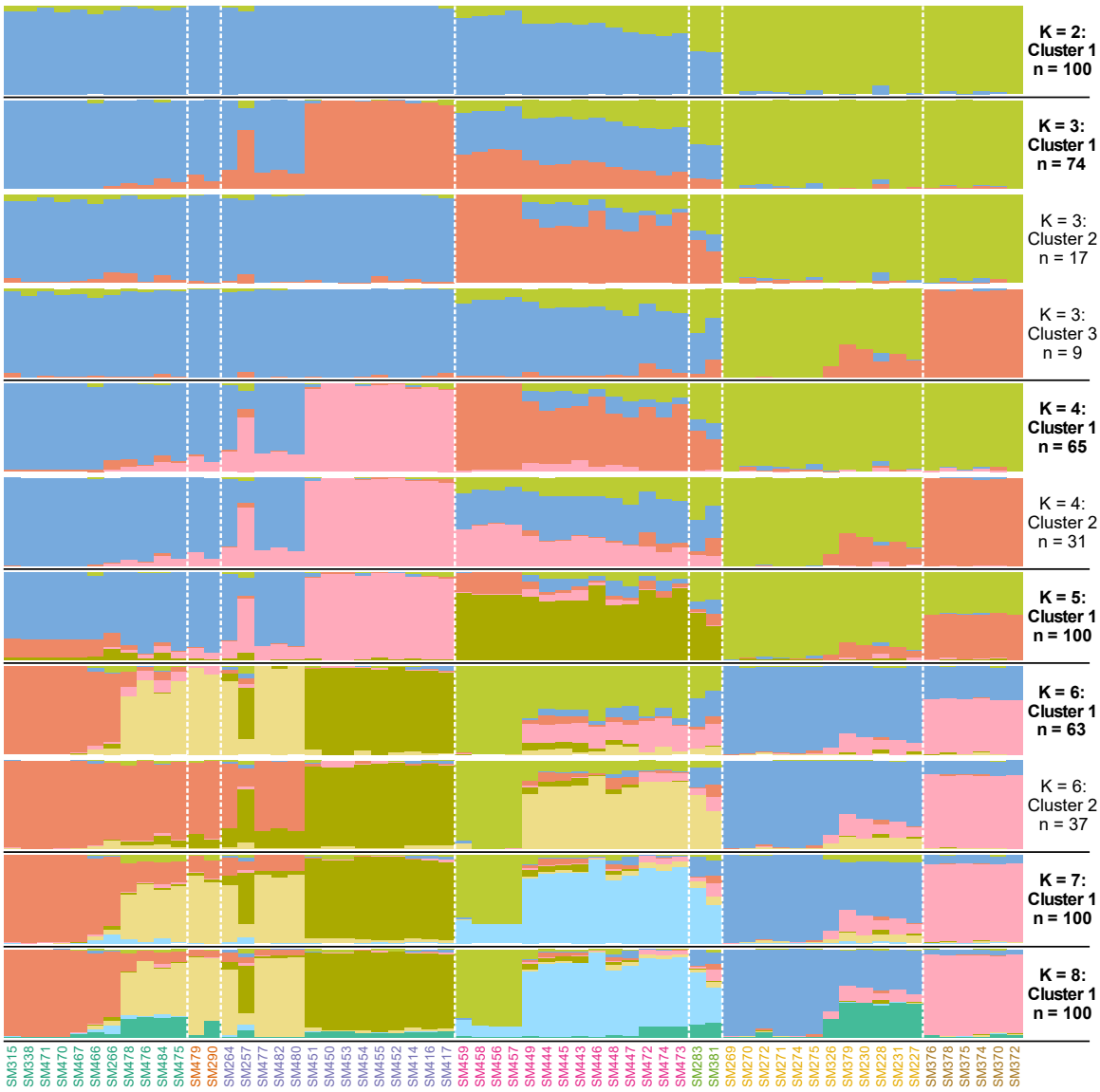


Fig. 4.5 Mean individual ancestry proportions estimated by sNMF for K = 2–8 with 100 runs for each K clustered by the CLUMPAK method. One cluster for $K = 3$ with $n = 4$ runs is not shown. Voucher numbers are shown below individuals. Text colours follow Fig. 4.4. Within groups, individuals are arranged by latitude from south to north. Note especially the presence of more than one potentially optimal solution for values of K = 3, 4, and 6; the increasingly mixed ancestry in *E. a. atrorosea* (North) from south to north; and the ability of the method to distinguish *E. a. diabolis* from *E. a. abietina* in many runs.

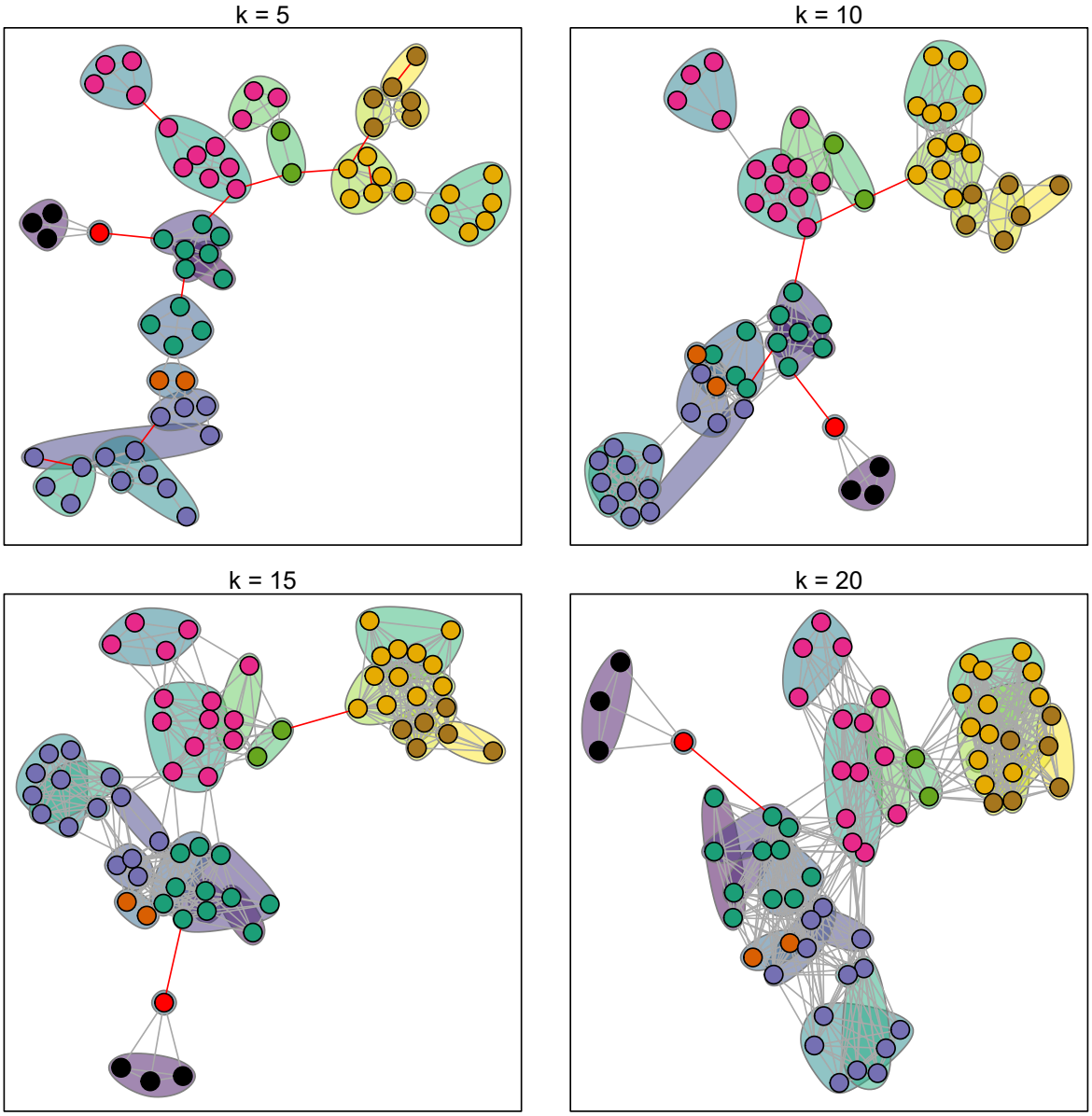


Fig. 4.6 NETVIEW results. Plots showing the NETVIEW networks for varying values of k (no. of allowed connections per node) based on d_{XY}. Nodes represent individuals and colours reflect prior population assignments (Fig. 4.4). The shaded regions envelope individuals located geographically close to each other, with the shading representing mean latitude (Viridis colour scale; lighter colours are more northerly). Red edges are unique to the minimum spanning tree. Note how the finer scale differences are more noticeable for small values of k, becoming obscured by the broader patterns as k increases.

4.2.6 Detecting recent hybrids

To investigate the presence of hybrids between genetically distinct clusters, I sequenced four individuals of putative hybrid origin which I identified based on a combination of morphological features and geographic location (Fig. 4.1). Firstly, I identified two individuals from the mid-elevation eastern slopes of Table Mountain (TM) from populations showing a range of intermediate flower colours (magenta to cerise) between *E. a. abietina* (light red; TM plateau) and *E. a. atrorosea* (North) (pink; TM lower eastern slopes). Secondly, I identified two individuals from Blackburn Ravine with floral tube lengths intermediate between *E. a. atrorosea* (South) (18–22 mm) and *E. a. constantiana* (8–11 mm). To test the hybrid origin of these individuals I employed NEWHYBRIDS (Anderson, 2008) in combination with SNMF. For each set of putative parents and hybrids I subset the SNP_m10_LD_-maf04 set to only the relevant individuals, removed resulting monomorphic sites, and ran sNMF with $K = 2$, again repeating the algorithm 100 times and summarising ancestry coefficients across runs as above. I then identified putatively non-admixed individuals as those whose maximum individual ancestry coefficient was ≥ 0.9, setting these as "P0" or "P1" in NEWHYBRIDS. To maximise the information content of the SNPs used, I calculated per-SNP F_{ST} (Weir and Cockerham, 1984) between P0 and P1 using the HIERFSTAT (Goudet, 2005) *basic.stats* function and chose the 500 SNPs with the highest F_{ST} while also only keeping one SNP per contig to avoid linkage effects. I used DARTR (Gruber et al., 2018) to convert the data into NEWHYBRIDS format. I then ran NEWHYBRIDS for 150,000 MCMC iterations, discarding the first 50000 as burn-in. Based on these analyses I identified individuals showing evidence of recent hybrid origin (including backcrosses). To refer to data sets excluding these individuals, I append the suffix "_noHybrids".

4.2.7 Evidence for recent and ongoing hybridization

There was widespread evidence for recent hybridisation within *E. abietina*. The two individuals originally suspected to have hybrid ancestry between *E. a. atrorosea* and *E. a. abietina* were consistently recovered by SNMF as sharing ancestry predominantly from clusters corresponding to *E. a. abietina* and *E. a. atrorosea* (North) (Fig. 4.5), and these were the only individuals inferred to be F2 hybrids between these groups by NEWHYBRIDS (Fig. 4.7). These individuals also fell between *E.*

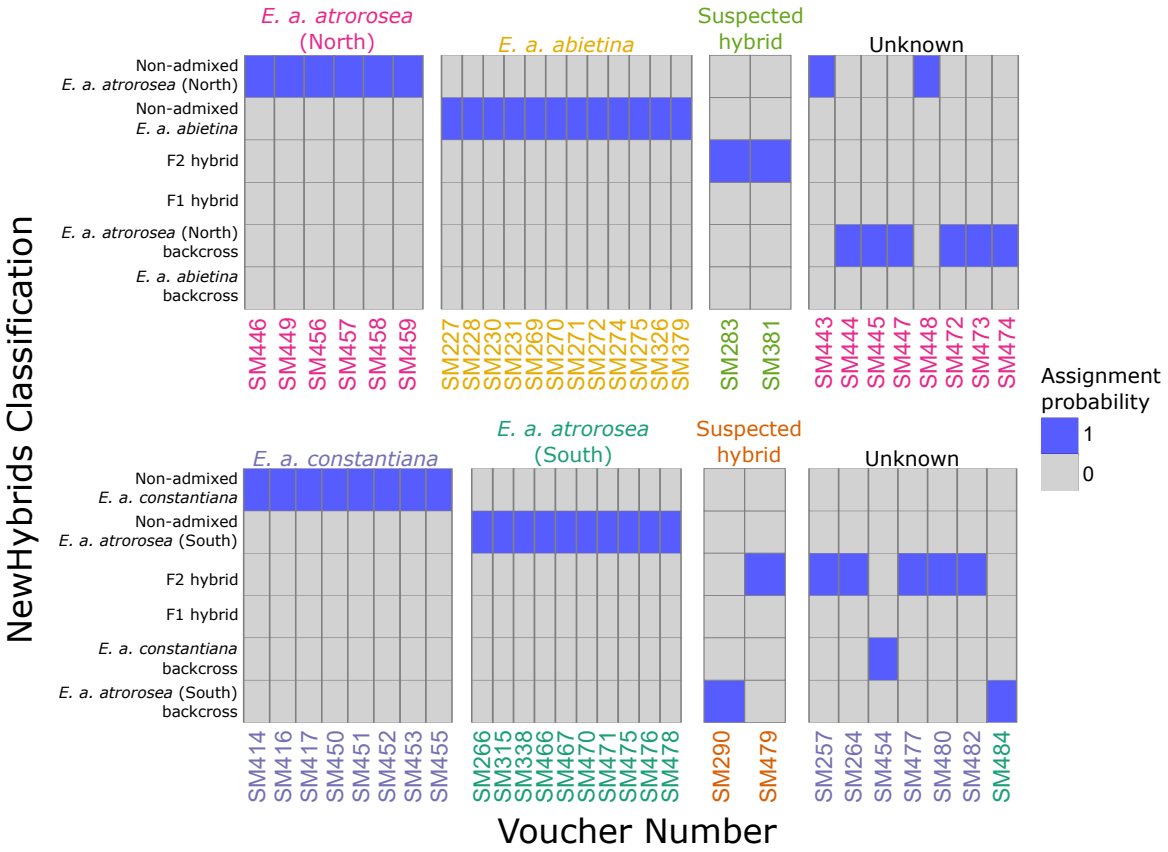


Fig. 4.7 Hybrid identification results: Posterior assignment probabilities of the various hybrid classes identifiable by NEWHYBRIDS. *Top*: *E. a. atrorosea* (North) x *E. a. abietina*; *Bottom*: *E. a. atrorosea* (South) x *E. a. constantiana*. Text colours follow Fig. 4.1.

a. abietina and *E. a. atrorosea* (North) in PCA space and in all NETVIEW graphs. Additionally, while sNMF strongly supported the distinctness of *E. a. abietina* from *E. a. atrorosea* (North), the reverse did not apply. Only six *E. a. atrorosea* (North) individuals had < 10% of their ancestry derived from the *E. a. abietina* parental population, and these consisted of the five southernmost *E. a. atrorosea* (North) samples plus one from Cecilia Forest in the central part of the *E. a. atrorosea* (North) range. The other eight *E. a. atrorosea* (North) individuals, which were not originally suspected of having mixed ancestry, had > 10% of their ancestry derived from the *E. a. abietina* parental population. Of these, six individuals – including three from the northernmost sampling site at Newlands Forest and three from Cecilia Forest – were inferred by NEWHYBRIDS to be *E. a. atrorosea* (North) backcrosses, while the remaining two individuals were inferred to be non-admixed *E. a. atrorosea* (North). Overall, these results point towards asymmetric gene flow from *E. a. abietina* into *E. a. atrorosea* (North).

The results regarding hybridisation between *E. a. atrorosea* (South) and *E. a. constantiana* were also complex. According to sNMF, only one out of 11 individuals originally identified as *E. a. atrorosea* (South) had mixed ancestry, which NEWHYBRIDS inferred to be an *E. a. atrorosea* (South) backcross. In contrast, six of the 14 individuals identified as *E. a. constantiana* had mixed ancestry. Five of these were inferred to be F2 hybrids (Fig. 4.7), including all from Blackburn Ravine and two from nearby Silvermine area just to the east, while one individual collected from the more northerly Vlakkenberg area was inferred to be an *E. a. constantiana* backcross. Of the two individuals originally suspected to be hybrids between the two subspecies, one was inferred to be an F2 hybrid and the other as an *E. a. atrorosea* (South) backcross.

4.2.8 Phylogenetic analysis (individual level)

I inferred two individual-level phylogenies, one including all individuals (n = 65) and one excluding putative hybrid individuals, including backcrosses (n = 45). I used IQ-TREE v.2.0.6 (Minh et al., 2020) using the full concatenated alignments. The ALL_m10 and ALL_m10_noHybrids VCF files were converted to fasta format using VCF2PHYLIP v.2.0 (Ortiz, 2019). For each analysis I chose the best-fitting substitution model using MODELFINDER (Kalyaanamoorthy et al., 2017) and the default Bayesian Information Criterion (best-fit model for all individuals = K3Pu+F+I+G4; for no hybrids = TPM2+F+R3). I estimated node certainty with 1000 ultrafast bootstrap replicates (Hoang et al., 2018) and 1000 SH approximate likelihood ratio test (SH-alrt; Guindon et al., 2010) replicates, and ran four independent runs to improve the search of the likelihood space. I plotted the maximum-likelihood trees with GGTREE v.3.2.1 (Yu et al., 2017) after rooting the tree at *E. quadrisulcata* with APE v.5.0 (Paradis and Schliep, 2019).

4.2.9 Phylogenetic analysis results

All individuals included. Fig. 4.8 depicts the results of the phylogenetic analysis with admixed individuals excluded. All four IQ-TREE runs returned virtually identical log-likelihood values. Overall, the tree showed a ladder-like pattern that appeared to be significantly influenced by the presence of individuals inferred to have mixed ancestry based on sNMF and NEWHYBRIDS results. At the same time, populations identified by the analyses of population structure were readily apparent

and often formed monophyletic clades. *E. a. abietina* and *E. a. diabolis* formed a well-supported clade and were confidently resolved as reciprocally monophyletic. Of the two individuals identified as *E. a. atrorosea* (North) x *E. a. abietina* F2 hybrids, one occupied a position clearly intermediate between the two populations, while the other fell within a clade containing most of the *E. a. atrorosea* (North) backcrosses and one non-admixed *E. a. atrorosea* (North) individual. The rest of the *E. a. atrorosea* (North) individuals (seven non-admixed and two backcrosses) formed an earlier-branching clade. The individuals identified as non-admixed *E. a. constantiana* formed a well-supported clade with the inclusion of the single *E. a. constantiana* backcross individual, while all the non-admixed *E. a. atrorosea* (South) individuals not collected from Blackburn Ravine also formed a well-supported clade. A relatively poorly supported clade (bootstrap = 90%, SH-alrt = 75%) lying between the non-admixed *E. a. atrorosea* (South) and *E. a. constantiana* clades consisted of all individuals from Blackburn Ravine regardless of prior identification. Within this clade, however, individuals identified as *E. a. atrorosea* (South) and *E. a. constantiana* each formed sub-clades that had good bootstrap support but poor SH-alrt support, and which each contained one of the individuals identified a priori as being of hybrid origin.

Admixed individuals excluded. Fig. 4.9 depicts the results of the phylogenetic analysis with admixed individuals excluded. The two best IQ-TREE runs returned similar log-likelihood values (run 2: -8,720,997.954, run 3: -8,721,638.908, difference = 640.954). Most nodes of the maximum-likelihood tree received high bootstrap and SH-alrt support, particularly at deeper phylogenetic levels at which populations were distinguished. The most notable exception was the branch subtending (*E. a. atrorosea* [North],(*E. a. abietina*,*E. a. diabolis*)), which had very low support values, meaning that the placement of *E. a. atrorosea* (North) could not be resolved. The three *E. a. atrorosea* (South) individuals from Blackburn Ravine that were not identified as being of recent hybrid origin were nevertheless recovered in an intermediate position between non-admixed *E. a. atrorosea* (South) and *E. a. constantiana*. Given that all other individuals collected from this locality were marked as putative hybrid-origin, this may indicate that these individuals contain mixed ancestry of too ancient origin to have been detected by the previous analyses.

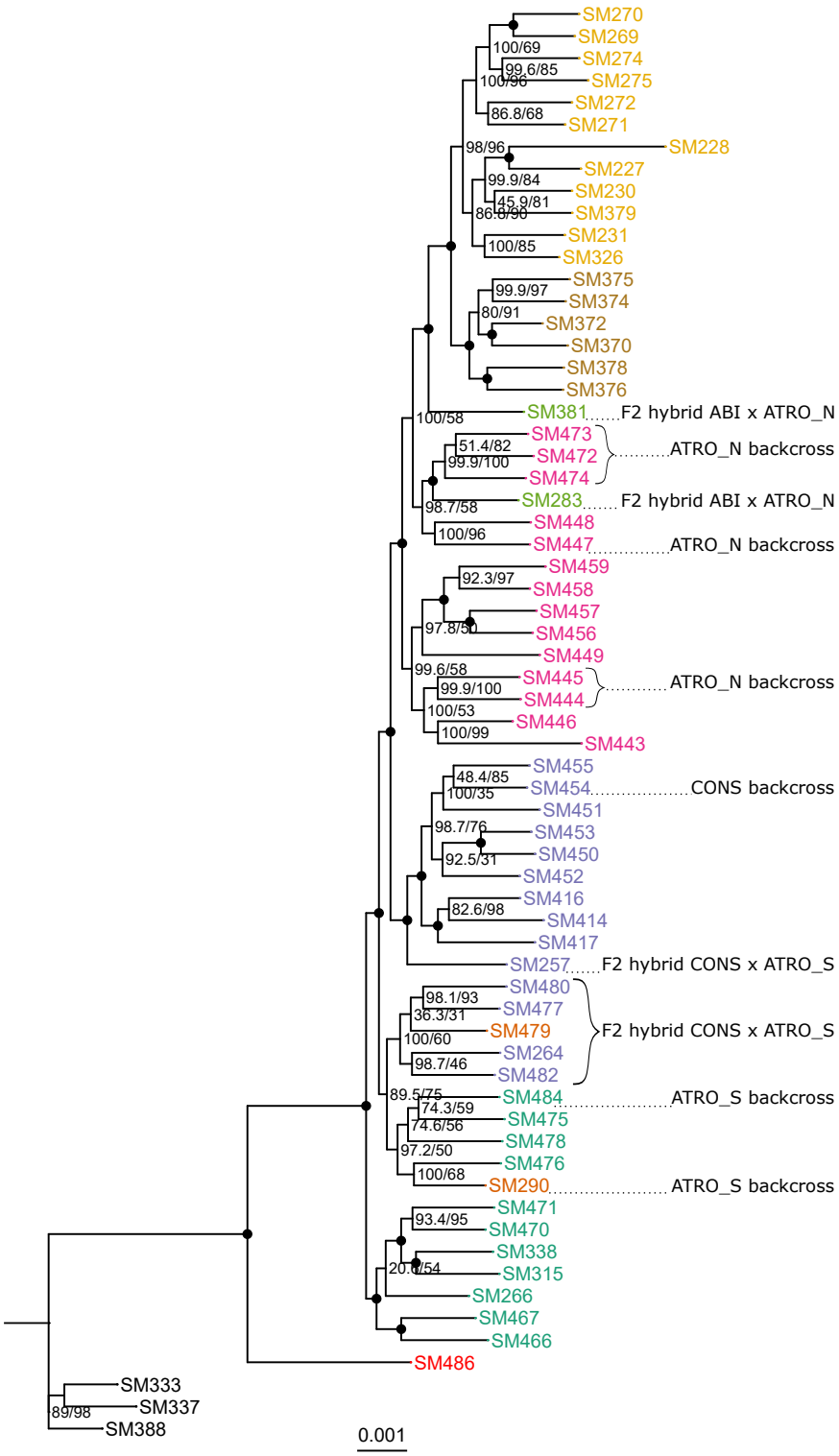


Fig. 4.8 The maximum-likelihood phylogeny inferred by IQ-TREE with all individuals included. Recent hybrids detected by NEWHYBRIDS are annotated. branch labels indicate bootstrap/SH-alrt support, and black circles indicate full support from both measures. Text colours follow Fig. 4.4.

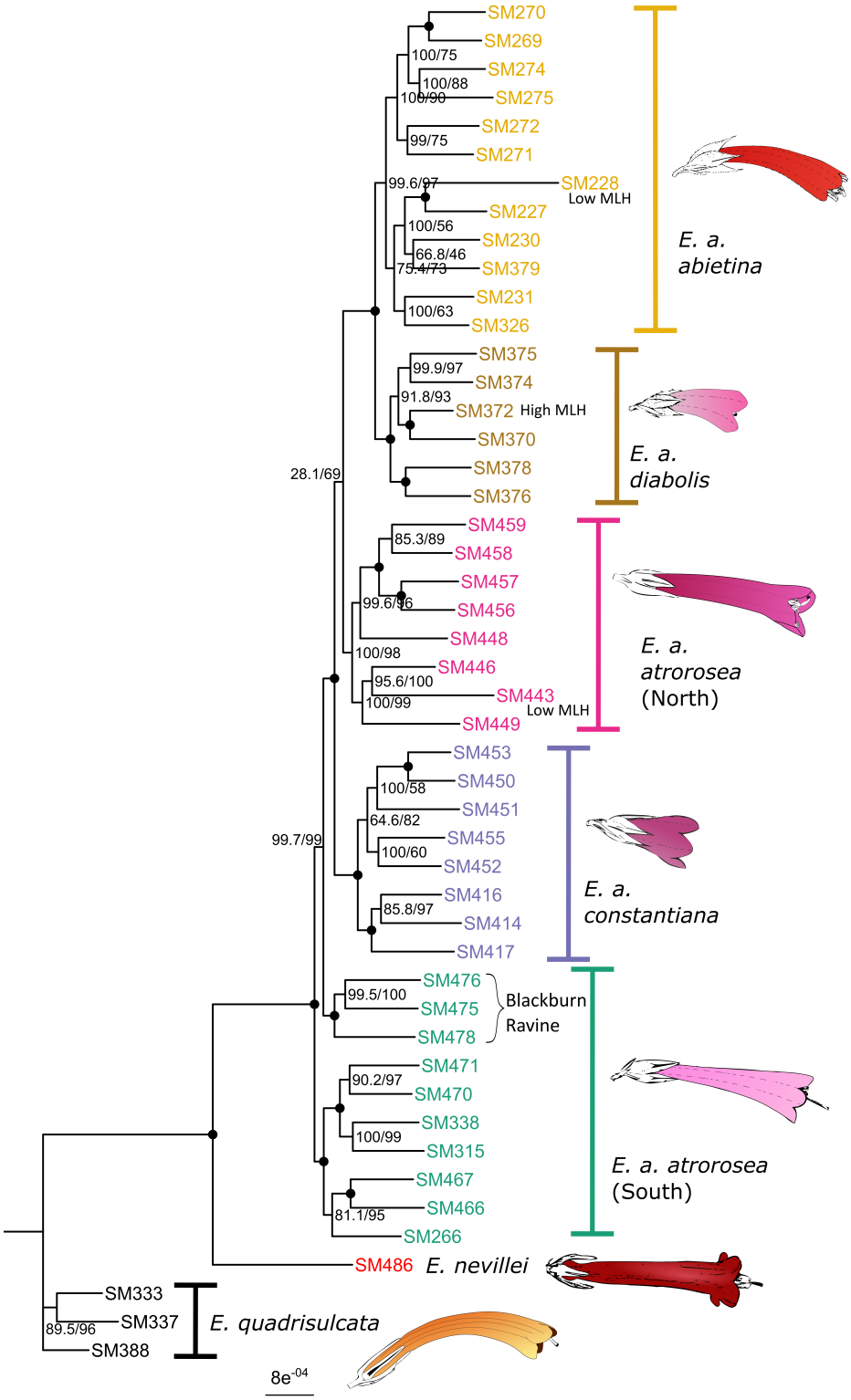


Fig. 4.9 The maximum-likelihood phylogeny inferred by IQ-TREE with admixed individuals excluded. Branch labels indicate bootstrap/SH-alrt support, and black circles indicate full support from both measures. Text colours follow Fig. 4.4.

Grouping individuals into populations. Confidently estimating population-level summary statistics typically necessitates grouping individuals into putatively monophyletic, panmictic and outbred "populations". Based on the previous analyses, I removed 17 individuals with potentially mixed ancestry and three with anomalous MLH, and assigned the remainder to five distinct populations within *E. abietina*. Finally, I split *E. a. atrorosea* (South) into two groups, creating a separate group for three individuals collected from Blackburn Ravine whose inclusion would make the population paraphyletic according to the phylogenetic analyses (see Fig. 4.9). I distinguish these populations from the previously named entities as follows:

- *E. a. abietina* = *ABI* (11 individuals)
- *E. a. diabolis* = *DIAB* (5 individuals)
- *E. a. constantiana* = *CONS* (8 individuals)
- *E. a. atrorosea* (South) = $ATRO_S$ (7 individuals)
- *E. a. atrorosea* (Blackburn Ravine) = $ATRO_{BLACKBURN}$ (3 individuals)
- *E. a. atrorosea* (North) = $ATRO_N$ (7 individuals)

All analyses in which individuals were grouped into populations used this assignment scheme.

4.2.10 Summary statistics

Individual-level summary statistics. To investigate the genetic diversity of all sampled individuals, I calculated per-sample multi-locus heterozygosity (MLH) using the R package INBREEDR (Stoffel et al., 2016) with the SNP_m10_LD_maf04 set. This revealed three anomalous samples within *E. abietina*, two with unusually low MLH which may be inbred (one from *E. a. abietina* and one from *E. a. atrorosea* [North]), and one with unusually high MLH (from *E. a. diabolis*) which may have been a chimeric sample stemming from a collection mishap in which two adjacent individuals were assumed to be one (Fig. 4.10). There was no clear relationship between latitude and MLH.

Population-level summary statistics. To estimate the magnitude of pairwise genomic differentiation between the five populations plus the two outgroup taxa, I calculated pairwise Hudson's F_{ST} (Bhatia et al., 2013; Hudson et al., 1992) using ADMIXTOOLS2 (Maier et al., 2022). F_{ST} was on

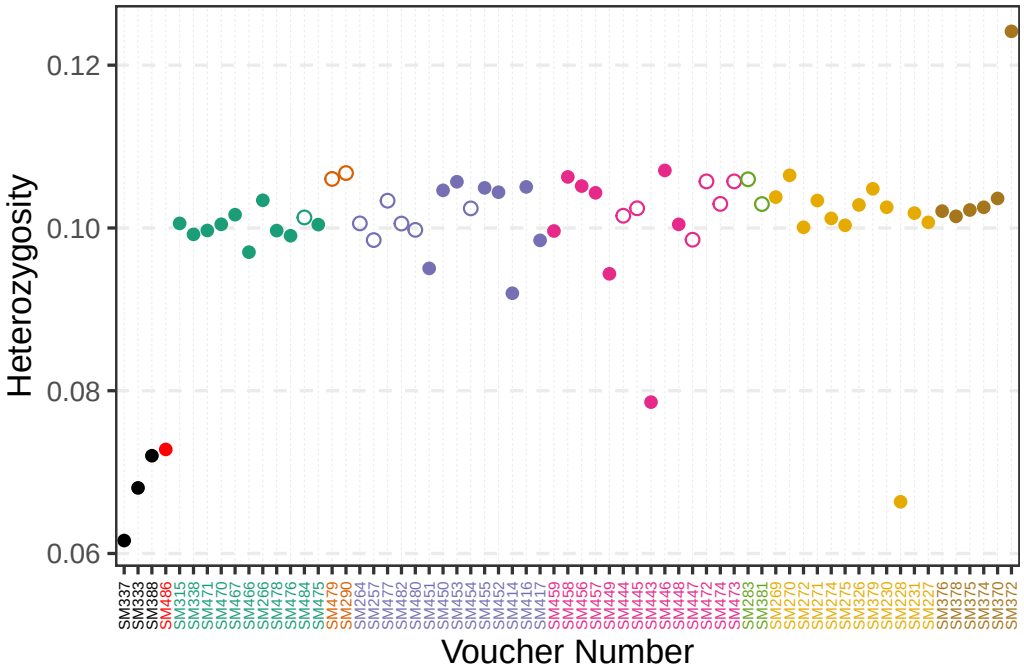


Fig. 4.10 Multilocus heterozygosity (MLH) of all individuals in the study. Text and point colours follow Fig. 4.4. Within groups, individuals are arranged by latitude from south to north.

average much higher at the species level than it was between populations within *E. abietina* (Fig. 4.11). Within *E. abietina*, differentiation was lowest between *ABI* and *DIAB* and highest between *CONS* and *DIAB*.

4.2.11 Population phylogeny

To estimate the population phylogeny while accounting for incomplete lineage sorting, I used a polymorphism-aware model (PoMo; Schrempf et al., 2019) implemented in IQ-TREE v.2.0.6 (Minh et al., 2020) with the SNP_m10_maf01 set. I set the substitution model to GTR+G4 (the GTR model [Tavaré 1986] with four discrete gamma rate categories), conducted 1000 ultrafast bootstrap replicates, and left all other parameters at their default values. The inferred PoMo tree (Fig. 4.11) had the same topology as the individual-level tree, except that $ATRO_N$ was recovered as sister to *CONS*, but with low support (bootstrap support = 72).

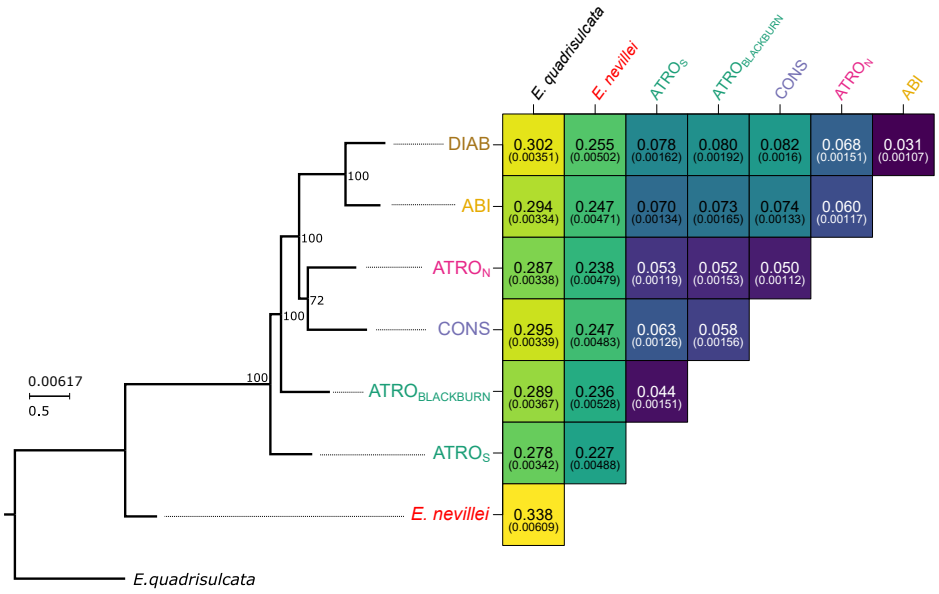


Fig. 4.11 *Left*: population-level phylogeny estimated by POMO, an ILS-aware method. Branch labels are UFBoot bootstrap. Branch lengths represent estimated number of mutations and frequency shifts per site (below scale) and approximate number of substitutions per site (above scale). *Right*: Pairwise F_{ST} between populations, with standard deviation in brackets. Darker colours indicate lower values.

4.2.12 Testing for reticulate evolution

I next aimed to test hypotheses of ancient introgression between populations. To search for evidence of ancient introgression, I used the *findGraphs* method implemented in ADMIXTOOLS2 (Maier et al., 2022) which conducts a heuristic search of the graph space by iteratively proposing modifications to the current graph – such as modifying the tree topology and adding admixed edges – each time evaluating the new graph's likelihood score, which is determined by comparing the $f3$-statistics (Peter, 2016) predicted by the graph to those estimated from the data. The algorithm attempts to find the best-fitting graph for the lowest number of allowed admixed edges (N_{ADMIX}) before adding more admixed edges one at a time.

I ran this analysis in two ways: first by specifying the population tree estimated by POMO as the starting tree, and second by setting *E. quadrisulcata* as the outgroup without specifying any other restrictions to the randomly generated starting tree. In each case I ran the analysis five times. I started the graph searches at $N_{ADMIX} = 0$ and set the maximum value of N_{ADMIX} to 2. To search the likelihood space more exhaustively, I set the total number of generations after which to stop to 5000; the number

of generations without improvement after which to stop to 100; the number of graphs evaluated in each generation to 30; and the *plusminus_generations* parameter (which helps to break out of local optima) to 20. All other parameters were kept at their default values. I manually inspected the ten resulting best graphs for each value of N_{ADMIX} to check for concordance in the graph topology, edge weights, and admixture proportions.

To identify the simplest graph that best fit the data, I compared alternative best-fitting graphs in a pairwise manner using a resampling procedure implemented in ADMIXTOOLS2 and described in Maier et al. (2022), which aims to test whether two graphs have similar predictive power. I used *qpgraph_resample_multi* to generate 100 replicate bootstrap resampled SNP block training and test sets and then evaluate each graph by estimating its weights using the training set and calculating its "out-of-sample" likelihood score using the (unseen) test set. This procedure allows to test the null hypothesis that both graphs have equivalent predictive power, or more specifically, that the differences between two graphs' out-of-sample scores for each bootstrap replicate are equal to zero. To estimate statistical significance I used *compare_fits*, which conducts a two-sided z-test on the score differences assuming a normal distribution of values and known standard deviation.

4.2.13 Evidence of ancient introgression

The ADMIXTOOLS2 graph search analysis results were effectively identical regardless of whether the starting tree was specified or not, with extremely slight differences in edge weights accounting for the lower scores of the best graphs found when no starting tree was specified (results not shown). When the number of allowed admixed edges (N_{ADMIX}) was zero the best graph in all runs matched the POMO tree topology (Fig. 4.12A). With $N_{ADMIX} = 1$, all runs converged on the same optimal topology and essentially identical edge weights. This graph showed the ancestor of *ABI* and *DIAB* as an admixed edge with most (76%) of its ancestry derived from the ancestor of (i.e., the edge subtending) $ATRO_N$ and the rest derived from the ancestor of *E. abietina* as a whole (Fig. 4.12B).

With $N_{ADMIX} = 2$, the independent runs found two distinct optimal graphs. The best-scoring graph (Fig. 4.12C), found in two runs, retained the admixed edge found with $N_{ADMIX} = 1$ and additionally recovered $ATRO_{BLACKBURN}$ as descending from an admixed edge with equal ancestry derived from the ancestors of $ATRO_S$ and $CONS$. The next-best graph (Fig. 4.12D) depicted a more complex history

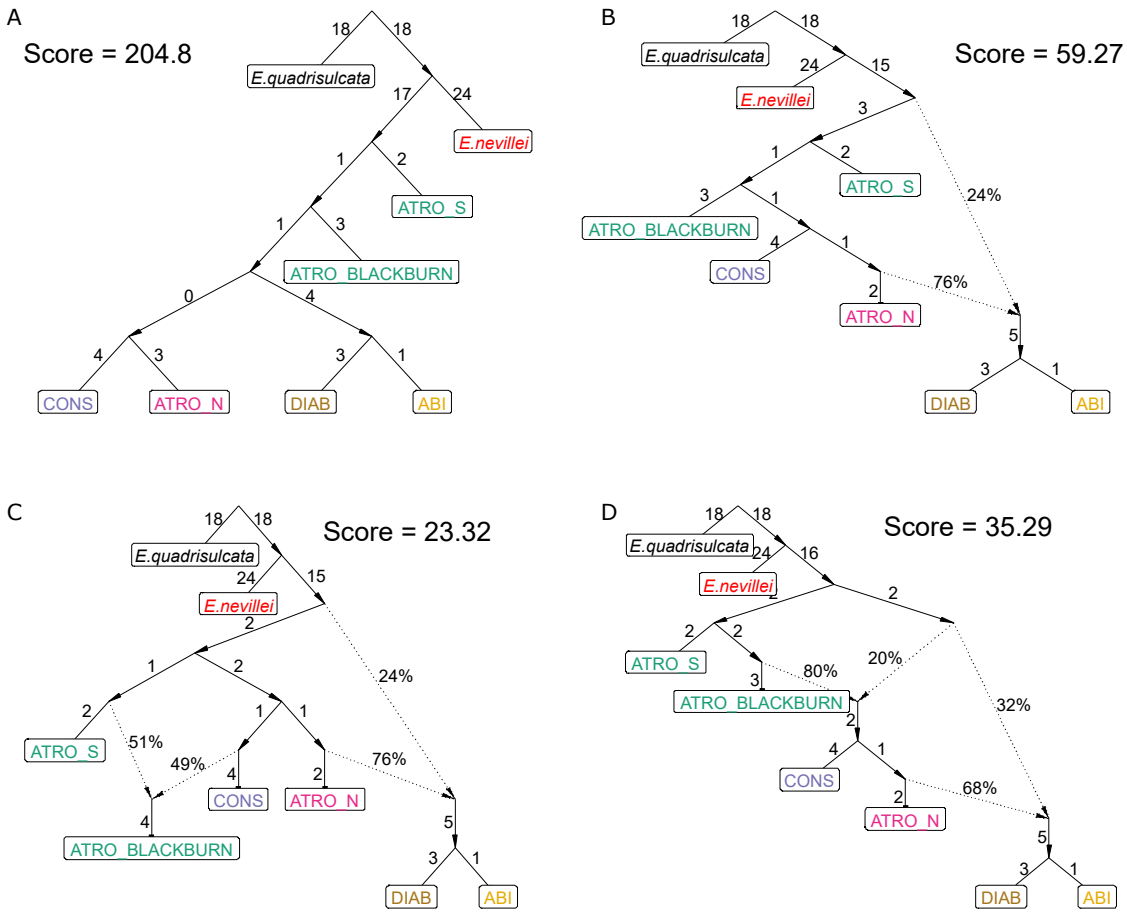


Fig. 4.12 The four unique top-scoring admixture graphs according to ADMIXTOOLS2. *A*: $N_{ADMIX} = 0$; *B*: $N_{ADMIX} = 1$; *C, D*: $N_{ADMIX} = 2$. Numbers above solid non-admixed edges are proportional to their weight, while percentages above dashed admixture edges indicate their contribution to the admixed edge. The graph in panel B had similar predictive power to those in C and D (see text for details), suggesting that one reticulation best fit the data.

of admixture, with one admixed edge subtending *CONS* and $ATRO_N$ deriving 80% of its ancestry from the ancestor of $ATRO_{BLACKBURN}$ and 20% from a "ghost" lineage that was sister to the rest of *E. abietina*. This same "ghost" lineage contributed 32% of its ancestry to an admixed edge subtending *ABI* and *DIAB*, which derived the rest of its ancestry from the ancestor of $ATRO_N$.

The best $N_{ADMIX} = 1$ graph had greater predictive power than the best $N_{ADMIX} = 0$ graph (mean score difference = -150.8 $\pm$ 39.5 SD, z = -3.82, p < 0.001). The two best $N_{ADMIX} = 2$ graphs had similar predictive power (mean score difference = -11.7 $\pm$11.3 SD, z = -1.04, p = 0.30), however, neither had better predictive power than the best $N_{ADMIX} = 1$ graph (best graph: mean score difference = -31.0 $\pm$17.7 SD, z = -1.75, p = 0.08; second-best graph: mean score difference = -19.3 $\pm$15.2 SD, z

= -1.27, p = 0.20). These results point to a single admixed edge as the most appropriate depiction of the reticulate history of the group.

4.3 Taxonomy and cryptic diversity

Given that genetic divergence between the three clearly differentiated species (*E. abietina*, *E. nevillei*, *E. quadrisulcata*) was found to be much higher than between the various populations of *E. abietina* (Fig. 4.11), it is perhaps reasonable to conclude that *E. abietina* constitutes a single – if, however, highly variable – species. On the other hand, there is mixed support for the current subspecific treatment of the complex. In particular, what is currently recognized as *E. a. atrorosea* appears to consist of two non-sister lineages that are geographically isolated but phenotypically very similar. The southern lineage, *E. a. atrorosea* (South), appears to be widespread on sandstone-derived soils south of Table Mountain, whereas the northern lineage, *E. a. atrorosea* (North), is seemingly restricted to the granite- and shale-derived soils of the eastern slopes of Table Mountain and Constantiaberg (see Fig. 4.1). None of the analyses suggested current or ancient hybridization between these two lineages, lending support to their independence. Such "cryptic" diversity is nowadays detected frequently in a range of organisms (Bickford et al., 2007), particularly since the advent of NGS-based genotyping in non-model organisms (e.g., Blair et al., 2019; Boucher et al., 2021; Daïnou et al., 2016; Hinojosa et al., 2019; Lutsak, 2020), and suggests that there is still a need for taxonomic studies in Cape *Erica*.

4.4 Hybridization and introgression

Although the populations of *Erica abietina* are genetically distinct, geographically separated and phenotypically recognizable, they are nevertheless only incompletely reproductively isolated. The discovery of several putative hybrids of recent origin – which occurred at localities where more than one subspecies was present or where their ranges met (Fig. 4.1, 4.7) – shows that gene flow between certain populations is ongoing and probably frequent. Interspecific hybridisation is well-known in Cape *Erica* (Oliver and Oliver, 2002, 2005), and does not always involve closely related species (e.g., Oliver, 1986). Interestingly, in this study none of the putative hybrid individuals was inferred to be first-generation and most were backcrosses. This implies that hybrids are fertile and liable to generate

"hybrid swarms", which may be the case at Blackburn Ravine where *E. a. atrorosea* (South) and *E. a. constantiana* appear to have formed a population that includes an abundance of admixed individuals that can freely backcross with either parental lineage.

In contrast, recent introgression between *E. a. abietina* and *E. a. atrorosea* (North) appears to have been unidirectional. This was made particularly evident in the sNMF analyses, in which all *E. a. abietina* individuals were inferred to be genetically non-admixed, whereas *E. a. atrorosea* (North) individuals showed a pattern of increasingly mixed ancestry from south to north. This could reflect a case of recent secondary contact between these lineages; however, this would imply recent range expansion in one or both of these populations, which seems unlikely given that genetic diversity (measured by heterozygosity) did not show a clear relationship with geography in either of them (Fig. 4.5). Nevertheless, denser geographic sampling of both populations coupled with formal analyses aimed at detecting range expansions (e.g., He et al., 2017; Peter and Slatkin, 2013) would be a valuable endeavour. Alternatively, the pattern may instead reflect long-standing gene flow from *E. a. abietina* into *E. a. atrorosea* (North) that has not been sufficient to cause gene swamping (Bridle and Vines, 2007; Lenormand, 2002). *Erica a. abietina* is likely to be adapted to extremely nutrient-poor sandstone-derived soils such as exist on Table Mountain's upper plateau (Compton, 2004), and such specialisation has been shown to limit the ability of fynbos plants to adapt to more nutrient-rich soils (Verboom et al., 2017). This may mean that *E. a. abietina* alleles are maladaptive for *E. a. abietina* x *E. a. atrorosea* (North) hybrids occupying the lower eastern slopes – whose soils are relatively rich in nutrients (Compton, 2004; Cramer et al., 2018, Fig. 4.1) – resulting in a "migration-selection equilibrium" that prevents genetic swamping (Lenormand, 2002). Studies investigating the factors that determine the geographic range limits of these populations (e.g., local adaptation, competition, dispersal limitation; Gaston, 2009) would help to illuminate these possibilities.

Evidence of present-day hybridisation does not necessarily mean that introgression played a role in a group's diversification (e.g., Jordan et al., 2017; Kessler et al., 2022; Westbury et al., 2019), however, the analysis based on $f3$-statistics provides good evidence that *E. abietina* does have a reticulate evolutionary history. The most likely interpretation of the best-fitting admixture graph (Fig. 4.12B) is that there was an early split between the ancestor of *E. a. abietina* (plus *E. a. diabolis*) and the ancestor of the rest of the *E. abietina* complex, which was followed much more recently

by the re-establishment of gene flow between their descendants. This implies the influence of a "ghost" lineage: one that is unknown, unsampled or extinct and which introgressed with an extant lineage (Ottenburghs, 2020).This phenomenon appears to be common across the tree of life (e.g., Barlow et al., 2018; Green et al., 2010; Maier et al., 2022; Meyer et al., 2012) and may be an important driver of diversification in rapidly evolving lineages (Ottenburghs, 2020). *Erica nevillei* and *E. quadrisulcata* differ from *E. abietina* in being narrow endemics confined to rocky outcrops at relatively high elevations, making the absence from Table Mountain's upper plateau of a related species with similar traits somewhat surprising especially given that conditions there are seemingly very similar and the area hosts several endemic *Erica* species with similar niches. It therefore seems possible that such a species did indeed exist, and represents the aforementioned "ghost" lineage. Such speculation could be tested with improved sampling of the relevant populations and a more thorough genotyping method (such as whole-genome sequencing) which would allow for more detailed and powerful analyses (e.g., Mondal et al., 2019), including, for example, those that are able to detect adaptive introgression of genomic regions (e.g., Racimo et al., 2015).

4.5 Floral trait evolution

Despite the close relationships and weak reproductive barriers between the populations of *Erica abietina*, shifts between pollination syndromes occurred at least twice in the complex. Given the phylogenetic results, the most parsimonious reconstruction of floral trait evolution is that long tubes are plesiomorphic and short tubes were derived in *E. a. constantiana* and *E. a. diabolis* independently. Such a pattern, combined with incomplete reproductive isolation between short- and long-tubed populations in the case of *E. a. constantiana* and *E. a. atrorosea* (South), implies that strong selective forces drive and maintain shifts to insect pollination at least in the *E. abietina* complex, and perhaps in Cape *Erica* more broadly.

The Cape Honeybee (*Apis mellifera* subsp. *capensis*) seems likely to be the primary pollinator of *E. a. constantiana* based on its scent and numerous personal observations (e.g., `https://www.inaturalist.org/observations/26399879`, `https://www.inaturalist.org/observations/131065480`). Going from bird to bee pollination may be interpreted as a shift to a less specialised

pollination system (Cronk and Ojeda, 2008). Johnson (1996), for example, suggested that frequent shifts to sunbird pollination in fynbos plants may have been driven by the relatively low abundance of insects especially at higher elevations, and that sunbirds are a more reliable pollinator in general. However, van der Niet et al. (2020) showed that individual honeybees visiting two co-occurring Cape *Erica* species exhibited remarkable floral constancy, in that during foraging bouts they tended to consistently prefer one or the other species rather than visiting both. Such constancy presumably explained the authors' finding that rates of interspecific pollen transfer were extremely low. Based on a detailed analysis of a community of co-occurring insect-pollinated *Erica*, Bouman et al. (2017) suggested that, rather than incurring a fitness cost, bee pollination may instead be beneficial to co-occurring *Erica* species because it enables them to collectively attract pollinators while simultaneously avoiding cross-species pollination. Such benefits may explain the apparently independent evolution of short-tubed flowers in *E. a. constantiana* and *E. a. diabolis*. Overall, these results highlight the variability of selection on floral traits and point to *Erica* as a whole being highly sensitive to such selection.

4.6 Conclusions

The results of this chapter add to the small but growing and much-needed body of research focused on understanding speciation in action in the Fynbos flora (Barraclough, 2006; Ellis et al., 2014; Lexer et al., 2014; Prunier et al., 2017; Prunier and Holsinger, 2010). I have shown that the subspecific classification of *E. abietina* is only a partial reflection of its evolutionary history. Notably, there is evidence of cryptic diversity in the complex which may have gone unrecognised due to a focus on floral traits in the taxonomic literature (Oliver and Oliver, 2002). Then, apart from ongoing gene flow in the complex revealed by the presence of late-generation subspecific hybrids, there is good evidence to suggest that ancient gene flow may have also influenced the complex's evolution in the form of "ghost" introgression from a now-extinct lineage. Finally, I suggest that floral trait divergence is likely to be driven by strong selective forces, but that it is unlikely to drive lineage divergence without the action of additional factors, particularly geographic isolation. Exploring the possibility of links between introgression and floral trait evolution in *Erica* may provide important insights

into the spectacular diversification of the genus (see e.g., Nelson et al., 2021). Overall these results paint a picture of a highly dynamic system whose evolutionary history has been shaped by diverse, interacting forces (Donoghue and Sanderson, 2015). Refining this picture will undoubtedly further our understanding of diversification in *Erica*.

Chapter 5

Synthesis — Drivers and modes of diversification in *Erica*

5.1 The role of gene flow

One of the most striking results of this thesis is the amount of evidence indicating incomplete reproductive isolation between taxa. Introgression has long been argued to be an important, if not essential, source of novel phenotypic variation that drives speciation by generating novel trait combinations essentially instantaneously (Anderson and Stebbins Jr, 1954). The adoption of genetic and genomic methods by the field of systematics has led to increasingly strong support for this idea (Baack and Rieseberg, 2007; Mallet, 2007; Nosil, 2008). However, gene flow is also often cited as a homogenising force that inhibits divergence and speciation (Lenormand, 2002; Levin, 1981), and for decades broad adherence to the "biological species concept" (Mayr, 1999) meant that reproductive isolation was widely regarded as *the* defining feature of species (de Queiroz, 2005; Mallet, 2001). This apparent paradox is resolved by the many modulating factors, such as natural selection and geography, that determine how gene flow influences diversification (Morjan and Rieseberg, 2004; Nosil, 2008; Rieseberg and Wendel, 1993).

It seems likely that throughout its history, as in the present day, many – if not all – of the *E. abietina/E. viscaria* clade's lineages have remained cross-compatible long after divergence. Although gene flow is itself a potential source of novel diversity and can even initiate speciation on its own

(Mallet, 2007), it seems unlikely that the level of diversity in Cape *Erica* could have arisen without additional modulating factors enabling lineages to become independent and form distinct species (de Queiroz, 2007). Therefore, while evidence of recent and historical gene flow between species suggests that introgression has played a role in the overall diversification dynamics of Cape *Erica*, the exact nature of that role cannot be understood without considering it in the broader context of natural selection and geography.

5.2 Speciation and floral trait evolution in context

Despite the evident lability of floral traits in *Erica*, their link to its diversification is less certain. In the CFR, there is some evidence that floral tube length variation correlates with intraspecific reproductive isolation in *Erica* (Newman and Johnson, 2021) and in other angiosperms (e.g., Minnaar et al., 2019). However, a global-scale meta-analysis showed that floral trait divergence on its own is a poor predictor of speciation, and that instead it almost always acts together with other factors, such as geographic isolation and habitat divergence, to drive speciation (Kay and Sargent, 2009).

The results of Chapter 3 supported previous work indicating that floral traits that reflect pollination mode are highly labile in Cape *Erica* (Pirie et al., 2011). Beyond that, despite the polytomous backbones of many clades, several populations with well differentiated floral phenotypes were found to emerge from those polytomies as highly distinct lineages. At the same time, in the "core-*viscaria*-clade", which presents the clearest example of this pattern, the phylogeny exhibits strong geographic structure and yet almost no correlation with phenotypic variation. This suggests that while floral trait changes have almost certainly occurred regularly, their specific geographic context might have been important in determining whether the change persisted.

The results of Chapter 4 indicated that differences in pollination syndrome cannot be assumed to signify reproductive isolation. For example, all but one individual that showed mixed ancestry between *E. a. constantiana* and *E. a. atrorosea* (South) occurred at the same locality (Blackburn Ravine), whereas non-admixed individuals occurred in areas where only one of the two floral types occurred. This suggests that floral trait divergence between these two populations (and perhaps also between *E. a. diabolis* and *E. a. abietina*) has resulted from the combined effect of pollination-driven divergent

selection and geographic isolation. The cryptic diversity that was found within *E. a. atrorosea* further illustrates that divergence can occur without a pollinator shift — instead, in this case it seems to have arisen in the context of geographic isolation coupled with the adaptation of the northern lineage to relatively nutrient-rich soils.

5.3 Budding speciation

The pattern of a small number of highly phenotypically and genetically distinct clades nested within an almost polytomous backbone, and whose closest relatives are usually their closest neighbours regardless of phenotype, is a characteristic feature of "budding speciation" (Grossenbacher et al., 2014). This occurs when an evolutionarily distinct population emerges from within a widespread and/or generalist species and develops into an independent evolutionary lineage while its progenitor lives on (Fig. 5.1; Crawford, 2010). Certain details of distribution and ecology in the *E. abietina/E. viscaria* clade add further support to the budding speciation model for at least some of its short-flowered taxa. *Erica petrusiana*, apart from being ecologically distinct from *E. viscaria* subsp. *longifolia*, also has a geographic range that is both peripheral and extremely small (see Chapter 3), features that are arguably essential to budding speciation due to their combined effect of limiting gene flow (Anacker and Strauss, 2014). *Erica latiflora* is a similar case with the exception that its tiny range is nested within that of *E. viscaria* subsp. *longifolia*. While some authors have argued that budding speciation is highly unlikely to occur without some degree of geographic isolation (Coyne and Orr, 2004), others have argued that ecological differentiation strong enough to induce selection against gene flow should be sufficient (Anacker and Strauss, 2014; Grossenbacher et al., 2014). This may be the case in *E. latiflora*. Although it is a highly localised and little-known species that has almost certainly suffered from habitat loss due to extensive agricultural development in its range, circumstantial evidence suggests that it is restricted to shale-derived soils in a small area otherwise surrounded by sandstone-derived soils. Within *E. abietina*, it could be argued that *E. a. constantiana* and *E. a. diabolis* represent budded lineages, especially considering their geographic context, close kinship with the more widespread populations of the species, and the uncertainty of their phylogenetic placement.

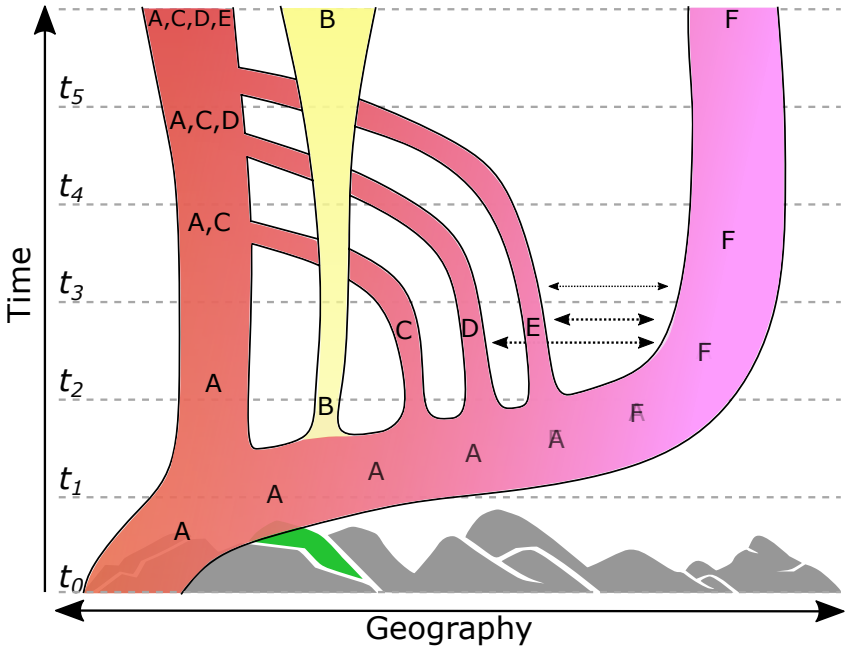


Fig. 5.1 A simple model to illustrate budding speciation. The colours indicate a trait value, for example, flower colour, while lineage location and width denote range and population size, respectively. At time t_0, only one relatively localised lineage exists (A), but by t_1 its range has begun to expand considerably in response to favourable climatic changes. By t_2, four small subpopulations have "budded" off as geographic isolates, but only lineage B has undergone a significant trait shift in response to a distinct habitat within its range. Eventually lineage B establishes itself as distinct and accumulates private genetic diversity, whereas most of the other lineages reintegrate with lineage A, while lineage F maintains occasional gene flow with its relatives but ultimately becomes distinct due to the combined effect of geographic isolation and a gradual trait shift. The present day may be anywhere along this continuum.

The present work suggests that budding speciation may be an ongoing feature of Cape *Erica* evolution. However, incorporating budding speciation as an addition to the standard model of tree-like evolution has also been shown to provide explanatory power regarding macroevolution at deeper phylogenetic levels (Crouch et al., 2021). Evidence suggesting a "hard" polytomy at the base of the *E. abietina/E. viscaria* (Chapter 3), along with extremely rapid diversification in Cape *Erica* (Pirie et al., 2016), could be interpreted as reflecting high rates of budding speciation in the past. Tank et al. (2015) demonstrated that the history of angiosperm diversification has been characterised by successively nested radiations, generating a highly asymmetric phylogeny. On a much smaller scale, this appears to have been the case in *Erica* in general (Pirie et al., 2016) and in the *E. abietina/E. viscaria* clade (Chapter 3). Interpreting this pattern is difficult. Tank et al. (2015) suggested a role for whole-genome duplication events, but this is highly unlikely to be the case in *Erica* given what we know about their genomes (Mugrabi De Kuppler, 2013; Nelson and Oliver, 2005). Instead, a

legacy of ancient budding speciation might be responsible. Polytomies are typically interpreted as representing a series of successive speciation events that occur so rapidly that they leave no molecular signal of the order in which they happened, thus appearing to have occurred simultaneously (e.g., Klak et al., 2013). However, a "multi-budding" model in which several lineages arise independently from the same common ancestor would have a very similar molecular signature, because no two budded lineages could be said to be more closely related to each other than to their single common ancestor. This is a more literal interpretation of polytomies. In the context of successive nested radiations, a multi-budding model would imply that each "wave" of diversification is preceded by the emergence of a single highly successful lineage that expands its range, and in doing so leaves behind a cohort of budded species (Anacker and Strauss, 2014; Brassac and Blattner, 2015; Grossenbacher et al., 2014; Otero et al., 2022). Just as this appears to best describe the phylogenetic patterns present in the "core-*viscaria*-clade", it may equally describe the state of affairs during the early evolution of the *E. abietina/E. viscaria* clade and, perhaps, in Cape *Erica* in general.

References

Adamson, R. (1959). Notes on the Phytogeography of the Flora of the Cape Peninsula. *Transactions of the Royal Society of South Africa*, 35(5):443–462.

Adamson, R. S. and Salter, T. M. (1950). Flora of the Cape Peninsula. *Flora of the Cape Peninsula.*

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402.

Anacker, B. L. and Strauss, S. Y. (2014). The geography and ecology of plant speciation: range overlap and niche divergence in sister species. *Proceedings of the Royal Society B: Biological Sciences*, 281(1778):20132980.

Anderson, E. and Stebbins Jr, G. L. (1954). Hybridization as an evolutionary stimulus. *Evolution*, pages 378–388.

Anderson, E. C. (2008). Bayesian inference of species hybrids using multilocus dominant genetic markers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1505):2841–2850.

Anisimova, M., Gil, M., Dufayard, J.-F., Dessimoz, C., and Gascuel, O. (2011). Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Systematic Biology*, 60(5):685–699.

Araya, Y. N., Silvertown, J., Gowing, D. J., McConway, K. J., Linder, H. P., and Midgley, G. (2010). A fundamental, eco-hydrological basis for niche segregation in plant communities. *New Phytologist*, 189(1):253–258.

Arcila, D., Hughes, L. C., Meléndez-Vazquez, B., Baldwin, C. C., White, W. T., Carpenter, K. E., Williams, J. T., Santos, M. D., Pogonoski, J. J., Miya, M., Ortí, G., and Betancur-R, R. (2021). Testing the Utility of Alternative Metrics of Branch Support to Address the Ancient Evolutionary Radiation of Tunas, Stromateoids, and Allies (Teleostei: Pelagiaria). *Systematic Biology*, 70(6):1123–1144.

Aristide, L. and Morlon, H. (2019). Understanding the effect of competition during evolutionary radiations: an integrated model of phenotypic and species diversification. *Ecology Letters*, 22(12):2006–2017.

Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., Reeb, C. A., and Saunders, N. C. (1987). INTRASPECIFIC PHYLOGEOGRAPHY: The Mitochondrial DNA Bridge Between Population Genetics and Systematics. *Annual Review of Ecology and Systematics*, 18(1):489–522.

Avise, J. C. et al. (2000). *Phylogeography: the history and formation of species*. Harvard university press.

Avise, J. C. and Robinson, T. J. (2008). Hemiplasy: A New Term in the Lexicon of Phylogenetics. *Systematic Biology*, 57(3):503–507.

Baack, E. J. and Rieseberg, L. H. (2007). A genomic view of introgression and hybrid speciation. *Current Opinion in Genetics & Development*, 17(6):513–518.

Bagley, J. C., Uribe-Convers, S., Carlsen, M. M., and Muchhala, N. (2020). Utility of targeted sequence capture for phylogenomics in rapid, recent angiosperm radiations: Neotropical *Burmeistera* bellflowers as a case study. *Molecular Phylogenetics and Evolution*, 152:106769.

Balkenhol, N., Waits, L. P., and Dezzani, R. J. (2009). Statistical approaches in landscape genetics: An evaluation of methods for linking landscape and genetic data. *Ecography*, 32(5):818–830.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19(5):455–477.

Barlow, A., Cahill, J. A., Hartmann, S., Theunert, C., Xenikoudakis, G., Fortes, G. G., Paijmans, J. L., Rabeder, G., Frischauf, C., Grandal-d'Anglade, A., et al. (2018). Partial genomic survival of cave bears in living brown bears. *Nature Ecology & Evolution*, 2(10):1563–1570.

Barnes, K., Nicolson, S. W., and Van Wyk, B. E. (1995). Nectar sugar composition in *Erica*. *Biochemical Systematics and Ecology*, 23(4):419–423.

Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P., and Marth, G. T. (2011). BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, 27(12):1691–1692.

Barraclough, T. G. (2006). What can phylogenetics tell us about speciation in the Cape flora? *Diversity & Distributions*, 12(1):21–26.

Bellstedt, D. U., Pirie, M. D., Visser, J. C., de Villiers, M. J., and Gehrke, B. (2010). A rapid and inexpensive method for the direct PCR amplification of DNA from plants. *American Journal of Botany*, 97(7).

Bergé, L. (2018). Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm. Technical report, Department of Economics at the University of Luxembourg.

Berv, J. S., Singhal, S., Field, D. J., Walker-Hale, N., McHugh, S. W., Shipley, J. R., Miller, E. T., Kimball, R. T., Braun, E. L., Dornburg, A., Parins-Fukuchi, C. T., Prum, R. O., Friedman, M., and Smith, S. A. (2022). Molecular early burst associated with the diversification of birds at the K–Pg boundary. *BioRxiv*.

Bhatia, G., Patterson, N., Sankararaman, S., and Price, A. L. (2013). Estimating and interpreting FST: The impact of rare variants. *Genome Research*, 23(9):1514–1521.

Bickford, D., Lohman, D. J., Sodhi, N. S., Ng, P. K. L., Meier, R., Winker, K., Ingram, K. K., and Das, I. (2007). Cryptic species as a window on diversity and conservation. *Trends in Ecology and Evolution*, 22(3):148–155.

Blair, C., Bryson, R. W., Linkem, C. W., Lazcano, D., Klicka, J., Mccormack, J. E., Klicka, J., Lazcano, D., Linkem, C. W., Bryson, R. W., and Blair, C. (2019). Cryptic diversity in the Mexican highlands: Thousands of UCE loci help illuminate phylogenetic relationships, species limits and divergence times of montane rattlesnakes (Viperidae: Crotalus). *Molecular Ecology Resources*, 19(2):349–365.

Borowiec, M. L. (2016). AMAS: A fast tool for alignment manipulation and computing of summary statistics. *Peerj*, 2016(1).

Boucher, F. C., Dentant, C., Ibanez, S., Capblancq, T., Boleda, M., Boulangeat, L., Smyčka, J., Roquet, C., and Lavergne, S. (2021). Discovery of cryptic plant diversity on the rooftops of the Alps. *Scientific Reports*, 11(1).

Bougie, T., Brelsford, A., and Hedin, M. (2021). Evolutionary impacts of introgressive hybridization in a rapidly evolving group of jumping spiders (F. Salticidae, *Habronattus americanus* group). *Molecular Phylogenetics and Evolution*, 161:107165.

Bouman, R. W., Steenhuisen, S. L., and Van Der Niet, T. (2017). The role of the pollination niche in community assembly of *Erica* species in a biodiversity hotspot. *Journal of Plant Ecology*, 10(4):634–648.

Bradshaw, P. L. and Cowling, R. M. (2014). Landscapes, rock types, and climate of the Greater Cape Floristic Region. In Allsopp, N., Colville, J. F., and Verboom, G. A., editors, *Fynbos: Ecology, evolution, and conservation of a Megadiverse Region*, chapter 2, pages 26–46. Oxford University Press, Oxford.

Brassac, J. and Blattner, F. R. (2015). Species-level phylogeny and polyploid relationships in *Hordeum* (Poaceae) inferred by next-generation sequencing and in silico cloning of multiple nuclear loci. *Systematic Biology*, 64(5):792–808.

Brewer, G. E., Clarkson, J. J., Maurin, O., Zuntini, A. R., Barber, V., Bellot, S., Biggs, N., Cowan, R. S., Davies, N. M., Dodsworth, S., et al. (2019). Factors affecting targeted sequencing of 353 nuclear genes from herbarium specimens spanning the diversity of angiosperms. *Frontiers in plant science*, 10:1102.

Bridle, J. R. and Vines, T. H. (2007). Limits to evolution at range margins: when and why does adaptation fail? *Trends in Ecology & Evolution*, 22(3):140–147.

Brusatte, S. L., Lloyd, G. T., Wang, S. C., and Norell, M. A. (2014). Gradual Assembly of Avian Body Plan Culminated in Rapid Rates of Evolution across the Dinosaur-Bird Transition. *Current Biology*, 24(20):2386–2392.

Burbrink, F. T. and Gehara, M. (2018). The biogeography of deep time phylogenetic reticulation. *Systematic Biology*, 67(5):743–755.

Chamala, S., García, N., Godden, G. T., Krishnakumar, V., Jordon-Thaden, I. E., Smet, R. D., Barbazuk, W. B., Soltis, D. E., and Soltis, P. S. (2015). MarkerMiner 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes. *Applications in Plant Sciences*, 3(4):1400115.

Chan, K. O., Hutter, C. R., Wood, P. L., Grismer, L. L., Das, I., and Brown, R. M. (2020). Gene flow creates a mirage of cryptic species in a Southeast Asian spotted stream frog complex. *Molecular Ecology*, 29(20):3970–3987.

Charif, D. and Lobry, J. (2007). SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In Bastolla, U., Porto, M., Roman, H., and Vendruscolo, M., editors, *Structural approaches to sequence evolution: Molecules, networks, populations*, Biological and Medical Physics, Biomedical Engineering, pages 207–232. Springer Verlag, New York. ISBN : 978-3-540-35305-8.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884–i890.

Chifman, J. and Kubatko, L. (2014). Phylogenetics Quartet Inference from SNP Data Under the Coalescent Model. *Bioinformatics*, 30(23):3317–3324.

Coetzee, A. (2016). *Nectar distribution and nectarivorous bird foraging behaviour at different spatial scales*. PhD thesis, Stellenbosch: Stellenbosch University.

Coetzee, A., Spottiswoode, C. N., and Seymour, C. L. (2020). Post-pollination barriers enable coexistence of pollinator-sharing ornithophilous *Erica* species. *Journal of Plant Research*, 133(6):873–881.

Compton, J. S. (2004). *The rocks and mountains of Cape Town*. Juta and Company Ltd.

Costa, L., Marques, A., Buddenhagen, C., Thomas, W. W., Huettel, B., Schubert, V., Dodsworth, S., Houben, A., Souza, G., and Pedrosa-Harand, A. (2021). Aiming off the target: recycling target capture sequencing reads for investigating repetitive DNA. *Annals of Botany*, 128(7):835–848.

Cowling, R. M. (1987). Fire and its role in coexistence and speciation in Gondwanan shrublands. *South African Journal of Science*, 83(2):106–112.

Cowling, R. M., MacDonald, I., and Simmons, M. (1996). The Cape Peninsula, South Africa: physiographical, biological and historical background to an extraordinary hot-spot of biodiversity. *Biodiversity & Conservation*, 5(5):527–550.

Cowling, R. M., Potts, A. J., Bradshaw, P. L., Colville, J., Arianoutsou, M., Ferrier, S., Forest, F., Fyllas, N. M., Hopper, S. D., Ojeda, F., Procheş, Ş., Smith, R. J., Rundel, P. W., Vassilakis, E., and Zutta, B. R. (2015). Variation in plant diversity in mediterranean-climate ecosystems: The role of climatic and topographical stability. *Journal of Biogeography*, 42(3):552–564.

Cowling, R. M., Procheş, Ş., and Partridge, T. C. (2009). Explaining the uniqueness of the Cape flora: Incorporating geomorphic evolution as a factor for explaining its diversification. *Molecular Phylogenetics and Evolution*, 51(1):64–74.

Coyne, J. A. and Orr, H. A. (2004). *Speciation*. Sinauer Associates.

Cramer, M. D., Power, S. C., Belev, A., Gillson, L., Bond, W. J., Hoffman, M. T., and Hedin, L. O. (2018). Are forest-shrubland mosaics of the Cape Floristic Region an example of alternate stable states? *Ecography*, 42(4):717–729.

Crawford, D. J. (2010). Progenitor-derivative species pairs and plant speciation. *TAXON*, 59(5):1413–1423.

Cronk, Q. and Ojeda, I. (2008). Bird-pollinated flowers in an evolutionary and molecular context. *Journal of Experimental Botany*, 59(4):715–727.

Crouch, N. M. A., Edie, S. M., Collins, K. S., Bieler, R., and Jablonski, D. (2021). Calibrating phylogenies assuming bifurcation or budding alters inferred macroevolutionary dynamics in a densely sampled phylogeny of bivalve families. *Proceedings of the Royal Society B: Biological Sciences*, 288(1964).

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *Interjournal*, Complex Systems:1695.

Daïnou, K., Blanc-Jolivet, C., Degen, B., Kimani, P., Ndiade-Bourobou, D., Donkpegan, A. S. L., Tosso, F., Kaymak, E., Bourland, N., Doucet, J.-L., and Hardy, O. J. (2016). Revealing hidden species diversity in closely related species using nuclear SNPs, SSRs and DNA sequences – a case study in the tree genus *Milicia*. *BMC Evolutionary Biology*, 16(1):259.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., and Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., et al. (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, 10(2):giab008.

de Queiroz, K. (2005). Ernst Mayr and the modern concept of species. *Proceedings of the National Academy of Sciences*, 102(suppl 1):6600–7.

de Queiroz, K. (2007). Species Concepts and Species Delimitation. *Systematic Biology*, 56(6):879–886.

De Smet, R., Adams, K. L., Vandepoele, K., Montagu, M. C. E. V., Maere, S., and de Peer, Y. V. (2013). Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences*, 110(8):2898–2903.

de Sousa, F., Bertrand, Y. J. K., Nylinder, S., Oxelman, B., Eriksson, J. S., and Pfeil, B. E. (2014). Phylogenetic Properties of 50 Nuclear Loci in *Medicago* (Leguminosae) Generated Using Multiplexed Sequence Capture and Next-Generation Sequencing. *PLoS ONE*, 9(10):e109704.

Degnan, J. H. and Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS genetics*, 2(5):e68.

Degnan, J. H. and Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution*, 24(6):332–340.

Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, 6(5):361–375.

Dolinay, M., Nečas, T., Zimkus, B. M., Schmitz, A., Fokam, E. B., Lemmon, E. M., Lemmon, A. R., and Gvoždík, V. (2021). Gene flow in phylogenomics: Sequence capture resolves species limits and biogeography of Afromontane forest endemic frogs from the Cameroon Highlands. *Molecular Phylogenetics and Evolution*, 163:107258.

Donoghue, M. J. (2008). A phylogenetic perspective on the distribution of plant diversity. *Proceedings of the National Academy of Sciences*, 105(Supplement 1):11549–55.

Donoghue, M. J. and Sanderson, M. J. (2015). Confluence, synnovation, and depauperons in plant diversification. *New Phytologist*, 207(2):260–274.

Dornburg, A., Fisk, J. N., Tamagnan, J., and Townsend, J. P. (2016). PhyInformR: phylogenetic experimental design and phylogenomic data exploration in R. *BMC Evolutionary Biology*, 16(1).

Douglas, J., Jiménez-Silva, C. L., and Bouckaert, R. (2022). StarBeast3: Adaptive Parallelized Bayesian Inference under the Multispecies Coalescent. *Systematic Biology*.

Ellis, A. G., Verboom, G. A., van der Niet, T., Johnson, S. D., Linder, H. P., Allsopp, N., Colville, J. F., and Marrone, P. (2014). Speciation and extinction in the greater Cape Floristic Region. In Allsopp, N., Colville, J. F., and Verboom, G. A., editors, *Fynbos: Ecology, evolution, and conservation of a Megadiverse Region*, chapter 6, pages 119–141. Oxford University Press, Oxford.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., and Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *Plos One*, 6(5):1–10.

Esterhuysen, E. E. (1963). Notes on South African species of *Erica*. *Journal of South African Botany*, 29:51–58.

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048.

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., and Glenn, T. C. (2012). Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales. *Systematic Biology*, 61(5):717–726.

Felice, F. D., Micheli, G., and Camilloni, G. (2019). Restriction enzymes and their use in molecular biology: An overview. *Journal of Biosciences*, 44(2).

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic zoology*, 19(2):99–113.

Folk, R. A., Mandel, J. R., and Freudenstein, J. V. (2015). A Protocol for Targeted Enrichment of Intron-Containing Sequence Markers for Recent Radiations: A Phylogenomic Example from *Heuchera* (Saxifragaceae). *Applications in Plant Sciences*, 3(8):1500039.

Folk, R. A., Soltis, P. S., Soltis, D. E., and Guralnick, R. (2018). New prospects in the detection and comparative analysis of hybridization in the tree of life. *American Journal of Botany*, 105(3).

Francis, R. M. (2017). pophelper: an R package and web app to analyse and visualize population structure. *Molecular Ecology Resources*, 17(1):27–32.

Frichot, E. and François, O. (2015). LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution*, 6(8):925–929.

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., and François, O. (2014). Fast and Efficient Estimation of Individual Ancestry Coefficients. *Genetics*, 196(4):973–983.

Gardner, E. (2021). HerbChomper: A bioinformatic tool for trimming poorly-aligned ends from DNA sequences. *Website https://github. com/artocarpus/HerbChomper [accessed 3 December 2021]*.

Gardner, E. M., Johnson, M. G., Pereira, J. T., Puad, A. S. A., Arifiani, D., Sahromi, Wickett, N. J., and Zerega, N. J. C. (2021). Paralogs and Off-Target Sequences Improve Phylogenetic Resolution in a Densely Sampled Study of the Breadfruit Genus (*Artocarpus*, Moraceae) . *Systematic Biology*, 70(3):558–575.

Garrison, E. (2012). Vcflib: A C++ library for parsing and manipulating VCF files. *Github Https://github. Com/ekg/vcflib*.

Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907*.

Gaston, K. J. (2009). Geographic range limits: achieving synthesis. *Proceedings of the Royal Society B: Biological Sciences*, 276(1661):1395–1406.

Gatesy, J., Sloan, D. B., Warren, J. M., Baker, R. H., Simmons, M. P., and Springer, M. S. (2019). Partitioned coalescence support reveals biases in species-tree methods and detects gene trees that determine phylogenomic conflicts. *Molecular Phylogenetics and Evolution*, 139(June):106539.

Gatesy, J. and Springer, M. S. (2014). Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Molecular Phylogenetics and Evolution*, 80(1):231–266.

Giarla, T. C. and Esselstyn, J. A. (2015). The challenges of resolving a rapid, recent radiation: Empirical and simulated phylogenomics of Philippine shrews. *Systematic Biology*, 64(5):727–740.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe, D. B., Lander, E. S., and Nusbaum, C. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, 27(2):182–189.

Goldblatt, P. and Manning, J. C. (2002). Plant diversity of the Cape region of southern Africa. *Annals of the Missouri Botanical Garden*, 89(2):281–302.

Gottlieb, L. (2004). Rethinking classic examples of recent speciation in plants. *New Phytologist*, 161(1):71–82.

Goudet, J. (2005). HIERFSTAT, a package for R to compute and test hierarchical F -statistics. *Molecular Ecology Notes*, 5:184–186.

Graybeal, A. (1994). Evaluating the phylogenetic utility of genes: a search for genes informative about deep divergences among vertebrates. *Systematic Biology*, 43(2):174–193.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., et al. (2010). A draft sequence of the Neandertal genome. *Science*, 328(5979):710–722.

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G., and Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics*, 43(10):1031–1034.

Grossenbacher, D. L., Veloz, S. D., and Sexton, J. P. (2014). Niche and range size patterns suggest that speciation begins in small, ecologically diverged populations in North American Monkeyflowers (*Mimulus* spp.). *Evolution*, 68(5):1270–1280.

Gruber, B., Unmack, P. J., Berry, O. F., and Georges, A. (2018). DartR: An R package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Molecular Ecology Resources*, 18(3):691–699.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59(3):307–321.

Hahn, M. W. and Nakhleh, L. (2016). Irrational exuberance for resolved species trees. *Evolution*, 70(1):7–17.

Hansma, J., Tohver, E., Schrank, C., Jourdan, F., and Adams, D. (2016). The timing of the Cape Orogeny: New 40 Ar/ 39 Ar age constraints on deformation and cooling of the Cape Fold Belt, South Africa. *Gondwana Research*, 32:122–137.

He, Q., Prado, J. R., and Knowles, L. L. (2017). Inferring the geographic origin of a range expansion: Latitudinal and longitudinal coordinates inferred from genomic data in an ABC framework with the program X-ORIGIN. *Molecular Ecology*, 26(24):6908–6920.

Helme, N. A. and Trinder-Smith, T. H. (2006). The endemic flora of the Cape Peninsula, South Africa. *South African Journal of Botany*, 72(2):205–210.

Hinojosa, J. C., Koubínová, D., Szenteczki, M., Pitteloud, C., Dincă, V., Alvarez, N., and Vila, R. (2019). A mirage of cryptic species: genomics uncover striking mito-nuclear discordance in the butterfly *Thymelicus sylvestris*. *Molecular Ecology*, 28(17):3857–3868.

Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, 35(2):518–522.

Hoffmann, V., Verboom, G. A., and Cotterill, F. P. (2015). Dated plant phylogenies resolve Neogene climate and landscape evolution in the Cape Floristic Region. *Plos One*, 10(9):e0137847.

Huang, J., Flouri, T., and Yang, Z. (2020). A Simulation Study to Examine the Information Content in Phylogenomic Data Sets under the Multispecies Coalescent Model. *Molecular Biology and Evolution*, 37(11):3211–3224.

Hudson, R. R., Slatkin, M., and Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics*, 132(2):583–589.

Hughes, C. and Eastwood, R. (2006). Island radiation on a continental scale: Exceptional rates of plant diversification after uplift of the Andes. *Proceedings of the National Academy of Sciences*, 103(27):10334–10339.

Inglis, P. W., Pappas, M. d. C. R., Resende, L. V., and Grattapaglia, D. (2018). Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. *Plos One*, 13(10):e0206085.

Jackman, S. D., Vandervalk, B. P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S. A., Jahesh, G., Khan, H., Coombe, L., Warren, R. L., and Birol, I. (2017). ABySS 2.0: resource-efficient assembly of large genomes using a bloom filter. *Genome Research*, 27(5):768–777.

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y. W., Faircloth, B. C., Nabholz, B., Howard, J. T., Suh, A., Weber, C. C., da Fonseca, R. R., Li, J., Zhang, F., Li, H., Zhou, L., Narula, N., Liu, L., Ganapathy, G., Boussau, B., Bayzid, M. S., Zavidovych, V., Subramanian, S., Gabaldón, T., Capella-Gutiérrez, S., Huerta-Cepas, J., Rekepalli, B., Munch, K., Schierup, M., Lindow, B., Warren, W. C., Ray, D., Green, R. E., Bruford, M. W., Zhan, X., Dixon, A., Li, S., Li, N., Huang, Y., Derryberry, E. P., Bertelsen, M. F., Sheldon, F. H., Brumfield, R. T., Mello, C. V., Lovell, P. V., Wirthlin, M., Schneider, M. P. C., Prosdocimi, F., Samaniego, J. A., Velazquez, A. M. V., Alfaro-Núñez, A., Campos, P. F., Petersen, B., Sicheritz-Ponten, T., Pas, A., Bailey, T., Scofield, P., Bunce, M., Lambert, D. M., Zhou, Q., Perelman, P., Driskell, A. C., Shapiro, B., Xiong, Z., Zeng, Y., Liu, S., Li, Z., Liu, B., Wu, K., Xiao, J., Yinqi, X., Zheng, Q., Zhang, Y., Yang, H., Wang, J., Smeds, L., Rheindt, F. E., Braun, M., Fjeldsa, J., Orlando, L., Barker, F. K., Jønsson, K. A., Johnson, W., Koepfli, K.-P., O'Brien, S., Haussler, D., Ryder, O. A., Rahbek, C., Willerslev, E., Graves, G. R., Glenn, T. C., McCormack, J., Burt, D., Ellegren, H., Alström, P., Edwards, S. V., Stamatakis, A., Mindell, D. P., Cracraft, J., Braun, E. L., Warnow, T., Jun, W., Gilbert, M. T. P., and Zhang, G. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–1331.

Jay, P., Whibley, A., Frézal, L., de Cara, M. Á. R., Nowell, R. W., Mallet, J., Dasmahapatra, K. K., and Joron, M. (2018). Supergene evolution triggered by the introgression of a chromosomal inversion. *Current Biology*, 28(11):1839–1845.e3.

Jiang, X., Edwards, S. V., and Liu, L. (2020). The Multispecies Coalescent Model Outperforms Concatenation Across Diverse Phylogenomic Data Sets. *Systematic Biology*, 69(4):795–812.

Johnson, M. G., Gardner, E. M., Liu, Y., Medina, R., Goffinet, B., Shaw, A. J., Zerega, N. J., and Wickett, N. J. (2016). HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences*, 4(7):1600016.

Johnson, M. G., Pokorny, L., Dodsworth, S., Botigué, L. R., Cowan, R. S., Devault, A., Eiserhardt, W. L., Epitawalage, N., Forest, F., Kim, J. T., Leebens-Mack, J. H., Leitch, I. J., Maurin, O., Soltis, D. E., Soltis, P. S., Wong, G. K. S., Baker, W. J., and Wickett, N. J. (2019). A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering. *Systematic Biology*, 68(4):594–606.

Johnson, S. D. (1996). Pollination, adaptation and speciation models in the Cape flora of South Africa. *TAXON*, 45(1):59–66.

Jombart, T. and Ahmed, I. (2011). adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics*, 27(21):3070–3071.

Jordan, C. Y., Lohse, K., Turner, F., Thomson, M., Gharbi, K., and Ennos, R. A. (2017). Maintaining their genetic distance: little evidence for introgression between widely hybridising species of *Geum* with contrasting mating systems. *Molecular Ecology*, 27:1214–1228.

Kadlec, M., Bellstedt, D. U., Le Maitre, N. C., and Pirie, M. D. (2017). Targeted NGS for species level phylogenomics: "made to measure" or "one size fits all"? *Peerj*, 5(e3569).

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., Von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14(6):587–589.

Kamada, T., Kawai, S., et al. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1):7–15.

Karin, B. R., Gamble, T., and Jackman, T. R. (2019). Optimizing Phylogenomics with Rapidly Evolving Long Exons: Comparison with Anchored Hybrid Enrichment and Ultraconserved Elements. *Molecular Biology and Evolution*, 37(3):904–922.

Karin, E. L., Mirdita, M., and Söding, J. (2020). MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome*, 8(1).

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780.

Kautt, A. F., Kratochwil, C. F., Nater, A., Machado-Schiaffino, G., Olave, M., Henning, F., Torres-Dowdall, J., Härer, A., Hulsey, C. D., Franchini, P., Pippel, M., Myers, E. W., and Meyer, A. (2020). Contrasting signatures of genomic divergence during sympatric speciation. *Nature*, 588(7836):106–111.

Kay, K. M. and Sargent, R. D. (2009). The Role of Animal Pollination in Plant Speciation: Integrating Ecology, Geography, and Genetics. *The Annual Review of Ecology Evolution and Systematics*, 40(September):637–656.

Kessler, C., Wootton, E., and Shafer, A. B. (2022). Speciation without gene-flow in hybridizing deer. *Biorxiv Evolutionary Biology*.

Klak, C., Bruyns, P. V., and Hanáček, P. (2013). A phylogenetic hypothesis for the recently diversified Ruschieae (Aizoaceae) in southern Africa. *Molecular Phylogenetics and Evolution*, 69:1005–1020.

Knaus, B. J. and Grünwald, N. J. (2017). VCFR: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources*, 17(1):44–53.

Knowles, L. L., Huang, H., Sukumaran, J., and Smith, S. A. (2018). A matter of phylogenetic scale: Distinguishing incomplete lineage sorting from lateral gene transfer as the cause of gene tree discord in recent versus deep diversification histories. *American Journal of Botany*, 105(3).

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., and Mayrose, I. (2015). CLUMPAK: A program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources*, pages 1179–1191.

Korunes, K. L. and Samuk, K. (2021). PIXY: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources*, 21(4):1359–1368.

Kuhl, H., Frankl-Vilches, C., Bakker, A., Mayr, G., Nikolaus, G., Boerno, S. T., Klages, S., Timmermann, B., and Gahr, M. (2020). An Unbiased Molecular Approach Using 3′-UTRs Resolves the Avian Family-Level Tree of Life. *Molecular Biology and Evolution*, 38(1):108–127.

Lanfear, R., Calcott, B., Kainer, D., Mayer, C., and Stamatakis, A. (2014). Selecting optimal partitioning schemes for phylogenomic datasets. *BMC evolutionary biology*, 14(1):1–14.

Le Maitre, N. C., Pirie, M. D., and Bellstedt, D. U. (2019a). An approach to determining anthocyanin synthesis enzyme gene expression in an evolutionary context: an example from *Erica plukenetii*. *Annals of Botany*, pages 1–9.

Le Maitre, N. C., Pirie, M. D., and Bellstedt, D. U. (2019b). Floral Color, Anthocyanin Synthesis Gene Expression and Control in Cape *Erica* Species. *Frontiers in Plant Science*, 10:1565.

Lenormand, T. (2002). Gene flow and the limits to natural selection. *Trends in Ecology & Evolution*, 17(4):183–189.

Levin, D. A. (1981). Dispersal Versus Gene Flow in Plants. *Annals of the Missouri Botanical Garden*, 68(2):233–253.

Lexer, C., Wüest, R. O., Mangili, S., Heuertz, M., Stölting, K. N., Pearman, P. B., Forest, F., Salamin, N., Zimmermann, N. E., and Bossolini, E. (2014). Genomics of the divergence continuum in an African plant biodiversity hotspot, I: Drivers of population divergence in *Restio capensis* (Restionaceae). *Molecular Ecology*, 23(17):4373–4386.

Li, C., Riethoven, J.-J. M., and Ma, L. (2010). Exon-primed intron-crossing (EPIC) markers for non-model teleost fishes. *BMC Evolutionary Biology*, 10(1):90.

Li, G., Figueiró, H. V., Eizirik, E., and Murphy, W. J. (2019). Recombination-Aware Phylogenomics Reveals the Structured Genomic Landscape of Hybridizing Cat Species. *Molecular Biology and Evolution*.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *Arxiv:1303.3997*.

Li, J.-N., He, C., Guo, P., Zhang, P., and Liang, D. (2017). A workflow of massive identification and application of intron markers using snakes as a model. *Ecology and Evolution*, 7(23):10042–10055.

Li, Q., Scornavacca, C., Galtier, N., and Chan, Y.-B. (2020). The Multilocus Multispecies Coalescent: A Flexible New Model of Gene Family Evolution. *Systematic Biology*, 70(4):822–837.

Lin, H. Y., Hao, Y. J., Li, J. H., Fu, C. X., Soltis, P. S., Soltis, D. E., and Zhao, Y. P. (2019). Phylogenomic conflict resulting from ancient introgression following species diversification in *Stewartia* s.l. (Theaceae). *Molecular Phylogenetics and Evolution*, 135.

Linder, H. P. (2003). The radiation of the Cape flora, southern Africa. *Biological Reviews*, 78(4):597–638.

Lu, M., Bond, W. J., Sheffer, E., Cramer, M. D., West, A. G., Allsopp, N., February, E. C., Chimphango, S., Ma, Z., Slingsby, J. A., and Hedin, L. O. (2022). Biome boundary maintained by intense belowground resource competition in world's thinnest-rooted plant community. *Proceedings of the National Academy of Sciences*, 119(9).

Lutsak, T. (2020). Coalescence-based species delimitation using genome-wide data reveals hidden diversity in a cosmopolitan group of lichens. *Organisms Diversity and Evolution*, 20(2):189–218.

Maddison, W. P. (1997). Gene Trees in Species Trees. *Systematic Biology*, 46(3):523–536.

Madeira, F., Pearce, M., Tivey, A. R. N., Basutkar, P., Lee, J., Edbali, O., Madhusoodanan, N., Kolesnikov, A., and Lopez, R. (2022). Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Research*, 50(W1):W276–W279.

Maier, R., Flegontov, P., Flegontova, O., Changmai, P., and Reich, D. (2022). On the limits of fitting complex models of population history to genetic data. *bioRxiv*.

Mallet, J. (2001). The speciation revolution. *Journal of Evolutionary Biology*, 14(6):887–888.

Mallet, J. (2007). Hybrid speciation. *Nature*, 446(7133):279–283.

Manning, J. and Goldblatt, P. (2012). *Plants of The Greater Cape Floristic Region 1: The Core Cape Flora*, volume 29. South African National Biodiversity Institute, Pretoria.

Matasci, N., Hung, L.-H., Yan, Z., Carpenter, E. J., Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Ayyampalayam, S., Barker, M., Burleigh, J. G., Gitzendanner, M. A., Wafula, E., Der, J. P., dePamphilis, C. W., Roure, B., Philippe, H., Ruhfel, B. R., Miles, N. W., Graham, S. W., Mathews, S., Surek, B., Melkonian, M., Soltis, D. E., Soltis, P. S., Rothfels, C., Pokorny, L., Shaw, J. A., DeGironimo, L., Stevenson, D. W., Villarreal, J. C., Chen, T., Kutchan, T. M., Rolf, M., Baucom, R. S., Deyholos, M. K., Samudrala, R., Tian, Z., Wu, X., Sun, X., Zhang, Y., Wang, J., Leebens-Mack, J., and Wong, G. K.-S. (2014). Data access for the 1,000 Plants (1KP) project. *GigaScience*, 3(1).

Mayr, E. (1999). *Systematics and the origin of species, from the viewpoint of a zoologist*. Harvard University Press.

McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C., and Brumfield, R. T. (2013). Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, 66(2):526–538.

McDade, L. (1990). Hybrids and phylogenetic systematics I. Patterns of character expression in hybrids and their implications for cladistic analysis. *Evolution*, 44(6):1685–1700.

McKain, M. R., Johnson, M. G., Uribe-Convers, S., Eaton, D., and Yang, Y. (2018). Practical considerations for plant phylogenomics. *Applications in Plant Sciences*, 6(3):e1038.

McLay, T. G., Birch, J. L., Gunn, B. F., Ning, W., Tate, J. A., Nauheimer, L., Joyce, E. M., Simpson, L., Schmidt-Lebuhn, A. N., Baker, W. J., et al. (2021). New targets acquired: Improving locus recovery from the Angiosperms353 probe set. *Applications in Plant Sciences*, 9(7).

Meier, R. (2017). Citation of taxonomic publications: the why, when, what and what not. *Systematic Entomology*, 42(2):301–304.

Meiklejohn, K. A., Faircloth, B. C., Glenn, T. C., Kimball, R. T., and Braun, E. L. (2016). Analysis of a rapid evolutionary radiation using ultraconserved elements: Evidence for a bias in some multispecies coalescent methods. *Systematic Biology*, 65(4):612–627.

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prüfer, K., De Filippo, C., et al. (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science*, 338(6104):222–226.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., Lanfear, R., and Teeling, E. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5):1530–1534.

Minnaar, C., de Jager, M. L., and Anderson, B. (2019). Intraspecific divergence in floral-tube length promotes asymmetric pollen movement and reproductive isolation. *New Phytologist*, 224(3):1160–1170.

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research*, 41(12):e121–e121.

Molloy, E. K. and Warnow, T. (2017). To include or not to include: The Impact of Gene Filtering on Species Tree Estimation Methods. *Systematic Biology*, 67(2):285–303.

Mondal, M., Bertranpetit, J., and Lao, O. (2019). Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. *Nature Communications*, 10(1):1–9.

Morgan, M., Pagès, H., Obenchain, V., and Hayden, N. (2021). *Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import*. R package version 2.10.0.

Morjan, C. L. and Rieseberg, L. H. (2004). How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. *Molecular Ecology*, 13(6):1341–56.

Morris, J. L., Puttick, M. N., Clark, J. W., Edwards, D., Kenrick, P., Pressel, S., Wellman, C. H., Yang, Z., Schneider, H., and Donoghue, P. C. J. (2018). The timescale of early land plant evolution. *Proceedings of the National Academy of Sciences*, 115(10).

Mořkovský, L., Janoušek, V., Reif, J., Rídl, J., Pačes, J., Choleva, L., Janko, K., Nachman, M. W., and Reifová, R. (2018). Genomic islands of differentiation in two songbird species reveal candidate genes for hybrid female sterility. *Molecular Ecology*, 27(4):949–958.

Mugrabi De Kuppler, A. L. (2013). Phylogenetics, flow-cytometry and pollen storage in *Erica* L. (Ericaceae). *Phd Thesis*, pages 1–76.

Mugrabi De Kuppler, A. L., Fagúndez, J., Bellstedt, D. U., Oliver, E. G. H., Léon, J., and Pirie, M. D. (2015). Testing reticulate versus coalescent origins of *Erica lusitanica* using a species phylogeny of the northern heathers (Ericeae , Ericaceae). *Molecular Phylogenetics and Evolution*, 88:121–131.

Muir, R. A., Bordy, E. M., Reddering, J. S. V., and Viljoen, J. H. A. (2017). Lithostratigraphy of the Enon Formation (Uitenhage Group), South Africa. *South African Journal of Geology*, 120(2):273–280.

Murat, F., Peer, Y. V. d., and Salse, J. (2012). Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes. *Genome Biology and Evolution*, 4(9):917–928.

Nelson, E. and Oliver, E. (2004). Cape heaths in European gardens: the early history of South African *Erica* species in cultivation, their deliberate hybridization and the orthographic bedlam. *Bothalia*, 34(2):127–140.

Nelson, E. C. and Oliver, E. G. H. (2005). Chromosome numbers in *Erica* - an updated checklist. In *Yearbook of the Heather Society 2005*, volume 53, pages 57–58. Heather Society, Ipswich.

Nelson, T. C., Stathos, A. M., Vanderpool, D. D., Finseth, F. R., wu Yuan, Y., and Fishman, L. (2021). Ancient and recent introgression shape the evolutionary history of pollinator adaptation and speciation in a model monkeyflower radiation *Mimulus* section *Erythranthe*. *PLoS Genetics*, 17(2):e1009095.

Newman, E. and Johnson, S. D. (2021). A shift in long-proboscid fly pollinators and floral tube length among populations of *Erica junonia* (Ericaceae). *South African Journal of Botany*, 142:451–458.

Nge, F. J., Biffin, E., Thiele, K. R., and Waycott, M. (2021). Reticulate Evolution, Ancient Chloroplast Haplotypes, and Rapid Radiation of the Australian Plant Genus Adenanthos (Proteaceae). *Frontiers in Ecology and Evolution*, 8.

Nosil, P. (2008). Speciation with gene flow could be common. *Molecular Ecology*, 17(9):2103–2106.

Nürk, N. M., Linder, H. P., Onstein, R. E., Larcombe, M. J., Hughes, C. E., Fernández, L. P., Schlüter, P. M., Valente, L., Beierkuhnlein, C., Cutts, V., Donoghue, M. J., Edwards, E. J., Field, R., Flantua, S. G. A., Higgins, S. I., Jentsch, A., Liede-Schumann, S., and Pirie, M. D. (2020). Diversification in evolutionary arenas — Assessment and synthesis. *Ecology and Evolution*, 10(12):6163–6182.

Ojeda, F., Budde, K. B., Heuertz, M., Segarra-Moragues, J. G., and González-Martínez, S. C. (2016). Biogeography and evolution of seeder and resprouter forms of *Erica coccinea* (Ericaceae) in the fire-prone Cape fynbos. *Plant Ecology*, 217(6):751–761.

Oliver, E. G. and Forshaw, N. (2012). *Erica* identification v. 3.00. Electronic Guide.

Oliver, E. G. H. (1977). The identity of *Erica flavisepala*. *Bothalia*, 12(2):195–197.

Oliver, E. G. H. (1986). The identity of *Erica vinacea* and notes on hybridization in *Erica*. *Bothalia*, 16(1):35–38.

Oliver, E. G. H. and Oliver, I. M. (2002). The genus *Erica* (Ericaceae) in southern Africa: Taxonomic notes 1. *Bothalia*, 32(1):37–61.

Oliver, E. G. H. and Oliver, I. M. (2005). The genus *Erica* (Ericaceae) in southern Africa: taxonomic notes 2. *Bothalia*, 35(2):121–148.

Ortiz, E. M. (2019). vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis. https://doi.org/10.5281/zenodo.2540861.

Otero, A., Vargas, P., Fernández-Mazuecos, M., Jiménez-Mejías, P., Valcárcel, V., Villa-Machío, I., and Hipp, A. L. (2022). A snapshot of progenitor–derivative speciation in *Iberodes* (Boraginaceae). *Molecular Ecology*, 31(11):3192–3209.

Ottenburghs, J. (2020). Ghost Introgression: Spooky Gene Flow in the Distant Past. *BioEssays*, 42(6):2000012.

Overcast, I., Ruffley, M., Rosindell, J., Harmon, L., Borges, P. A. V., Emerson, B. C., Etienne, R. S., Gillespie, R., Krehenwinkel, H., Mahler, D. L., Massol, F., Parent, C. E., Patiño, J., Peter, B., Week, B., Wagner, C., Hickerson, M. J., and Rominger, A. (2021). A unified model of species abundance, genetic diversity, and functional diversity reveals the mechanisms structuring ecological communities. *Molecular Ecology Resources*, 21(8):2782–2800.

Paradis, E. (2013). Molecular dating of phylogenies by likelihood methods: A comparison of models and a new information criterion. *Molecular Phylogenetics and Evolution*, 67(2):436–444.

Paradis, E. and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35:526–528.

Peter, B. M. (2016). Admixture, Population Structure, and F-Statistics. *Genetics*, 202(4):1485–1501.

Peter, B. M. and Slatkin, M. (2013). Detecting range expansions from genetic data. *Evolution*, 67(11):3274–3289.

Pirie, M. D., Blackhall-Miles, R., Bourke, G., Crowley, D., Ebrahim, I., Forest, F., Knaack, M., Koopman, R., Lansdowne, A., Nürk, N. M., Osborne, J., Pearce, T. R., Rohrauer, D., Smit, M., and Wilman, V. (2022). Preventing species extinctions: A global conservation consortium for *Erica*. *PLANTS, PEOPLE, PLANET*, 4(4):335–344.

Pirie, M. D., Kandziora, M., Nürk, N. M., Le Maitre, N. C., Mugrabi De Kuppler, A., Gehrke, B., Oliver, E. G., and Bellstedt, D. U. (2019). Leaps and bounds: geographical and ecological distance constrained the colonisation of the Afrotemperate by *Erica*. *BMC Evolutionary Biology*, 19(1):0–26.

Pirie, M. D., Oliver, E. G., Gehrke, B., Heringer, L., De Kuppler, A. M., Le Maitre, N. C., and Bellstedt, D. U. (2017). Underestimated regional species diversity in the Cape Floristic Region revealed by phylogenetic analysis of the *Erica abietina/E. viscaria* clade (Ericaceae). *Botanical Journal of the Linnean Society*, 184(2):185–203.

Pirie, M. D., Oliver, E. G. H., and Bellstedt, D. U. (2011). A densely sampled ITS phylogeny of the Cape flagship genus *Erica* L. suggests numerous shifts in floral macro-morphology. *Molecular Phylogenetics and Evolution*, 61(2):593–601.

Pirie, M. D., Oliver, E. G. H., Kuppler, A. M. D., Gehrke, B., Maitre, N. C. L., Kandziora, M., and Bellstedt, D. U. (2016). The biodiversity hotspot as evolutionary hot-bed: spectacular radiation of *Erica* in the Cape Floristic Region. *Bmc Evolutionary Biology*.

Porter, C. K., Confer, J. L., Aldinger, K. R., Canterbury, R. A., Larkin, J. L., and McNeil, D. J. (2021). Strong yet incomplete reproductive isolation in *Vermivora* is not contradicted by other lines of evidence: A reply to Toews et al. *Ecology and Evolution*, 11(15):1–7.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155:945–959.

Prum, R. O., Berv, J. S., Dornburg, A., Field, D. J., Townsend, J. P., Lemmon, E. M., and Lemmon, A. R. (2015). A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature*, 526(7574):569–573.

Prunier, R., Akman, M., Kremer, C. T., Aitken, N., Chuah, A., Borevitz, J., and Holsinger, K. E. (2017). Isolation by distance and isolation by environment contribute to population differentiation in *Protea repens* (Proteaceae L.), a widespread South African species. *American Journal of Botany*, 104(5):674–684.

Prunier, R. and Holsinger, K. E. (2010). Was it an explosion? Using population genetics to explore the dynamics of a recent radiation within *Protea* (Proteaceae L.). *Molecular Ecology*, 19(18):3968–3980.

Puntambekar, S., Newhouse, R., Chauhan, R., and Willis, T. (2020). Rapid speciation of cichlids fishes may be explained by evolutionary divergence of novel open reading frames. *Biorxiv Evolutionary Biology*.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., and Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3):559–575.

Puritz, J. B., Hollenbeck, C. M., and Gold, J. R. (2014a). *dDocent*: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *Peerj*, 2:e431.

Puritz, J. B., Matz, M. V., Toonen, R. J., Weber, J. N., Bolnick, D. I., and Bird, C. E. (2014b). Demystifying the RAD fad. *Molecular Ecology*, 23(24):5937–5942.

Pyron, R. A., O'Connell, K. A., Lemmon, E. M., Lemmon, A. R., and Beamer, D. A. (2022). Candidate-species delimitation in *Desmognathus* salamanders reveals gene flow across lineage boundaries, confounding phylogenetic estimation and clarifying hybrid zones. *Ecology and Evolution*, 12(2).

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabosky, D. L., Slater, G. J., and Alfaro, M. E. (2012). Clade Age and Species Richness Are Decoupled Across the Eukaryotic Tree of Life. *Plos Biol*, 10(8):1001381.

Racimo, F., Sankararaman, S., Nielsen, R., and Huerta-Sánchez, E. (2015). Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, 16(6):359–371.

Raxworthy, C. J. and Smith, B. T. (2021). Mining museums for historical DNA: advances and challenges in museomics. *Trends in Ecology & Evolution*, 36(11):1049–1060.

Rebelo, A., Siegfried, W., and Crowe, A. (1984). Avian pollinators and the pollination syndromes of selected mountain fynbos plants. *South African Journal of Botany*, 3(5):285–296.

Rebelo, A., Siegfried, W., and Oliver, E. (1985). Pollination syndromes of *Erica* species in the south-western Cape. *South African Journal of Botany*, 51(4):270–280.

Reddy, S., Kimball, R. T., Pandey, A., Hosner, P. A., Braun, M. J., Hackett, S. J., Han, K.-L., Harshman, J., Huddleston, C. J., Kingston, S., Marks, B. D., Miglia, K. J., Moore, W. S., Sheldon, F. H., Witt, C. C., Yuri, T., and Braun, E. L. (2017). Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. *Systematic Biology*, 66(5):857–879.

Revell, L. J. (2012). phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3:217–223.

Richards, E. J., Brown, J. M., Barley, A. J., Chong, R. A., and Thomson, R. C. (2018). Variation across mitochondrial gene trees provides evidence for systematic error: How much gene tree variation is biological? *Systematic Biology*, 67(5):847–860.

Rieseberg, L. H. and Wendel, J. F. (1993). Introgression and its consequences in plants. *Hybrid zones and the evolutionary process*, 70:109.

Rieseberg, L. H. and Willis, J. H. (2007). Plant speciation. *Science*, 317(5840):910–914.

Roberts, A., Hockey, P., Dean, W., and Ryan, P. (2005). *Roberts Birds of Southern Africa*. Trustees of the J. Voelcker Bird Book Fund.

Roch, S. and Warnow, T. (2015). On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. *Systematic Biology*, 64(4):663–676.

Rochette, N. C., Rivera-Colón, A. G., and Catchen, J. M. (2019). Stacks 2: analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular Ecology*, 28(21):4737–4754.

Rodríguez, A., Burgon, J. D., Lyra, M., Irisarri, I., Baurain, D., Blaustein, L., Göçmen, B., Künzel, S., Mable, B. K., Nolte, A. W., Veith, M., Steinfartz, S., Elmer, K. R., Philippe, H., and Vences, M. (2017). Inferring the shallow phylogeny of true salamanders (Salamandra) by multiple phylogenomic approaches. *Molecular Phylogenetics and Evolution*, 115:16–26.

Roycroft, E. J., Moussalli, A., and Rowe, K. C. (2019). Phylogenomics Uncovers Confidence and Conflict in the Rapid Radiation of Australo-Papuan Rodents. *Systematic Biology*, 69(3):431–444.

Salichos, L. and Rokas, A. (2013). Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, 497(7449):327–331.

Sayyari, E. and Mirarab, S. (2016). Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. *Molecular Biology and Evolution*, 33(7):1654–1668.

Scharf, T. E., Codilean, A. T., de Wit, M., Jansen, J. D., and Kubik, P. W. (2013). Strong rocks sustain ancient postorogenic topography in southern Africa. *Geology*, 41(3):331–334.

Schley, R. J., Pennington, R. T., Pérez-Escobar, O. A., Helmstetter, A. J., de la Estrella, M., Larridon, I., Sabino Kikuchi, I. A. B., Barraclough, T. G., Forest, F., and Klitgård, B. (2020). Introgression across evolutionary scales suggests reticulation contributes to Amazonian tree diversity. *Molecular Ecology*, 29(21):4170–4185.

Schnitzler, J., Barraclough, T. G., Boatwright, J. S., Goldblatt, P., Manning, J. C., Powell, M. P., Rebelo, T., and Savolainen, V. (2011). Causes of plant diversification in the Cape biodiversity hotspot of South Africa. *Systematic Biology*, 60(3):343–357.

Schrempf, D., Minh, B. Q., von Haeseler, A., and Kosiol, C. (2019). Polymorphism-Aware Species Trees with Advanced Mutation Models, Bootstrap, and Rate Heterogeneity. *Molecular Biology and Evolution*, 36(6):1294–1301.

Sedlazeck, F. J., Rescheneder, P., and Von Haeseler, A. (2013). NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics*, 29(21):2790–2791.

Shah, N., Nute, M. G., Warnow, T., and Pop, M. (2019). Misunderstood parameter of NCBI BLAST impacts the correctness of bioinformatics workflows. *Bioinformatics*, 35(9):1613–1614.

Shone, R. and Booth, P. (2005). The Cape Basin, South Africa: A review. *Journal of African Earth Sciences*, 43(1-3):196–210.

Shrestha, M., Dyer, A. G., Boyd-Gerny, S., Wong, B. B., and Burd, M. (2013). Shades of red: Bird-pollinated flowers target the specific colour discrimination abilities of avian vision. *New Phytologist*, 198(1):301–310.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212.

Simmons, M. and Cowling, R. (1996). Why is the Cape Peninsula so rich in plant species? An analysis of the independent diversity components. *Biodiversity & Conservation*, 5(5):551–573.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., and Birol, İ. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research*, 19(6):1117–1123.

Slater, G. and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6(1):31.

Slingsby, J. a., Britton, M. N., and Anthony Verboom, G. (2014). Ecology limits the diversity of the Cape flora: Phylogenetics and diversification of the genus *Tetraria*. *Molecular Phylogenetics and Evolution*, 72:61–70.

Smith, M. L. and Hahn, M. W. (2021). New approaches for inferring phylogenies in the presence of paralogs. *Trends in Genetics*, 37(2):174–187.

Sobel, J. M., Chen, G. F., Watt, L. R., and Schemske, D. W. (2010). The biology of speciation. *Evolution*, 64(2):295–315.

Solís-Lemus, C., Bastide, P., and Ané, C. (2017). PhyloNetworks: a package for phylogenetic networks. *Molecular Biology and Evolution*, 34(12):3292–3298.

Soltis, D. E. and Kuzoff, R. K. (1995). Discordance between nuclear and chloroplast phylogenies in the *Heuchera* group (Saxifragaceae). *Evolution*, 49(4):727–742.

Soltis, P. S., Marchant, D. B., de Peer, Y. V., and Soltis, D. E. (2015). Polyploidy and genome evolution in plants. *Current Opinion in Genetics & Development*, 35:119–125.

Soltis, P. S. and Soltis, D. E. (2020). Plant genomes: Markers of evolutionary history and drivers of evolutionary change. *PLANTS, PEOPLE, PLANET*, 3(1):74–82.

Sonah, H., Bastien, M., Iquira, E., Tardivel, A., Légaré, G., Boyle, B., Normandeau, É., Laroche, J., Larose, S., Jean, M., and Belzile, F. (2013). An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping. *PLoS ONE*, 8(1):e54603.

Soza, V. L., Lindsley, D., Waalkes, A., Ramage, E., Patwardhan, R. P., Burton, J. N., Adey, A., Kumar, A., Qiu, R., Shendure, J., and Hall, B. (2019). The *Rhododendron* Genome and Chromosomal Organization Provide Insight into Shared Whole-Genome Duplications across the Heath Family (Ericaceae). *Genome Biology and Evolution*, 11(12):3353–3371.

Springer, M. and Gatesy, J. (2018). Delimiting coalescence genes (c-genes) in phylogenomic data sets. *Genes*, 9(3):123.

Springer, M. S. and Gatesy, J. (2016). The gene tree delusion. *Molecular Phylogenetics and Evolution*, 94:1–33.

Springer, M. S., Molloy, E. K., Sloan, D. B., Simmons, M. P., and Gatesy, J. (2019). ILS-aware analysis of low-homoplasy retroelement insertions: Inference of species trees and introgression using quartets. *Journal of Heredity*, 111(2):147–168.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

Steenwyk, J. L., Buida III, T. J., Li, Y., Shen, X.-X., and Rokas, A. (2020). ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS biology*, 18(12):e3001007.

Steinig, E. J., Neuditschko, M., Khatkar, M. S., Raadsma, H. W., and Zenger, K. R. (2016). NETVIEW P: a network visualization tool to unravel complex population structure using genome-wide SNPs. *Molecular Ecology Resources*, 16(1):216–227.

Stock, W. D. and Verboom, G. A. (2012). Phylogenetic ecology of foliar N and P concentrations and N:P ratios across mediterranean-type ecosystems. *Global Ecology and Biogeography*, 21(12):1147–1156.

Stoffel, M. A., Esser, M., Kardos, M., Humble, E., Nichols, H., David, P., and Hoffman, J. I. (2016). inbreedR: an R package for the analysis of inbreeding based on genetic markers. *Methods in Ecology and Evolution*, 7(11):1331–1339.

Svardal, H., Quah, F. X., Malinsky, M., Ngatunga, B. P., Miska, E. A., Salzburger, W., Genner, M. J., Turner, G. F., and Durbin, R. (2019). Ancestral Hybridization Facilitated Species Diversification in the Lake Malawi Cichlid Fish Adaptive Radiation. *Molecular Biology and Evolution*, 37(4):1100–1113.

Szarmach, S. J., Brelsford, A., Witt, C. C., and Toews, D. P. (2021). Comparing divergence landscapes from reduced-representation and whole genome resequencing in the yellow-rumped warbler (*Setophaga coronata*) species complex. *Molecular Ecology*, 30(23):5994–6005.

Tan, G., Muffato, M., Ledergerber, C., Herrero, J., Goldman, N., Gil, M., and Dessimoz, C. (2015). Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference. *Systematic Biology*, 64(5):778–791.

Tange, O. (2020). GNU Parallel 20200522 ('Kraftwerk'). Zenodo: https://doi.org/10.5281/zenodo.3841377.

Tank, D. C., Eastman, J. M., Pennell, M. W., Soltis, P. S., Soltis, D. E., Hinchliff, C. E., Brown, J. W., Sessa, E. B., and Harmon, L. J. (2015). Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytologist*, 207(2):454–467.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17(2):57–86.

Templeton, A. R. (1981). Mechanisms of Speciation – A Population Genetic Approach. *Annual Review of Ecology and Systematics*, 12:23–48.

The Marie Curie SPECIATION Network (2012). What do we need to know about speciation? *Trends in Ecology and Evolution*, 27(1):27–39.

Townsend, J. P., Su, Z., and Tekle, Y. I. (2012). Phylogenetic signal and noise: predicting the power of a data set to resolve phylogeny. *Systematic Biology*, 61(5):835.

Trinder-Smith, H., Cowling, R., and Linder, H. (1996). Profiling a besieged flora: endemic and threatened plants of the Cape Peninsula, South Africa. *Biodiversity & Conservation*, 5(5):575–589.

Ufimov, R., Gorospe, J. M., Fér, T., Kandziora, M., Salomon, L., van Loo, M., and Schmickl, R. (2022). Utilizing paralogues for phylogenetic reconstruction has the potential to increase species tree support and reduce gene tree discordance in target enrichment data. *Molecular Ecology Resources*.

van der Niet, T., Pires, K., and Steenhuisen, S. L. (2020). Flower constancy of the Cape honey bee pollinator of two co-flowering *Erica* species from the Cape Floristic Region (South Africa). *South African Journal of Botany*, 132:371–377.

van Der Niet, T., Pirie, M. D., Shuttleworth, A., Johnson, S. D., and Midgley, J. J. (2014). Do pollinator distributions underlie the evolution of pollination ecotypes in the Cape shrub *Erica plukenetii*? *Annals of Botany*, 113(2):301–315.

van Santen, M. and Linder, H. P. (2020). The assembly of the Cape flora is consistent with an edaphic rather than climatic filter. *Molecular Phylogenetics and Evolution*, 142:106645.

Verboom, G. A., Archibald, J. K., Bakker, F. T., Bellstedt, D. U., Conrad, F., Dreyer, L. L., Forest, F., Galley, C., Goldblatt, P., Henning, J. F., Mummenhoff, K., Linder, H. P., Muasya, A. M., Oberlander, K. C., Savolainen, V., Snijman, D. A., van der Niet, T., and Nowell, T. L. (2009). Origin and diversification of the Greater Cape flora: Ancient species repository, hot-bed of recent radiation, or both? *Molecular Phylogenetics and Evolution*, 51(1):44–53.

Verboom, G. A., Bergh, N. G., Haiden, S. A., Hoffmann, V., and Britton, M. N. (2015). Topography as a driver of diversification in the Cape Floristic Region of South Africa. *The New Phytologist*, 207(2):368–376.

Verboom, G. A., Stock, W. D., and Cramer, M. D. (2017). Specialization to Extremely Low-Nutrient Soils Limits the Nutritional Adaptability of Plant Lineages. *The American Naturalist*, 189(6):684–699.

Via, S. (2009). Natural selection in action during speciation. *Proceedings of the National Academy of Sciences*, 106(supplement_1):9939–9946.

Weber, M. G., Wagner, C. E., Best, R. J., Harmon, L. J., and Matthews, B. (2017). Evolution in a Community Context: On Integrating Ecological Interactions and Macroevolution. *Trends in Ecology and Evolution*, 32(4):291–304.

Weir, B. S. and Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, 38(6):1358–1370.

Westbury, M. V., Petersen, B., and Lorenzen, E. D. (2019). Genomic analyses reveal an absence of contemporary introgressive admixture between fin whales and blue whales, despite known hybrids. *Plos One*, 14(9):e0222004.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Yang, F. S., Nie, S., Liu, H., Shi, T. L., Tian, X. C., Zhou, S. S., Bao, Y. T., Jia, K. H., Guo, J. F., Zhao, W., An, N., Zhang, R. G., Yun, Q. Z., Wang, X. Z., Mannapperuma, C., Porth, I., El-Kassaby, Y. A., Street, N. R., Wang, X. R., Van de Peer, Y., and Mao, J. F. (2020). Chromosome-level genome assembly of a parent species of widely cultivated azaleas. *Nature Communications*, 11(1).

Yang, Z. (1998). On the best evolutionary rate for phylogenetic analysis. *Systematic Biology*, 47(1):125–133.

Yu, G., Smith, D., Zhu, H., Guan, Y., and Lam, T. T.-Y. (2017). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8:28–36.

Zhang, C. and Mirarab, S. (2022). Weighting by Gene Tree Uncertainty Improves Accuracy of Quartet-based Species Trees. *Molecular Biology and Evolution*. Advance online publication. https://doi.org/10.1093/molbev/msac215.

Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018). ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19(Suppl 6):15–30.

Zhang, C., Scornavacca, C., Molloy, E. K., and Mirarab, S. (2020). ASTRAL-Pro: Quartet-Based Species-Tree Inference despite Paralogy. *Molecular Biology and Evolution*, 37(11):3292–3307.

Zhang, L., Xu, P., Cai, Y., Ma, L., Li, S., Li, S., Xie, W., Song, J., Peng, L., Yan, H., Zou, L., Ma, Y., Zhang, C., Gao, Q., and Wang, J. (2017). The draft genome assembly of *Rhododendron delavayi* Franch. var. *delavayi*. *GigaScience*, 6(10).

Zhang, Y., Deng, S., Liang, D., and Zhang, P. (2019). Sequence capture across large phylogenetic scales by using pooled PCR-generated baits: A case study of Lepidoptera. *Molecular Ecology Resources*, 19(4):1037–1051.

Zhao, T., Zwaenepoel, A., Xue, J.-Y., Kao, S.-M., Li, Z., Schranz, M. E., and Van de Peer, Y. (2021). Whole-genome microsynteny-based phylogeny of angiosperms. *Nature Communications*, 12(1):1–14.

Zhou, W., Soghigian, J., and Xiang, Q.-Y. (2022). A new pipeline for removing paralogs in target enrichment data. *Systematic Biology*, 71(2):410–425.

Appendix A

DNA extraction protocol for *Erica* leaf material

Note

This protocol was adapted from that of Inglis et al. (2018). It works best on young leaf material (in *Erica* the youngest leaves are located at the branch tips). If using fresh material, it is best to grind the leaves using liquid N, but they can also be ground in SWB (see below) in 2 mL Eppendorf tubes with steel beads using a TissueLyser or a similar product.

The primary modifications to the Inglis et al. (2018) protocol are as follows:

1. Instead of CTAB, SDS is used as the detergent for cell lysis. This allows for a purification step in which K Acetate is added after lysis, causing proteins and polysaccharides to precipitate along with SDS during a cooling step.
2. A combination of NaCl and Na Acetate is used to inhibit the co-precipitation of polysaccharides with DNA after the addition of isopropanol. NaCl increases the solubility of polysaccharides but not DNA, but can also cause the isopropanol to come out of solution. By experimentation, the addition of Na Acetate was found to prevent the latter from occurring.

Materials

- SWB: Sorbitol wash buffer (2,000 µL per sample): 100 mM Tris-HCl pH 8.0, 0.35 M Sorbitol, 1% PVP, 10 mM EDTA pH 8.0. **STORE AT 4°C** for up to 6 months.
- 20% SDS (Sodium dodecyl sulphate) (80 µL per sample)
- Extraction buffer (800 µL per sample): 500mM NaCl, 100mMTris-HCl, 50mM EDTA, adjusted to pH 8.0.
- RNase A: 100 mg/mL (Qiagen Cat. No. / ID: 19101)
- 5M K Acetate (KAc; **STORE AT 4°C.**) (80 µL per sample)
- CIA: Chloroform:Isoamyl alcohol (24:1) (600 µL per sample).
- 5M NaCl: (150 µL per sample).
- 3M Na Acetate, pH 5.2 (50 µL per sample).
- Eppendorf tubes: 1 x 2 mL tube, 2 x 1.5 mL tubes.
- Isopropanol: 500 µL per sample.
- Ethanol (70%): 1000 µL per sample.
- TE elution buffer: 105 µL per sample.

Protocol

1. Prepare a working solution of Extraction buffer, adding 2 µL RNase A per sample and 1% β-mercaptoethanol.
2. Prepare a working solution of SWB, adding 1% β-mercaptoethanol.
3. Prepare water bath or oven at 65°C.
4. Weigh out 25-35 mg dry leaf material per sample.
5. Add dry leaf material to 2 mL Eppendorf tube with at least two grinding beads. I used two 2.5 mm and two to three 1 mm diameter steel grinding beads.
6. Grind in TissueLyser for 5 min. @ 30 Hz.
7. Add 1,000 µL SWB, then spin @ 6,000 *x g* for 5 min. **Important**: work quickly, as some DNA degradation can occur if this step is prolonged. Take note of the supernatant's viscosity colour. Discard the supernatant and **repeat** the wash if the supernatant was dark and/or viscous.
8. Add 800 µL Extraction buffer, followed by 80 µL 20% SDS.

9. Mix well and place on heat at 65°C for 45 min. to 1 hour, turning every 10-15 min.
10. After lysis, add 80 µL KAc and mix well.
11. Place in -20°C freezer for 10-15 min. **Important**: do not leave in the freezer for longer! The solution needs to cool but must not freeze. Alternatively, the tubes can be placed on ice for a longer period (> 30 min.)
12. While waiting, prepare new 1.5 mL Eppendorf tubes and add 600 µL CIA to each.
13. Remove samples from freezer and spin @ 6,000 *x g* for 5 min. Recover 800 µL supernatant into the Eppendorf tubes with CIA.
14. Place in tissuelyser and shake for 1 min. @ 10 Hz (Note: make sure tubes are securely shut! I usually avoid using the outer wells of our tube holders as their lids don't sit as flush on the tube lids at the edges).
15. Spin @ 17,000 *x g* for 8 min.
16. Prepare new 1.5 mL Eppendorf tubes and add 150 µL 5M NaCl and 50 µL Na Acetate to each.
17. Recover *ca.* 700 µL of the aqueous phase into the new tubes.
18. Add 500 µL cold isopropanol and place in -20°C freezer for *ca.* 15 min. **Important**: do not extend this step beyond *ca.* 30 min. as the salts may precipitate.
19. Spin @ 17,000 *x g* for 10 min. Discard liquid and place inverted on two layers of kimwipe. (**Note**: take care not to lose the pellets at this stage!) Allow to dry for 8 min. Turn tubes on their side if the pellet seems loose.
20. Add 1,000 µL 70% EtOH.
21. Flick/shake tubes to loosen pellet.
22. Spin @ 17,000 *x g* for 3 min.
23. Discard ethanol (carefully!) and invert briefly on fresh kimwipe, taking care not to lose the pellet. Seal tube immediately after inversion and place back in centrifuge.
24. Briefly spin down (10 seconds).
25. Pipette out residual ethanol and leave tubes open in rack for 5 min.
26. Check that no droplets remain and elute in 105 µL TE (or as desired – enough for qubit, nanodrop and gel electrophoresis + volume required for sequencing).

Appendix B

Voucher tables

Table B.1 Voucher information of samples with target capture data (Chapters 2 and 3). Unless otherwise noted, collections were made by the author. Specimens have been deposited at NBG.

Voucher No.	Sample No.	Organism	Latitude	Longitude	iNaturalist*	Collector	Note
SM227	S379B	*E. abietina abietina*	-33.955299	18.424203	25265963	-	-
SM228	S380	*E. abietina abietina*	-33.955805	18.424214	25265904	-	-
SM231	S382	*E. abietina abietina*	-33.955778	18.427173	25265913	-	-
SM269	S384	*E. abietina abietina*	-33.989500	18.413416	28425132	-	-
SM270	S385	*E. abietina abietina*	-33.989421	18.413170	28425133	-	-
SM271	S386	*E. abietina abietina*	-33.988890	18.412935	28425137	-	-
SM272	S387	*E. abietina abietina*	-33.988911	18.411541	28425139	-	-
SM466	S353pB	*E. abietina atrorosea*	-34.090397	18.421659	62543295	-	-
SM475	TC082	*E. abietina atrorosea*	-34.056584	18.372305	63159621	-	-
SM479	TC083	*E. abietina atrorosea* x *E. a. constantiana*	-34.058591	18.374384	63162993	-	-
SM403	TC124	*E. abietina atrorosea* x *E. viscaria viscaria*	-34.101185	18.394171	39853366	-	-
SM415	TC057	*E. abietina constantiana*	-33.999237	18.400214	40627237	-	-
SM416	TC072z	*E. abietina constantiana*	-33.998168	18.400628	40647963	-	-
SM451	TC155B	*E. abietina constantiana*	-34.022580	18.401466	57652605	-	-
SM453	TC063	*E. abietina constantiana*	-34.022344	18.404076	57653130	-	-
SM480	TC065	*E. abietina constantiana*	-34.058591	18.374384	63160250	-	-
SM371	TC043	*E. abietina diabolis*	-33.952053	18.446026	37291642	-	-
SM372	TC123	*E. abietina diabolis*	-33.951908	18.446146	37291782	-	-
SM373	TC053	*E. abietina diabolis*	-33.951878	18.445099	37452995	-	-
SM374	TC028	*E. abietina diabolis*	-33.953216	18.439782	37482859	-	-
SM375	TC184	*E. abietina diabolis*	-33.953356	18.439866	37482860	-	-
SM376	TC044	*E. abietina diabolis*	-33.954363	18.438808	37482861	-	-
SM377	TC116	*E. abietina diabolis*	-33.954329	18.438171	37482863	-	-
SM378	TC045	*E. abietina diabolis*	-33.954038	18.437654	37482864	-	-
SM497	TC052	*E. amphigena*	-34.283811	19.111704	63750814	-	-
SM568	TC211	*E. anguliger*	-34.050745	19.629348	139097180	-	-
EO12619	MP45	*E. arborea*	-	-	-	Ojeda, F (Oliver, EGH)	Sierra del Aljibe, ESP

Continued on next page

Table B.1 – continued from previous page

Voucher No.	Sample No.	Organism	Latitude	Longitude	iNaturalist*	Collector	Note
SM173	TC192	*E. articularis*	-33.995062	18.412987	141363041	Merry, C	-
MP1383	MP15	*E. australis*	-	-	-	Pirie, MD	-
SM214	S277	*E. axilliflora*	-34.698898	19.609428	24777607	-	-
SM436	TC061	*E. axilliflora*	-34.609039	19.560601	54915649	-	-
SM437	S361	*E. axilliflora*	-34.609088	19.560573	54916103	-	-
SM481	TC066	*E. baccans*	-34.058623	18.374230	63160607	-	-
EO12873	MP10	*E. banksii banksii*	-34.2275	19.155139	-	Pirie, MD	-
SM554	TC169pC	*E. brachialis*	-34.352590	18.488490	70803379	-	-
SM611	TC256	*E. bruniifolia*	-34.677857	19.747620	139109850	-	-
SM525	TC101	*E. caffra*	-34.084271	19.056075	68001687	-	-
SM498	TC174z	*E. calycina*	-33.936163	19.162276	64503411	-	-
SM556	TC171	*E. capensis*	-34.258076	18.386052	70803528	-	-
SM137	TC129	*E. cf. borbonifolia*	-34.060626	19.849437	139409785	-	-
SM561	TC204	*E. cf. ericoides*	-34.313326	19.413970	138235130	-	-
SM560	TC203	*E. cf. exleeana*	-34.313943	19.412749	138234900	-	-
SM545	TC160	*E. cf. imbricata*	-34.291926	18.829246	69253541	-	-
SM509	TC180	*E. cf. imbricata*	-33.353301	19.626286	65012535	-	-
SM569	TC212	*E. cf. maritima*	-34.038539	19.623474	139097181	-	-
SM538	TC150	*E. cf. pellucida*	-33.697468	19.114552	69252915	-	-
SM532	TC108	*E. cf. racemosa*	-34.015286	19.109067	68009422	-	-
SM605	TC250	*E. cf. russakiana*	-34.805438	20.036618	139109840	-	-
EO12845	MP58	*E. chrysocodon*	-33.955433	19.174194	-	Oliver, EGH	
MP1377	MP42	*E. ciliaris*	-	-	-	Fagundez, J (Pirie, MD)	Matas de Faja, PRT
SM440	TC081	*E. coccinea coccinea*	-34.639347	19.572571	54967439	-	-
SM570	TC213	*E. coccinea coccinea*	-34.151710	18.926250	139097184	-	-
SM576	TC218pB	*E. coccinea uniflora*	-34.524564	19.449894	139098686	-	-
SM577	TC219pB	*E. coccinea uniflora*	-34.552617	19.416942	139098689	-	-
SM578	TC220pB	*E. coccinea uniflora*	-34.552617	19.416942	139098690	-	-
SM604	TC249	*E. coccinea uniflora*	-34.803242	20.049640	139109837	-	-
SM461	S350	*E. corifolia*	-34.086724	18.423703	58059263	-	-
EO12832	MP4	*E. coventryi*	-	-	-	Oliver, EGH	Fernkloof NR, RSA
SM544	TC159	*E. cristata*	-34.292132	18.829087	69253527	-	-
SM176	TC136z	*E. cruenta*	-33.901626	19.275208	139870023	-	-
SM306	TC141	*E. cruenta*	-34.226385	18.993429	30927319	-	-
SM464	TC097B	*E. curviflora*	-34.093000	18.422442	60396831	-	-
SM603	TC248	*E. curvirostris*	-34.670754	20.042404	139109835	-	-
SM340	TC197pC	*E. curvistyla*	-32.150804	19.027138	32097120	-	-
SM550	TC165pB	*E. cygnea*	-34.286207	18.836168	69253660	-	-
SM551	TC166pB	*E. cygnea*	-34.286174	18.836189	69253697	-	-
SM193	TC088pTC088B	*E. desmantha*	-34.010113	19.005026	21742142	-	-
CM19	MP8	*E. diosmifolia*	-33.969111	18.409444	-	Merry, C	-
SM565	TC208	*E. discolor*	-34.317892	19.405846	138235926	-	-
SM392	TC049	*E. doliiformis*	-33.641167	19.132226	37642759	-	-
SM393	TC033	*E. doliiformis*	-33.641478	19.132155	37642862	-	-
SM540	TC152	*E. doliiformis*	-33.689805	19.095086	69252997	-	-
SM541	TC153	*E. doliiformis*	-33.689791	19.095094	69253114	-	-
SM537	TC149	*E. altevivens*	-33.693812	19.148721	68763878	-	-
SM131	TC185pC	*E. embothriifolia longiflora*	-34.064841	19.842018	21954768	-	-
SM496	TC173	*E. eriocephala*	-34.278909	19.118210	63750746	-	-
SM553	TC168	*E. fascicularis*	-34.288848	18.833429	69253740	-	-
SM141	TC086	*E. fascicularis imperialis*	-34.097513	19.849418	139869374	-	-

Continued on next page

Table B.1 – continued from previous page

Voucher No.	Sample No.	Organism	Latitude	Longitude	iNaturalist*	Collector	Note
SM196	TC193pC	*E. fastigiata* (Jonkershoek)	-34.006142	19.008541	21742153	-	-
SM241	TC023pB	*E. filamentosa*	-34.068215	20.482596	26246409	-	-
SM242	TC024	*E. filamentosa*	-34.067938	20.482795	26246416	-	-
SM369	TC095	*E. filiformis*	-34.241425	18.981499	35655599	-	-
SM488	TC119z	*E. flacca*	-32.148443	19.060606	63581885	-	-
SM536	TC148	*E. glauca elegans*	-33.695129	19.148883	69252884	-	-
SM197	TC194pC	*E. glutinosa*	-34.003577	19.011870	21820356	-	-
SM132	TC283	*E. goatcheriana petrensis*	-34.061413	19.844987	21958746	-	-
SM397	TC035	*E. grandiflora grandiflora*	-33.615834	19.099722	38458259	-	-
SM505	TC001	*E. grandiflora grandiflora*	-33.737935	19.077051	65004983	-	-
SM510	TC004z	*E. grandiflora grandiflora*	-33.383728	19.289508	65012667	-	-
SM511	TC036	*E. grandiflora grandiflora*	-33.880461	19.162320	65013572	-	-
SM323	TC115	*E. grandiflora perfoliosa*	-33.992405	18.982022	30184643	-	-
SM171	TC087pTC087B	*E. gysbertii*	-34.365374	18.830066	139419340	-	-
SM547	TC162	*E. gysbertii*	-34.291211	18.831392	69253586	-	-
SM353	TC026	*E. hibbertia*	-33.968671	19.167279	34270237	-	-
SM354	TC143	*E. hibbertia*	-33.968162	19.169098	34270277	-	-
SM363	TC041	*E. hibbertia*	-33.890484	19.333280	35371936	-	-
SM502	TC084	*E. hibbertia*	-33.969402	19.167532	64506172	-	-
SM503	TC068	*E. hibbertia*	-33.968590	19.167484	64506330	-	-
SM167	TC135	*E. imbricata*	-34.356252	18.838414	139419335	-	-
SM303	TC140	*E. imbricata*	-34.112073	18.461566	30366250	-	-
SM409	TC071	*E. imbricata*	-32.721242	18.574251	40598990	-	-
SM490	TC183	*E. imbricata*	-34.293235	19.117652	63582359	-	-
SM533	TC300	*E. imbricata*	-34.015888	19.108541	68009503	-	-
SM558	TC201	*E. imbricata*	-34.313976	19.412756	72014000	-	-
SM527	TC103	*E. intervallaris*	-34.014616	19.108630	68001883	-	-
SM579	TC221pB	*E. irregularis*	-34.524689	19.450186	139098687	-	-
SM580	TC222	*E. irregularis*	-34.524689	19.450167	139098688	-	-
SM507	TC178	*E. junonia minor*	-33.369133	19.657260	65012138	-	-
SM557	TC172	*E. laeta*	-34.272490	18.452146	70865822	-	-
SM368a	TC287	*E. latiflora*	-34.241455	18.981607	35601894	-	-
SM368b	TC288	*E. latiflora*	-34.241455	18.981607	35601894	-	-
SM368c	TC042	*E. latiflora*	-34.241455	18.981607	35601894	-	-
SM342	TC199	*E. limosa*	-34.062023	18.388241	32097556	-	-
SM559	TC202	*E. longiaristata*	-34.313976	19.412756	138234782	-	-
EO12658	MP54	*E. madagascariensis*	-22.162556	46.895194	-	Oliver, EGH	Andringitra N.P., MDG
MP1378	MP43	*E. maderensis*	-	-	-	Fagundez, J (Pirie, MD)	Pico do Areeiro, PRT
SM574	TC217	*E. magnisylvae*	-34.540227	19.429491	139098683	-	-
SM534	TC146	*E. mammosa gilva*	-34.174599	18.387182	69050877	-	-
SM572	TC215	*E. massonii*	-34.151419	18.926541	139097187	-	-
SM535	TC147	*E. melastoma*	-33.694708	19.144099	69051000	-	-
SM341	TC198	*E. mollis*	-34.062117	18.388337	32097555	-	-
SM542	TC157	*E. monadelphia*	-34.296682	18.827416	69252836	-	-
SM530	TC106	*E. multumbellifera*	-34.011308	19.109605	68009210	-	-
EO12747	TC085	*E. nematophylla*	-33.999408	21.283511	-	Oliver, EGH	-
SM348	TC025z	*E. nevillei*	-34.078571	18.371302	34114117	-	-
SM486	TC067	*E. nevillei*	-34.050899	18.366696	63581343	-	-
SM524	TC100	*E. nevillei*	-34.077968	18.371745	66243717	-	-
SM350	TC200	*E. nivea*	-34.078457	18.371172	34176331	-	-
SM160	TC132z	*E. obliqua*	-34.316911	19.008306	139414847	-	-

Continued on next page

Table B.1 – continued from previous page

Voucher No.	Sample No.	Organism	Latitude	Longitude	iNaturalist*	Collector	Note
SM567	TC210	*E. pannosa*	-34.050952	19.627951	138236178	-	-
SM382	TC046	*E. parilis*	-32.957989	19.056368	37581928	-	-
SM395	S289pB	*E. parilis*	-33.879889	19.324957	37642932	-	-
SM506	TC002z	*E. parilis*	-33.353227	19.626078	65005076	-	-
SM501	TC176	*E. penicilliformis*	-33.957409	19.174084	63874419	-	-
SM155	TC187	*E. perspicua*	-34.313096	19.008777	139869377	-	-
EO12844	MP55	*E. perspicua latifolia*	-	-	-	Oliver, EGH	Hermanus area, RSA
SM419	TC073pTC293	*E. petrusiana*	-34.206847	18.841577	40650868	-	-
SM420	TC074pTC294	*E. petrusiana*	-34.206629	18.840715	43550523	-	-
SM421	TC295	*E. petrusiana*	-34.206628	18.840673	43550983	-	-
SM422	TC076pTC296	*E. petrusiana*	-34.206937	18.840749	43551418	-	-
SM398	TC096	*E. phillipsii*	-33.636825	19.150734	38267892	-	-
SM400	TC051	*E. phillipsii*	-33.636755	19.149796	38459653	-	-
SM407	TC069	*E. phillipsii*	-32.710507	18.559438	40280913	-	-
SM408	TC070	*E. phillipsii*	-32.725311	18.575945	40593733	-	-
SM161	TC188pC	*E. pillansii*	-34.319722	19.002768	21958778	-	-
SM152	TC285	*E. pinea*	-34.330780	19.015105	139414845	-	-
SM181	S290pB	*E. pinea*	-33.899825	19.268644	21742428	-	-
SM391	TC032	*E. pinea*	-33.627247	19.138071	37615603	-	-
SM394	TC117	*E. pinea*	-33.628470	19.141591	37641574	-	-
SM499	TC009	*E. pinea*	-33.936236	19.162540	64506045	-	-
SM489	TC182	*E. placentiflora*	-34.293235	19.117652	63582250	-	-
SM495	TC099	*E. placentiflora*	-34.278360	19.115821	63711996	-	-
SM331	TC196pC	*E. plukenetii lineata*	-34.657805	19.564903	30757955	-	-
SM531	TC107	*E. plukenetii penicillata*	-34.014398	19.109166	68009345	-	-
SM308	TC195	*E. plumigera*	-34.226597	18.993102	30927320	-	-
SM539	TC151pD	*E. praecox*	-33.690638	19.101909	69252930	-	-
SM508	TC179z	*E. pseudocalycina*	-33.374998	19.665178	65012324	-	-
SM469	S329	*E. pyxidiflora*	-34.179231	18.374584	62543511	-	-
SM333	TC040	*E. quadrisulcata*	-34.238763	18.463073	32097429	-	-
SM387	TC031	*E. quadrisulcata*	-34.213013	18.451676	37615109	-	-
SM388	TC054	*E. quadrisulcata*	-34.213810	18.451488	37615208	-	-
SM389	TC291	*E. quadrisulcata*	-34.214035	18.451405	37615247	-	-
SM390	TC292	*E. quadrisulcata*	-34.214149	18.451327	37615561	-	-
SM218	TC089	*E. regia casta*	-34.705132	19.703732	24777665	-	-
SM612	TC257	*E. regia casta*	-34.705122	19.703773	139109851	-	-
SM441	S364	*E. regia mariae*	-34.423345	20.411664	55131216	-	-
SM608	TC253	*E. regia mariae*	-34.639227	19.925561	139109844	-	-
SM609	TC254	*E. regia mariae*	-34.639229	19.925748	139109847	-	-
SM220	TC022	*E. regia regia*	-34.632403	19.719696	24777684	-	-
SM610	TC255	*E. regia regia*	-34.673109	19.751302	139109849	-	-
SM615	TC260	*E. regia regia*	-34.632246	19.720293	139109856	-	-
SM165	TC134	*E. retorta*	-34.331658	19.009356	139419334	-	-
SM168	TC189	*E. rhopalantha*	-34.361010	18.838370	139419338	-	-
201410903	MP16	*E. scoparia*	-	-	-	Pirie,MD	-
SM504	TC177	*E. cf. imbricata*	-33.969650	19.168905	64026836	-	-
SM617	TC262	*E. cf. placentiflora*	-34.632282	19.720372	139109857	-	-
SM520	TC017	*E. cf. involvens*	-34.365488	18.830055	65705579	-	-
SM521	TC018	*E. cf. involvens*	-34.365588	18.829596	65705827	-	-
SM516	TC015	*E. cf. placentiflora*	-34.523786	19.491117	139508880	-	-
SM500	TC175z	*E. serrata*	-33.957592	19.177049	64028147	-	-

Continued on next page

Table B.1 – continued from previous page

Voucher No.	Sample No.	Organism	Latitude	Longitude	iNaturalist*	Collector	Note
SM571	TC214z	*E. serrata*	-34.151659	18.926727	139097186	-	-
SM438	TC062B	*E. sessiliflora*	-34.631078	19.578888	54916713	-	-
SM566	TC209	*E. sessiliflora*	-34.317884	19.405938	138236031	-	-
Amsn	MP57	*E. sicula sicula*	38.112398	12.665409	-	Pirie, MD	-
SM383	TC029	*E. situshiemalis*	-32.959160	19.070036	37581993	-	-
SM384	TC047	*E. situshiemalis*	-32.959156	19.070643	37614602	-	-
SM385	TC030	*E. situshiemalis*	-32.963753	19.054453	37614694	-	-
SM386	TC048	*E. situshiemalis*	-32.963753	19.054417	37614795	-	-
HLA188	MP51	*E. spiculifolia*	43.368316	22.602508	-	Andersen, HL	-
SM519	TC016pTC297R	*E. stokoei*	-34.364735	18.831755	65705309	-	-
SM485	S377	*E. strigosa*	-34.057282	18.379094	63580879	-	-
SM491	TC005	*E. suffulta*	-34.292374	19.118079	63583296	-	-
SM494	TC008	*E. suffulta*	-34.277872	19.116293	63584397	-	-
SM178	TC137	*E. taxifolia*	-33.899378	19.267509	139417716	-	-
SM170	TC190z	*E. tenella*	-34.364621	18.835220	139419339	-	-
SM156	TC131	*E. tenuifolia*	-34.319827	19.001432	139869375	-	-
2004.0948	MP29	*E. terminalis*	-	-	-	Pirie, MD	Ex. Hort.
SM546	TC161	*E. thomae pink*	-34.291822	18.829595	69253557	-	-
SM425	TC058	*E. thomae tenax*	-34.330193	19.028444	53738251	-	-
SM523	TC299	*E. thomae thomae*	-34.364796	18.831655	65706231	-	-
SM555	TC170	*E. tristis*	-34.352425	18.488271	70803436	-	-
KB_108/01	TC282	*E. turgida*	-	-	-	Lansdowne, A	Ex. Hort.
KB_286/70	TC281	*E. turgida*	-	-	-	Lansdowne, A	Ex. Hort.
SM552	TC167	*E. urceolata*	-34.286912	18.836904	69253722	-	-
MP1376	MP41	*E. vagans*	-	-	-	Fagundez, J (Pirie, MD)	Uzal Capelada, ESP
SM179	TC138	*E. ventricosa*	-33.902820	19.268900	139419341	-	-
SM182	TC139B	*E. ventricosa*	-33.902820	19.268900	139419342	-	-
SM487	TC098pTC098B	*E. verecunda*	-32.148278	19.060391	63581509	-	-
KB_273/12	TC273	*E. verticillata* Adonis	-	-	-	Lansdowne, A	Ex. Hort.
SM583	TC225	*E. verticillata* Adonis	-	-	-	-	Rondevlei, Cape Town
SM592	TC234	*E. verticillata* Adonis	-	-	-	-	Rondevlei, Cape Town
SM595	TC237	*E. verticillata* Adonis	-	-	-	-	Rondevlei, Cape Town
KB_109/01	TC266	*E. verticillata* Belvedere	-	-	-	Lansdowne, A	Ex. Hort.
SM584	TC226	*E. verticillata* Belvedere	-	-	-	-	Rondevlei, Cape Town
KB_549/06	TC269	*E. verticillata* Cherise	-	-	-	Lansdowne, A	Ex. Hort.
KB_14/12	TC272	*E. verticillata* Dresden	-	-	-	Lansdowne, A	Ex. Hort.
KB_657/06	TC270	*E. verticillata* Harry Wood	-	-	-	Lansdowne, A	Ex. Hort.
SM581	TC223	*E. verticillata* Pretoria	-	-	-	-	Rondevlei, Cape Town
KB_12/12	TC271	*E. verticillata* Rot	-	-	-	Lansdowne, A	Ex. Hort.
KB_543/06	TC267	*E. verticillata* Tresco	-	-	-	Lansdowne, A	Ex. Hort.
KB_548/06	TC268	*E. verticillata* Violet Gray	-	-	-	Lansdowne, A	Ex. Hort.
KB_AL-A	TC274	*E. verticillata* F1	-	-	-	Lansdowne, A	Self-germinated in cult.
KB_AL-B	TC275	*E. verticillata* F1	-	-	-	Lansdowne, A	Self-germinated in cult.
KB_AL-C	TC276	*E. verticillata* F1	-	-	-	Lansdowne, A	Self-germinated in cult.
KB_AL-D	TC277	*E. verticillata* F1	-	-	-	Lansdowne, A	Self-germinated in cult.
KB_AL-E	TC278	*E. verticillata* F1	-	-	-	Lansdowne, A	Self-germinated in cult.
KB_AL-F	TC279	*E. verticillata* F1	-	-	-	Lansdowne, A	Self-germinated in cult.
KB_AL-G	TC280	*E. verticillata* F1	-	-	-	Lansdowne, A	Self-germinated in cult.
SM582	TC224	*E. verticillata* F1	-	-	-	-	Rondevlei, Cape Town
SM585	TC227	*E. verticillata* F1	-	-	-	-	Rondevlei, Cape Town
SM586	TC228	*E. verticillata* F1	-	-	-	-	Rondevlei, Cape Town

Continued on next page

Table B.1 – continued from previous page

Voucher No.	Sample No.	Organism	Latitude	Longitude	iNaturalist*	Collector	Note
SM587	TC229	*E. verticillata* F1	-	-	-	-	Rondevlei, Cape Town
SM588	TC230	*E. verticillata* F1	-	-	-	-	Rondevlei, Cape Town
SM589	TC231	*E. verticillata* F1	-	-	-	-	Rondevlei, Cape Town
SM590	TC232	*E. verticillata* F1	-	-	-	-	Rondevlei, Cape Town
SM593	TC235	*E. verticillata* F1	-	-	-	-	Rondevlei, Cape Town
SM594	TC236	*E. verticillata* F1	-	-	-	-	Rondevlei, Cape Town
SM596	TC238	*E. verticillata* F1	-	-	-	-	Rondevlei, Cape Town
SM597	TC239	*E. verticillata* F1	-	-	-	-	Rondevlei, Cape Town
SM598	TC240	*E. verticillata* F1	-	-	-	-	Rondevlei, Cape Town
SM599	TC241	*E. verticillata* F1	-	-	-	-	Rondevlei, Cape Town
SM600	TC242	*E. verticillata* F1	-	-	-	-	Rondevlei, Cape Town
SM601	TC243	*E. verticillata* F1	-	-	-	-	Rondevlei, Cape Town
SM602	TC244	*E. verticillata* F1	-	-	-	-	Rondevlei, Cape Town
SM219	TC090pTC090B	*E. vestita*	-34.654854	19.694728	24777679	-	-
SM252	TC091	*E. vestita*	-33.952555	20.706150	26246442	-	-
SM253	TC092B	*E. vestita*	-33.950263	20.701460	26246444	-	-
SM512	TC125pTC013	*E. vestita*	-34.534607	19.503297	139508878	-	-
SM515	TC014	*E. vestita*	-34.546746	19.447222	139508879	-	-
SM606	TC251	*E. vestita*	-34.801965	20.037946	139109843	-	-
SM607	TC252	*E. vestita*	-34.801839	20.037688	139508082	-	-
SM613	TC258	*E. vestita*	-34.650245	19.701244	139109852	-	-
SM109	TC110z	*E. viscaria cf. pendula*	-34.167457	19.136064	21592760	-	-
SM427	TC078	*E. viscaria cf. pustulata*	-34.401311	19.282172	53772719	-	-
SM428	TC060	*E. viscaria cf. pustulata*	-34.400735	19.283052	54724558	-	-
SM429	TC265	*E. viscaria cf. pustulata*	-34.396519	19.292284	54724874	-	-
SM431	TC080	*E. viscaria gallorum*	-34.434176	19.575653	54726962	-	-
SM432	S356	*E. viscaria gallorum*	-34.434025	19.575638	54727221	-	-
SM433	S357	*E. viscaria gallorum*	-34.433953	19.575611	54914917	-	-
SM528	TC145	*E. viscaria gallorum*	-34.014667	19.108790	68001972	-	-
SM221	S304pB	*E. viscaria longifolia*	-34.547511	19.634692	24777685	-	-
SM322	TC114	*E. viscaria longifolia*	-34.008756	19.005956	30185492	-	-
SM367	TC027	*E. viscaria longifolia*	-34.242307	18.986215	35619906	-	-
SM396	TC034	*E. viscaria longifolia*	-33.892868	19.342263	37643139	-	-
SM418	TC263	*E. viscaria longifolia*	-34.210136	18.846308	40648767	-	-
SM423	TC264	*E. viscaria longifolia*	-34.195473	18.876924	43552181	-	-
SM430	TC079	*E. viscaria longifolia*	-34.399793	19.277399	54726552	-	-
SM434	S358	*E. viscaria longifolia*	-34.533163	19.529632	54915123	-	-
SM435	S359pB	*E. viscaria longifolia*	-34.533183	19.529095	54915252	-	-
SM526	TC144	*E. viscaria longifolia*	-34.083606	19.056065	68001735	-	-
SM562	TC205	*E. viscaria longifolia*	-34.313078	19.415174	138235357	-	-
SM563	TC206	*E. viscaria longifolia*	-34.313010	19.415478	138235421	-	-
SM573	TC216	*E. viscaria longifolia*	-34.149797	18.927455	139097191	-	-
SM616	TC261	*E. viscaria longifolia*	-34.531837	19.622385	140661283	-	-
SM111	TC021	*E. viscaria macrosepala*	-34.218808	19.185306	21593175	-	-
SM112	TC111	*E. viscaria macrosepala*	-34.218510	19.185201	21593173	-	-
SM150	TC037	*E. viscaria macrosepala*	-34.330859	19.017926	21595097	-	-
SM162	TC113	*E. viscaria macrosepala*	-34.321200	18.995330	139417717	-	-
SM217	TC038	*E. viscaria macrosepala*	-34.699619	19.611001	24777646	-	-
SM424	TC154A	*E. viscaria macrosepala*	-34.329505	19.027246	53736552	-	-
SM426	TC059	*E. viscaria macrosepala*	-34.218796	19.185254	53738686	-	-
SM439	S363	*E. viscaria macrosepala*	-34.639183	19.572540	54916867	-	-

Continued on next page

Table B.1 – continued from previous page

Voucher No.	Sample No.	Organism	Latitude	Longitude	iNaturalist*	Collector	Note
SM614	TC259	*E. viscaria macrosepala*	-34.649675	19.700340	139109854	-	-
SM309	TC039	*E. viscaria pendula*	-34.226613	18.992817	30927323	-	-
SM310	TC127	*E. viscaria pendula*	-34.226667	18.992743	30927328	-	-
SM492	TC006	*E. viscaria pendula*	-34.293777	19.117615	63582891	-	-
SM493	TC007	*E. viscaria pendula*	-34.287821	19.107834	63584232	-	-
SM460	TC064	*E. viscaria viscaria*	-34.086622	18.424147	58058820	-	-
SM462	S351pB	*E. viscaria viscaria*	-34.086711	18.423764	58059688	-	-
SM463	S352pB	*E. viscaria viscaria*	-34.086430	18.423797	60396570	-	-
SM468	S328	*E. viscaria viscaria*	-34.181278	18.370693	62543403	-	-
SM564	TC207	*E. xeranthemifolia*	-34.312461	19.417026	138235606	-	-
W-2013.0655-01	MP53	*Calluna vulgaris*	60.498116	4.915009	-	Moe, B	-
W-1999.0498	MP27	*Daboecia cantabrica*	-	-	-	Pirie, MD	León, ESP
W-1996.0626	MP24	*Rhododendron rex fictolacteum*	-	-	-	Pirie, MD	Beima Shan, CHN

*iNaturalist observations can be viewed at `inaturalist.org/observations/<iNaturalistID>`.

Table B.2 Voucher information of samples with GBS data (Chapter 4). In bold is the voucher number of the individual that was sequenced in both sequencing batches (SM484). All collections were made by the author and specimens have been deposited at NBG.

Voucher No.	Sample No.	Batch	Latitude	Longitude	Taxon	iNaturalist ID*
SM337	S166z	2	-34.25393600	18.46956497	*E. quadrisulcata*	32097433
SM333	S214z	2	-34.23876297	18.46307300	*E. quadrisulcata*	32097429
SM388	S279B	2	-34.21381006	18.45148847	*E. quadrisulcata*	37615208
SM486	S378	1	-34.05089909	18.36669613	*E. nevillei*	63581343
SM269	S384	2	-33.98950000	18.41341600	*E. a. abietina*	28425132
SM270	S385	1	-33.98942100	18.41317000	*E. a. abietina*	28425133
SM272	S387	1	-33.98891100	18.41154097	*E. a. abietina*	28425139
SM271	S386	2	-33.98888997	18.41293500	*E. a. abietina*	28425137
SM274	S388	2	-33.98807800	18.41380200	*E. a. abietina*	28425141
SM275	S249	2	-33.98794697	18.41441700	*E. a. abietina*	28425144
SM326	S160	2	-33.96921497	18.38995997	*E. a. abietina*	30927835
SM379	S189	2	-33.95646097	18.43214497	*E. a. abietina*	37482865
SM230	S226zA	2	-33.95631900	18.42578100	*E. a. abietina*	25265906
SM228	S380	1	-33.95580500	18.42421424	*E. a. abietina*	25265904
SM231	S382	2	-33.95577800	18.42717300	*E. a. abietina*	25265913
SM227	S379B	1	-33.95529900	18.42420300	*E. a. abietina*	25265963
SM376	S186	2	-33.95436300	18.43880800	*E. a. diabolis*	37482861
SM378	S188	2	-33.95403797	18.43765397	*E. a. diabolis*	37482864
SM375	S185	2	-33.95335597	18.43986597	*E. a. diabolis*	37482860
SM374	S184	2	-33.95321597	18.43978200	*E. a. diabolis*	37482859
SM370	S180	2	-33.95213320	18.44620116	*E. a. diabolis*	37011954
SM372	S275	2	-33.95190820	18.44614618	*E. a. diabolis*	37291782
SM315	S142	2	-34.31851300	18.41953600	*E. a. atrorosea*	30927503
SM338	S167	2	-34.26035300	18.46235797	*E. a. atrorosea*	32097434
SM471	S331	2	-34.18790687	18.37421503	*E. a. atrorosea*	62543624
SM470	S330	2	-34.18193497	18.37504584	*E. a. atrorosea*	62543568
SM467	S354	2	-34.09040187	18.42164788	*E. a. atrorosea*	62543342
SM466	S353B	1	-34.09039687	18.42165895	*E. a. atrorosea*	62543295
SM266	S197	2	-34.07493800	18.39962600	*E. a. atrorosea*	26399907

Continued on next page

Table B.2 – continued from previous page

Voucher No.	Sample No.	Batch	Latitude	Longitude	Taxon	iNaturalist ID*
SM478	S370	2	-34.05868184	18.37431192	*E. a. atrorosea*	63160143
SM476	S368z	2	-34.05853602	18.37298490	*E. a. atrorosea*	63159779
SM484	S376,S376D	1,2	-34.05843102	18.37782595	*E. a. atrorosea*	63163263
SM475	S366	2	-34.05658386	18.37230496	*E. a. atrorosea*	63159621
SM459	S348	2	-34.02265593	18.40658594	*E. a. atrorosea*	58057956
SM458	S347	2	-34.02245503	18.40642802	*E. a. atrorosea*	58057565
SM456	S342	2	-34.02231108	18.40626206	*E. a. atrorosea*	57753114
SM457	S343	2	-34.02225300	18.40622485	*E. a. atrorosea*	58057054
SM449	S335	2	-34.01266297	18.41973715	*E. a. atrorosea*	56335274
SM444	S326	1	-34.00541992	18.41663785	*E. a. atrorosea*	55685470
SM445	S327	1	-34.00364943	18.41361299	*E. a. atrorosea*	56165232
SM443	S325	2	-34.00298209	18.42126902	*E. a. atrorosea*	55685271
SM446	S332	2	-34.00104589	18.41553614	*E. a. atrorosea*	56214476
SM448	S334	1	-33.99812511	18.42417587	*E. a. atrorosea*	56334836
SM447	S333	1	-33.99289091	18.42598803	*E. a. atrorosea*	56214830
SM472	S344	2	-33.97574005	18.44338987	*E. a. atrorosea*	62697322
SM474	S346	2	-33.97562800	18.44313707	*E. a. atrorosea*	63159540
SM473	S345	2	-33.97562494	18.44316792	*E. a. atrorosea*	63159457
SM264	S92zA	2	-34.06371900	18.38555600	*E. a. constantiana*	26399903
SM257	S110zA	2	-34.06253000	18.39633600	*E. a. constantiana*	26399851
SM477	S369	2	-34.05895711	18.37424889	*E. a. constantiana*	63159964
SM482	S374	2	-34.05861101	18.37419994	*E. a. constantiana*	63160382
SM480	S372	2	-34.05859102	18.37438401	*E. a. constantiana*	63160250
SM451	S337	2	-34.02258007	18.40146594	*E. a. constantiana*	57652605
SM450	S336	2	-34.02241807	18.40491090	*E. a. constantiana*	57651903
SM453	S338	2	-34.02234387	18.40407606	*E. a. constantiana*	57653130
SM454	S340	2	-34.02230691	18.40522405	*E. a. constantiana*	57653495
SM455	S341	2	-34.02205710	18.40527602	*E. a. constantiana*	57654128
SM452	S339	2	-34.02176587	18.40137105	*E. a. constantiana*	57652774
SM414	S314	2	-33.99948100	18.40007991	*E. a. constantiana*	40627133
SM416	S316	2	-33.99816791	18.40062842	*E. a. constantiana*	40647963

Continued on next page

Table B.2 – continued from previous page

Voucher No.	Sample No.	Batch	Latitude	Longitude	Taxon	iNaturalist ID*
SM417	S317	1	-33.99532737	18.40196483	*E. a. constantiana*	40648239
SM283	S120	2	-33.98199100	18.43313500	*E. a. abietina x E. a. atrorosea*	28933719
SM381	S191	2	-33.96038300	18.44319197	*E. a. abietina x E. a. atrorosea*	37482868
SM479	S371	2	-34.05859102	18.37438401	*E. a. atrorosea x E. a. constantiana*	63162993
SM290	S367	2	-34.05852000	18.37296500	*E. a. atrorosea x E. a. constantiana*	29499592

*iNaturalist observations can be viewed at `inaturalist.org/observations/<iNaturalistID>`.

Acknowledgements

First and foremost I would like to thank my supervisors, Dr. Nicolai Nürk, Assoc. Prof. Michael Pirie, and Prof. Tony Verboom, for their unwavering support, patience, generosity, and compassion, and for sharing their vast knowledge with me.

Primary funding for the project was provided by the Deutsche Forschungsgemeinschaft (PI 1169/1-2).

The Genomics Core Facility (GCF) at the University of Bergen, which is a part of the NorSeq consortium, provided services on whole-genome sequencing of three *Erica* species; GCF is supported in part by major grants from the Research Council of Norway (grant no. 245979/F50), Bergen Research Foundation (BFS) (grant no. BFS2017TMT04 and BFS2017TMT08), and Trond Mohn Foundation (TMS).

All sequence data resulting from this project will be deposited in the GenBank Sequence Read Archive.

I would like to thank Louise Maria Lindblom for providing DNA extractions of European *Erica* and outgroup species.

Computing facilities were provided by the Centre for High Performance Computing at the University of Bayreuth (`bzhpc.uni-bayreuth.de`) and the University of Cape Town's ICTS High Performance Computing team (`hpc.uct.ac.za`).

Collections were made under permits from CapeNature (CN35-31-8281) and SANParks (CRC/2019-2020/004–2019/V1). Voucher specimens were deposited at the Compton Herbarium (NBG) with the kind assistance of Shaun Pieterse. Access and permission to collect on private land was kindly provided by Antonie Botha of Stettynskloof. I would like to thank the many people who assisted with fieldwork and/or accompanied me in the field: Dirk Bellstedt, Magriet Brink, Campbell Fleming,

Martha Kandziora, Gabriella Leighton, Benjamin van Niekerk, Timo van der Niet, Corinne Merry, Robert Muir, Ted Oliver, Thaabiet Parker, Hana Petersen, Helen Pickering, Kervin Prayag, Paula Strauss, Hugh Verboom, Chris Vynbos, Gemma Walker, Suzanne Wilson-Smith, Joseph White, and of course my supervisors.

(Eidesstattliche) Versicherungen und Erklärungen

(§5 Nr. 4 PromO)

Hiermit erkläre ich, dass keine Tatsachen vorliegen, die mich nach den gesetzlichen Bestimmungen über die Führung akademischer Grade zur Führung eines Doktorgrades unwürdig erscheinen lassen.

(§8 S. 2 Nr. 5 PromO)

Hiermit erkläre ich mich damit einverstanden, dass die elektronische Fassung meiner Dissertation unter Wahrung meiner Urheberrechte und des Datenschutzes einer gesonderten überprüfung hinsichtlich der eigenständigen Anfertigung der Dissertation unterzogen werden kann.

(§8 S. 2 Nr. 7 PromO)

Hiermit erkläre ich eidesstattlich, dass ich die Dissertation selbständig verfasst und keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt habe.

(§8 S. 2 Nr. 8 PromO)

Ich habe die Dissertation nicht bereits zur Erlangung eines akademischen Grades anderweitig eingereicht und habe auch nicht bereits diese oder eine gleichartige Doktorprüfung endgültig nicht bestanden.

(§8 S. 2 Nr. 9 PromO)

Hiermit erkläre ich, dass ich keine Hilfe von gewerblichen Promotionsberatern bzw. -vermittlern in Anspruch genommen habe und auch künftig nicht nehmen werde.

.....

Ort, Datum, Unterschrift