Tracking thicket through space and time

Insights into the evolutionary history of the Albany Subtropical Thicket from comparative phylogeography and distribution modelling

Alastair John Potts

September 2011

Thesis Presented for the Degree of
Doctor of Philosophy
in the
Department of Botany,
University of Cape Town

Supervisors:
T. A. Hedderson and R. M. Cowling

To my parents,
for guiding me along the path,
teaching me the trees along the way,
and letting me explore further.
Also to Reda,
for walking with me,
and making my steps feel lighter.
To my daughters,
I wish you endless curiosity.

Declarations: I, Alastair John Potts, hereby:

1. grant the University of Cape Town free license to reproduce the above thesis in whole or in part, for the purpose of research;
2. declare that:
    a) the above thesis is my own unaided work, both in conception and execution, and that apart from the normal guidance of my supervisor, I have received no assistance apart from that stated below;
    b) neither the substance nor any part of the above thesis has been submitted in the past, or is being, or is to be submitted for a degree at this University or at any other University, except as stated below;

I am now presenting the thesis for examination for the Degree of PhD.

Alastair Potts, September 2011.

Acknowledgements

To Professor Richard Cowling (Nelson Mandela Metropolitan University), for setting my imagination free to 'see' the forever shifting vegetation, climate and landscape through time and space. To Professor Terry Hedderson, for showing me how to find my own research path and guiding me back from many dead-ends. To Jan Vlok, for sharing the near limitless knowledge of plant species and their distribution that resides in his head; this project would not have been possible without his key insights.

To my parents, who shared their love of the bushveld and taught me about the trees; there are very few children who could use scientific tree names to curse their classmates in Latin. Thank you also for giving me the freedom and the opportunity to follow my interests.

To Reda, my wife, for your patience, understanding, caring, sharing and prodding. Thanks for putting up with my vacant stares and sometimes unfathomable questions when problems were hassling me.

Thank you to Guido Grimm (Swedish Museum of Natural History) for many entertaining discussions on ITS and phylogeny reconstruction, as well as valuable edits of Chapter 3. Thank you to Ernesto Ismail for helping me typeset this thesis in LaTeX 2_ε. I am grateful to Guy Midgley, Kathryn Lannas and two anonymous reviewers for valuable comments on Chapter 2 and Alexis Stamatakis for his comments on Chapter 3.

I am incredibly grateful to the various foundations and funding bodies that have supported me throughout my years of study, particularly the Darwin Initiative, the National Research Foundation, the University of Cape Town, and the Dorothy Cameron Trust.

To the many people that collected samples for this project; phylogeography involves sampling widely through an enormous landscape which is impossible for one person to traverse in search of specific species. The time generously spent collecting samples by volunteers gives a far richer insight into phylogeographic patterns than

can be done by any one researcher alone. I am indebted to the following people for providing samples (in alphabetical order): Rauri Alcock, Nigel Barker, Warren Bass, Tarik Bodasing, An van Cauter, Corli Coetsee, Tony Dold, Gary and Wendy Holburn, Sam Jack, Colleen Mannheimer, Muthama Muasya, Carla Stava, David Styles, Syd Ramdhani, Karin van der Walt, Julia Wakeling, and Ben Wigley.

Abstract

Albany Subtropical Thicket (AST) is a species-rich biome restricted to the coastal lowlands of the southern Cape region of South Africa. Its Quaternary history is poorly understood, but climatic changes associated with Pleistocene glacial cycles may have profoundly affected the distributions, gene flows, and demographies of species. The glacial refugia hypothesis predicts that AST retracted into fragmented refugia during glacial cycles. The evolutionarily discrete drainage basin (EDDB) hypothesis suggests that the prevailing topography played an important population-structuring role. I evaluate these two hypotheses by combining community and species distribution models with multigene comparative phylogeography of three AST species *Pappea capensis*, *Nymania capensis*, and *Schotia afra*. Distribution models support the glacial refugia hypothesis, with highly reduced and fragmented distributions postdicted for the Last Glacial Maximum. These models, projected onto two climate scenarios for 2050, give a positive outlook for the future of AST, with no dramatic shifts or reduction in appropriate climate. Chloroplast and nuclear genomes were used for phylogeographic analyses. Intra-individual site polymorphisms (2ISPs) in nuclear DNA have traditionally hindered phylogeny reconstruction. I outline an approach that incorporates the variation present in 2ISPs that improves phylogenetic reconstruction across a range of methods. Phylogeographic structure in *N. capensis* and *P. capensis* corresponds to primary drainage basins, which supports the EDDB hypothesis. In contrast, *S. afra* comprises a single meta-population, spanning drainage basins with limited structuring. These contrasting patterns may relate to reproductive ecology. *Nymania capensis* and *P. capensis* are bird- and wind-dispersed, respectively, whilst *S. afra* has large pods that are eaten by large mammals including mega-herbivores (e.g. elephants). Long distance wind-dispersal is likely hampered by vegetation and watersheds, and dispersal by birds limited by territoriality and short gut-retention times. However, migrating mega-herbivores with poor digestion and large intestinal tracts retaining seeds for long periods may broadly disperse seeds, leading to lack of phylogeographic structure. These patterns also support the use of drainage basins as

a surrogate for biodiversity conservation efforts in an AST mega-conservancy network. The interaction between topography, palaeoclimatic history and seed dispersal ecology seems to have predictable influences on population structuring of AST flora. Thus, this thesis offers new insights into the evolutionary history and ecology of the Albany Subtropical Thicket.

Contents

List of Tables

List of Figures

Acronyms

ARZ – annual rainfall zone
AST – Albany Subtropical Thicket
BI – Bayesian Inference
CDM - community distribution model
cpDNA – chloroplast DNA
DNA – *D*eoxy*r*ibonucleic *a*cid
EDDB – evolutionarily distinct drainage basin
GCM – global climate model
GPS – global positioning system
ITS – internal transcribed spacer (specifically used for the ITS-1–5.8S–ITS-2 region of the ribosomal cistron)
LDD – long distance dispersal
LGM – Last Glacial Maximum
ML – Maximum Likelihood
MP – Maximum Parsimony
MSS – maximum test sensitivity plus specificity
nDNA – nuclear DNA
NJ – Neigbour Joining
NN – neighbour-net **PCR** – polymerase chain reaction
SDM – species distribution model
sPCA – spatial Principle Component Analysis
SP – statistical parsimony
SRZ – summer rainfall zone
2ISP – *i*ntra-*i*ndividual *s*ite *p*olymorphism

1. General Introduction

Species have limited distributions. Discovering *where* species do and do not occur, coupled with the questions of *why* and *why not*, has been a cornerstone to investigating evolution and speciation since the explorations of Charles Darwin (Darwin 1859), Alfred Wallace (Wallace 1880) and Sir Joseph Hooker (Hooker 1867*a*,*b*). The study of the distribution of species and the causes of their distributional breaks is termed biogeography. This field of research has provided invaluable insights into the distributional histories of both species and environments across the earth (Brown & Lomolino 1998). However, the level of detail that biogeography can provide has traditionally been hindered by its smallest unit of measure: the species. The distribution of a single or multiple co-occurring species does not provide a picture of their independent, and often, divergent evolutionary histories. Biogeography investigates the patterns and causal factors only after speciation has occurred, and thus is unable to detect recent evolutionary events or fine-scale environmental barriers.

Individuals within a species rarely form a single population with equal connectivity between all individuals. They can be isolated simply by distance, or in more complex scenarios by extrinsic environmental features that influence their dispersal. With the recent advent of the molecular era, the species as the smallest biogeographical unit was refined and often replaced by the individual, or rather an individual's genome. Tracking genetic units has provided a previously unattainable degree of resolution in exploring the distributional history and population dynamics of individual species (e.g. Avise 2000). The investigation of the distribution of genetic lineages within a species across aquatic and terrestrial landscapes has been termed phylogeography (Avise *et al.* 1987). This burgeoning discipline has opened a window into the evolutionary history of species and continues to provide answers to the questions of *where* and *why*, but at a far more detailed level.

Inferring the history of distributional shifts of a species traditionally relied on tracking the species through time by mapping and dating preserved remains, such as fossils or pollen. Unfortunately, this information is usually highly fragmented in space

and time due to the serendipitous nature of deposition, and appropriate fossils may be sparse to non-existent for many taxa. In addition, climates or landscapes conducive to the formation and preservation of such depositional archives have not occurred in all regions of the world, particularly in southern Africa (Chase & Meadows 2007). Nevertheless, another very recent avenue for exploring the distributional history of a species has arisen in the form of species distribution models (SDM), coupled with its historical extension, palaeodistribution modelling. Species distribution modelling involves two steps: 1) predicting an environmental envelope from known localities and the environmental conditions at those localities, and 2) extrapolating this envelope onto a spatially interpolated environmental surface for an area. These envelope models can be projected onto environmental surfaces that represent different time periods. Initially, palaeoclimatic surfaces were reconstructed using proxies such as palynology (e.g. Kershaw & Nix 1988). More recently, global climate models provide hypotheses of past and future climate surfaces (e.g. Richards *et al.* 2007, Waltari *et al.* 2007). Thus, SDM offers an alternative avenue by which the shifts in species distributions in time may be investigated, especially in regions where palaeoarchives are rare. Furthermore, these methods can be applied, with additional caveats, to larger community-level units such as biomes or vegetation units. Community distribution models (CDM) have provided valuable hypotheses of the past and future distributional shifts of vegetation communities (Carnaval & Moritz 2008, Midgley *et al.* 2002, VanDerWal *et al.* 2009*a*). Both SDM and CDM are based on a wide range of assumptions (reviewed in Svenning *et al.* 2011) and these are discussed in detail in Chapter 2 ('Modelling Approach', Pg. 46). Nonetheless, CDM offers an alternative to SDM when extensive locality information is deficient for most species within a community. The lack of sufficient locality information is a global problem that also extends to most plant species within South Africa's biomes.

The Albany Subtropical Thicket (AST) biome is found along the southern coastal lowlands of South Africa. It has historically been a poorly understood and documented vegetation unit (reviewed in Vlok *et al.* 2003) and has only been elevated to the status of biome relatively recently (Low & Rebelo 1996). In hindsight, its biome status seems obvious given its unique mix of growth forms, climate and ecology. Structurally, the AST vegetation is a closed shrubland of evergreen, sclerophyllous or succulent trees, shrubs and vines, without a conspicuous grass layer. This, coupled with an absence of clear strata, distinctive climatic conditions and the absence of fire, prompted Low &

Rebelo (1996) to elevate this vegetation's status to a biome. The AST is characterised by slow-growing trees and shrubs (Pierce & Cowling 1984) that are resistant to drought (e.g. Holmes & Cowling 1993, Ting & Hanscom 1977). It also has a surprisingly high living biomass for a semi-arid region (Mills & Cowling 2006, Mills *et al.* 2005) with very low fluctuations in biomass in response to drought cycles (Hoare & Frost 2004). Characteristically, it has a deep surface litter which is eroded when thicket is transformed by intensive pastoralism (Mills & Fey 2004, Mills *et al.* 2005). The ecology of the AST is unique: it is not a fire-driven nor a drought-driven system (Vlok *et al.* 2003) - these are the predominant drivers in neighbouring biomes (Mucina & Rutherford 2006). Historically, the dominant disturbance was provided by large herbivores, specifically African Savannah Elephants and Black Rhinoceros (Kerley *et al.* 1995).

The phytosociological and ecological aspects of the AST vegetation have been relatively well explored (Vlok *et al.* 2003), given that its origin and affinities have been misunderstood and neglected until fairly recently. Based on fossil and phylogenetic data, Cowling *et al.* (2005) conclude that the AST is an ancient formation, extending back at least to the Eocene (33.9 - 55.8 Ma), which was once widespread during the Palaeogene (23.0 - 65.5 Ma). Cowling *et al.* (2005) suggest that the Pleistocene glacial periods, with their lower global temperatures, would have been trying times for thicket based on observations of frost intolerance in many AST species, especially succulents. Thus, they hypothesise that the AST retracted into frost-free refugia during these periods. Beyond this, the history of this vegetation during the Quaternary remains unexplored.

In this thesis, I to explore the Quaternary history of the Albany Subtropical Thicket. I will use a combination of species distribution modelling, community distribution modelling and the phylogeography of three dominant and widespread plant species found in the Albany Subtropical Thicket. Below I provide brief overviews of the current state of plant phylogeography and phylogeography in southern Africa. I also describe the coastal lowlands and the AST in terms of an ideal model study system. Finally, I give the prospectus of my thesis.

1.1. The current state of plant phylogeography

The study of plant phylogeography has considerably lagged behind that of animals (Avise 1998, 2000, Beheregaray 2008), largely due to the role of mitochondrial DNA (mtDNA) as a driving force behind the discipline. The overall reliance on mtDNA for phylogeographic studies in animals stems from the following desirable properties (Avise 2000): i) it tends to evolve faster than nuclear DNA (nDNA), ii) it has a stable gene order and polymorphisms are usually nucleotide substitutions, and iii) it is inherited asexually through the maternal lineage without recombination (but see Barr *et al.* 2005). In contrast, both plant mtDNA and chloroplast DNA (cpDNA) exhibit significantly lower rates of nucleotide substitution which results in far less phylogeographic resolution. Plant mtDNA is further hampered by extensive intramolecular recombination (Palmer 1992), which is not observed in animal mtDNA. Furthermore, hybridisation and polyploidisation are far more prevalent in plants than in animals (Mallet 2007, Moyle *et al.* 2004, Muller 1925). Thus, although the mode of inheritance of cpDNA is generally uniparental, there is high potential for chloroplast movements across species boundaries (e.g. Álvarez & Wendel 2006, Belahbib *et al.* 2001, King & Ferris 2000, Rieseberg *et al.* 2003, Vriesendorp & Bakker 2005). Such horizontal transfer blurs the relationship between gene history and true organismal history (Rieseberg & Soltis 1991). Thus, a significant cause of the lag in plant phylogeography are the difficulties involved with obtaining sufficient genetic variation (Newton *et al.* 1999, Schaal *et al.* 1998) and the additional complexity that is introduced by species boundaries that are far more permeable to genetic exchange than those observed in animals. In a recent review, Beheregaray (2008) demonstrates that the distribution of plant phylogeography studies is also severely biased in terms of continents and hemispheres. He reveals that studies on species from North America and Europe comprise 66% of all phylogeographic studies on terrestrial plants, whereas Africa, Australia and South America comprise only 4%, 5% and 4%, respectively. Furthermore, 88% of studies of terrestrial plants are set exclusively in the northern hemisphere, whilst only 8% are from the southern hemisphere (the remainder span both hemispheres). Plant phylogeography has provided valuable insights into the historical forces that have affected species, such as the locations of refugia and the corridors and direction of migration in response to the effects of Pleistocene glacial cycles, primarily in the northern hemisphere (Brunsfeld *et al.* 2001, Hewitt 2000, Taberlet *et al.* 1998), but also in other parts of the world (e.g. Australia's arid

zone biota, Byrne 2007, Byrne *et al.* 2008). In South Africa, phylogeography offers excellent potential for providing a deeper understanding of the history of the many unique and biodiversity-rich vegetation communities (Cowling *et al.* 2005), which include three biodiversity hotspots: the Succulent Karoo, the Cape Floristic Region and the Maputaland-Pondoland-Albany region (Myers *et al.* 2000, Steenkamp *et al.* 2004). However, at present there are only a handful of phylogeographic studies of plants that offer tantalising glimpses into the recent vegetation history of the region. For example, coastal species of *Streptocarpus* (a forest floor herb) have far higher genetic diversity than a highland species, suggesting that coastal populations were far bigger during the Pleistocene glacial periods than highland populations (Hughes *et al.* 2005). Two other studies of common and widespread South African plants find surprisingly little genetic structuring across large distances, complex landscapes and the recent Pleistocene climate cycles: *Elytropappus rhinocerotis* (Bergh *et al.* 2007), a shrub that is widely distributed in the winter-rainfall region, and the genus *Schotia* (Ramdhani *et al.* 2010) which are tree species restricted to the southern African summer and annual rainfall regions. In contrast, Prunier and Holsinger (Prunier & Holsinger 2010) detected limited gene flow between most populations within species of white proteas (*Protea* section *Exsertae*) and suggest that geographical isolation in the complex Cape Floristic Region landscape was responsible for the diversification observed in this group. Therefore, phylogeography is an underutilised field of research in investigations of South African vegetation history.

Nuclear DNA has not been widely used for phylogeography analysis in animals or plants due to its complex intrinsic properties. This biparentally inherited genome is subject to recombination, longer coalescent times, and heterozygosity. These can be significant problems as recombination can mislead the reconstruction of phylogenetic relationships (Buckler *et al.* 1997), longer coalescent times decrease the degree of phylogeographic resolution as genetic patterns require a longer evolutionary period to emerge (Avise 1998, Avise & Wollenberg 1997), and determining the correct haplotypes for heterozygous individuals can be expensive as cloning is often required. There is an added problem of polyploidy in the nuclear genome, which is more prominent in plants than in animals (Levin 1983, Mallet 2007, Moyle *et al.* 2004, Muller 1925). This process multiplies the number of chromosomes and genomic content within lineages, which is often followed by massive silencing and elimination of duplicated genes (Adams & Wendel 2005, Otto 2003). This can lead to dramatic re-arrangements that can

mislead attempts to reconstruct phylogenetic relationships (Wendel 2000). However, the evolution of polyploidy may also increase the likelihood of reproductive barriers which can, given enough time, lead to the development of spatial structure (e.g Eidesen *et al.* 2007, Popp *et al.* 2008, Thompson & Whitton 2006, Trewick *et al.* 2002).

Despite these problems, nDNA offers a critical contrast to the generally uniparentally inherited organelles (mtDNA and cpDNA) that may give a biased representation of a species genetic coherence (e.g. population fragmentation detected by cpDNA may be caused by extremely localised seed dispersal, whereas in reality extensive gene flow may occur between populations via pollen) or evolutionary history as the smaller effective population size of uniparentally-inherited markers render them more susceptible to stochastic processes (e.g. Edwards & Beerli 2000). Thus, the biparentally-inherited nDNA may be more representative of a species history and less prone to the vagaries of a species natural history or the stochastic nature of the coalescent (Avise 2000, 2004). The most widespread nDNA region used in plant systematics or phylogeography is the non-coding internal transcribed spacer (ITS) region of the 18S-5.8S-25S nuclear ribosomal cistron (Álvarez & Wendel 2003, Baldwin *et al.* 1995, Feliner & Rosselló 2007). The ITS region (ITS-1, 5.8S, and ITS-2) offers a valuable source of information for plant phylogeography (Chiang & Schaal 1999, Feliner *et al.* 2004, Rosselló *et al.* 2007) due to its often higher rate of mutation compared with, for example, the chloroplast genome (Schaal *et al.* 1998). However, it has often proved difficult to extract a clear signal from ITS for phylogeography and lower level phylogenies due to the presence of multiple ITS variants within an individual (Feliner & Rosselló 2007, King & Roalson 2008). The ribosomal cistrons that include ITS form a multi-gene family arranged in tandem arrays. These arrays are confined to one or more chromosomal loci (termed nucleolus organiser region(s), reviewed in Volkov *et al.* 1999). Thus, there are hundreds to thousands of copies of ITS within any given individual. Although these copies are homogenised through a process of concerted evolution, numerous unique variants of ITS usually remain within an individual. Thus, a number of processes including incomplete concerted evolution, recombination (crossing-over), interbreeding/hybridisation, and autopolyploidisation can etch conflicting phylogenetic signals onto ribosomal DNA such as ITS (Bailey *et al.* 2003). The result of these processes is the presence of intra-individual site polymorphisms (2ISPs, pronounced 'twisps') within ITS sequences. Currently, the presence of 2ISPs severely hampers phylogenetic reconstructions as

the majority of the currently available algorithms and software treat potentially informative polymorphisms as ambiguous characters or missing data. However, intra-individual site polymorphisms can offer additional phylogenetic or phylogeographic information (e.g. Fama *et al.* 2000, Feliner *et al.* 2004, Grimm *et al.* 2007). Therefore, Chapter 3 is devoted to exploring an informative treatment of 2ISPs and comparing this with the standard treatment.

Plant phylogeography is an exciting but understudied field, especially in southern Africa. It has revealed the historic responses of lineages to dramatic environmental changes elsewhere in the world, and has been most useful when the histories of multiple species are studied (e.g. Garrick *et al.* 2004, 2007, 2008, Sunnucks *et al.* 2006). Here I use comparative phylogeography of three plant species and molecular data (cpDNA and nDNA) to explore the history of the Albany Subtropical Thicket.

1.2. The current state of southern African phylogeography

Phylogeography in Africa has lagged behind in comparison to the northern hemisphere (Beheregaray 2008). Furthermore, there is a great disparity in the African taxa studied using these techniques. The Greater Cape Floristic Region (G-CFR, Born *et al.* 2007) encompasses two biodiversity hotspots the Cape Floristic Province and Succulent Karoo. Although this area is a small proportion of the southern African region (Figure 1.1.A), the number of articles investigating the phylogeography of taxa restricted to the G-CFR outnumbers that from taxa restricted to southern Africa (Figure 1.1.B). Furthermore, articles based on taxa restricted to southern Africa (including the G-CFR) also marginally outnumbers those from the rest of Africa. There are two factors likely to be driving this: 1) much of the phylogeographic research in southern Africa has been spear-headed by researchers from South Africa, where funding is available for such research, and 2) the logistical difficulties surrounding sampling and obtaining the relevant permits to obtain samples of a target taxon increase dramatically with each new country that the distribution spans. The Greater Cape Floristic Region offers an ideal study system that is geographically small, and with few geopolitical boundaries as it largely encompassed within two provinces of a single county, South Africa. The AST offers a similarly ideal study system as it is geographically restricted (∼47,000 km^2; Vlok *et al.* 2003) with few geopolitical boundaries as it spans only the Western and Eastern Cape Provinces of South Africa.

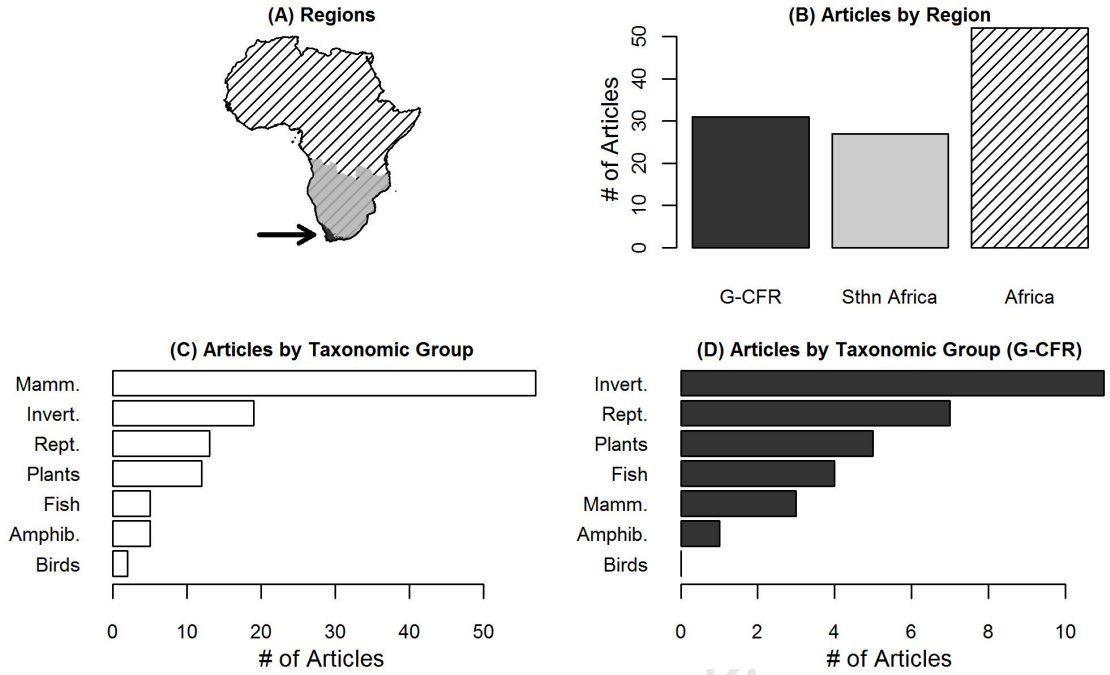

Figure 1.1. The regional and taxonomic coverage of 113 articles on of phylogeography of African terrestrial taxa published between 1987 and 2011. Articles were compiled by searching for 'phylogeography' or 'phylogeographic' in ISI Web of Knowledge and Google Scholar and filtering the title, abstract and keywords using all of the country names in Africa. Each article was checked manually to ensure that African terrestrial taxa were the focus of the study. (A) Articles are divided into three nested terrestrial regions: the Greater Cape Floristic Region (G-CFR, highlighted with an arrow), Southern Africa, and Africa. An article was classified within whichever region contained all the samples of the taxon or taxa of interest. (B) The number of studies restricted to each region. The number of articles on terrestrial taxa by taxonomic group is shown (C) for all articles, (D) and those restricted to the G-CFR.

A study of the literature reveals that mammals have by far received the most attention of phylogeographers in Africa (Figure 1.1.C). Unsurprisingly, this has been driven by a focus on the charismatic large mammals on the continent, many of which have been the target of numerous rounds of phylogeographic study (e.g. elephants, black rhinos, lions and cheetahs). The ordering of major groups remain unchanged when ranked by taxonomic group and focuses on only those articles that study species within the G-CFR, with one interesting exception (Figure 1.1.D): mammals drop to fifth place on the list. This is likely because there are few large mammals that are solely restricted to this region, whereas there are numerous endemics from the other

taxonomic groups. It is indicative of the difficulties surrounding plant phylogeography that the number of articles on plants still lags behind those on invertebrates and reptiles despite the renowned floral diversity of the region.

1.3. The coastal lowlands and the Albany Subtropical Thicket as a study system

The topography of South Africa is characterized by a narrow coastal plain of low relief, which is separated from an unusually high interior plateau by the Great Escarpment (Figure 1.2.A). The escarpment has formed through a series of uplift events, likely related to periods of reorganisation and spreading of mid-oceanic ridges (Moore *et al.* 2009), the last of which occurred approximately five million years ago (Ma) during the late Miocene (Partridge & Maud 1987). By the end of the Pliocene (∼2.6 Ma), the coastal lowland landscape very closely resembled the topographically and edaphically heterogeneous ones of today (Cowling *et al.* 2009). The coastal plain is interrupted in the south-west by the Cape Folded Belt which also displays a wide range of topographic relief. The orogeny of the Cape Folded Belt may be ancient (280 – 215 Ma), but it has shared periods of uplift with the Great Escarpment (Partridge & Maud 2000). The close proximity of the Great Escarpment to the coast resulted in an extended series of parallel, relatively short and deeply incised drainage basins along the coast of South Africa; the AST spans eight of these drainage basins. On the basis of a floristic assessment of the AST, Vlok *et al.* (2003) suggested that these drainage basins are discrete biogeographical units. In this thesis, this proposal is termed the 'evolutionarily discrete drainage basin hypothesis' (EDDB). Based on this hypothesis, drainage basins have been treated as unique entities in large-scale conservation planning that aims to ensure the persistence of evolutionary processes for the AST biota (Rouget *et al.* 2006). This hypothesis is supported by phylogeographic research on freshwater redfins (Swartz *et al.* 2009) and terrestrial cicadas (Price *et al.* 2010), but there is limited data on whether it applies to terrestrial plants.

There are two dominant rainfall regimes across South Africa (Carr *et al.* 2006, Schulze *et al.* 1997): a winter-rainfall zone (WRZ) in the west and a summer-rainfall zone (SRZ) in the east. Between these two regimes lies the annual-rainfall zone (ARZ) that receives both summer and winter rainfall (Figure 1.2.B). The present-day WRZ

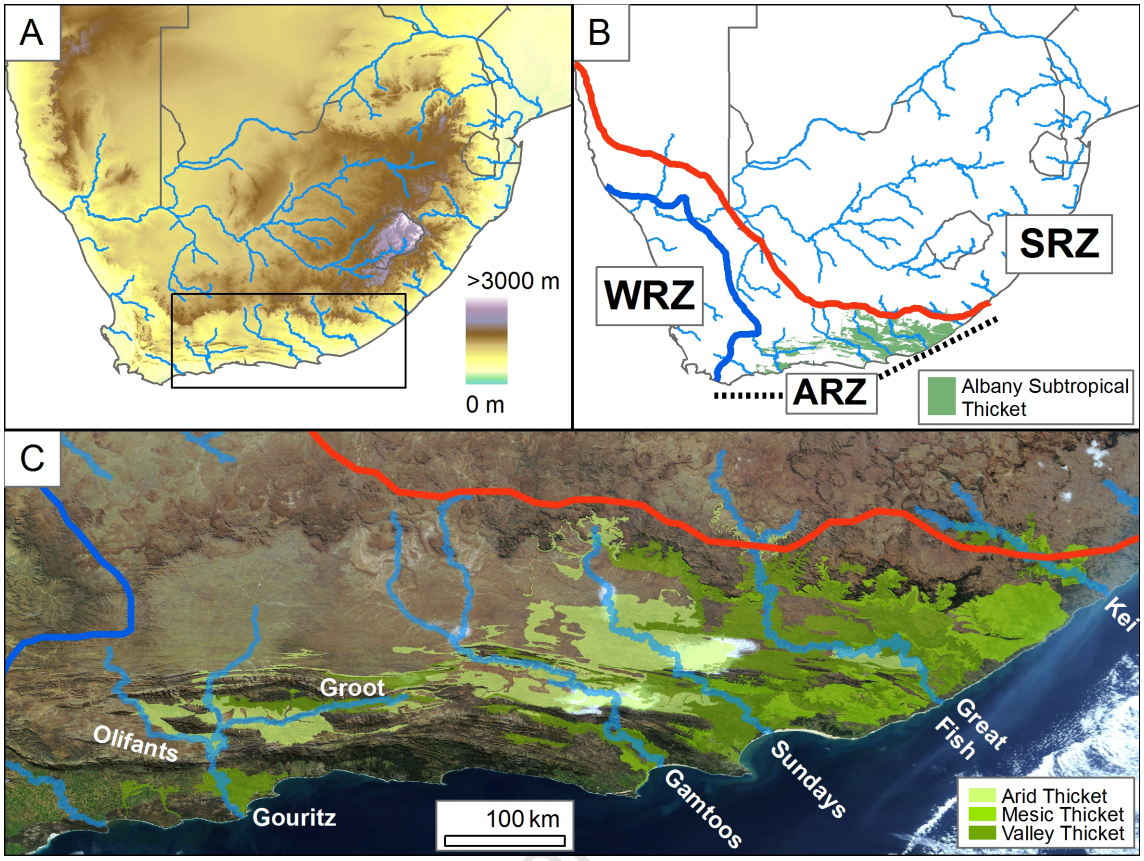

Figure 1.2. The (A) topography and (B) major rainfall regimes of southern Africa, and (C) the three inland subtypes of the Albany Subtropical Thicket. The major rainfall regimes follow Carr *et al.* (2006) where the winter-rainfall zone (WRZ), annual-rainfall zone (ARZ) and summer-rainfall zone receive >66% winter rain, 66% to 33% winter rain, and <33% winter rain, respectively.

was initiated by the increase in Antarctic glaciation (~14 Ma; Zachos *et al.* 2001) and the associated increase in the strength of the South Atlantic high-pressure cell. Summer aridity in the west and interior was exacerbated by the Late Miocene uplift of the Great Escarpment and interior plateau due to the rainshadow effects that prevented convective uplift precipitation systems from reaching the west (Tyson 1986). This has resulted in, or contributed to, the climatic isolation of coastal lowland vegetation types, such as the AST, that do not extend into the interior plateau (Mucina & Rutherford 2006).

Southern Africa has been spared the Quaternary glaciations (Partridge 1997)

that have had a marked and catastrophic effect on plant distributions in the northern hemisphere (e.g. Taberlet *et al.* 1998). However, the exact nature of Pleistocene climatic fluctuations across southern Africa remains largely speculative because of the limited number of reliable palaeoenvironmental records (Chase & Meadows 2007). During the Last Glacial Maximum (LGM, 24 – 18 ka) temperatures were lower (Partridge *et al.* 1999) by as much as 5 – 6°C (Talma & Vogel 1992), with increased aridity and heterogenous shifts in the seasonality of rainfall (Chase & Meadows 2007, van Zinderen Bakker 1976). Lowered temperatures likely caused a reduction in the distribution of the AST vegetation as most of its dominant species are frost intolerant (Cowling *et al.* 2005). Thus, the AST is postulated to have retracted into refugia during glacial periods (Cowling *et al.* 2005); in this thesis, this postulation is termed the 'glacial refugia hypothesis'. This hypothesis is only supported by limited palaeodata that suggest that the thicket was highly fragmented during glacial periods (Scholtz 1986).

The AST is dominant in, and largely restricted to, the ARZ (Figure 1.2.B) and is most typical (i.e. solid swathes of thicket rather than mosaics with other vegetation) in a semi-arid climate where rainfall is between 200 mm and 800 mm per year (Vlok *et al.* 2003). In this rainfall zone, four AST subtypes have been identified based on geography, floristics, structure and grain (Vlok *et al.* 2003): i) arid thicket, ii) valley thicket, iii) mesic thicket and iv) dune thicket. Only the first three are examined in this thesis (Figure 1.2.C). Dune thicket vegetation is floristically distinct from the subtypes on other substrata (Low & Rebelo 1996), occupies a very narrow belt along the coast, and requires an understanding of coastal dune field movements through the Quaternary which is currently unavailable and beyond the scope of this thesis. The three mainland subtypes are arranged along a gradient of increasing moisture availability (arid thicket to valley thicket to mesic thicket) but are differentiated based on ecology (Cowling *et al.* 2005) and structural characteristics, such as the relative cover of woody, grass and succulent species, as well as the incidence of spinescence and woody lianas (Vlok *et al.* 2003). More than 50% of the AST exists as a mosaic where thicket occurs in isolated clumps forming a coarse-grained patchwork (10 – 500 ha stands of continuous thicket) with neighbouring vegetation such as fynbos, karoo, forest or grassland.

The AST vegetation is resistant to browsing by natural herbivores and mega-herbivores provide the main natural disturbance in dense thicket facilitating access

for smaller herbivores (Kerley *et al.* 1995, Stuart-Hill 1992). However, with the advent of commercial livestock farming approximately 12% of the AST has experienced severe degradation (Rouget *et al.* 2006) as sustained browsing, primarily by goats, has transformed the dense, close-canopy, shrubland into an open community with scattered and degraded thicket clumps and isolated trees in a matrix of ephemeral shrubs (Figure 1.3; Lechmere-Oertel *et al.* 2005*a*,*b*, Stuart-Hill 1992). A restrictive set of AST plant species are able to persist in degraded landscapes, either in thicket clumps or as isolated trees (Lechmere-Oertel *et al.* 2005*b*). I selected three species that are widespread and common within the AST, but which also persist in degraded landscapes, specifically *Nymania capensis* (Chapter 4), *Pappea capensis* and *Schotia afra* (Chapter 5; Figure 1.3). Thus, the distributions of these species reflect the pre-colonial extent of the AST, i.e. prior to degradation, and the phylogeographic patterns are likely to be free from possible interference caused by the reduction in thicket.

As noted by Price *et al.* (2010), the coastal plains of South Africa offer an ideal model to investigate the role of drainage basins and watersheds on the diversification of terrestrial biota for a number of reasons: (a) there are numerous drainage basins that span a wide range of climatic and topographic complexity, (b) a relatively stable geology through the Quaternary, and c) a reasonable understanding of the climatic history of the region. Furthermore, the AST is an ideal vegetation to investigate because of its high plant diversity; this biome forms the western part of the Maputaland-Pondoland-Albany biodiversity hotspot (Steenkamp *et al.* 2004), which is delimited, in part, on the basis of high levels of plant endemism.

Figure 1.3. Intact and degraded Albany Subtropical Thicket. (A) A fence-line contrast of intact thicket (left) and degraded thicket (right) on the upper slopes of a hill. (B) Completely degraded thicket where individual trees of *Pappea capensis* persist in a matrix of ephemeral shrubs. Both photos were taken near Steytlerville, Eastern Cape South Africa.

1.4. Prospectus of thesis

It is my thesis to explore the Quaternary history of the Albany Subtropical Thicket. I will use a multidisciplinary approach that combines species distribution modelling, community distribution modelling and phylogeography. I will explore the distributional shifts in suitable climate for the AST subtypes and examine the phylogeography of three dominant and widespread AST plant species. The questions I aim to answer are thus: Were suitable climatic conditions for AST vegetation reduced during the LGM and, if so, where were potential refugial areas? [Chapter 2]; Have the Pleistocene climatic cycles and the complex topography of the coastal lowlands had an effect

on species cohesion within the Albany Subtropical Thicket? [Chapters 4 and 5]; And finally, do differences in seed dispersal ecology affect how both climate and landscape influence a species cohesion within the AST? [Chapter 5]. As addressing phylogeographic questions using nuclear DNA sequence data from plants has been problematic, this thesis includes a consideration of the methods used to analyse such data [Chapter 3].

2. The past and future distributions of Albany Subtropical Thicket: insights from community distribution modelling

2.1. Abstract

Southern Africa lacks the necessary palaeo-archives to infer regional vegetation history. Here spatially-explicit community distribution models (CDMs) are used to explore Last Glacial Maximum (LGM) distributions of the mega-diverse Albany Subtropical Thicket (AST) subtypes. The potential distributional changes in these subtypes under projected future climate for 2050 was also assessed. I generated CDMs for AST subtypes using fine-scale vegetation maps. These CDMs were projected onto various global climate models for the LGM and two scenarios for 2050. The changes in subtype distribution were estimated as the gain, loss, change and turnover in range between the present and past or future scenarios. These range changes were also analysed in the context of an ambitious mega-conservancy network proposed for the Albany Subtropical Thicket. The results indicate a dramatic range reduction during the LGM for all AST subtypes; arid and valley thickets experienced the most severe reductions with an overall decline in altitude and concomitant fragmentation, whereas mesic thicket had a reduced but fairly continuous range that included a shift onto the exposed continental shelf. Across all global climate models and scenarios of future climate, both arid and valley thicket are predicted to have relatively low levels of range loss, which are generally compensated by predicted range gains. The mesic thicket is projected to undergo range loss with very little range gain. The spatially explicit hypotheses suggest that the Pleistocene glacial periods saw dramatically reduced distributions of AST subtypes. Last Glacial Maxima refugial areas are identified and extreme bottlenecks are postdicted for arid and valley thicket species. The distributions of subtypes under future climate scenarios are optimistic, with limited range shifts, and some expansion. Finally, I suggest that the AST species are likely to persist under

projected climate change within the proposed mega-conservancy network.

2.2. Introduction

The severe glacial-interglacial climatic shifts of the Pleistocene, driven by variations in the Earth's orbit (Zachos *et al.* 2001), significantly altered geographical distributions of species and communities (Hewitt 2004, Jansson & Dynesius 2002, VanDerWal *et al.* 2009*a*). These climate oscillations resulted in widespread local, and sometimes global, extinctions of species and communities, as well as fragmentation, reduction or expansion of ranges, and dispersal to novel areas or survival in refugia (Dynesius & Jansson 2000). The altered climate predicted for the near future may also greatly affect distributions (Hannah *et al.* 2005, Thomas *et al.* 2004), as already evident in recent changes in species' distributions and abundance attributed to human-induced climate change (Parmesan 2006).

Until relatively recently, studies reconstructing vegetation changes have relied solely on fossil and pollen profiles. However, several intrinsic and sometimes rather severe limitations (e.g. selective preservation regimes, reworking of deposits) curtail the range of species, communities, and areas to which such techniques can be applied. The development of global climatic models (GCMs) that reconstruct potential palaeoclimatic conditions or predict future climate changes, coupled with a flurry of new techniques for modelling biotic distributions under different climatic scenarios, provides the potential to gain new insights into the temporal distribution of vegetation types. Such modelling has been widely applied to single species (species distribution modelling; Alsos *et al.* 2009, Pearman *et al.* 2008) and suites of species that share a biogeographical unit (community distribution modelling; Loarie *et al.* 2008, Riordan & Rundel 2009). Climate is considered a driving factor in setting the broad limits to vegetation distributions at regional and global scales (Mather & Yoshioka 1968, Woodward 1987); however, both the concentration of atmospheric CO_2 and fire are also critical factors determining the distribution of vegetation (Scheiter & Higgins 2009). Nonetheless, modelling the climatic niche of a floristic grouping may provide valuable hypotheses of the historical or future changes of the distribution of plant communities. Biomes are broad vegetation types that are largely defined by life form, ecophysiology, and climate (Woodward *et al.* 2004), and thus may be suitable units for modelling. Community distribution models (CDM) have been used for tropical forests

in Australia (Hilbert *et al.* 2007, VanDerWal *et al.* 2009*a*) and Brazil (Carnaval & Moritz 2008), and fynbos and succulent karoo in South Africa (Midgley & Roberts 2003).

The subtropical thicket vegetation in southern Africa, locally termed Albany Thicket (Mucina & Rutherford 2006) or Albany Subtropical Thicket (hereafter referred to as AST), forms the western part of the Maputaland-Pondoland-Albany biodiversity hotspot (Steenkamp *et al.* 2004), and is delimited on the basis of high levels of plant endemism, especially succulents and bulbs (Vlok *et al.* 2003). This biome, occurring on the southern coastal lowlands of the Western Cape and Eastern Cape Provinces of South Africa, is characterised as a dense, woody, semi-succulent and thorny vegetation, with an average height of 2–5 m, that is relatively impenetrable in an unaltered condition (Mucina & Rutherford 2006). In this study, the units of interest are three well-defined subtypes of AST - arid thicket, valley thicket and mesic thicket (Vlok *et al.* 2003). These subtypes are differentiated by climate, structural characteristics and ecology (Cowling *et al.* 2005). The distribution of AST is thought to have been greatly affected by rapid fluctuations in environmental conditions during the late Quaternary glacial cycles (Cowling *et al.* 2005). The impact on this biome from projected climate change remains largely unexplored.

Subtropical Thicket is thought to have once been widespread across southern Africa, but a combination of subcontinental uplift, declining global temperatures and the establishment of fire-driven ecosystems, dominated by grass or fynbos, during the Neogene (23–1.8 Ma) resulted in wide-scale contraction of its distribution (Cowling *et al.* 2005, Vlok *et al.* 2003). The subsequent saw-tooth climatic fluctuations between Pleistocene glacial-interglacial periods are thought to have resulted in the contraction and expansion of AST vegetation (Cowling *et al.* 2005). Changes in precipitation and temperature, as well as land area due to the exposure of the Agulhas Bank as a consequence of fluctuating sea levels, are likely to be important factors responsible for driving cycles of AST expansion and contraction.

Southern Africa, like many parts of the globe, lacks sufficient palaeoevironmental archives to provide a clear picture of how climate and vegetation, including the AST, have changed during the Quaternary (Chase & Meadows 2007, Lewis 2008). In this chapter, I develop regional-scale CDMs based on climatic envelopes for vegetation subtypes of the AST, and then project these envelope characterisations onto past and

future climate scenarios in order to determine the broad changes in distribution. I focus on the Last Glacial Maximum (LGM), around 21 000 years before present, as this represents the coldest period in the recent past with expanded and thickened ice sheets in the high latitudes as well as a large dip in atmospheric CO_2 (Zachos *et al.* 2001). I also focus on the projected climate for 2050, as this represents a world affected by anthropogenic climate change. The LGM has been the focus of the Paleoclimate Modelling Intercomparison Project (Braconnot *et al.* 2007), while scenarios of possible climate outcomes for 2050 have been produced by the Intergovernmental Panel on Climate Change (Nakicenovic & Swart 2000); thus, global climate models are readily available for these periods.

The aim of this chapter is to gain insight into the history and future of the AST biome and to provide a framework for integrating future analyses of species and genetic diversity across the biome. My objectives are:

1. To propose spatially explicit hypotheses for the LGM distribution of thicket subtypes.
2. To investigate potential distributional shifts in the climate envelopes of AST subtypes associated with projected anthropogenic climate change. Developing such an understanding is important for the development of long-term conservation and management strategies for thicket (Araújo *et al.* 2004).
3. To assess the extent to which the proposed AST mega-conservancy network (Rouget *et al.* 2006) captures a significant proportion of postdicted refugial areas.
4. To assess the potential shifts of the modelled climate envelopes for each subtype in relation to a proposed mega-conservancy network.

2.3. Methods

2.3.1. Study area and location data

The study area has already been outlined in Chapter 1 (Pg. 19). In brief, the coastal lowlands of South Africa are a long series of relatively short drainage basins found between the coast and Great Escarpment, which can reach over 3000 m in places and separates the coastal lowlands from the interior plateau. The AST covers an area of approximately 47 000 km^2 (Vlok *et al.* 2003) along these lowlands in the Western and

Eastern Cape provinces of South Africa (Figure 2.1). The geology, topography and climate of the southern coastal lowlands are remarkably varied. The AST lies within a bimodal rainfall belt where the most reliable rainfall is recorded in spring and autumn but copious rain may fall – and pronounced dry spells may occur – at any time of the year.

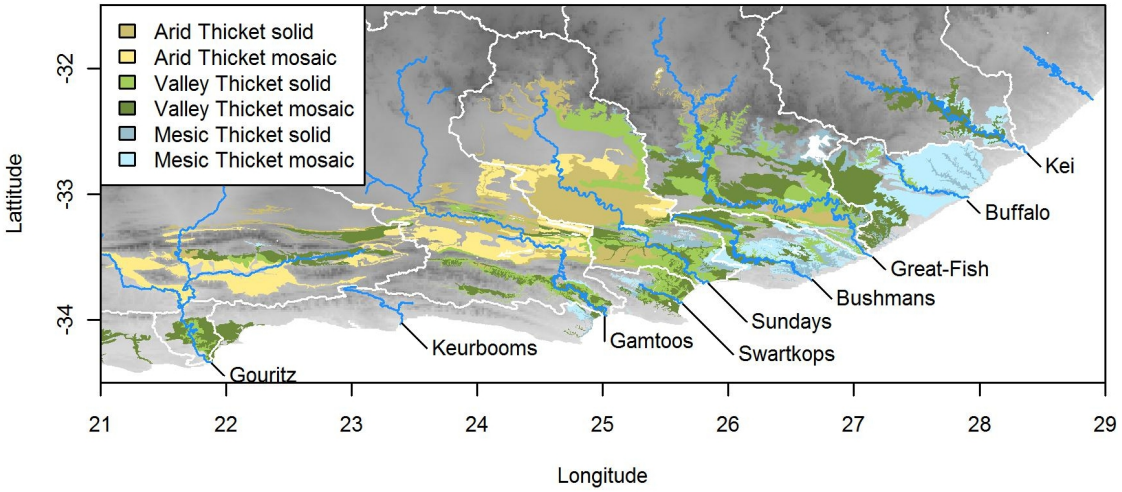

Figure 2.1. Distribution of vegetation subtypes within the Albany Subtropical Thicket study region following Vlok *et al.* (2003). The altitude, rivers (blue) and watersheds between primary catchments (white) are shown.

The AST has been split into structurally and compositionally different subtypes that are clustered floristically into four broad vegetation groupings - arid thicket, valley thicket, mesic thicket and dune thicket (Vlok *et al.* 2003). Only the first three are used for the purposes of this study. As mentioned in Chapter 1, dune thicket vegetation is floristically distinct from the subtypes on other substrata (Low & Rebelo 1996), occupies a very narrow belt along the coast, and would require inclusion of coastal dune field movements which is beyond the scope of this thesis. The three mainland subtypes are arranged along a gradient of increasing moisture availability (arid thicket to valley thicket to mesic thicket) but are differentiated based on ecology (Cowling *et al.* 2005) and structural characteristics, such as the relative cover of woody, grass and succulent species, as well as the incidence of spinescence and woody lianas (Vlok *et al.* 2003). More than 50% of the AST exists as a mosaic where thicket occurs in isolated clumps forming a coarse-grained patchwork (10–500 ha stands of continuous thicket) with neighbouring vegetation such as fynbos, karoo, forest or grassland.

2.3.2. Locality and environmental data

The AST has been extensively surveyed through field observations, which were mapped onto LANDSAT Thematic Mapper imagery and converted to GIS polygons (Vlok *et al.* 2003). For this study, a surface reference data (SRD) raster map was created for each AST subtype using a grid with the same scale (2.5×2.5 arc-minutes) and alignment as the unprojected environmental data. A mask from $-31.5°$ S to $-35.0°$ S and $20.0°$ E to $30.0°$ E was used to sample background data from the environmental layers (Appendix Figure A.1, Pg. 221). The mask covered the distribution of the AST, plus contained representative environmental conditions of the western winter-rainfall regime, the eastern summer-rainfall regime, and the interior plateau. The grid cells are approximately 3.9 km $\times$ 4.6 km, and this remains unchanged across the sampling mask, therefore controlling for the changing area size of cells was unnecessary. A thicket subtype (continuous or mosaic) was considered present in a grid cell if more than 33% of that cell contained the subtype. These SRD maps were used to train and test the community distribution models.

Current climate information was obtained from the WorldClim database (version 1.4, Hijmans *et al.* 2005, http://www.wordlcim.org). These climatic layers are based on weather conditions recorded over 50 years from 1950 to 2000, which were then interpolated using thin plate smoothing splines (Hutchinson 1995) on a 30 arc-second resolution grid. Here I use 19 bioclimatic variables (Table 2.1) derived from the monthly layers, which have been rescaled to 2.5 arc-minute grids (via mean aggregation) as this is the only resolution readily available for LGM global climate models. The bioclimatic variables characterise the dimensions of climate considered pertinent in determining species distributions and represent summaries of annual trends for temperature and precipitation, aspects of seasonality, and extreme or potentially limiting environmental factors.

Table 2.1. Variable description, clustering and selection of the 19 Bioclim variables (Busby 1991). The selected variables were used for community distribution modelling of the impacts of past and future climates on the distribution of Albany Subtropical Thicket vegetation subtypes. See Figure 2.2 (Pg. 34) for the hierarchical clustering of variables.

BIOCLIM code	Selected variables	Climate variables
		Cluster 1
BIO12	X	Annual precipitation (mm)
BIO18		Precipitation in the warmest quarter (mm)
BIO13		Precipitation in the wettest month (mm)
BIO16		Precipitation in the wettest quarter (mm)
		Cluster 2
BIO19		Precipitation in the coldest quarter (mm)
BIO14		Precipitation in the driest month (mm)
BIO17	X	Precipitation in the driest quarter (mm)
		Cluster 3
BIO02		Mean diurnal range (Mean of monthly maximum temperature minus minimum temperature)
BIO04		Temperature seasonality (standard deviation of of annual mean temperature 100)
BIO07	X	Temperature annual range
		Cluster 4
BIO06	X	Minimum temperature of the coldest month (°C)
BIO11		Mean temperature of the coldest quarter (°C)
		Unclustered
BIO05	X	Maximum temperature of the warmest month (°C)
BIO08	X	Mean temperature of the wettest quarter (°C)
BIO15	X	Precipitation seasonality (standard deviation of monthly precipitation values)
BIO01	X	Annual mean temperature
BIO10	X	Mean temperature in the warmest quarter (°C)
BIO09	X	Mean temperature in the driest quarter (°C)
BIO03	X	Isothermality

The LGM climate scenarios were derived from simulations from the CCSM and MIROC GCMs (see Table 2.2 for details). Future climate scenarios are based on the CCCMA, CSIRO-Mk2 and HADCM3 GCMs under the *A2a* and *B2a* emission scenarios (Nakicenovic & Swart 2000) for the year 2050 (Table 2.2). All GCMs have a resolution of 2.5 arc-minutes. Nakicenovic & Swart (2000) provide detailed descriptions of the different climate scenarios. In brief, the *A2a* emission scenario describes a world with continued population growth, slow advances in technological solutions, and fragmented and independently operating nations with regionally orientated economic development. The *B2a* scenario is similar to the *A2a* scenario but emphasizes a more environmentally conscious, but still divided global society and a slower rate of population growth. Both the future and LGM climate estimates have been statistically downscaled using the WorldClim data set (Hijmans *et al.* 2005) and GCM data from the fourth IPCC Assessment Reports and the Paleoclimate Modelling Intercomparison Project II (PMIP2), respectively.

Table 2.2. Details of Last Glacial Maximum (LGM) and 2050 global climate models (GCM) used to project modelled climate envelopes of Albany Subtropical Thicket vegetation subtypes. The LGM and 2050 bioclimatic datasets are available for download from the WorldClim website and the International Center for Tropical Agriculture website, respectively.

Period	Scenarios	GCM	Reference	Source
LGM	-	CCSM	(Collins *et al.* 2004)	http://www.worldclim.org/past
LGM	-	MIROC	(Hasumi & Emori 2004)	http://www.worldclim.org/past
2050	*A2a, B2a*	CCCMA	(Flato *et al.* 2000)	http://gisweb.ciat.cgiar.org/GCMPage/
2050	*A2a, B2a*	CSIRO-Mk2	(Hirst *et al.* 1996)	http://gisweb.ciat.cgiar.org/GCMPage/
2050	*A2a, B2a*	HADCM3	(Gordon *et al.* 2000)	http://gisweb.ciat.cgiar.org/GCMPage/

2.3.3. Community distribution modelling

The environmental variables used to generate CDMs of the AST subtypes are based exclusively on biologically meaningful aspects of climate variation (e.g. Hugall *et al.* 2002). Thus, the CDMs used here should only be interpreted in terms of climatic suitability rather than total environmental suitability. Community distribution modelling, and species distribution modelling, follows a three-stage process. Firstly, data of known presences, in this case the SRD maps, are used in combination with environmental data of current conditions to construct a statistical model of the climatic

envelope. Secondly, this model is then extrapolated over a landscape surface using raster grids of environmental data representing past, present or future conditions. These two steps were performed with the maximum entropy and machine-learning algorithm implemented in MAXENT version 3.3.3a (Phillips & Dudík 2008, Phillips *et al.* 2006), where the model output is usually interpreted as the 'probability of occurrence'. MAXENT performs well in comparison to other methods (Elith *et al.* 2006) for predicting the past and future distribution of species or communities (Carnaval & Moritz 2008, Hijmans & Graham 2006, Pearman *et al.* 2008). Thirdly, a presence-absence prediction is required, so a threshold value was applied to reclassify the 'probability' map to binary data ('present' or 'absent'). The threshold value was calculated using the maximum test sensitivity plus specificity criterion, which has been shown to perform well in comparison with other possible threshold criteria (Liu *et al.* 2005).

Climate variables are often highly inter-correlated, and this multicollinearity may affect the modelling of climate envelopes (Dormann 2007). Therefore, I reduced the potential redundancy by selecting single variables that are representative of a group of correlated variables. To achieve this, correlated climate variables were clustered using hierarchical clustering with a dissimilarity measure of $1 - R^2$. Variables were considered clustered if the group had a significant correlation value greater than 0.8. The hierarchical clustering results and the selected subset of variables used for all modelling are shown in Figure 2.2 and Table 2.1, respectively. MAXENT also uses l_1 regularization which forces the program to focus on the most important climatic layers and thus further avoids overfitting of the model (Phillips et al. 2006). The default settings in MAXENT were used as these have been optimised across a wide range of data sets and automatically selects suitable regularization values and functions of environmental variables (Phillips & Dudík 2008).

Model evaluation was performed using 10-fold (K-fold) cross-validation. This evaluation method randomly partitions the locality data into K subsamples, each of which is in turn used as test data, while the remaining K-1 subsamples are used for training data (i.e. the cross-validation process is repeated K times). At each of the 10 iterations, the cells in the generated probability of occurrence maps were converted to binary present-absent values, and only cells that were predicted as present by all iterations were used to produce the final depiction of the climate envelope. Model performance was evaluated at each iteration using the standard statistical measure of

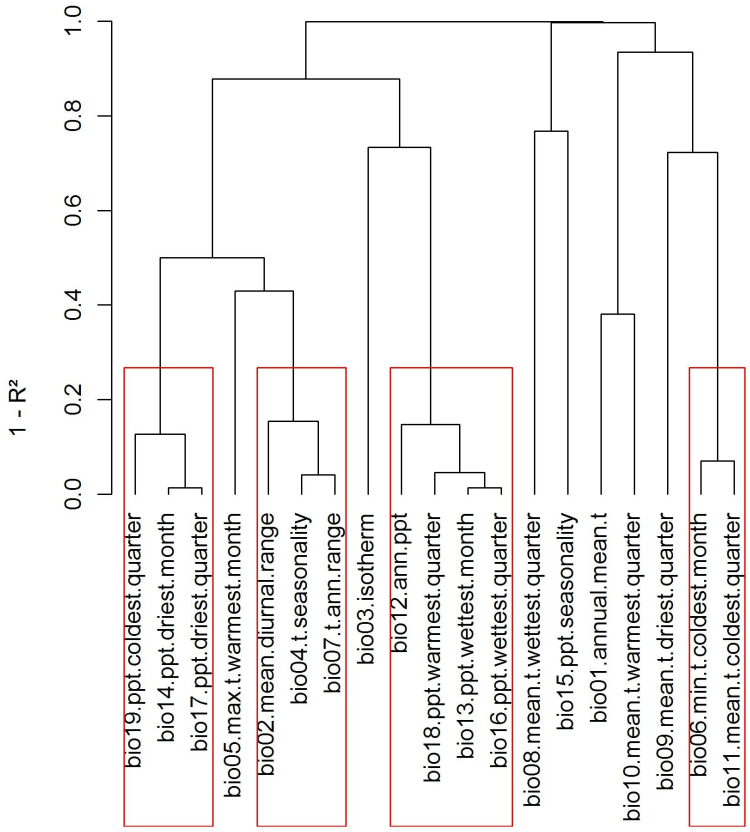

Figure 2.2. Hierarchical clustering of the 19 climate variables using a dissimilarity dendrogram. Clusters shown in red boxes are highly inter-correlated ($R^2 \geq 0.8$). Variable descriptions and codes are shown in Table 2.1 (Pg. 31).

predictive ability, the area under the receiver operating characteristic curve (AUC). The AUC statistic ranges from 0.5 (model prediction is no better than random) to 1.0 (perfect model prediction of presence versus absence).

Projecting the CDM onto novel climatic conditions that the model has not experienced during training can produce misleading results (Elith *et al.* 2010, Pearson *et al.* 2006). In order to determine areas that may be affected by novel climatic conditions I used clamping and the 'multivariate environmental similarity surface' (MESS) as implemented in MAXENT (Elith *et al.* 2010). MAXENT limits or 'clamps' variable values that fall outside the training range to the limit of the training range; areas of novel climate are identified by comparing the absolute difference between predictions with and without clamping. The MESS statistic measures the similarity of any given point to a reference set of points, with respect to the predictor variables

(Elith *et al.* 2010). Positive values indicate cells that are similar to the environmental values used for training, whereas negative values indicate novel climate.

In order to determine the overlap between climate envelopes of the different subtypes I used two measures introduced by Warren *et al.* (2008), as implemented in the PHYLOCLIM library version 1.0 in R version 2.13 (R Development Core Team 2011): Schoener's D and a measure derived from the Hellinger distance called I. Both of these similarity measures range from 0, when the two predicted envelopes do not overlap, to 1, when the two predicted envelopes are identical. These two measures were used to test two hypotheses: 1) the niche equivalency test asks whether the CDMs of two subtypes are more different than expected if they are drawn from the same underlying distribution, and 2) the background similarity test asks whether CDMs drawn from subtypes with partially or entirely non-overlapping distributions are any more different from one another than expected by chance. The null distributions for these two tests were generated using the PHYLOCLIM library and 100 replications.

2.3.4. Spatial analysis of subtype distribution under altered climates

The modelled effects of altered climates on the distribution of the AST subtypes were assessed by calculating the gain, loss, change and turnover in range. Here I use the term 'range' in the context of CDMs, thus referring to the modelled distribution of climatic suitability for each of the subtypes, which must not to be mistaken with the entire biological envelope or functional niche. As MAXENT has been noted to be prone to over-prediction (Hijmans & Graham 2006) and this is likely to lead to over-prediction in projections under altered climate, I calculate all of the following statistics in relation to the present day CDM rather than the actual subtype distribution as represented by the surface reference data. The percentage of overall predicted change in range size (C) of the climate envelope was estimated using:

$$C = 100 \times \frac{(RG - RL)}{CR}$$

where RG is the range gain by cell, RL is the range loss by cell, and CR is the current predicted range by cell. A negative C value indicates a loss in overall range, whereas a positive value indicates an increase in overall range size. The turnover (T) by cell

of the climate envelope range was estimated using:

$$T = 100 \times \frac{(RG + RL)}{(CR + RG)}$$

A turnover value of 0 indicates no shift in range, whereas a turnover of 100 indicates a complete turnover of the range.

The percentage of current SRD and past, future and current CDM coverage within the proposed megaconservancy network was calculated in order to determine the extent to which the altered climates may shift the climate envelopes within this network. Megaconservancy networks are linked areas that span major climatic gradients which are designated for conservation management in such a way as to facilitate resilience of the component biota to anticipated climate change (Rouget *et al.* 2006).

2.4. Results

The cross-validation results of CDMs for the different AST subtypes all showed 'good' AUC values (≥0.80; Table 2.3), following previously given definitions (Taylor & Hellberg 2006), indicating that the models have high specificity (true positive rate) and sensitivity (false positive rate; but see Lobo *et al.* 2008 regarding problems with this statistic when used for distribution modelling). I do not report the relative contribution of each variable to the various models or the jacknife variable analysis results as these may be misleading when environmental variables are correlated. Although the most highly correlated variables were removed, a number still remain which display moderate correlation. Under the selected threshold criterion, the CDMs greatly overpredicted the actual distribution of the different thicket subtypes; however, the models did predict the majority of the actual distributions with very low levels of omission (Table 2.3; Figure 2.3). The niche equivalency and background similarity tests indicate that the CDMs developed for each subtype were significantly more different from one another than would be expected if they came from the same underlying distribution or expected by random chance (Table 2.4).

The altered conditions for the LGM and the 2050 GCMs relative to present day climate are reported in Appendix Figure A.2 (Pg. 222). The LGM climates estimated by the CCSM and the MIROC simulations were broadly similar over the study region,

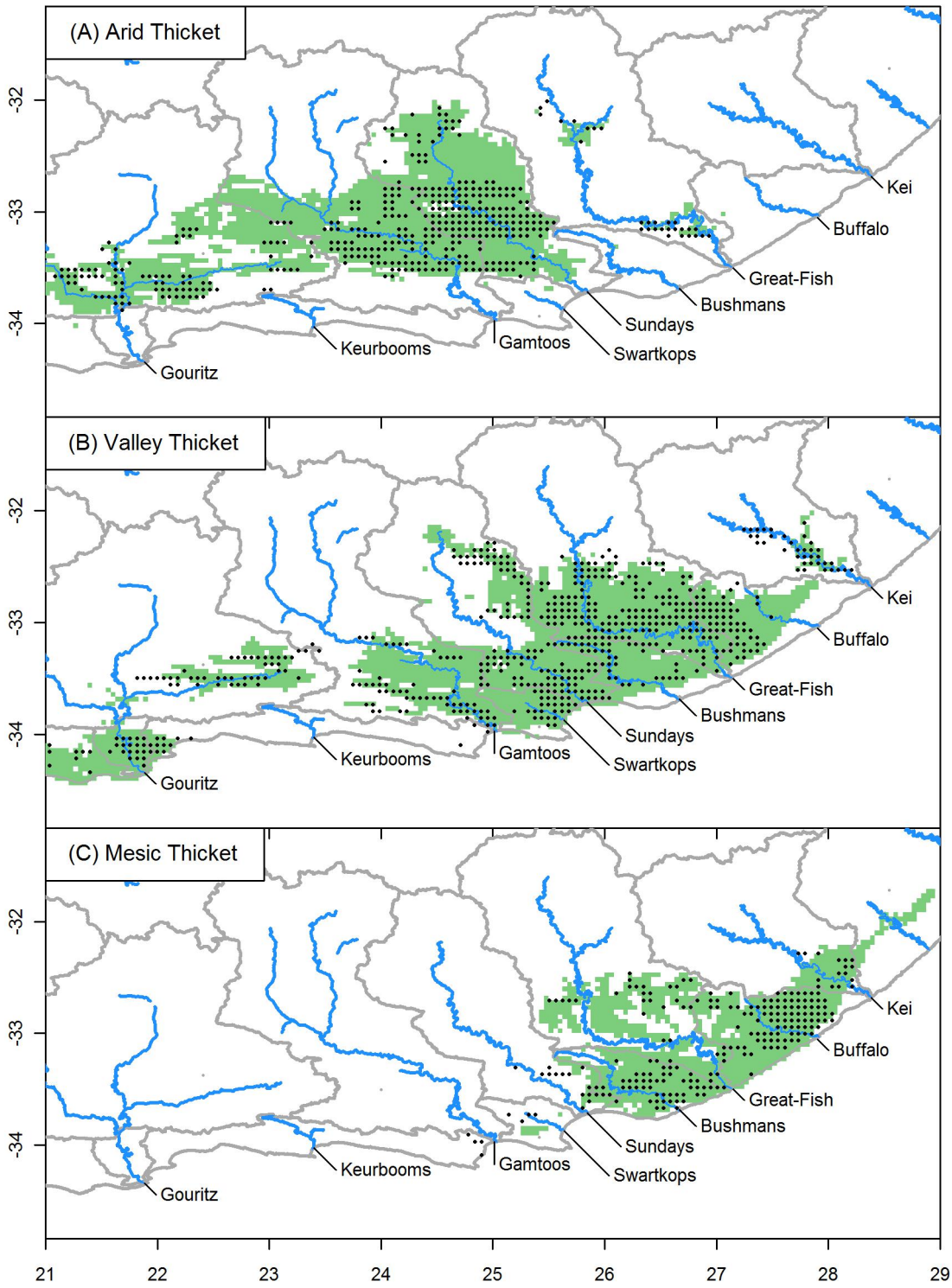

Figure 2.3. The distribution of the three mapped inland Albany Subtropical Thicket subtypes (dots; Vlok *et al.* 2003) and the community distribution models (green) under current climatic conditions.

Table 2.3. Comparison of predicted community distribution model (CDM) distribution of Albany Thicket subtypes and their present day distributions as represented by the surface reference data (SRD). The threshold criterion used to convert the probability landscape into a binary presence-absence distribution was maximum test sensitivity plus specificity. SRD and CEM are expressed as number of grid cells, whereas Common, Omissions, Commissions are expressed as percentage of the SRD. The model validation statistic of area under the curve (AUC) is included.

Thicket type	Arid	Valley	Mesic
SRD	1018	1448	584
CDM	2297	2663	1561
Area Difference (%)	226	184	267
Common (%)[a]	92	88	93
Commissions (%)[b]	134	95	175
Omissions (%)[c]	8	12	7
AUC (SD)	0.846 (0.007)	0.809 (0.008)	0.901 (0.005)

[a]Common = cells shared by SRD and the model.
[b]Commissions = cells predicted by the model but not actually present in the SRD.
[c]Omissions = cells present in the SRD but not predicted by the model.

Table 2.4. Results for the niche similarity indices *I* and *D*. All *I* and *D* values were significant for the niche equivalency test and niche similarity test (see Methods).

I \D	Arid	Valley	Mesic
Arid	-	0.5047	0.1231
Valley	0.6657	-	0.6078
Mesic	0.4138	0.725	-

both cooler and drier, but there were some differences. Notably, MIROC was colder than CCSM for all temperature variables, and although the medians were equivalent for the precipitation variables, MIROC has a far greater range of precipitation differences in comparison with the current climate than the CCSM. The 2050 GCMs under the two scenarios contain many differences, although all indicate warmer than present climates and most models indicated drier precipitation regimes. Isothermality (BIO03) and temperature seasonality (BIO04) varied greatly between models and scenarios for 2050.

Under the cooler and drier climate conditions of the LGM, all subtypes were projected to have a reduced range, with arid and valley thickets contracting into

fragmented refugial areas (Figure 2.4). However, there were notable differences between the subtype distribution reconstructions from the two models, with MIROC generally postdicting a much smaller distribution area than the CCSM models. The distribution of arid and valley thickets under MIROC was largely nested within the larger distributions postdicted by the CCSM. However, the two GCMs postdicted very different distributions of mesic thicket, with MIROC postdicting the greatest LGM distribution east of the Kei River and CCSM postdicting the greatest distribution west of the Kei River. Due to the limited number of GCMs used here I considered any cells predicted by either model as possible refugial areas for the thicket subtypes. The arid thicket displayed a contraction into the lower reaches of the catchments with the largest refugial areas occurring in the Gamtoos and Sundays catchments, and smaller refugia in the Gouritz, Swartkops and Great-Fish catchments (Figure 2.4.A). The valley thicket displayed a complete range loss in the western Gouritz catchment, but had a large refugium in the eastern Great-Fish and Buffalo catchments. Smaller refugia are postdicted in the lower reaches of the Gamtoos, Swartkops and Sundays Rivers (Figure 2.4.B). The postdicted range loss experienced by the mesic thicket was not as severe or as fragmented as the other subtypes, and included a range shift onto the exposed continental shelf and a shift eastward along the coast. A fairly continuous refugium for mesic thicket was predicted near the coast from the Bushmans catchment to the Kei catchment and beyond (Figure 2.4.C). There was a large percentage of range loss for all thicket subtypes across both LGM GCMs (Table 2.5). The percentage range loss was ordered arid thicket > valley thicket > mesic thicket, and only the mesic thicket displayed any range gain.

For the warmer and drier future climate of the 2050 CDMs there was a large degree of agreement in projected distribution for each of the subtypes under the different models for each scenario, and the two scenarios showed qualitatively similar results (Figures 2.5 and 2.6). For the arid thicket, most of the current CDM distribution fell within the distribution predicted by three or two models, with the most noticeable decline in range in the Gouritz catchment and a sizeable expansion in the Great-Fish catchment (Figures 2.5.A and 2.6.A). For the valley thicket the majority of the areas predicted by two or three models fell within the current CDM for the subtype. Range loss was predicted largely at the margins of the current CDM in the Sundays and Kei catchments, with the greatest range gain occurring in the Gouritz catchment (Figures 2.5.B and 2.6.B). Projected future distribution of mesic thicket showed a

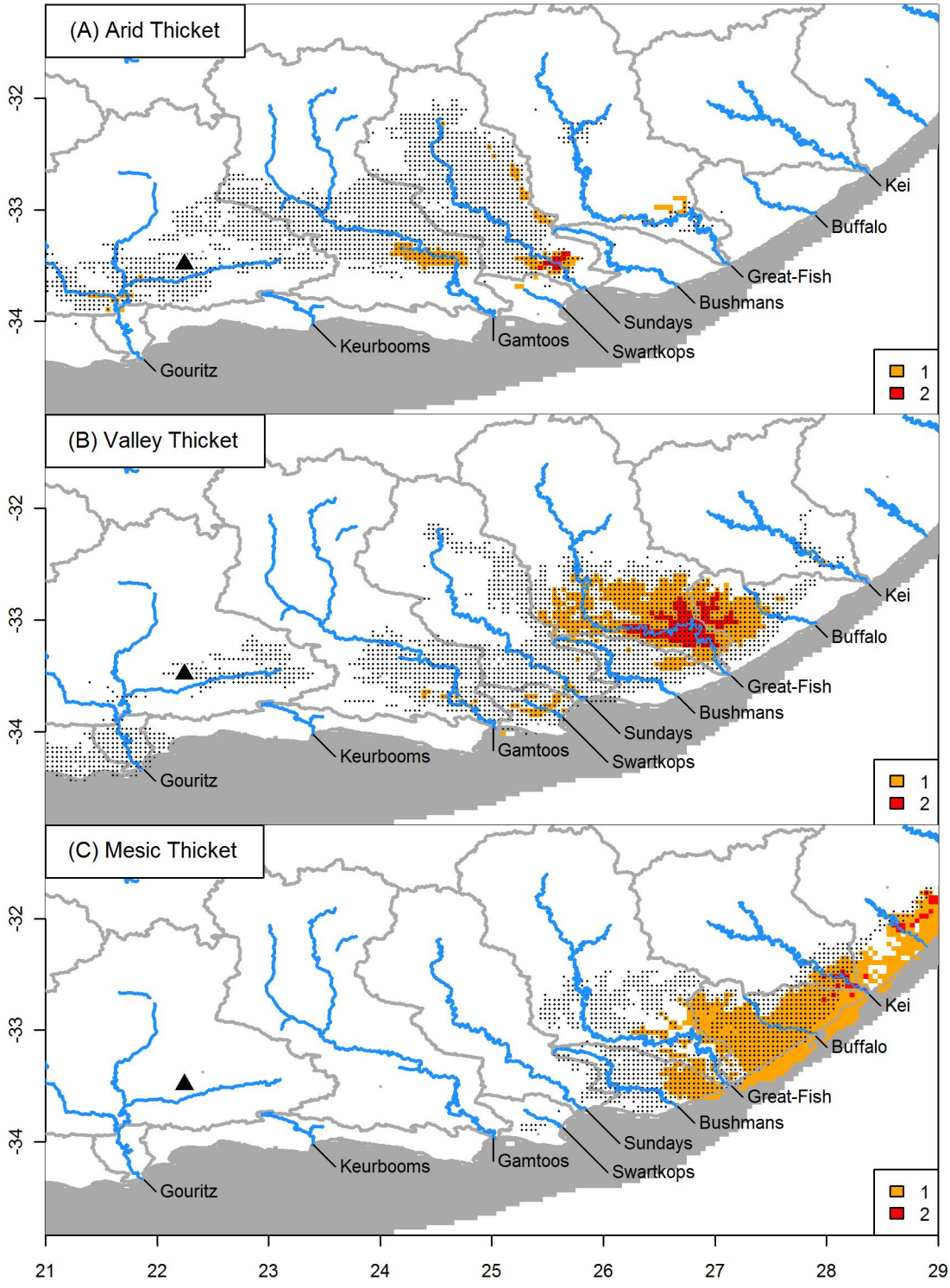

Figure 2.4. The community distribution models of the three Albany Subtropical Thicket subtypes under current (dots) and projected Last Glacial Maximum (shaded) climatic conditions. Two global climate models were used to derive the LGM distribution: (1) CCSM and (2) MIROC. Grey shading shows the continental shelf exposed due to lower sea levels during the Last Glacial Maximum. The Boomplaas Cave is shown (black triangle; see text for details, Pg. 49).

large proportional loss of range, with a retraction towards the coast, and a potential fragmentation of the subtype (Figures 2.5.C and 2.6.C). An area of high climatic suitability is predicted in the Keurbooms catchment, but mesic thicket vegetation is unlikely to reach this area due to its distance and isolation. The projected MAXENT CDMs onto 2050 conditions under the two scenarios suggested an expansion of arid thicket, with little range loss, a relatively stable overall area of valley thicket with little range loss or range gain, and a substantial decline in overall area for the mesic thicket, with little range gain (Table 2.5). The percentage range loss was ordered mesic thicket > valley thicket > arid thicket.

No areas of novel climate were identified by clamping that coincided with predicted CDM distributions for any of the past and future climate conditions under the different GCMs. The MESS maps for each AST subtype projected onto LGM and 2050 GCMs are shown in Appendix Figures A.3 (Pg. 226), A.4 (Pg. 227), and A.5 (Pg. 228). The MESS maps for the two LGM GCMs did not identify any novel climatic areas that coincided with the postdicted CEM distribution for this period. In contrast, the MESS maps did identify areas of overlap between predicted distributions of the CDM under 2050 conditions and novel climate. For the arid thicket, areas of novel climate were identified in the middle reaches of the Sundays and the lower reaches of the Great-Fish catchments for both *A2a* and *B2a* scenarios. The MESS maps found minimal overlap between novel climate and the 2050 CDM distribution of mesic and valley thicket for either scenario.

The extent to which plants were able to colonise the continental shelf during periods of lowered sea levels remains unknown. In this study, only mesic thicket is predicted to have extended onto this shelf. However, this and the fact that arid and valley thicket subtypes did not, should be treated with some scepticism as the extrapolation and downscaling of LGM GCM climate onto the shelf may be misleading. The statistical downscaling procedure uses fine-scaled current land-surface climate layers. The present-day climate over the shelf is controlled by the sea-surface, rather than a land-surface, thus it is not analogous and can affect the downscaling.

With regards to the mega-conservancy network, the current CDMs over-predicted the actual distribution of the subtypes within the network; however the overprediction is relatively constant between the subtypes (144 to 160% of the actual distribution; Table 2.6). This overprediction causes some overlap between the CDMs, and so the

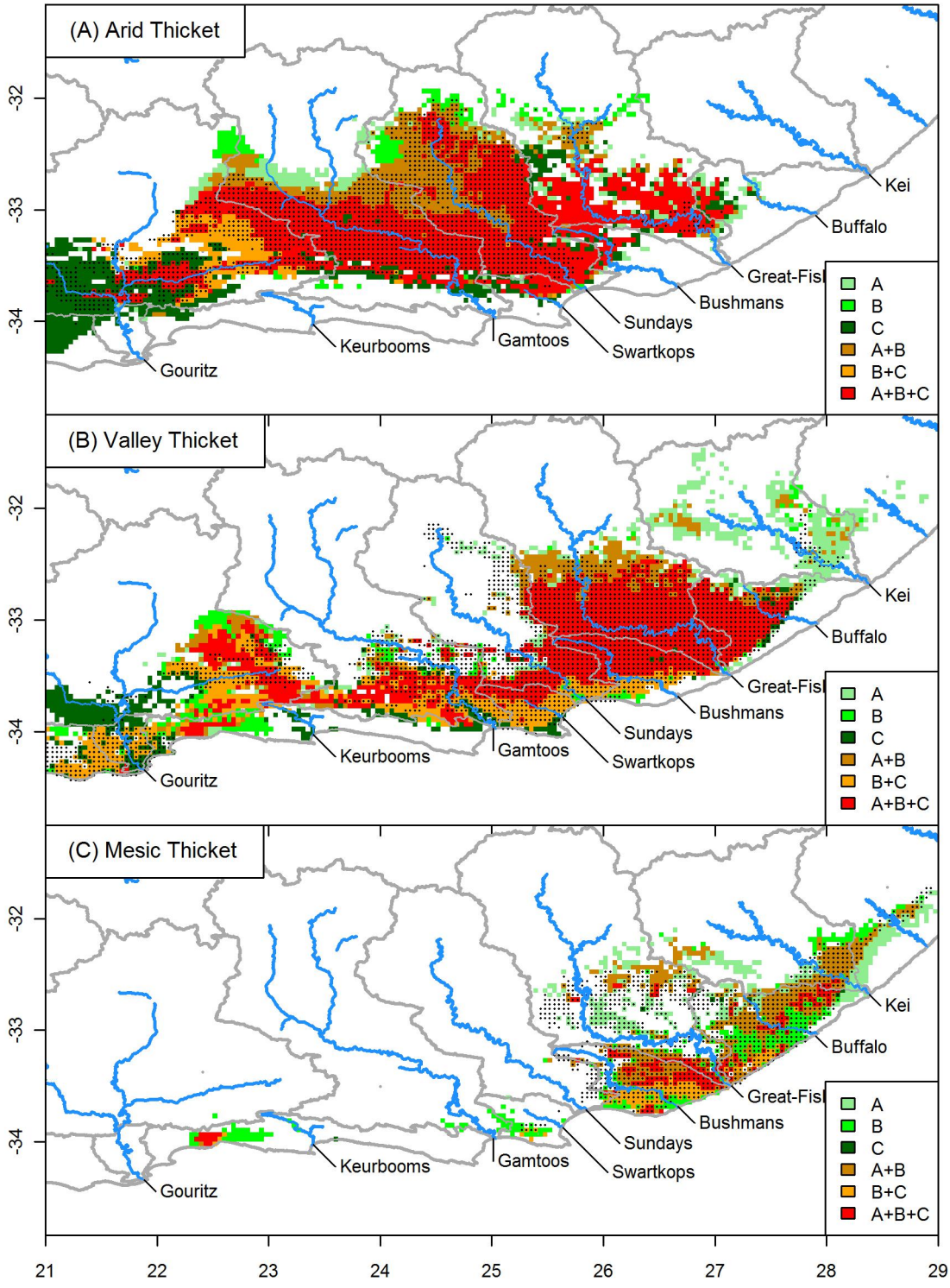

Figure 2.5. The community distribution models of the three Albany Subtropical Thicket subtypes under current (dots) and projected 2050 scenario *A2a* climatic conditions (shaded). The GCMs used were: (A) CCCMA, (B) CSIRO, and (C) HCCPR.

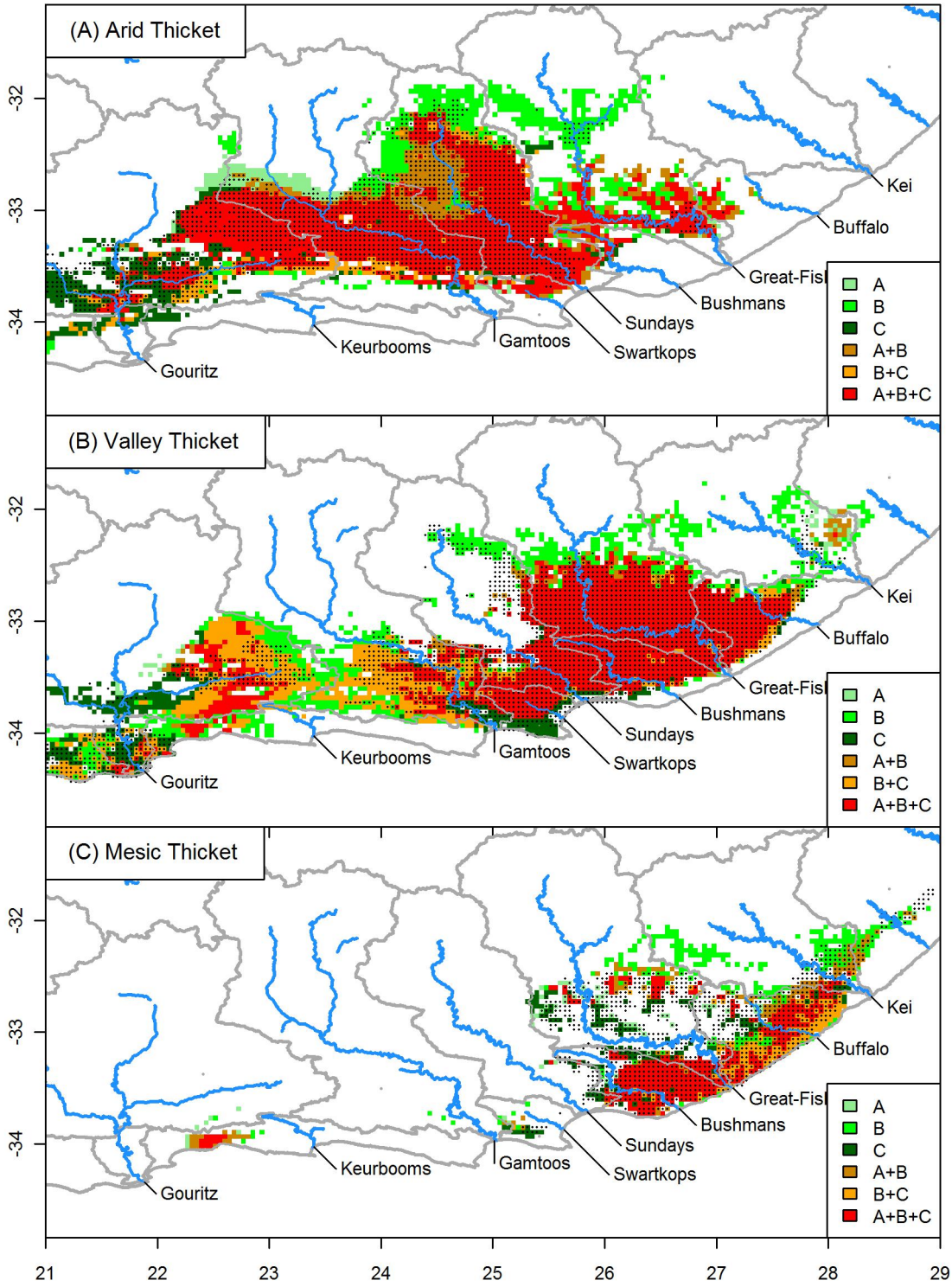

Figure 2.6. The community distribution models of the three Albany Subtropical Thicket subtypes under current (dots) and projected 2050 scenario *B2a* climatic conditions (shaded). The GCMs used were: (A) CCCMA, (B) CSIRO, and (C) HCCPR.

sum of the network coverage can be greater than 100%. Valley thicket had the greatest percentage present in the network during the LGM, followed by mesic thicket and then arid thicket. Under future scenarios, arid thicket expanded its current percentage of the network, valley thicket declined slightly, and mesic thicket almost halves the current percentage of its climate envelope in the network. The distributions of each subtype during the LGM and two future scenarios of 2050 in relation to the network are shown in Appendix Figures A.6 (Pg. 229), A.7 (Pg. 230) and A.8 (Pg. 231), respectively.

Table 2.5. Comparison of the range changes of community distribution models for Albany Thicket subtypes between current climatic conditions and Last Glacial Maximum or 2050 *A2a* and *B2a* scenario conditions. The area, range gain (RG), range loss (RL), range change (C), and range turnover (T) are presented as the percentage of past or future climate envelope relative to the present-day climate envelope.

Scenario	Thicket Subtype	Model	Area	RG	RL	C	T
LGM	Arid	MIROC	7	0	93	-93	93
		CCSM	1	0	99	-99	99
	Valley	MIROC	21	0	79	-79	79
		CCSM	6	0	94	-94	94
	Mesic	MIROC	80	14	43	-28	50
		CCSM	25	17	96	-79	96
2050: *A2a*	Arid	CCCMA	142	55	13	42	44
		CSIRO	133	44	11	33	38
		HCCPR	137	56	19	37	48
	Valley	CCCMA	139	56	16	40	46
		CSIRO	127	39	11	28	36
		HCCPR	107	28	20	8	38
	Mesic	CCCMA	70	17	45	-28	52
		CSIRO	85	23	36	-13	48
		HCCPR	25	6	78	-73	79
2050: *B2a*	Arid	CCCMA	108	29	21	8	38
		CSIRO	134	47	13	34	41
		HCCPR	113	30	17	13	36
	Valley	CCCMA	92	17	24	-7	35
		CSIRO	148	55	6	49	39
		HCCPR	109	23	13	10	29
	Mesic	CCCMA	67	11	42	-31	48
		CSIRO	82	25	41	-15	53
		HCCPR	60	7	45	-38	48

Table 2.6. The percentage of the mega-conservancy network covered by the surface reference data (SRD) and the community distribution models (CDM) of current, Last Glacial Maximum (LGM) and 2050 climate. The 2050 climate models were based on two scenarios of climate change (*A2a* and *B2a*). The percentage sum of network coverage can be greater than 100 due to overlapping climate envelopes.

	Current		LGM		2050: A2a			2050: B2a		
Subtype	SRD	CDM	CCSM	MIROC	CCCMA	CSIRO	HCCPR	CCCMA	CSIRO	HCCPR
Arid	27	41	5	1	60	54	57	44	56	50
Valley	43	62	15	4	43	48	48	39	58	55
Mesic	15	24	11	1	13	14	4	11	11	12
Non-thicket	15	15	50	77	2	4	7	9	4	6

2.5. Discussion

The modelling and projection of climate envelopes onto altered climates has provided valuable, and sometimes testable, insights for vegetation communities ranging from tropical rainforests to succulent deserts (Midgley & Roberts 2003, VanDerWal *et al.* 2009*a*). Very little is known regarding the historical or possible future distributional shifts of the AST vegetation. Here I have developed significantly different CDMs for each of the well-defined AST subtypes (Table 2.4). These CDMs are used to propose hypotheses regarding subtype distribution during the LGM and under two future scenarios of climate change for 2050.

2.5.1. Modelling approach

The modelling of the climate envelope, or niche, assumes that the current species distribution is in equilibrium with its environment and that its realised niche is equivalent to the fundamental niche (Phillips *et al.* 2006). Projecting these models onto altered climates involves many further assumptions, such as unchanging precedence of limiting factors, unchanging biotic interactions, unchanging natural processes such as disturbance regimes, and negligible effects from genetic variability, phenotypic plasticity or evolutionary changes (Dormann 2007, Harte *et al.* 2004). Also, extrapolation errors can occur when models are projected onto novel climates (Elith *et al.* 2010). Furthermore, modelling floristic communities, such as biomes, also assumes that species composition remains relatively static. There is some evidence for 'biome conservatism' within species (Crisp *et al.* 2009), but this likely only involves broad vegetation structure and does not apply to the composition of local or even regional communities (e.g. Williams *et al.* 2004). The widely used 'biomization' procedure used to reconstruct historical biome distribution through pollen core records also assumes a largely static species composition (Prentice *et al.* 1996).

These assumptions can be seen as fairly significant stumbling blocks when assessing the potential distributional changes of a biome (and of a single species) as we are unable to incorporate uncertainty in how these will affect the distribution of species beyond 'simple' changes in climate. Additional uncertainties surround the method used to derive the distribution model (Elith *et al.* 2006, Hijmans & Graham 2006), as well as in the development of past or future climate conditions using GCMs

(Beaumont *et al.* 2008). Despite these compounded uncertainties, I feel that niche modelling can still provide a valuable, if not precise, insight into both the history and potential future of AST for the following reasons:

1. Climatic niche conservatism of plant species has been demonstrated over moderate time periods (i.e. LGM to present) despite profound changes in climate and environmental conditions (Martínez-Meyer & Peterson 2006).
2. Biome compositional change observed in pollen archives appears to be largely affected by the shift to non-analogous climates (Webb *et al.* 2003, Williams *et al.* 2001). Conversely, biome composition in altered environmental conditions is more likely to remain relatively constant in areas that are similar to the current climate envelope of vegetation communities. Thus, the projection of CDMs based on climate will indicate where communities have the greatest potential to remain constant.
3. The CDM approach has been used with the Brazilian Atlantic and Australian Wet Tropic forests with positive, but limited, validation through pollen records and congruence with phytogeographic or phylogeography data (Carnaval & Moritz 2008, Moussalli *et al.* 2009, VanDerWal *et al.* 2009*a*).
4. The projected CDM distributions of the different subtypes onto both past and present climatic conditions do not overlap with areas of novel climate, with the exception of a small area of the arid thicket CDM distribution predicted for 2050. Thus, extrapolation errors should be minimal.

I also interpret the results not as the shifting distribution of the vegetation, but rather the shifting distribution of the vegetation's current climate envelope. It is within this envelope that the theoretical assumptions of biome niche modelling are likely to remain true. Thus, while the community modelling approach will most likely not accurately predict the distributional changes in individual species, it will provide an overall mean distributional response of characteristic species of a biome in altered climate conditions.

2.5.2. AST subtypes during the LGM

Despite a lack of glaciation on the continent, the glacial-interglacial climatic oscillations greatly affected the distribution of biomes in Africa (e.g Hamilton & Taylor

1991, Scott *et al.* 2008). The CDM results indicate dramatic range reductions during the LGM for all AST subtypes; arid and valley thickets suffered the most severe reductions with an overall decline in altitude and concomitant fragmentation, whereas mesic thicket had a reduced but fairly continuous range that included a shift onto the exposed continental shelf. These results provide the first regional support for the glacial refugia hypothesis, which postulates that the AST contracted and expanded in response to glacial and interglacial periods, respectively (Cowling *et al.* 2005).

The physiological and competitive effects of atmospheric [CO_2] and fire are not included in the community distribution models. These may have had significant consequences on the distribution of the thicket subtypes besides the changes in climate. For example, range contraction of the thicket subtypes during the LGM may have been further exacerbated by the low atmospheric CO_2 favouring neighbouring C_4 fire-driven ecosystems (Bond 2008, Bond *et al.* 2003) as thicket is largely fire intolerant (Cowling *et al.* 2005, Trollope 1974); however, the effects of lower temperatures may have reduced the intensity and frequency of fires (Govender *et al.* 2006). It is difficult to predict how atmospheric [CO_2] and fire may alter the results; however the results mirror the loss of woody biomass in the neighbouring fire-driven grasslands (Bond *et al.* 2003), so it is unlikely that the thicket vegetation increased in range.

The GCMs indicate a decrease in mean precipitation during the LGM (Appendix Figure A.2, Pg. 222), yet the arid thicket is postdicted to have been severely reduced in range during the same period. This potential contradiction is because the distribution of the arid thicket is not determined only by precipitation. Although this subtype is called the 'arid' thicket, a more correct name would be the 'arid and warm' thicket. A key threshold for plant physiology is air temperature near or at freezing point (Woodward 1987). The arid thicket is dominated by succulent species, such as *Portulacaria afra*, which are particularly sensitive to frost (Cowling *et al.* 2005, Von Willert *et al.* 1990). The cooler temperatures and resultant increase in frost events found in the CCSM and MIROC GCMs (Appendix Figure A.2, Pg. 222) are the likely drivers for the range decline of the arid thicket climate envelope despite the increased aridity during the LGM (Von Willert *et al.* 1990).

The decreased mean temperature and precipitation of LGM bioclimatic variables derived from the CCSM and MIROC GCMs (Appendix Figure A.2, Pg. 222) correspond with a recent regional synthesis of palaeo-archives. Chase & Meadows

(2007) suggest that the study area would have been both colder and drier during the LGM; however quantitative estimates for the degree of cooling or drying are not given due to methodological uncertainties and conflicts that surround the archives (reviewed in Chase & Meadows 2007). Therefore, the LGM bioclimatic variables are in accordance with the current knowledge regarding the LGM climate over the study area.

The charcoal and macrofaunal deposits found at Boomplaas Cave (Figure 2.4) provide the only inland validation point for the predicted LGM distributions. The present-day vegetation surrounding Boomplaas Cave is comprised of continuous arid thicket and arid thicket/succulent karoo mosaics (Moffett & Deacon 1977, Vlok *et al.* 2005). The models postdict that no AST subtypes occurred near this site during the LGM. This is supported by charcoal deposits, which suggest the presence of asteraceous shrubland during the LGM and a lack of thicket species in the surrounding area, with a transition to valley thicket occurring towards the Holocene (Deacon *et al.* 1984, Scholtz 1986). Also, macromammalian remains provide further evidence for a lack of thicket as a dominance of grazers is observed during the LGM suggesting open grassland in the surrounding area (Faith in press, Klein 1980). A replacement by browsers towards the wetter-than-present Holocene is thought to indicate an expanding thicket biome (Cowling *et al.* 2005).

The results provide spatially explicit hypotheses regarding the refugial areas of the AST during the altered environmental conditions of the LGM. The CDMs identify potential LGM refugia for the subtypes, as well as areas of AST expansion. The arid and valley thicket refugia are near or coincide with present-day centres of continuous thicket for the respective subtypes (compare Figures 2.1 and 2.4). This suggests that the present-day distribution of continuous and mosaic thickets may have been largely influenced by the proximity to and size of the LGM refugial areas. Range reduction and fragmentation of these two subtypes resulted in a downward retraction into the catchment valleys. Populations would have been separated by high watersheds with unsuitable climatic conditions. Thus, the watersheds along the coastal lowlands may have been significant barriers to gene flow for thicket species. Such barriers to dispersal can be expected to leave genetic fingerprints within the component species of the AST (e.g. Médail & Diadema 2009, Tribsch & Schonswetter 2003).

Hypotheses of the LGM distribution of thicket subtypes can in principle be tested

with further studies of palaeoenvironmental archives (e.g. pollen cores). However, these are rare and temporally limited in the study region (Lewis 2008). Another, possibly more viable approach, given this dearth of suitable palaeoenvironmental data, is to search for evidence of the postdicted range shifts and changes in population size in genetic data from dominant species (Médail & Diadema 2009, Soltis *et al.* 2006, Taberlet *et al.* 1998). A number of expectations regarding genetic diversity can be derived from these CDM results:

1. The highest genetic diversity of the different thicket types should be found in the postdicted refugia. Therefore, genetic diversity of arid thicket species should largely reside in the Gamtoos and Sundays catchments, whereas the Great-Fish and neighbouring Buffalo catchments should be repositories for valley thicket species, and genetic diversity for mesic thicket species should reside near the coast across the Bushmans and Kei catchments.
2. Clade subdivision of closely related species and population subdivision of widespread species should be evident between catchments for arid and valley thickets, with far less subdivision evident for mesic thicket species. The greatest levels of subdivision for arid and valley thicket species should be between the Gouritz catchment and those eastward for arid and valley thicket. Limited subdivision in the mesic thicket should occur between populations on either side of the Great-Fish River.
3. Signatures of population and range expansion should be evident from refugial areas into the expanded current distribution for the AST subtypes.

There is already some evidence for the second hypothesis. A small clade of species within the genus *Euphorbia*, usually associated with the AST, displays a pattern of restricted species distributions, usually only within single or neighbouring catchments; also there is a phylogenetic divide between species in the Gouritz catchment and those found in the Gamtoos catchment and eastwards (Ritz *et al.* 2003). The barrier effects on genetic diversity of the series of watersheds between catchments are also evident in the *Platypleura plumosa* (Hemiptera: Cicadidae) complex (Price *et al.* 2010). Clades within this complex are largely restricted to single or neighbouring primary catchments, suggesting that watersheds have been strong barriers to gene flow. The cicadas most likely had to track their host species, which may have experienced fragmentation and retraction into catchment valleys as plant distributions shifted

during glacial periods.

2.5.3. Potential effects of projected 2050 climate change

Climate change is predicted to have a significant impact on the distribution of species (Loarie *et al.* 2008, Thomas *et al.* 2004) and biomes (Eeley *et al.* 1999), and recent range shifts have already been attributed to anthropogenic climate change (Parmesan 2006). Understanding how species or communities will shift in a landscape – and tracking their niches – is important for the conservation and management of biodiversity (Araújo & Rahbek 2006, Araújo *et al.* 2004). This study follows on others trying to model the distribution of vegetation at the biome scale as a result of climate change (Eeley *et al.* 1999, Midgley *et al.* 2002). Although I have used climate change projections for 2050, much of the projected climate change has not yet occurred according to the climate models, especially for the A emission scenarios (see Fig. 10.4 in Solomon *et al.* 2007). Thus, these results describe an intermediate future, depending on the emissions pathways that are to be set over the next few years through international negotiations. Nonetheless, the results reveal a fairly optimistic outlook for the AST, at least for the arid and valley thickets, in comparison to other studies modelling vegetation units (Midgley *et al.* 2002) or syntheses of individually modelled species within a biome (Broennimann *et al.* 2006, Loarie *et al.* 2008). Across three GCMs and two scenarios of future climate, both arid and valley thicket are predicted to have relatively low levels of range loss, which is generally compensated by predicted range gain. Very similar results were obtained when the distributional shifts of *Portulacaria afra*, a dominant species in the arid thicket and a common component of many valley thicket communities (Vlok *et al.* 2003), were modelled under similar scenarios of future climate (Robertson & Palmer 2002). Importantly, the areas of range loss are outside of the postdicted LGM refugial areas. If the refugial areas postdicted here are confirmed to be havens of species and genetic diversity, then this suggests that future climate change may not have a substantial effect on the biodiversity of these subtypes. However, in contrast to the arid and valley thicket, the mesic thicket is projected to undergo range loss with very little range gain. This suggests that mesic thicket species may experience the greatest impact of climate change.

Again, both atmospheric [CO_2] and fire may play important roles and alter the results. Elevated CO_2 levels are predicted to affect the distribution of plant

species, especially C_4 plants (generally grasses and sedges) which are at a competitive disadvantage at higher levels of CO_2 (Ehleringer *et al.* 1997, Kgope *et al.* 2010). As AST plants are predominantly C_3 (generally woody shrubs and trees) and CAM species (generally succulents; Vlok *et al.* 2003), elevated CO_2 may give thicket a competitive advantage to expand both in overall range and in mosaics, especially in C_4 grassland found in the east and possibly into the fynbos and renosterveld in the west (Bond *et al.* 2003, Cowling *et al.* 1997, Kgope *et al.* 2010). The AST boundaries with fire-driven biomes such as savannas and fynbos are thought to be largely controlled by fire as AST is considered to be fire-intolerant (Cowling *et al.* 1997, Trollope 1974). However, bush encroachment has already been observed in fire-driven savannas (Bond 2008, Wigley *et al.* 2010), and similar mechanisms may result in the expansion of AST into these neighbouring biomes (Kgope *et al.* 2010).

Validation of the projected range changes is necessary but difficult (Dormann 2007). The spatially explicit projections of range change can be used to focus on areas that should be the first to show the effects of climate change. For example, the results suggest that mesic thicket will experience the greatest impact of climate change; therefore, evidence of the negative climatic effects, such as population declines and shifting species distribution, may be the greatest, earliest and most easily measurable in the mesic thicket. Evidence of expanding thicket subtypes may already exist in the form of expanding thicket clumps in mosaics from the late 20th century to the present. This could be tested using similar methods to Wigley *et al.* (2009). Furthermore, I recommend that the predictions of the models be tested by monitoring the growth and population changes of key components of the thicket subtypes, as has been done for the desert (Foden *et al.* 2007) and succulent karoo biomes (Musil *et al.* 2009). In this regard, relatively short-lived species of arborescent *Aloe* and *Euphorbia* would be the most appropriate subjects for monitoring.

2.5.4. Conservation implications

Community distribution modelling of floristic groups has not been extensively tested, and it remains largely unknown whether this over- or underestimates distributional changes of their component species. Midgley *et al.* (2002) suggest that biome-level modelling can serve as a preliminary attempt to estimate the potential distributional shifts of suites of species where species distributional data are poor, as is the case for

the AST.

The results provide insights into potential future range shifts and contractions of the AST subtypes. However, this does not provide an accurate assessment of the potential levels of biodiversity loss for AST species in the face of climate change as it lacks information regarding individual species, and range restricted endemics from the arid and valley thickets may still be at risk despite the positive results for these subtypes (Midgley *et al.* 2002). Using climate modelling of endemic species of the AST, Broennimann *et al.* (2006) predict a high turnover of 200 AST endemics (60-72%). However, they model their species distribution models at a resolution of quarter degree squares ($\sim 25 \times 25$ km). This averages a wide range of the environmental space, especially in such a topographically and climatically complex landscape as the coastal lowlands, which most likely creates a bias towards high turnover rates. A more realistic estimate of range change and species turnover should involve georeferenced locality data, much higher resolution climate datasets and also include different scenarios of dispersal (e.g. Loarie *et al.* 2008). This should be prioritized for future study as it will provide a far more nuanced and accurate perspective on the threat to AST biodiversity under future climate projections and provide a further test of the biome modelling approach.

Despite the uncertainties regarding the future of component species, the results provide an initial assessment of the effects of climate change on each of the AST subtypes, with a ranking of negative impacts on subtypes as arid thicket < valley thicket < mesic thicket. The overall results suggest that the ambitious megaconservancy network proposed by Rouget *et al.* (2006) will be fairly robust to the impacts of projected climate change as the amount of combined climate envelopes for the AST subtype increases slightly. However, the projected range changes will affect the proportion of each subtype conserved in the network, with an expansion of arid thicket, a small reduction of valley thicket, and quite a substantial reduction in mesic thicket. This suggests that the network may need to be expanded towards the east to accommodate the projected range reduction of mesic thicket, possibly focussing on the Bushmans and Buffalo catchments.

2.5.5. Conclusions

In this chapter I provide spatially explicit hypotheses of past and future distributions of the AST subtypes. I show that different subtypes likely experienced different distributional histories with an overall dramatic change in extent of AST in the region. Refugial areas were identified, and extreme bottlenecks were postdicted for arid and valley thickets. These factors are expected to have had a dramatic impact on the historical dynamics and population connectivity of flora and fauna associated with the Albany Subtropical Thicket. The results for future climate suggest that the subtypes will have contrasting responses to projected changes in future global climate, with arid thicket increasing in area, valley thicket remaining fairly constant, and mesic thicket declining.

3. Constructing phylogenies in the presence of intra-individual site polymorphisms (2ISPs) with a focus on the nuclear ribosomal cistron

3.1. Abstract

Nuclear DNA is widely used to establish phylogenetic relationships among organisms but remains underutilised in phylogeographic studies. The internal transcribed spacers of the nuclear ribosomal cistron are the most commonly used nuclear regions for phylogenetic inference. Numerous ITS variants may be present in an individual's genome; these result in **i**ntra-**i**ndividual **s**ite **p**olymorphisms (2ISP; pronounced 'twisp') in direct or clone consensus sequences. Dealing with 2ISPs in this, and other, nuclear regions has been problematic as phylogeny reconstruction algorithms generally do not take such variation into account. Here I demonstrate the improvements offered by an approach that treats 2ISPs as informative, rather than ambiguous, characters implemented in three widely used methods of phylogenetic tree-inference: Neighbour Joining, Maximum Parsimony and Maximum Likelihood. Simulation, real-world and case study datasets were used to compare the 2ISP-ambiguous versus the 2ISP-informative approaches across these phylogenetic methods. The simulation results show that both approaches remain vulnerable to the reticulate signal present in hybrid samples, but that the 2ISP-informative approach offers a dramatic improvement in terms of tree accuracy and branch support when 2ISP variation is inherited within species or population boundaries. Improved bootstrap support is also observed for the majority of real-world datasets using the 2ISP-informative approach, especially in 2ISP-rich datasets. Furthermore, appropriate placement of hybrids in a case study is observed in distance-based planar networks based on the 2ISP-informative polymorphism p-distance. Lastly, the two detailed case studies demonstrate the improved resolution and support of the 2ISP-informative approach. These results demonstrate that treating once problematic 2ISPs as informative characters can

dramatically improve resolution in phylogeny reconstruction. I envisage that this method should greatly aid phylogenetic inference using any nuclear DNA regions that contain 2ISPs, especially at the intra-generic or intra-specific level including phylogeographic studies.

3.2. Introduction

Nuclear DNA plays a critical role in establishing phylogenetic relationships among organisms. In plant systematics, biparentally inherited nuclear data offer a crucial contrast to the generally uniparentally inherited chloroplast or mitochondrial genomes. The internal transcribed spacers (ITS-1 and ITS-2) of the nuclear ribosomal 18S-5.8S-25S cistron (35S rDNA) are the most widely used nuclear regions for phylogenetic inference in plants (Álvarez & Wendel 2003, Baldwin *et al.* 1995, Feliner & Rosselló 2007). Furthermore, ITS offers a valuable source of information for plant phylogeography (Chiang & Schaal 1999, Feliner *et al.* 2004, Rosselló *et al.* 2007) due to its often higher rate of mutation compared with, for example, the chloroplast genome (Schaal *et al.* 1998). However, it has often proved difficult to extract a clear signal from the ITS region (ITS-1, 5.8S, and ITS-2) for phylogeography and lower level phylogenies (Feliner & Rosselló 2007, King & Roalson 2008). A number of processes, such as incomplete concerted evolution, recombination (crossing-over), interbreeding/hybridisation, and autopolyploidisation can etch conflicting phylogenetic signals onto the 35S rDNA (Bailey *et al.* 2003). The result of these processes is the presence of intra-individual ITS polymorphism, which hampers phylogenetic reconstructions as the majority of the currently available algorithms and software treat potentially informative polymorphisms as ambiguous characters or missing data.

The 35S rDNA cistrons, including the genes encoding for the 18S, 5.8S and 25S rRNA, form a multi-gene family arranged in tandem arrays. The arrays are confined to one or several chromosomal loci, called nucleolus organiser region(s) (NOR; reviewed in Volkov *et al.* 1999), each comprising hundreds to thousands of copies (Hemleben *et al.* 1988, Rogers & Bendich 1987). All 35S rDNA copies within an individual were for a long time thought to be virtually identical due to the process of concerted evolution (reviewed in Nei & Rooney 2005). Concerted evolution permanently homogenises copies of identical genes to the same sequence via a combination of unequal crossing over and gene conversion (Elder & Turner 1995, Zimmer *et al.* 1980). However,

efficiency of concerted evolution in the case of 35S rDNA is decreased by the high number of copies (Dubcovsky & Dvorak 1995, Volkov *et al.* 2007) and the number of NORs (Copenhaver & Pikaard 1996, Schlötterer & Tautz 1994). Thus, individuals may contain different ITS variants. Intra-individual polymorphism may arise from a number of mechanisms (Arnheim *et al.* 1980, Buckler *et al.* 1997, Pikaard 2001, Volkov *et al.* 1999): (1) incomplete concerted evolution where rates of mutation and recombination outstrip the homogenizing effects of unequal crossing over and gene conversion within an array and between the arrays of orthologous NORs, (2) duplication and translocation of the NOR, leading to non-orthologous (paralogous) NORs, (3) autopolyploidy, (4) hybridisation, often connected with allopolyploidy in plants (homoeology in a strict sense), (5) deficiencies in the gene repair mechanism (retention of pseudogene copies within an array), and (6) intragenomic competition between orthologous and/or paralogous NORs, and subsequent (gradual) loss of functionality and elimination of NORs.

Independent of the source of intra-individual polymorphism, the result will be overlapping and incompatible signal that interferes with phylogenetic reconstructions. There are two possible approaches to incorporate intra-individual ITS variation into phylogenetic inferences (Baldwin *et al.* 1995): (1) cloning with multiple ITS clones per individual, or (2) directly sequenced PCR products with polymorphisms coded. Using clones for phylogeny reconstruction has a number of problems including (i) the cost and labour involved, (ii) that cloning is not error-free, (iii) obtaining a fully representative population of clones for an individual is virtually impossible, and (iv) clone phylogenies are often difficult to interpret, for example copies associated with specific individuals may be spread throughout a phylogenetic tree (e.g. King & Roalson 2008, Rosselló *et al.* 2007). With regards to the last problem, analytical approaches exist that attempt to use the clone population to reconstruct an organismal phylogeny (phylogeny of individuals; Göker & Grimm 2008, Joly & Bruneau 2006), similar to using species composition across plant communities to identify the relationships between communities (Legendre & Legendre 1998). However, this is also problematic in terms of either assuming diploidy (Joly & Bruneau 2006), which may be hard to establish, or may require extensive sampling of the clone population within each individual.

A less cost- and labour-intensive alternative is to use direct sequencing. Sites with multiple peaks at a base within an individual have been widely observed (Álvarez &

Wendel 2003, Baldwin *et al.* 1995). These have been termed super-imposed nucleotide additive patterns (SNAP, Whittall *et al.* 2000) or additive polymorphic sites (APS, Aguilar & Feliner 2003). Unfortunately, neither of these terms encompasses the entire range of polymorphism that can occur at a site. The definition of SNAP is restricted to parsimony-informative nucleotide polymorphisms and does not encompass the presence of indels; indels are detectable and may be informative (e.g. Baldwin *et al.* 1995, Rosselló *et al.* 2007, Whittall *et al.* 2000). The APS definition very specifically refers only to those polymorphisms where the two bases involved in a polymorphic site are also found separately in other accessions in the dataset, thereby ignoring polymorphisms where only one base is found elsewhere in the dataset. Thus, polymorphisms under both definitions are conditional on sampling, which may very well be incomplete. Here I define **i**ntra-**i**ndividual **s**ite **p**olymorphisms (2ISP, pronounced 'twisp'), which includes both SNAP and APS, but extends to any site that contains a polymorphism, including indels.

In general, 2ISPs reflect either intra- or inter-array heterogeneity within an individual. The heights of multiple peaks observed at a base position from direct sequencing have been observed to be proportional to the underlying frequency of the ITS variants (Rauscher *et al.* 2002). Furthermore, 2ISPs can be verified, either by detection in the forward and reverse sequences and/or cloning a subset of individuals for confirmation. Therefore 2ISPs, after verification, offer informative data that can aid phylogenetic inferences (Fama *et al.* 2000, Göker & Grimm 2008, Grimm *et al.* 2007, Joly & Bruneau 2006). However, this approach has been rarely used. Intra-individual site polymorphisms are coded using IUPAC nomenclature (International Union of Pure and Applied Chemistry, http://www.iupac.org); for example, a polymorphism involving cytosine ('C') and thymine ('T') will be represented as 'Y' (pyrimidine) in a clone consensus sequence or a directly sequenced chromatogram. Most implementations in widely used software (PAUP*, Swofford 2002)(MrBayes, Ronquist & Huelsenbeck 2003)(distance module in Mesquite, Maddison & Maddison 2010) treat IUPAC codes, however, as 'ambiguous' characters or missing data, for example 'Y' is treated as possibly 'C' or 'T', not what is meant by the coding of a 2ISP which is 'C' and 'T'. Depending on the optimality criterion and settings, IUPAC codes are either averaged (e.g. PAUP*) or ignored (e.g. MrBayes). In contrast, RAxML (Stamatakis 2006) treats all IUPAC codes as polymorphisms, since the probability of substituting an 'A' by 'Y' equals that of 'A' by 'C' and a 'T'; nonetheless, this can

still lead to a flattening of the likelihood surface making it more difficult to determine the best-known tree and reducing support values (Felsenstein 2004). The treatment of IUPAC codes may have profound implications for the phylogeny produced by current methods, especially when intra-individual variability exceeds intraspecific or generic variability. Nonetheless, a number of authors have suggested that ITS polymorphisms offer valuable phylogenetic and phylogeographic information, yet have had no clear-cut means to include the polymorphisms from directly sequenced PCR products into phylogenetic analyses (e.g. Fama *et al.* 2000, Feliner *et al.* 2004).

Here I propose a straightforward approach to include 2ISPs into phylogenetic inference by using a step-matrix which treats shifts between monomorphic and polymorphic sites as a single step (Figure 3.1). This may seem an overly simplistic interpretation of 2ISPs as the exact nature of different evolutionary mechanisms within ITS is largely unexplored (Elder & Turner 1995). For example, mutations may be linked (e.g. Grimm *et al.* 2007) and occur in tandem to maintain secondary structure (Coleman 2003). Also, the degree of homogeneity within individuals and populations, i.e. the fixation of the mutation in the gene pool, may be influenced by intra- and inter-array competition (in particular in polyploids; Chen & Pikaard 1997, Komarova *et al.* 2004, Volkov *et al.* 1999) and historical demography (de Sousa Queiroz *et al.* 2011). A methodological problem linked to parsimony using the Sankoff algorithm is that character changes observed among lineages must segregate from a polymorphic ancestor ('polymorphism parsimony'; see Felsenstein 2004, p. 76). With many conflicting processes, the extent of which are difficult to determine, I suggest that 2ISPs be treated as simple fixation events as a starting point (see Grimm *et al.* 2007). This treatment explicitly assumes that IUPAC codes are polymorphisms, not ambiguities. Thus, it provides a means to include the information present in polymorphisms into phylogeny reconstruction.

This study demonstrates that treating 2ISPs as informative characters provides additional information for inferring evolutionary relationships among individuals. In order to do so, I firstly use simulations to explore the different phylogenetic outcomes when the source of 2ISPs is either hybridisation (joining of two evolutionary lineages introducing new 2ISPs) or inheritance (2ISPs inherited from polymorphic parent/s). Secondly, I explore the phylogenetic outcomes using real-world datasets under the 2ISP-informative versus the 2ISP-ambiguous treatments. Lastly, I present two detailed case studies to highlight both the limitations and advantages of the 2ISP-informative

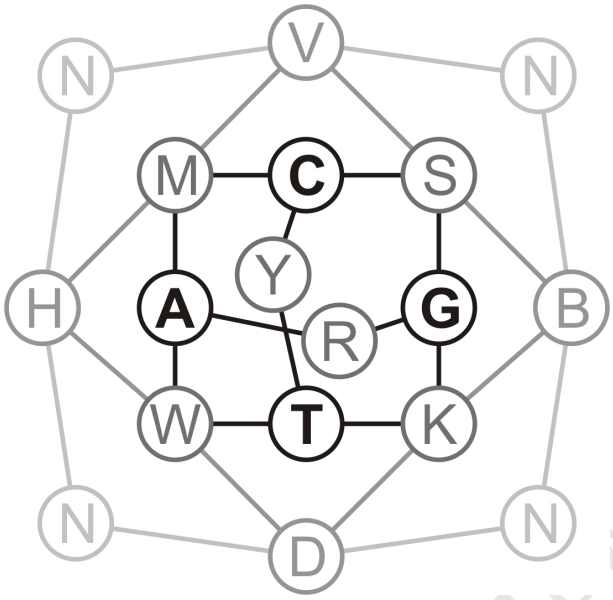

Figure 3.1. Pictogram of proposed step matrix representing mutations for a single site between DNA bases and/or base polymorphisms coded using IUPAC codes. Each line represents a single step. This step matrix is also used to calculate the polymorphism p-distance.

approach.

3.3. Methods

3.3.1. Treating 2ISPs as informative

To treat 2ISPs as informative we use a simple model. In the case of single-copy genes the shift from a C to a T, for example, involves a single mutation event, which, subsequently, is fixed in the genome of the individual and part of its offspring. In diploids, where most genes are present at two homologous sites (alleles), the shift from a C to a T in trace files representing the individuals of a population can occur in two ways: (1) if the mutation is propagated via asexual reproduction, two mutation events are needed: C–C to C–T to T–T, or (2) the fixation of the T variant in a population via sexual reproduction. In high copy nuclear gene regions, such as ITS, the matter is more complicated due to the additional requirement that a mutation and the corresponding ITS variant must spread via concerted evolution onto enough copies to be detected in the sampling of copy sequences via cloning or direct sequencing of the PCR product. Thus a shift from a C to a T in a direct PCR sequence of a high copy nuclear gene will require not only a mutation of C to T on an ITS copy, but a spread of this mutation as well as a subsequent loss of the C across ITS copies within individuals (intra- and inter-array homogenisation) plus a spread of the fixed T variant through the population (allelic homogenisation). Assuming that any mutation has equal probability of reaching detectable levels or fixation, I suggest that the mutation from one monomorphic state to another through a polymorphic state should be treated as a simple mutation event (i.e. C⇔Y⇔T). Thus, a step matrix can be defined representing the fixed mutations required to shift from one state to another (Figure 3.1). This step matrix can be considered an extension of the allele sharing distance (ASD, Bowcock 1994) which is widely used for single nucleotide polymorphism data; however, the ASD can only deal with a maximum of two bases and the polymorphism of these bases at any single site.

Such a step matrix can easily be incorporated into both distance and parsimony algorithms used to infer phylogenetic trees, as well as distance-based splits networks. Given a sufficient amount of data, this matrix could also be estimated for model-based methods such as maximum likelihood. This step matrix can be used to calculate a modified p-distance that takes into account polymorphisms (hereafter termed polymorphism p-distance).

3.3.2. Simulations

Two simulation studies were performed in order to determine the effect of 1) the presence of a single hybrid sample in a dataset, and 2) inherited polymorphism via independent heterogeneous alleles (or variants) on phylogenetic inference with intra-individual polymorphisms treated as either ambiguous or informative characters. Using the PHANGORN library version 1.3.1 (Schliep 2011) in R version 2.13.0 (R Development Core Team 2011), a random tree with 20 tips was generated. A DNA dataset with only monomorphic bases (i.e. A, G, C and T) was simulated onto this tree with mutation rates of 0.0025, 0.0050, 0.0100, 0.0250 and 0.0500 per unit of branch length; this represents a range of signal-poor to signal-rich datasets (Tables 3.1 and 3.2). The starting sequence used for each dataset simulation was based on the ITS-1 and ITS-2 regions (490 bases in total) of a sequence from *Nymania capensis* (Meliaceae) extracted from Genbank (DQ861633). The second lowest rate (0.0500) aimed to include at least one parsimony informative character in the dataset for each branching event, irrespective of whether the parsimony informative character was associated with a branching event or not; in the case of a dichotomous tree with 20 tips there will be 19 branching events. This was to ensure that inferring the relationships between samples, for this and faster rates, was not affected by a lack of signal.

Table 3.1. Summary statistics of hybrid-free (HF) datasets from hybridisation simulations across different mutation rates (see text for details). The mean and standard deviation (in brackets) of the number of variable sites (VS) and parsimony informative sites (PI) are shown.

Rate	VS	PI
0.0025	24.4 (4.8)	11.5 (3.0)
0.0050	46.8 (7.5)	23 (3.9)
0.0100	88.9 (11.9)	44.9 (7.3)
0.0250	186.2 (19.8)	103.8 (14.8)
0.0500	304.9 (22.1)	191.4 (21.9)

Table 3.2. Summary statistics of datasets from independent variant evolution simulations across different mutation rates, specifically the A_1, A_2 and $A_{1\&2}$ datasets (see text for details). The mean and standard deviation (in brackets) of the number of variable sites (VS) and parsimony informative sites (PI) are shown. The $A_{1\&2}$ dataset contains intra-individual site polymorphisms (2ISPs), thus the variable and parsimony informative sites are split into standard DNA characters (-STD) and 2ISP characters (-2ISP).

	A_1		A_2		$A_{1\&2}$			
Rate	VS	PI	VS	PI	VS-STD	VS-2ISP	PI-std	PI-2ISP
0.0025	23.8 (5.1)	11.2 (3.6)	22.8 (5.1)	11 (3.6)	0.2 (0.5)	45.4 (7.9)	0.1 (0.3)	21.9 (5.5)
0.0050	44.8 (7.8)	21.5 (5.7)	44.1 (7.8)	21.6 (5.6)	0.3 (0.5)	84.5 (11.4)	0.1 (0.3)	42.4 (8.7)
0.0100	86.7 (12.1)	43.2 (9.3)	84.4 (13)	43.3 (9.2)	1.3 (1.4)	155.8 (17.6)	0.5 (0.7)	83.9 (14.2)
0.0250	187.2 (17.1)	102.7 (16.7)	185.4 (19.6)	101.9 (16)	7.5 (3.5)	301.2 (20.8)	2.7 (1.9)	187.1 (23.5)
0.0500	302.1 (20.5)	187.9 (23.8)	299.6 (22.5)	186.3 (24.2)	26.9 (9.1)	414.9 (16.6)	10.6 (4.8)	308.3 (27.4)

Phylogenetic trees for all datasets were inferred using neighbour joining (NJ), maximum parsimony (MP) and maximum likelihood (ML). Bayesian inference was not included as there is currently no software that can incorporate a 2ISP-informative approach. Furthermore, current Bayesian inference software explicitly deals with 2ISPs by removing any columns in a dataset with such polymorphisms (e.g. MrBayes and Beast). Neighbour joining was implemented using the ape library version 2.7.1 (Paradis *et al.* 2004) in R. The ape library treats 2ISPs as missing characters, an option that is widely used in other software (e.g. PAUP*, MrBayes). Genetic distances were calculated using either uncorrected p-distances (the ambiguous treatment, hereafter referred to as NJ-A) or polymorphism p-distances (the informative treatment, hereafter referred to as NJ-I) with pairwise deletion of missing characters. Neighbour joining bootstrapping used 100 replicates. Maximum parsimony was implemented in PAUP* version 4b10. Heuristic searches for most parsimonious trees included 100 replicates, each with random addition of sequences, and the default branch swapping and character optimisation options (TBR and ACCTRAN, respectively). Intra-individual polymorphic sites are treated as either ambiguities (hereafter referred to as MP-A), the default treatment, or informative sites (hereafter referred to as MP-I), set using a cost matrix. For computational efficiency, no more than 100 equally parsimonious trees were stored per replicate (NCHUCK = 100); this prohibited time-consuming searches of equally parsimonious trees which was a problem predominantly with the MP-A treatment when the dataset contained 2ISPs. Branch support was evaluated with 100 bootstrap replicates, each with 10 replicates of random sequence addition and the same options as used above for topology reconstruction. RAxML 7.2.6 was used to compute trees and perform bootstrap analyses under maximum likelihood (Stamatakis 2006). A standard RAxML analysis includes polymorphic base calls into the analysis; however, this can still lead to a flattening of the likelihood surface making it more difficult to determine the best-known tree and reducing support values. Thus, the standard analysis was treated as the analogue to the ambiguous treatment under NJ and MP (ML-A). RAxML includes a multi-state analysis for any kind of categorical data; this would treat each IUPAC code as a unique character, thereby estimating the rate of transitions between characters (i.e. estimating the rate of transitions between states instead of the single steps shown in Figure 3.1). This was considered the informative treatment of intra-individual site polymorphisms (ML-I). The GTR-Γ model was used for all datasets and 100 rapid bootstraps were used to assess branch support.

The tree topologies between datasets were compared using the topological distance defined by Penny & Hendy (1985) as implemented in the APE library in R. This topological distance is measured as twice the number of internal branches defining different partitions of the tips between two trees. Under this definition, the MP consensus tree will be biased towards lower topological distances than the other methods as polytomies are not penalised as much as incongruent relationships between samples that may be observed in the ML and NJ trees (with ties broken randomly).

Simulation 1: presence of hybrids (heterogeneity due to hybridisation)

For the hybrid analysis, a simulated dataset was considered the hybrid-free (HF) dataset, and a second dataset (hybrid-present, HP) was created by adding a single hybrid sample to the HF dataset. The hybrid sample was created by combining two sequences selected at random from the dataset, with any variable sites coded using IUPAC nomenclature. Twenty HF datasets were simulated per mutation rate, each on a different random tree, and 5 HP datasets were created for each HF dataset, totalling 100 HP datasets per mutation rate.

The topological distances between NJ, MP or ML trees estimated from the HF and HP datasets were calculated where 2ISPs were treated as ambiguous or informative. Each hybrid sample was removed from the tree derived from a HP dataset prior to calculating the topological distance. The topological distance between trees inferred from the HF and HP datasets was compared to the phylogenetic distance between the two 'parent' samples that were used to form the hybrid sample. Phylogenetic distance was measured using Faith's phylogenetic diversity (PD) (Faith 1994) as implemented in the PICANTE library version 1.2.0 (Kembel *et al.* 2010) in R. In addition, the topological distance between the original tree used to simulate the data and the NJ, MP and ML trees derived from the HF and HP datasets was calculated. In order to assess the effect of hybrid presence on branch support, the percentage of branches with bootstrap values above a low, medium or high threshold (>50%, >70%, and >90%, respectively) was calculated. The bootstrap values were extracted using the APE library in R from NJ, MP and ML trees inferred from the HF and HP datasets with 2ISPs treated as ambiguous or informative.

Simulation 2: independent variants (heterogeneity due to independent evolution)

To simulate polymorphism induced by two sets of independently evolving ITS sequences (ITS homoeologues in a broad sense, Cronn *et al.* 2002), two independent variant datasets (A_1 and A_2) were simulated on the same tree with the same ancestral sequence but a different random starting seed (this ensures that each dataset is unique). The two DNA datasets can be considered independent histories of two dominant variants present in the common ancestor, either representing intra-array (ITS paralogy in a broad sense), allelic (between homologous NORs), homoeologous s.str. (between orthologous NORs) or paralogous variation (in a strict sense, i.e. between NORs originated by duplication and translocation). The tip sequences were combined across the two datasets, creating a polymorphic combined variants dataset ($A_{1\&2}$) representing the mutation and inheritance of polymorphisms along branches and branching events. This is a highly simplified model of multicopy inheritance as it does not include the effects of concerted or reticulate evolution. For each mutation rate, 200 datasets were simulated (two independent datasets per tree), which combined made 100 additional polymorphic datasets.

The topological distances between NJ, MP or ML trees inferred from the $A_{1\&2}$ datasets and the two variant (A_1 and A_2) datasets were calculated where 2ISPs were treated as ambiguous or informative. Topological distance was also compared between the original tree used to simulate the data and the trees derived using the different methods from the A_1, A_2 or $A_{1\&2}$ datasets. The percentage of branches per tree with bootstrap support above a low, medium or high threshold were calculated across the NJ, MP or ML trees derived from the A_1, A_2 or $A_{1\&2}$ datasets.

3.3.3. Published datasets

The 2ISP-informative and 2ISP-ambiguous treatments were investigated using 13 previously published dataset alignments and a novel alignment generated for this study (see section 3.3.4 below; Table 3.3). As 2ISPs occur on other nuclear gene regions, the selection of datasets was not limited to ITS. The aligned datasets were either obtained directly from the authors or downloaded from TreeBase (www.treebase.org). If indel coding was included in the dataset, then these were kept for subsequent analyses. Two

datasets were used as case study examples and were further explored to demonstrate the advantages and limitations of the 2ISP-informative approach (see section 3.3.4). Two datasets were included twice, with adaptations. A second reduced *Hieracium* dataset with hybrids removed was included (Fehrer *et al.* 2009; see section 3.3.4). The Amaryllidaceae dataset (Meerow *et al.* 2006) contained two anomalous consensus sequences of clones; these two samples contained unusually high number of 2ISPs (74 and 29) and were very unlike any of the other samples in the dataset. This may be due to the inclusion of pseudogenes or issues with making the consensus sequence. Meerow *et al.* (2006, p. 43) state 'phylogenetic signal was still present despite the ambiguous base calls'. Uncertain as to whether these ambiguous base calls actually represent intra-individual polymorphism or just uncertainty, I also included a reduced Amaryllidaceae dataset that excluded these two samples.

Table 3.3. Summary statistics for published datasets.

Region	Taxa	N	Align. chars	PI STD	PI ALL	PI 2ISP	PU 2ISP	*P* index	2ISPs					Source
									N_{2ISP}	Mean	Med.	Min.	Max.	
ITS	*Acer*	27	392	36	55	30	96	0.09	25	5.15	3	0	33	(Göker & Grimm 2008)
	Amaryllidaceae (A)[1]	29	659	317	319	5	115	0.97	6	4.17	0	0	74	(Meerow *et al.* 2006)
	Amaryllidaceae (B)[1]	27	659	296	296	0	18	1	4	0.67	0	0	15	
	Erythronium	28	627	121	121	3	21	0.95	9	0.86	0	0	5	(Allen *et al.* 2003)
	Fagus	20	773	4	74	74	145	-0.9	19	18.2	17	0	38	(Göker & Grimm 2008)
	Hieracium (A)[2]	62	521	44	106	90	179	-0.34	60	11.03	7	0	33	(Fehrer *et al.* 2009)
	Hieracium (B)[2]	45	521	44	90	63	150	-0.18	43	6.51	6	0	25	
	Krigia	28	627	121	121	3	21	0.95	9	0.86	0	0	5	(Kim & Jansen 1994)
	Nymania	30	638	12	33	33	31	-0.47	30	5.73	6	1	11	This study
	Platanus	28	627	121	121	3	21	0.95	9	0.86	0	0	5	(Feng *et al.* 2005)
	Sambucus	22	634	146	148	5	47	0.93	16	2.77	2	0	10	(Eriksson & Donoghue 1997)
	Tequila	21	767	98	114	28	76	0.56	20	5.43	5	0	18	(Moore *et al.* 2006)
	Zelkova	9	703	7	16	14	73	-0.33	9	10.67	8	5	17	(Göker & Grimm 2008)
PHYC	*Puya*	59	1048	98	101	16	76	0.72	26	1.61	0	0	16	(Jabaily & Sytsma 2010)
WAXY	*Tequila*	41	676	160	201	86	138	0.3	32	13.17	2	0	53	(Moore *et al.* 2006)
5S	Machaerantherinae	34	157	54	70	21	86	0.44	28	3.62	3	0	14	(Morgan *et al.* 2009)

N, the number of samples; Align. chars., the numbers of characters in the alignment; PI, the number of parsimony informative sites for standard characters (STD), all characters (ALL) and intra-individual polymorphisms (2ISP); PU-2ISP, the number of parsimony uninformative 2ISPs; *P* Index, the parsimony informative sites index (the *P* index ranges from -1 to 1, where all parsimony informative sites are only 2ISPs or only standard DNA characters, respectively); the number of samples that contain 2ISPs (N_{2ISP} Present), and the mean, median, minimum and maximum number of 2ISPs per sample.
[1] Two instances of the Amaryllidaceae dataset were analysed: (A) the full dataset, and (B) a reduced dataset where two anomalous consensus sequences with an unusually high number of 2ISPs (74 and 29) were removed.
[2] Two instances of the *Hieracium* dataset were analysed: (A) the full dataset, and (B) with putative hybrids removed (see text for details and Figure 3.9 for samples removed).

All datasets were analysed using NJ, MP and ML. NJ-A and NJ-I analyses were conducted using the APE library in R with 1000 bootstrap replicates to asses branch support. MP-A and MP-I analyses were undertaken in PAUP* with heuristic searches performed with 100 random sequence addition replicates. Tree bisection-reconnection branch swapping with the ambiguous or informative treatment was used for each replicate. Also, to ensure feasible computational times no more than 1000 trees of length greater or equal to one were retained per replicate. Assessment of MP bootstrap support followed the suggestions of Müller (2005), with 10,000 bootstrap replicates composed of a single random sequence replicate and TBR branch swapping. RAxML 7.2.6 was used to compute trees and perform bootstrap analyses under ML-A and ML-I treatments. The GTR-Γ model was used for all datasets and 1000 rapid bootstraps were used to assess branch support. Summary information for each dataset and the phylogenies were calculated using the APE library in R; however the number of parsimony informative characters for standard bases and 2ISPs was calculated using a custom script. In order to compare the NJ, MP and ML support values for the ambiguous versus informative approach against information content within standard DNA and 2ISPs across the datasets we did the following: (1) the percentage of branches supported above low, medium and high bootstrap support thresholds (>50%, >70% and >90%) for each tree under the informative approach was deducted from the ambiguous approach, and (2) the information content within each dataset was characterised using a parsimony informative sites index (P index), which was calculated as follows:

$P = \frac{PI_{STD} - PI_{2ISP}}{PI_{STD} + PI_{2ISP}}$,

where PI_{STD} and PI_{2ISP} are the number of parsimony informative sites for standard DNA and 2ISP characters, respectively. The P index ranges from -1 to 1, where all parsimony informative sites are only 2ISPs or only standard DNA characters, respectively.

3.3.4. Case studies: *Hieracium* and *Nymania*

I used direct sequence data from datasets that contained both intra- and inter-individual ITS variability and where the limits for polymorphic site identification in the trace file were clearly outlined. Both datasets are based on angiosperm genera, *Hieracium* (Asteraceae; Fehrer *et al.* 2009) and *Nymania capensis* (Meliaceae; this

study). The *Hieracium* dataset contained 60 sequences from the 5′ external transcribed spacer (ETS) of the 35S rDNA. In Astereaceae (Linder *et al.* 2000), and *Hieracium*, the ETS region, found in the same cistron as ITS, generally evolves much faster but lacks a conserved region such as the 5.8S region found between ITS-1 and ITS-2. This dataset was selected because it contains large number of putative hybrid samples, identified by 2ISPs, between two fairly divergent clades. Thus, this dataset was used to explore the effects of hybrid samples on the different phylogeny reconstruction methods. A reduced dataset was created by removing those putative hybrid samples based on their 2ISP patterns (i.e. a large proportion of 2ISPs present where sequences from the two clades contain segregating sites) and their central location in the splits graph. This reduced dataset is meant to represent a hybrid-free scenario (at least between the two major clades).

The *Nymania capensis* dataset comprised 30 individuals sampled across three primary drainage basins in the Albany Subtropical Thicket biome which spans the Western and Eastern Cape Provinces of South Africa. Ten individuals were sampled per drainage basin. Two additional individuals from the disjunct northern distribution of the species were sampled from herbarium material as outgroup taxa (BOL48535 and BOL60966). Full collection details for these individuals, and those used in Chapter 4, are given in Appendix Table A.1 (Pg. 238).

Genomic DNA was extracted from silica-dried leaf material using a modified version of the method specified by Gawel & Jarret (1991). Polyvinylpyrrolidone-40 (PVP) was added when grinding the leaf material in liquid nitrogen using a mortar and pestle. Nuclear variation was sampled for the ITS-1, 5.8S and ITS-2 region using the primers ITS5m (Sang *et al.* 1995) and ITS4 (White *et al.* 1990). PCR reactions were performed in 25 μl, with 5 μl 1 $\times$ KAPA HiFi Buffer, 0.75 mM dNTPs, 0.75 mM forward primer, 0.75 mM reverse primer, 0.4 μl of the proofreading KAPA HiFi DNA polymerase (2 Units) and 1.2 μl template DNA ($\sim$ 15 ng). PCR was conducted using a GeneAmp 2700 PCR System thermocycler (Applied Biosystems, USA) under the following conditions: initial denaturation and polymerase activation at 98°C for 20 seconds (s) followed by 30 cycles of 94°C for 45 s, 58°C for 30 s, 72°C for 30 s; and a final extension at 72°C for 1 minute. Eight samples were cloned to verify the presence of 2ISPs observed in direct sequences. Cloning was performed using the pGEM-T Easy Vector System II (Promega) following the manufacturers instructions, but downscaled to half reactions. To facilitate cloning, Kapa HiFi

PCR products were incubated at 72°C for 10 minutes with Kapa Taq polymerase to provide 5′ terminal thymidine overhangs. Eight clones were sequenced per sample. All sequences were aligned using Codon Code Aligner version 3.5.7 (Codon Code Corp, http://www.codoncode.com). The following steps were followed in order to identify polymorphic sites across and within sequences: (1) each base-call within every sequence was assigned a quality score using the automated base-calling program phred (Ewing *et al.* 1998), (2) sites that contained secondary peaks that were greater than 20% of the primary peaks were scored as polymorphic using the 'Call second peaks' option in Codon Code Aligner, and (3) all polymorphic sites were verified by eye. Following Fehrer *et al.* (2009), overlapping and non-overlapping peaks were coded in capitals letters and small letters, respectively. In order to determine if pseudogenes were present in ITS I confirmed the presence or lack of mutations in four conserved angiosperm motifs within the dataset, one in ITS-1 (Liu & Schardl 1994) and three within 5.8S (Harpke & Peterson 2008).

In order to detect potentially incompatible or ambiguous signals in the datasets, such as those that may be caused by hybridisation or allopolyploidisation events, I generated and visually inspected neighbour-net (NN) splits graphs (Bryant & Moulton 2004) using SplitsTree version 4.8 (Huson & Bryant 2006). As SplitsTree does not treat polymorphisms as informative, the polymorphism p-distance was calculated with a script written for, and implemented in, R and used the resulting distance matrix to produce the NN splits graphs. The NJ, MP and ML analyses were conducted using the same settings for PAUP and RAxML given above for the published dataset analyses.

To determine whether direct PCR sequence successfully detected variable sites observed in the clone sequences, the direct sequence 2ISPs were compared with the variable sites found in the clones for each individual. These results were compared with those reported in Rosselló *et al.* (2007) and Yamaji *et al.* (2007)

3.4. Results

Simulation results remained relatively unchanged across the datasets simulated under different mutation rates per unit of branch length, thus I report the results for those simulated under the mutation rate of 0.0050 per unit of branch length, but refer to appendix figures that contain the results from all datasets.

3.4.1. Simulation 1: presence of hybrids

Hybrid 2ISPs either represent crosses (diploid hybrids) or allopolyploids. The number of topological changes between the HP and HF trees increases proportionally with the parental PD of the hybrids under NJ-A, NJ-I or MP-A (Figures 3.2.A and 3.2.B), irrespective of the mutation rate (Appendix Figure A.9, Pg. 232). In contrast, across the majority of comparisons there were no topology changes between MP-I HF and HP trees (Figure 3.2.C). The topological distances between ML-A and ML-I do not show any clear relationship with the parental PD of the hybrid (Figure 3.2.C); however, ML-A has overall a lower topological distance between HF and HP trees than ML-I. When HF and HP trees generated from the different methods are compared using topological distance with the original tree used to simulated the data (Figure 3.3; Appendix Figure A.10, Pg. 233), I found that: i) the ambiguous and informative treatments produced identical results under the HF dataset, ii) the topological distance between NJ-A and NJ-I were equally increased, iii) MP-I performed better than MP-A, and iv) ML-A performed better than ML-I.

There is a significant decline in the percentage of branches with support above the low, medium and high bootstrap support thresholds in NJ, MP and ML trees inferred from the HF dataset to the HP dataset (Figure 3.4), irrespective of treatment or mutation rate (Appendix Figure A.11, Pg. 234). However, the NJ-I and MP-I show significantly smaller declines in the percentage of supported branches than for NJ-A and MP-A.

3.4.2. Simulation 2: independent variants

Intra-individual site polymorphisms from two different variants represent inherited polymorphism and heterogeneity due to the independent evolution of different alleles or paralogues (within or across NORs) of the same gene (region). The topological distance of NJ, MP and ML trees between $A_{1\&2}$ and A_1/A_2 datasets depends greatly on whether 2ISPs are treated as informative or ambiguous (Figure 3.5). Far less topological change is observed under the 2ISP-informative treatments than the 2ISP-ambiguous treatment across all methods irrespective of mutation rate (Appendix Figure A.12, Pg. 235). As the datasets increase in richness, the trees inferred using the 2ISP-informative approach from the $A_{1\&2}$ datasets show an increase in topological congruence with the

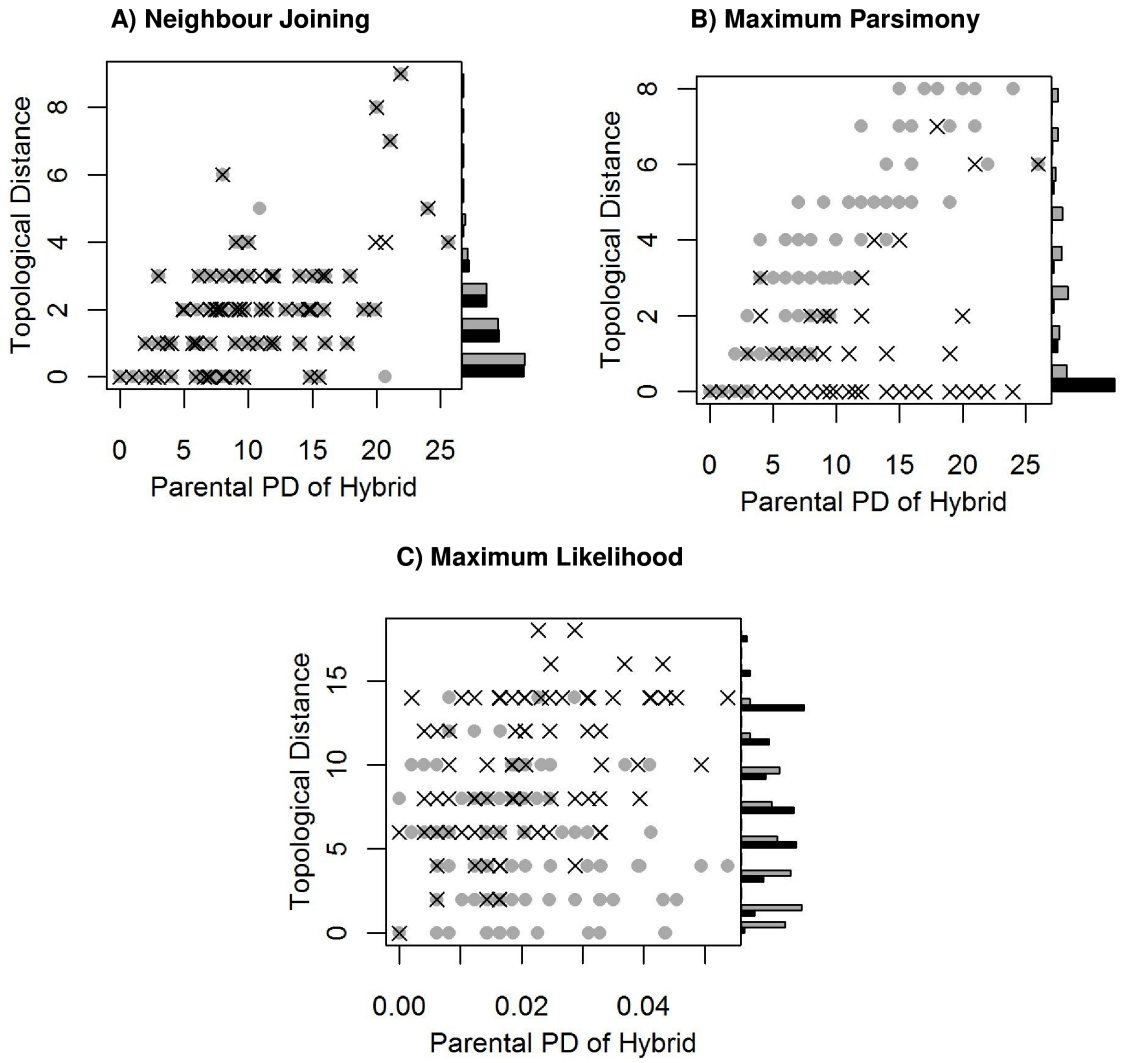

Figure 3.2. Parental phylogenetic diversity (PD) of a single hybrid sample and the topological distance between trees inferred from hybrid-free (HF) or hybrid-present (HP) datasets using different phylogenetic methods with intra-individual site polymorphisms treated as either ambiguous (grey circles) or informative (black crosses). Three different methods are used: (A) Neighbour Joining, (B) Maximum Parsimony, and (C) Maximum Likelihood. Data were simulated with a mutation rate of 0.0050 (see text for details, Pg. 62). The frequencies of the topological differences for the informative (black bars) and ambiguous (grey bars) treatments are given on the right-hand side y-axis. Note that topological distances are not directly comparable between the three methods (see text for details).

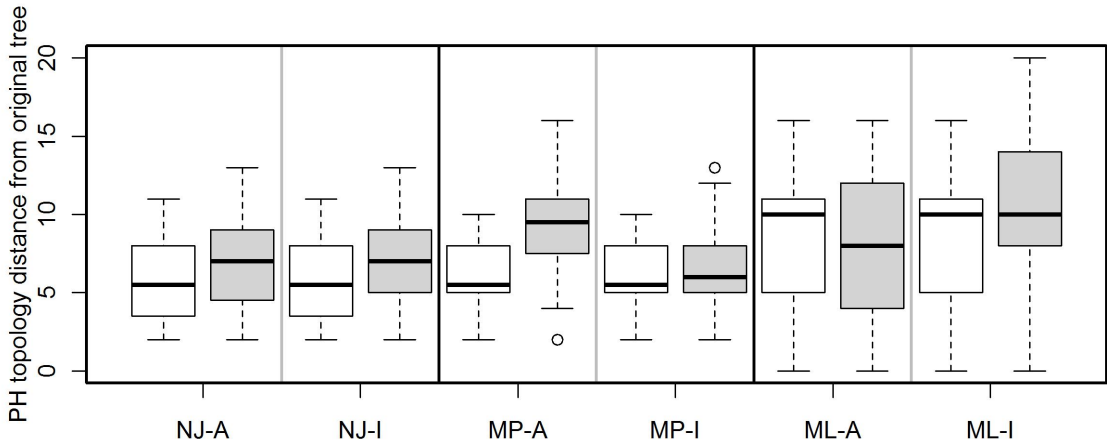

Figure 3.3. The topological distance between the original tree used to simulate the data and the trees inferred from the hybrid-free (HF; white) or hybrid-present (HP; light grey) datasets using different phylogenetic methods with intra-individual site polymorphisms treated as ambiguous (-A) or informative (-I). Three different methods are used: Neighbour Joining (NJ), Maximum Parsimony (MP), and Maximum Likelihood (ML). Data were simulated with a mutation rate of 0.0050 (see text for details, Pg. 62). Note that the topological distance of polytomies in the MP consensus trees are not penalised as much as the incorrect relationships in the NJ and ML trees.

trees inferred from the A_1 and A_2 datasets. This is to be expected given that lower mutation rates have greater signal stochasticity when sequences are simulated onto a tree. These results are maintained when the trees generated from the three datasets are compared using topological distances with the original tree used to simulate the data (Figure 3.6; Appendix Figure A.13, Pg. 236), as i) the NJ-A and MP-A trees have significantly higher topological distances when estimated on the $A_{1\&2}$ dataset in comparison to the A_1 or A_2 datasets, ii) the ML-A trees have a far greater variance when estimated on the $A_{1\&2}$ dataset than the A_1 or A_2 datasets, and iii) all 2ISP-informative approaches display a significantly lower topological distance to the original tree when estimated using the $A_{1\&2}$ dataset than either the A_1 or A_2 datasets.

The comparison between the percentage of branches above the low, medium and high bootstrap support thresholds and trees inferred from the A_1, A_2 or $A_{1\&2}$ datasets

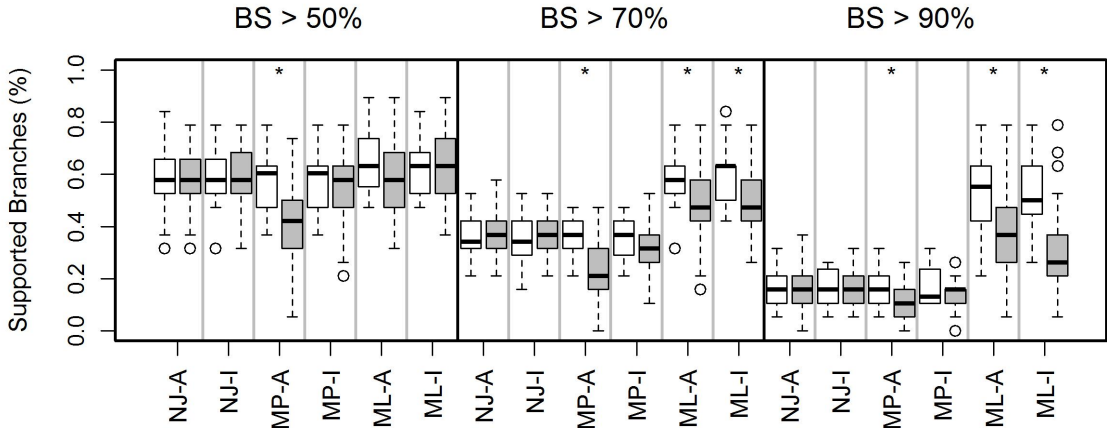

Figure 3.4. The percentage of branches with bootstrap values greater than 50%, 70% and 90% for hybrid-free (HF; white) or hybrid-present (HP; grey) datasets analysed using Neighbour-Joining (NJ), Maximum Parsimony (MP) and Maximum Likelihood (ML) with intra-individual site polymorphisms treated as ambiguous (-A) or informative (-I). The HF and HP datasets were simulated using a mutation rate of 0.0050 (see text for details, Pg. 62). Significant differences, ascertained using a Student's t-test, between HF and HP datasets for a given method are indicated with stars.

reveals marked and significant differences between the 2ISP-ambiguous and 2ISP-informative approaches across the different methods (Figure 3.7; Appendix Figure A.14, Pg. 237). Under the ambiguous approach, all methods have significant *declines* in the percentage of supported branches from the A_1 and A_2 to the $A_{1\&2}$ datasets. In contrast, all methods have significant *increases* in the percentage of supported branches under the informative approach.

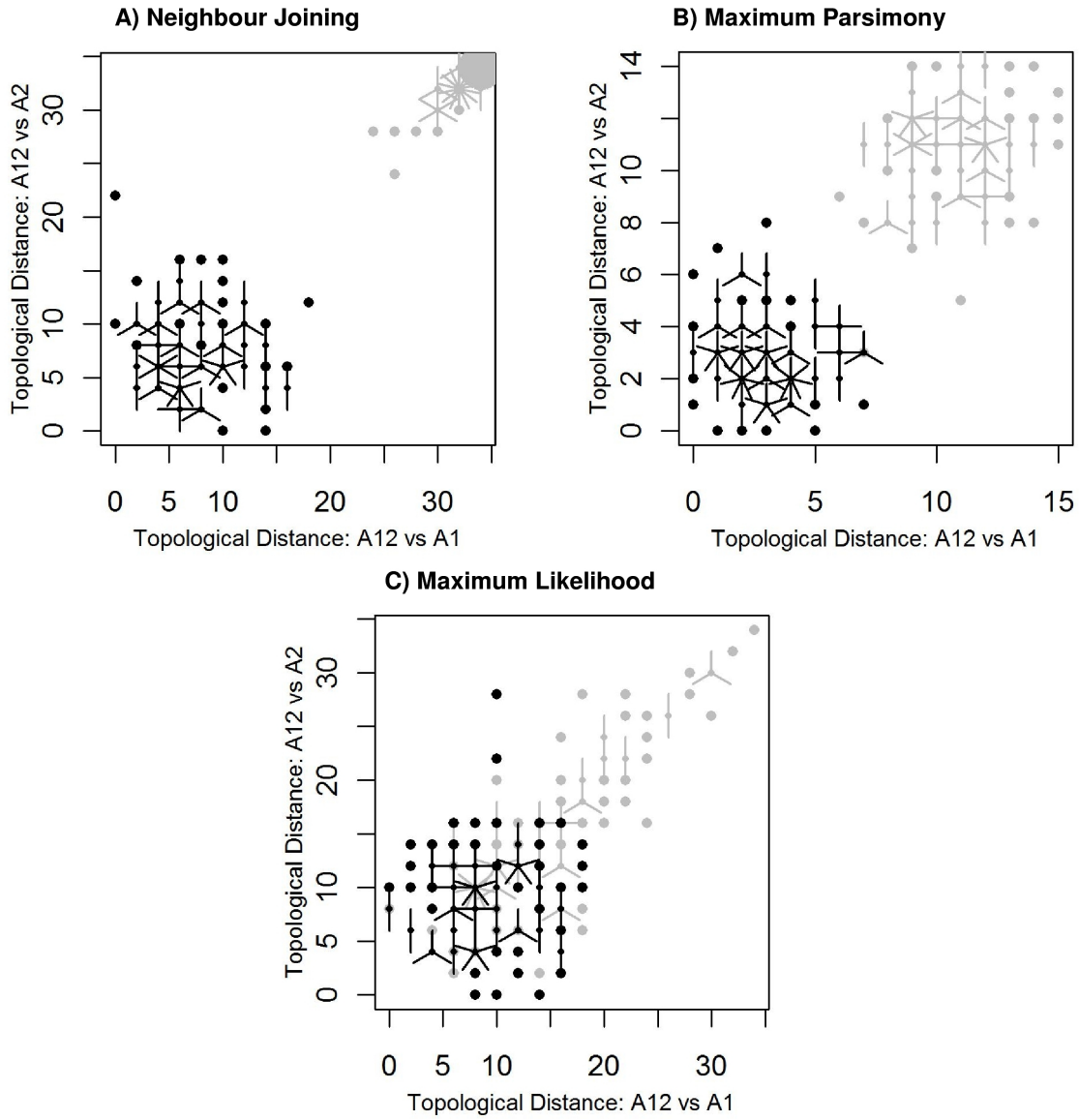

Figure 3.5. Topological distance between trees inferred from combined variants ($A_{1\&2}$) and individual variant datasets (A_1 and A_2) using (A)Neighbour Joining, (B) Maximum Parsimony, or (C) Maximum Likelihood with intra-individual site polymorphisms treated as ambiguous ('grey') or informative ('black'). The number of samples that overlap on a given point correspond to the number of petals shown. The mutation rate per branch length used to simulate the data was 0.0050 (see text for details, Pg. 62).

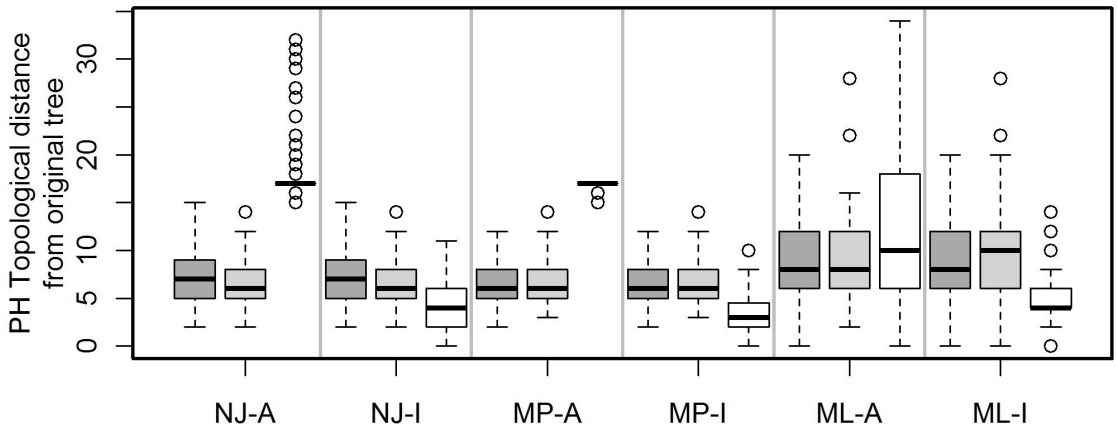

Figure 3.6. The topological distance between the original tree used to simulate the data and the trees inferred from the combined variants ($A_{1\&2}$, white) and individual variant (A_1, dark grey; A_2, light grey) datasets using Neighbour Joining (NJ), Maximum Parsimony (MP) or Maximum Likelihood (ML) with intra-individual site polymorphisms treated as ambiguous (-A) or informative (-I). The mutation rate per branch length used to simulate the data was 0.0050 (see text for details, Pg. 62).

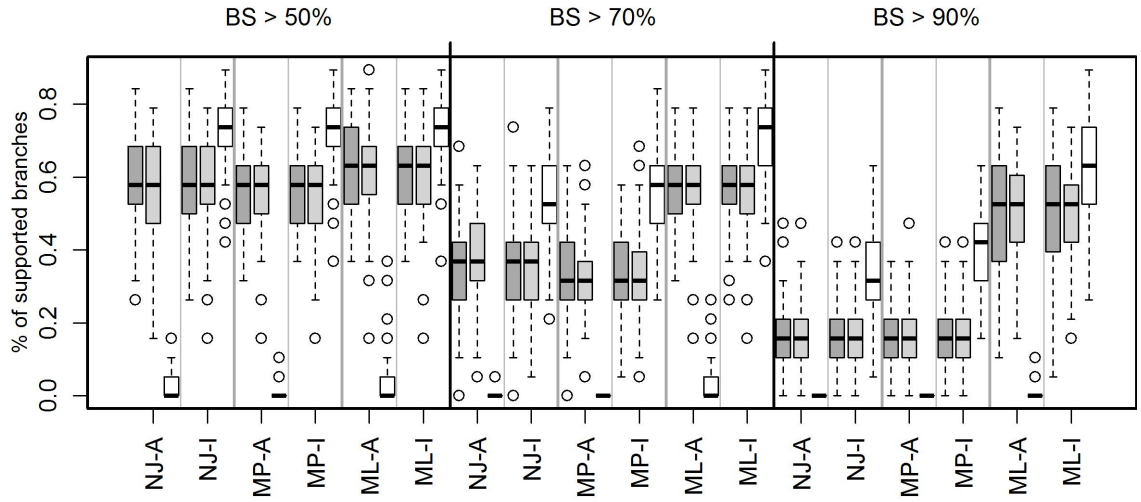

Figure 3.7. The percentage of nodes with low, medium and high bootstrap support values (>50%, >70% and >90%, respectively) from trees inferred from two independent variant datasets (A_1, dark grey; A_2, light grey boxes) and the combined variants dataset ($A_{1\&2}$, white) analysed using Neighbour Joining (NJ), Maximum Parsimony (MP) or Maximum Likelihood (ML) with intra-individual site polymorphisms treated as ambiguous (-A) or informative (-I). The datasets were simulated under a rate of 0.0050 mutations per unit of branch length (see text for details, Pg. 62).

3.4.3. Published datasets

The published datasets represent real-world data where multiple, sometimes conflicting, processes lead to the presence of 2ISPs. The majority of studies had P index values greater than 0, indicating that standard DNA parsimony-informative sites outweigh 2ISPs (Table 3.3); this is to be expected given that 2ISP-rich datasets have been traditionally difficult to analyse and therefore publish.

Across the majority of datasets the percentage of supported branches increased under the 2ISP-informative treatment compared with the 2ISP-ambiguous treatment (Figure 3.8). A minority of datasets (one to four) showed a decline in the percentage of supported nodes at the different bootstrap thresholds; these declines were also fairly minor (less than 10%) in comparison to the gain in support for some datasets (15 - 50%). Particularly, all data sets with a P index < 0 gained support. A negative correlation between the percentage of supported branches and the P index was observed across all combinations of methods and support thresholds; furthermore a high (42 - 75%) and significant correlation between these two variables was observed between some comparisons (e.g. NJ and MP for BS $> 50\%$ and BS $> 70\%$).

The presence of two anomalous sequences with high numbers of 2ISPs in the Amaryllidaceae dataset does not dramatically change the percentage of supported branches between the informative or ambiguous treatments under the NJ, MP or ML methods (see open diamond symbols in Figure 3.8). This is most likely because the majority of 2ISPs in these sites are not shared with other samples and therefore are treated as autapomorphies, and thus do not provide any misleading phylogenetic signals. The percentage of branches supported in the *Hieracium* datasets is far greater under the informative approach, irrespective of whether putative hybrids samples were present or not (see open triangles in Figure 3.8). This is discussed further below.

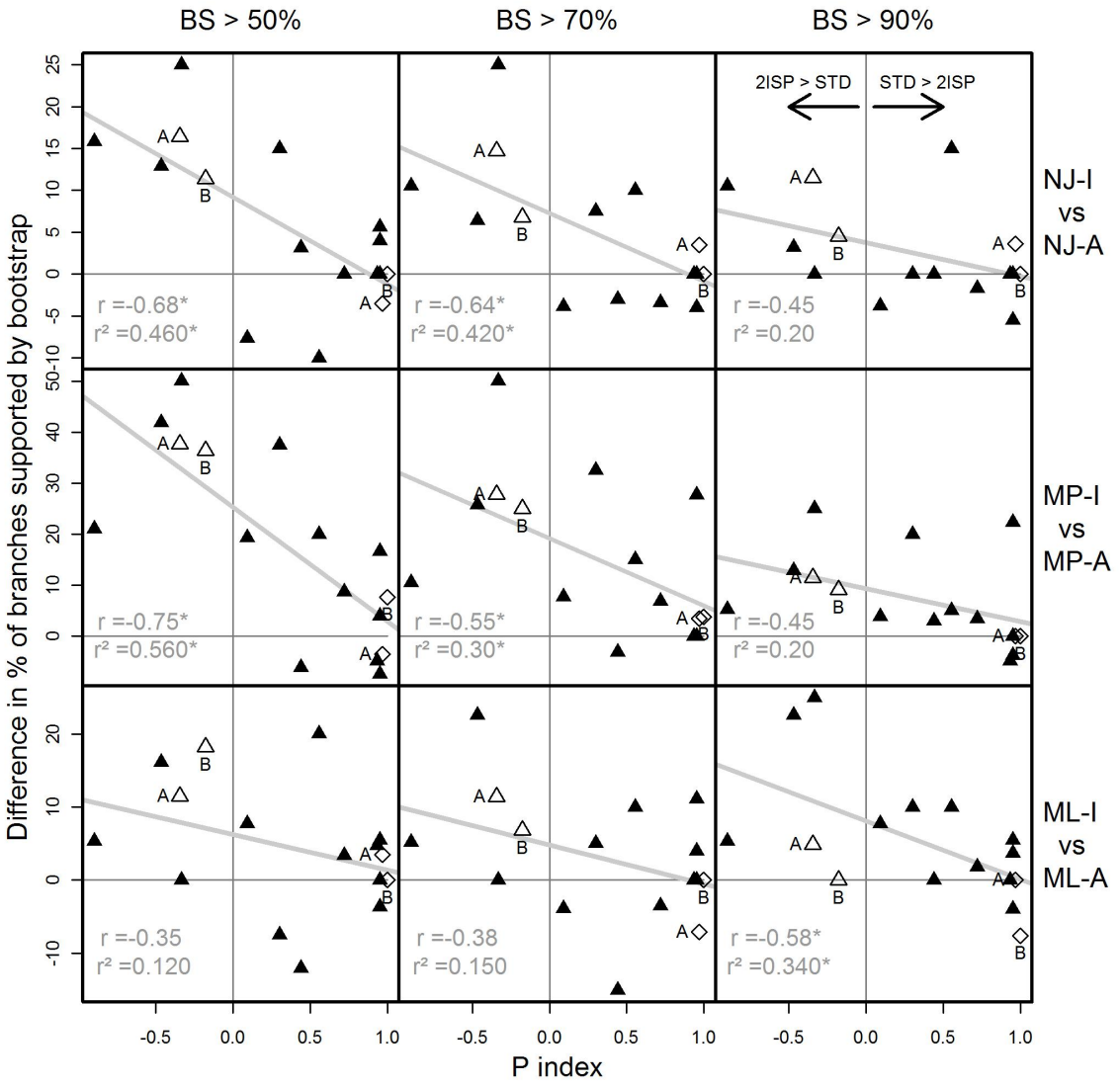

Figure 3.8. The *P* index (Pg. 69) compared to the change in the percentage of branches supported between informative to ambiguous phylogenetic treatments of intra-individual site polymorphisms (2ISPs) using real-world datasets. Three different bootstrap thresholds are considered low, medium and high (>50%, >70% and >90%, respectively). Treatments of 2ISPs are compared within three different phylogenetic methods: i) Neighbour Joining informative (NJ-I) versus Neighbour Joining ambiguous (NJ-A), ii) Maximum Parsimony informative (MP-I) versus the Maximum Parsimony ambiguous (MP-A), and iii) Maximum Likelihood informative (ML-I) versus Maximum Likelihood ambiguous (ML-A). Variation of two datasets (A and B; see text for details) are shown using open triangles (*Hieracium*) and diamonds (Amaryllidaceae). Pearson correlation coefficients (r) and coefficients of determination (r^2) are shown, with significant values ($p < 0.05$) indicated with a *.

3.4.4. Case study 1: dataset including hybrids

Fehrer *et al.* (2009) identified 23 samples in the Hieracium dataset as hybrids based on 2ISP patterns. Seventeen of these hybrid samples contained 2ISP patterns shared between two divergent clades, both in genetic and geographic terms, and were centrally located in the NN splits graph (Figure 3.9). Only these 17 samples were removed for the reduced 'hybrid-free' *Hieracium* dataset.

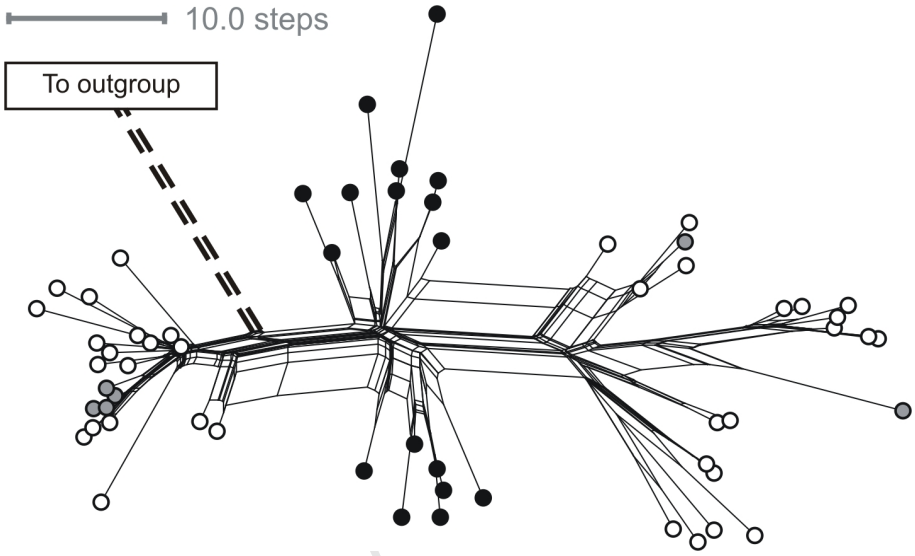

Figure 3.9. NeighborNet splits network of the complete *Hieracium* dataset with intra-individual site polymorphisms treated as informative. The white circles represent accessions used in the original analysis by Fehrer *et al.* (2009), while the black and grey circles were considered to represent putative hybrids identified by the presence of certain 2ISPs; these putative hybrids were removed from the phylogenetic inference in their study. The samples represented by black circles were considered putative hybrids removed for phylogenetic inference in this study and were removed to form a second *Hieracium* subset for other analyses. The distance to the outgroup samples (dashed branches) is 37.7 steps.

Under the full dataset none of the methods could resolve both clades with support >70% (Figure 3.10). Under NJ, both the ambiguous and informative treatments separated the samples from the different clades with a high bootstrap support for clade 2. The strict consensus tree under the MP ambiguous approach reconstructed all in-group samples to be part of a single polytomy. The MP informative approach resolved clade 2 with medium bootstrap support, but the samples from clade 1 formed

a basal polytomy. The ML-A approach resolved the two clades but without support. However, this was not the case for the ML-I approach as three samples from clade 1 formed a distinct clade, placed as sister to clade 2, although without support.

Under the reduced, hybrid-free, *Hieracium* dataset nearly all methods resolve both clades and the split between them with some degree of support (Figure 3.11). Again, the NJ-I phylogram contains more resolution than the NJ-A phylogram as there is far more relative genetic distance between samples, but otherwise the topologies are fairly similar. Both clades form polytomies in the MP-A tree, whereas the MP-I tree contains far more supported resolution. Support for clade 1 drops from 89% under MP-A to 69% under MP-I. The ML-A tree resolves both clades with high BS support. The ML-I tree fails to find support for the clade 1 samples as these samples form a basal gradation to clade 2. A comparison between the full and reduced *Hieracium* datasets (Figures 3.10 and 3.11) demonstrates the effect of the conflicting signals caused by putative hybrids which greatly affect tree reconstruction irrespective of the method applied.

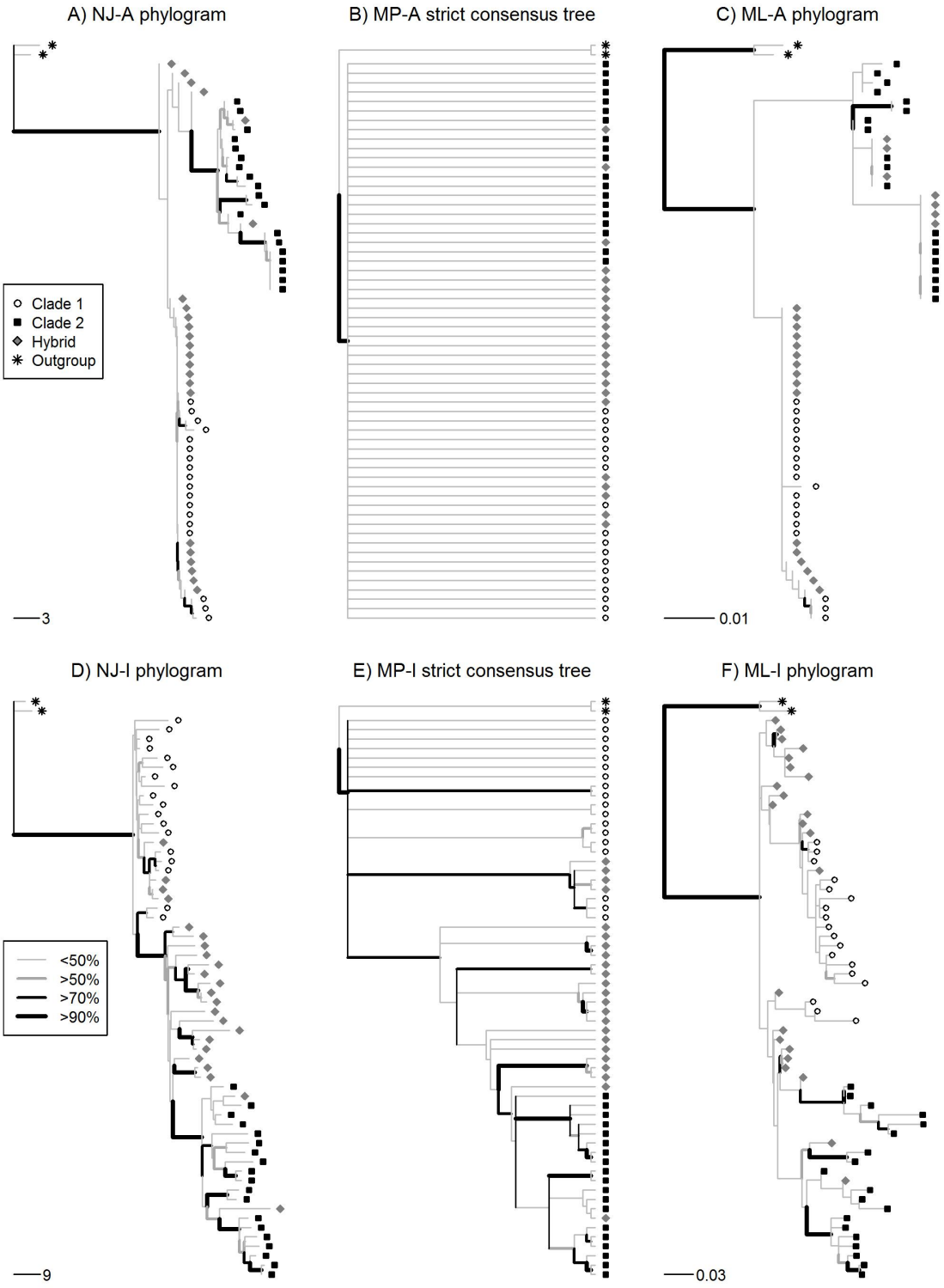

Figure 3.10. The complete *Hieracium* dataset analysed under Neighbour Joining (NJ), Maximum Parsimony (MP) and Maximum Likelihood (ML) treating intra-individual site polymorphisms as either ambiguous (-A) or informative (-I). Branch support is shown using a combination of line thickness and colour.

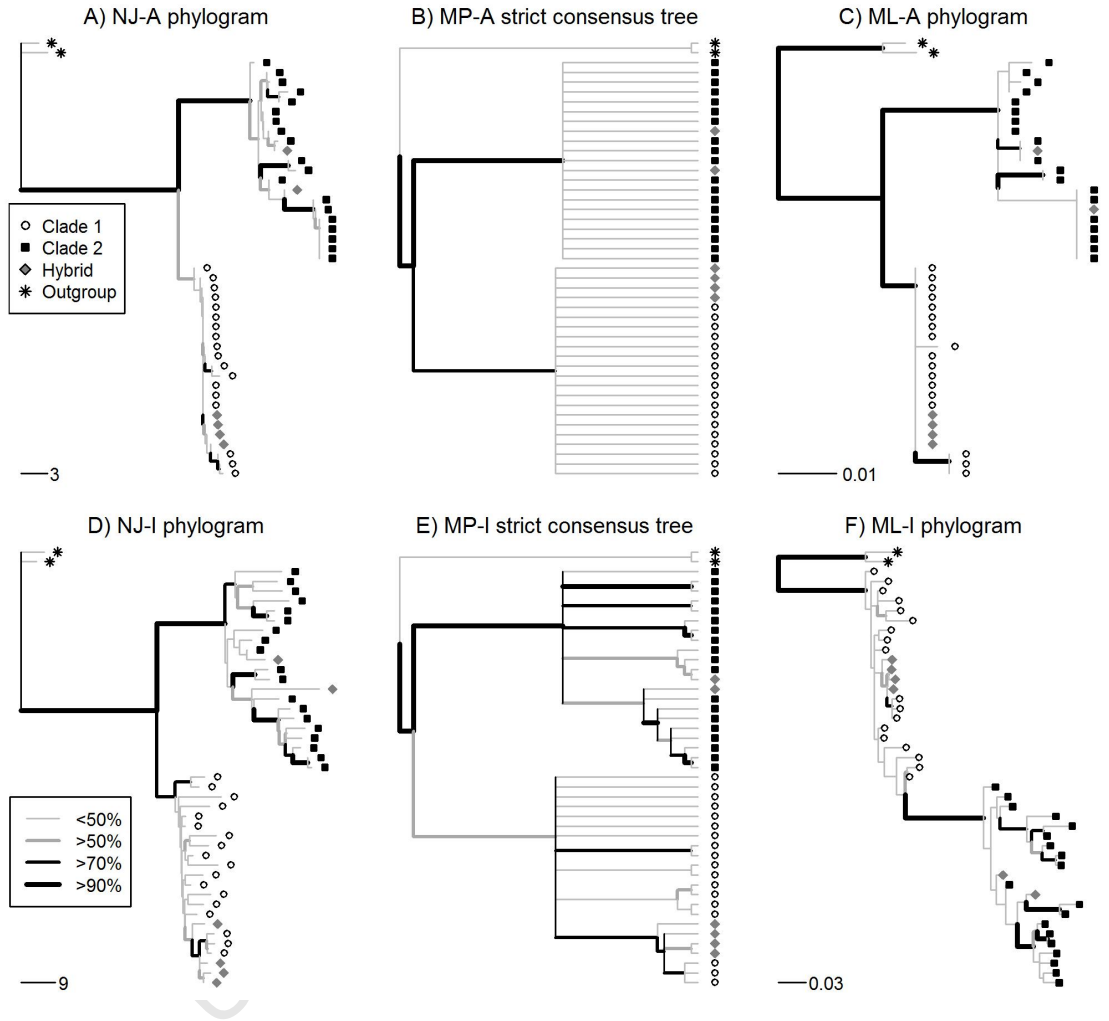

Figure 3.11. The reduced *Hieracium* dataset analysed under Neighbour Joining (NJ), Maximum Parsimony (MP) and Maximum Likelihood (ML) treating intra-individual site polymorphisms as either ambiguous (-A) or informative (-I). Hybrid samples were removed from the complete dataset that were centrally located in the splits graph (Figure 3.9), which also contained 2ISP patterns consistent with hybridisation (see text). Branch support is shown using a combination of line thickness and colour.

3.4.5. Case study 2: intraspecific dataset

The *Nymania* dataset contained 32 parsimony-informative sites, 17 of which were standard DNA characters and 29 2ISP characters, with some overlap (Tables 3.3, 3.4). Three 2ISP indels were observed, which were coded as simple mutations. Each 2ISP indel was confined to samples from a specific drainage basin. No mutations were observed from four conserved motifs nor in the 5.8S region in the direct sequences, therefore I considered the dataset free of pseudogene ITS variants.

In the *Nymania NN* splits graph (2ISPs treated as informative), three distinct clusters were resolved that correspond directly to the drainage basin from which individuals were sampled (Figure 3.12). The only exception to this is individual AJP0532 which was collected in the Gamtoos basin but is grouped with individuals from the Sundays basin; however, this collection locality was very close to the watershed boundary between the Gamtoos and the Sundays basins, and given its clustering within the Sundays, I subsequently treated it as a member of the Sundays population (this is supported by chloroplast sequence data; Chapter 4).

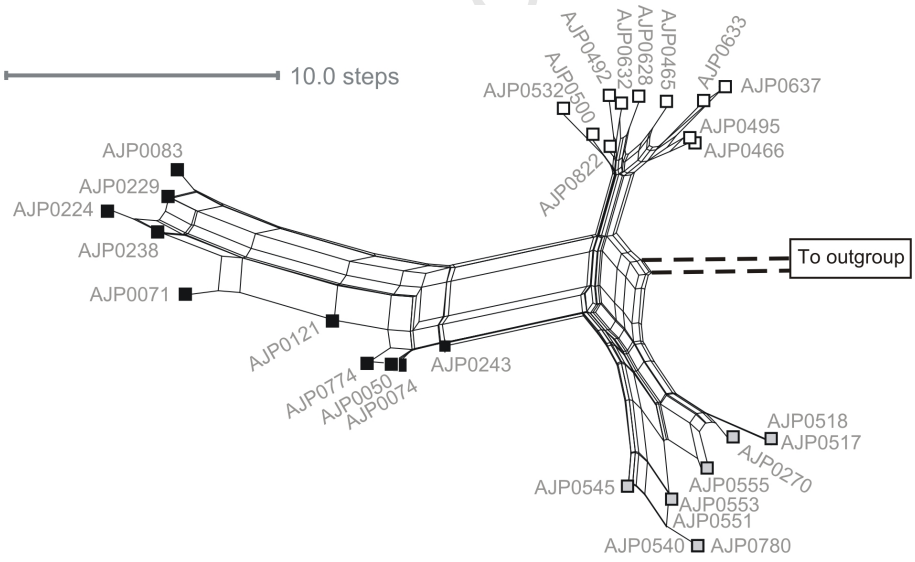

Figure 3.12. NeighbourNet splits graph of *Nymania capensis* ITS sequences with intra-individual site polymorphisms treated as informative characters. Ingroup sampling localities are shown: Sundays (white squares), Gamtoos (grey squares) and Gouritz (black squares) drainage basins. The distance to outgroup samples is 82.5 steps.

Phylogenetic tree-building using this 2ISP-rich dataset generally supported the

groupings observed in the splits graph treating the 2ISPs as informative (Figure 3.13). The NJ-I phylogram separates the drainage basin samples into well supported clades; in contrast, all Sundays samples do not form a supported clade in the NJ-A phylogram. The MP-A strict consensus tree approaches a star tree, i.e. is entirely unresolved, whereas the Gouritz, Gamtoos and Sundays samples received high, medium and low support as clades, respectively, in the MP-I tree. The ML-A tree does not contain any samples in supported or unsupported clades that exclusively match to drainage basins. However, the Gamtoos and Gouritz samples form well-supported clades in the ML-I tree.

The lack of support above 70% for a Sundays clade in the MP-I tree and the basal grading of the Sundays samples in the ML-I tree may be because this is the least differentiated and divergent group; the Sundays samples cluster in the splits graph (Figure 3.12), connected by bundles of short edges to the centre of the graph, which can be considered to represent the putative root or common ancestor. Thus, although these individuals are more closely related to one another than to either of the other two clusters, which would be indicative of a common origin, there is simply not enough signal in the data to support their own clade under the conditions used by these two algorithms and with respect to the out-group determined in-group root.

The comparison between the presence and absence of 2ISPs observed in the direct sequencing and the variable sites observed in the clones for the eight cloned samples of *N. capensis* demonstrate that the majority of variability is detected by direct sequencing (Table 3.5). A large proportion (77%) of 2ISPs observed in the direct sequences of *Nymania capensis* match up to the variable sites observed in clones (Table 3.6). Also, only a small proportion of 2ISPs observed in the direct sequences are not supported by the clones (5.5%; Table 3.6); however, this is possibly due to insufficient sampling of the intra-individual clone population rather than false 2ISP detection in the direct sequencing. These results are similar to those reported by Rosselló *et al.* (2007) and Yamaji *et al.* (2007; Table 3.6).

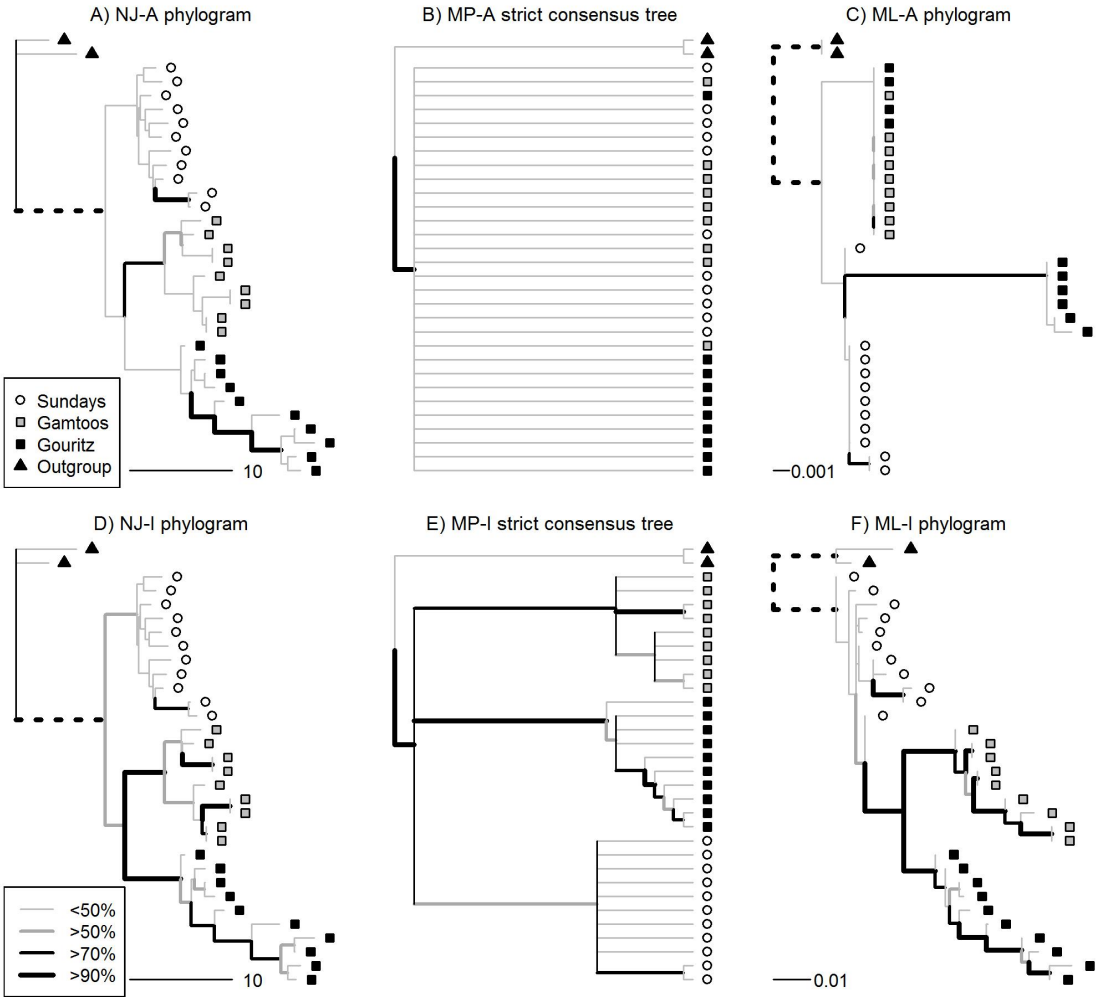

Figure 3.13. The *Nymania* dataset analysed under Neighbour Joining (NJ), Maximum Parsimony (MP) and Maximum Likelihood (ML) treating intra-individual site polymorphisms as either ambiguous (-A) or informative (-I). Dashed branches have been reduced by a factor of 10. Branch support is shown using a combination of line thickness and colour.

Table 3.4. Variable sites in direct-PCR ribosomal ITS sequences from *Nymania capensis* accessions. Intra-individual individual polymorphisms are coded using IUPAC nomenclature, with capital letters indicating complete overlap of bases in the trace file and small letters indicating one base dominant over another in the trace file. Intra-individual polymorphisms involving a base and an indel are coded using 'X'. All sequences are compared to the reference consensus sequence.

Nucleotide positions — ITS-1

Drainage Basin	Sample	25	51	61	63	64	72	74	79	86	89	92	98	106	114	130	145	207	213	225	235	253	261
	Consensus	**C**	**A**	**T**	**G**	**A**	**G**	**T**	**G**	**C**	**G**	**T**	**A**	**C**	**G**	**C**	**G**	**C**	**C**	**G**	**C**	**C**	**C**
Gamtoos	AJP0270	A	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	m
	AJP0517	A	.	.	.	C	.	.	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	.
	AJP0518	A	.	.	.	C	.	.	.	.	.	.	.	y	.	.	.	.	.	.	.	.	.
	AJP0540	A	.	.	.	M	.	.	.	X	.	.	M	.	.	.	.	Y	.	R	Y	.	m
	AJP0545	m	.	.	.	m	.	.	.	X	k	.	m	.	.	.	.	Y	.	R	Y	.	m
	AJP0551	m	.	.	.	m	.	.	.	X	.	.	m	.	.	.	.	y	.	R	Y	.	.
	AJP0553	m	.	.	.	m	.	.	.	X	.	.	M	.	.	.	.	Y	.	r	Y	.	.
	AJP0555	A	.	.	.	m	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	M
	AJP0780	A	.	.	.	M	.	.	.	X	.	.	M	.	.	.	.	y	.	r	y	.	m
Gouritz	AJP0050	m	W	w	.	m	r	.	R	.	k	.	m	.	k	y	k	.	.	.	.	.	.
	AJP0071	m	w	A	k	.	A	.	A	.	T	.	m	.	t	y	K	.	X	X	.	.	.
	AJP0074	m	W	W	.	M	R	.	R	.	K	.	M	.	K	Y	K	.	X	X	.	.	.
	AJP0083	.	T	A	.	.	A	y	A	.	T	.	C	.	T	T	.	.	.	.	.	.	.
	AJP0121	.	w	w	.	.	r	.	r	.	k	.	m	.	T	y	K	.	m	.	.	.	.
	AJP0224	.	T	A	.	.	A	.	A	.	T	.	c	.	T	T	T	.	X	X	.	.	.
	AJP0229	.	T	A	.	.	A	.	A	.	T	.	C	.	T	T	K	.	.	.	.	.	.
	AJP0238	.	w	A	.	.	A	.	A	.	T	.	C	.	T	T	K	.	X	X	.	.	.
	AJP0243	m	W	w	.	m	r	.	r	.	k	.	m	.	k	y	.	.	.	.	.	.	.
	AJP0774	m	w	w	.	m	r	.	r	.	k	.	m	.	K	y	k	.	.	.	.	.	.
Sundays	AJP0465	m	.	.	.	.	.	.	.	.	k	.	.	.	k	.	.	.	.	.	.	m	.
	AJP0466	.	.	.	.	.	.	.	.	.	k	.	m	.	k	.	.	.	.	.	.	m	.
	AJP0492	m	.	.	.	m	.	.	.	.	k	y	m	.	k	.	.	.	.	.	.	m	.
	AJP0495	.	.	.	.	.	.	.	.	.	k	y	m	.	k	.	.	.	.	.	.	m	.
	AJP0500	m	.	.	.	m	.	.	.	.	k	.	m	.	k	.	.	.	.	.	.	m	.
	AJP0532	.	.	.	.	m	.	.	.	.	k	.	m	.	k	.	.	.	.	.	.	m	.
	AJP0628	m	.	.	.	m	.	.	.	.	k	.	m	.	k	.	.	.	.	.	.	m	.
	AJP0632	m	.	.	.	m	.	.	.	.	k	.	m	.	k	.	.	.	.	.	.	m	.
	AJP0633	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.
	AJP0637	.	.	.	.	.	.	.	.	.	.	y	.	.	.	.	.	.	.	.	.	A	.
	AJP0822	m	.	.	.	M	.	.	.	.	K	.	m	.	K	.	.	.	.	.	.	M	.

Nucleotide positions — ITS-2

Drainage Basin	Sample	437	449	452	475	477	518	522	523	524	525	531	532	536	540	550	571	577	601
	Consensus	**C**	**C**	**C**	**G**	**G**	**C**	**A**	**T**	**T**	**C**	**C**	**T**	**G**	**T**	**A**	**C**	**T**	**C**
Gamtoos	AJP0270	.	.	.	.	.	.	r	w	.	.	.	.	.	C	.	.	.	.
	AJP0517	.	.	.	.	.	.	R	W	Y	.	.	.	.	C	.	.	y	.
	AJP0518	.	.	.	.	.	.	r	w	y	.	.	.	.	C	.	.	y	.
	AJP0540	.	.	.	.	.	.	r	w	y	.	.	.	r	C	.	.	.	.
	AJP0545	.	.	.	.	.	.	r	w	y	.	.	.	.	y	.	.	.	.
	AJP0551	.	.	.	.	.	.	R	W	Y	.	.	.	.	C	.	.	.	.
	AJP0553	.	.	.	.	.	.	R	W	Y	.	.	.	.	C	.	.	.	.
	AJP0555	.	.	.	.	.	.	r	w	y	.	.	.	.	C	.	.	.	.
	AJP0780	.	.	.	.	.	.	r	w	y	.	.	.	r	C	.	.	.	.
Gouritz	AJP0050	.	.	.	.	.	.	r	.	y	.	.	k	.	y	.	.	.	.
	AJP0071	.	.	.	.	.	M	r	.	y	.	.	.	.	y	.	.	.	.
	AJP0074	.	.	.	.	.	.	R	.	Y	.	.	.	.	y	.	.	.	.
	AJP0083	.	.	.	.	.	M	.	.	.	.	.	.	.	.	.	.	.	.
	AJP0121	.	.	.	.	.	m	r	.	y	.	.	.	.	y	.	.	.	.
	AJP0224	.	.	.	.	.	M	r	.	y	.	.	.	.	.	.	.	.	.
	AJP0229	.	.	.	.	.	m	.	.	.	.	.	.	.	.	.	.	.	.
	AJP0238	.	.	.	.	.	m	r	.	y	.	.	.	.	.	.	.	.	.
	AJP0243	.	.	.	.	.	.	r	.	y	.	.	.	.	y	.	.	.	.
	AJP0774	.	.	.	.	.	m	r	.	y	.	.	k	.	y	.	.	.	.
Sundays	AJP0465	X	.	.	.	R	.	.	.	.	.	y	T	.	.	m	.	.	.
	AJP0466	X	.	.	r	r	.	.	.	.	.	y	.	.	.	.	.	.	.
	AJP0492	X	.	.	.	R	.	.	.	.	.	y	.	.	.	.	s	.	.
	AJP0495	X	.	.	.	R	.	.	.	.	.	Y	.	.	.	.	.	.	.
	AJP0500	X	.	.	.	R	.	.	.	.	m	.	.	.	.	.	.	.	.
	AJP0532	X	.	.	.	R	.	.	.	.	m	Y	.	.	.	.	.	.	y
	AJP0628	X	y	y	.	R	.	.	.	.	.	y	.	.	.	.	.	.	.
	AJP0632	.	y	.	.	R	.	.	.	.	.	y	.	.	.	.	s	.	.
	AJP0633	X	.	.	.	r	.	.	.	.	.	y	.	.	.	.	.	.	.
	AJP0637	X	.	.	.	r	.	.	.	.	.	y	.	.	.	.	.	.	.
	AJP0822	X	.	.	.	r	.	.	.	.	.	y	.	.	y	.	.	.	.

Table 3.5. Comparison of variable sites between direct-PCR sequences and cloned sequences from eight *Nymania capensis* accessions. Intra-individual individual site polymorphisms are coded using IUPAC nomenclature, with capital letters indicating complete overlap of bases in the trace file and small letters indicating one dominant base dominant over another in the trace file. Intra-individual site polymorphisms detected in the direct sequences involving a base and an indel are coded using 'X'. Clone sequences are compared to the direct sequence from a given accession. A dash (-) denotes a nucleotide indel that may be 1 to 5 bases in length.

	Nucleotide positions																																			
	ITS-1																				5.8S		ITS-2													
	25	51	61	64	72	79	86	89	92	98	106	114	130	145	207	213	225	235	253	261	308	323	437	475	477	518	522	523	524	525	531	536	540	571	577	601
LKP0121	C	**w**	**w**	A	**r**	**r**	C	**k**	T	**m**	C	T	**y**	**K**	C	**m**	G	C	C	C	G	C	C	G	G	**m**	**r**	T	**y**	C	C	G	**y**	C	T	C
2	.	T	A	.	A	A	.	T	.	C	.	.	T	T	.	C	.	.	.	.	.	.	.	.	.	A	A	.	T	.	.	.	T	.	.	.
3	.	T	A	.	A	A	.	T	.	C	.	.	T	G	.	C	.	.	.	.	.	.	.	.	.	C	A	.	T	.	.	.	T	.	.	.
1	.	T	A	.	A	A	.	T	.	C	.	.	T	T	.	C	.	.	.	.	.	.	.	.	.	C	A	.	T	.	.	.	T	.	.	.
1	A	A	T	T	G	G	.	G	.	A	.	G	C	G	.	C	.	T	.	.	.	.	.	.	.	C	G	A	C	.	.	.	C	.	.	.
LKP0495	C	A	T	A	G	G	C	**k**	**y**	**m**	C	**k**	C	G	C	C	G	C	**m**	C	**r**	C	**X**	**R**	**R**	C	A	T	T	C	**Y**	G	T	C	T	C
2	.	.	.	.	.	.	.	G	T	A	.	G	.	.	.	.	.	.	A	.	G	.	C	G	G	.	.	.	.	.	C	.	.	.	.	.
2	A	.	.	C	.	.	.	T	T	C	.	T	.	.	.	.	.	.	C	.	G	T	C	G	G	.	.	.	.	.	C	.	.	G	.	.
2	.	.	.	.	.	.	.	G	T	A	.	G	.	.	.	.	.	.	A	.	G	.	-	G	A	.	.	.	.	.	C	.	.	.	.	.
1	A	.	.	C	.	.	.	G	T	A	.	G	.	.	.	.	.	.	A	.	G	.	C	G	G	.	.	.	.	.	C	.	.	.	.	.
1	.	.	.	.	.	.	.	G	T	A	.	G	.	.	.	.	.	.	A	.	G	.	-	G	G	.	.	.	.	.	T	.	.	.	.	.
LKP0517	A	A	T	C	G	G	C	G	T	A	**Y**	G	C	G	C	C	G	C	C	C	G	C	C	G	G	C	**R**	**W**	**Y**	C	C	G	C	C	**y**	C
1	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	A	T	T	.	.	.	.	.	C	.
2	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	-	.	.	.	A	T	T	.	.	.	.	.	.	.
1	C	.	.	A	.	.	.	.	.	.	C	.	.	.	.	.	.	.	A	.	.	.	-	.	A	.	A	T	T	A	.	.	T	.	.	.
3	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	G	A	C	.	.	.	.	.	.	.
1	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	T	.	.	.	.	.	.	C	.	.	.	G	A	C	.	.	.	.	.	.	.
LKP0532	C	A	T	**m**	G	G	C	**k**	T	**m**	C	K	C	G	C	C	G	C	**m**	C	**r**	C	**X**	G	**R**	C	A	T	T	**m**	**Y**	G	T	C	T	**y**
2	.	.	.	A	.	.	.	G	.	A	.	G	.	.	.	.	.	.	A	.	G	C	-	.	G	.	.	.	.	C	T	.	.	.	.	C
2	.	.	.	A	.	.	.	G	.	A	.	G	.	.	.	.	.	.	A	.	G	C	-	.	A	.	.	.	.	C	C	.	.	.	.	C
1	.	.	.	A	.	.	.	G	.	A	.	G	.	.	.	.	.	.	A	.	G	C	C	.	A	.	.	.	.	A	C	.	.	.	.	C
1	A	.	.	C	.	.	.	T	.	C	.	T	.	.	.	.	.	.	C	.	G	T	-	.	G	.	.	.	.	C	C	.	.	.	.	T
1	A	.	.	C	.	.	.	T	.	C	.	T	.	.	.	.	.	.	C	.	G	C	C	.	G	.	.	.	.	C	C	.	.	.	.	C

Table 3.5. continued.

		Nucleotide positions																																			
		ITS-1																				5.8S		ITS-2													
		25	51	61	64	72	79	86	89	92	98	106	114	130	145	207	213	225	235	253	261	308	323	437	475	477	518	522	523	524	525	531	536	540	571	577	601
LKP0553		**m**	A	T	**m**	G	G	**X**	G	T	**M**	C	G	C	G	**Y**	C	**r**	**Y**	C	C	G	C	C	G	G	C	**R**	**W**	**Y**	C	C	G	C	C	T	C
	2	A	.	.	C	.	.	-	.	.	A	.	.	.	.	C	.	G	C	.	.	.	.	.	.	.	.	A	T	T	.	.	G	.	.	.	.
	2	A	.	.	C	.	.	-	.	.	A	.	.	.	.	C	.	G	C	.	.	.	.	.	.	.	.	G	A	C	.	.	A	.	.	.	.
	1	C	.	.	A	.	.	C	.	.	C	.	.	.	.	T	.	A	T	.	.	.	.	.	.	.	.	G	A	C	.	.	A	.	.	.	.
	1	A	.	.	C	.	.	-	.	.	A	.	.	.	.	C	.	G	C	.	.	.	.	.	.	.	.	G	A	C	.	.	G	.	.	.	.
	1	C	.	.	A	.	.	C	.	.	C	.	.	.	.	T	.	A	T	.	.	.	.	.	.	.	.	A	T	T	.	.	G	.	.	.	.
	1	C	.	.	A	.	.	C	.	.	A	.	.	.	.	C	.	G	C	.	.	.	.	.	.	.	.	G	A	C	.	.	G	.	.	.	.
LKP0637		C	A	T	A	G	G	C	G	**y**	A	C	G	C	G	C	C	G	C	A	C	**r**	C	C	G	**r**	C	A	T	T	C	**y**	G	T	C	T	C
	2	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	G	.	.	.	.	.	C	.	.	.	.	.
	2	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	A	.	.	.	.	.	C	.	.	.	.	.
	2	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	A	.	.	.	.	.	C	.	.	.	.	.
	1	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	G	.	.	.	.	.	T	.	.	.	.	.
	1	A	.	.	C	.	.	.	.	T	.	.	.	.	.	.	.	.	.	C	.	G	.	.	.	G	.	G	A	C	.	C	.	C	.	.	.
LKP0780		A	A	T	**M**	G	G	**X**	G	T	**M**	C	G	C	G	**y**	C	**r**	**y**	C	**m**	G	C	C	G	G	C	**r**	**w**	**y**	C	C	**r**	C	C	T	C
	3	.	.	.	C	.	.	-	.	.	A	.	.	.	.	C	.	G	C	.	C	.	.	.	.	.	.	G	A	C	.	.	A	.	.	.	.
	1	.	.	.	C	.	.	-	.	.	A	.	.	.	.	C	.	G	C	.	C	.	.	.	.	.	.	A	T	T	.	.	G	.	.	.	.
	2	C	.	.	A	.	.	C	.	.	C	.	.	.	.	T	.	A	T	.	C	.	.	.	.	.	.	A	T	T	.	.	G	.	.	.	.
	1	.	.	.	C	.	.	-	.	.	A	.	.	.	.	C	.	G	C	.	C	.	.	.	.	.	.	G	A	C	.	.	G	.	.	.	.
	1	.	.	.	C	.	.	T	.	.	A	.	.	.	.	C	.	G	C	.	A	.	.	.	.	.	.	A	T	T	.	.	G	.	.	.	.
LKP0822		**m**	A	T	**M**	G	G	C	**K**	T	**m**	C	**K**	C	G	C	C	G	C	**M**	C	G	C	**X**	G	**r**	C	A	T	T	C	**y**	G	**y**	C	T	C
	2	A	.	.	C	.	.	.	T	.	C	.	T	.	.	.	.	.	.	C	.	.	T	C	.	G	.	.	.	.	T	C	.	T	.	.	T
	1	C	.	.	A	.	.	.	G	.	A	.	G	.	.	.	.	.	.	A	.	.	.	-	.	G	.	.	.	.	.	T	.	T	.	.	.
	1	A	.	.	C	.	.	.	G	.	A	.	G	.	.	.	.	.	.	C	.	.	.	C	.	G	.	G	A	C	.	C	.	C	.	.	.
	1	C	.	.	A	.	.	.	G	.	A	.	G	.	.	.	.	.	.	A	.	.	.	-	A	A	.	.	.	.	.	C	.	T	.	.	.
	1	C	.	.	C	.	.	.	T	.	C	.	T	.	.	.	.	.	.	A	.	.	.	C	.	G	.	.	.	.	.	C	.	T	.	.	.
	1	C	.	.	A	.	.	.	G	.	A	.	G	.	.	.	.	.	.	A	.	.	.	C	.	G	.	.	.	.	.	C	.	T	.	.	.
	1	A	.	.	C	.	.	.	G	.	A	.	G	.	.	.	.	.	.	C	.	.	.	C	.	G	.	.	.	.	.	C	.	C	.	.	.

Table 3.6. Comparison of intra-individual site polymorphisms detected from direct-PCR sequencing versus cloning across multiple sequences from three plant taxa. The number of samples and number of clones per sample are shown in brackets.

	Direct sequences					
	Nymania capensis (8; 8)		*Buxus balearica*[1] (5; 3-11)		*Asarum* sect. *Asiasarum*[2] (30; 13-31)	
Clone sequences	Present	Absent	Present	Absent	Present	Absent
Present	69	20	29	4	177	54
Absent	4	-	3	-	1	-

[1]Rosselló *et al.* (2007)
[2]Yamaji *et al.* (2007)

3.5. Discussion

I present a method that incorporates 2ISPs as informative characters in NJ and MP phylogeny reconstruction using a simple step matrix (Figure 3.1). I also test the performance of standard nucleotide-data ML analyses versus treating 2ISPs as additional character states. I demonstrate that the presence of samples with a hybrid origin negatively affects all tree reconstruction methods, irrespective of whether the ambiguous or informative approaches are used (Figures 3.2, 3.3 and 3.4). However, I also demonstrate that the informative approach offers significant improvements for the topological resolution (based on the number of supported branches) and reliability (regarding the distance to the real tree) of trees constructed from simulated nuclear datasets (Figures 3.5, 3.6 and 3.7) when 2ISPs represent inherited variation. Furthermore, I demonstrate that the 2ISP-informative approach also offer significant improvements for real-world datasets, including those with putative hybrids (Figures 3.8, 3.10, 3.11 and 3.13). Data from the nuclear genome provides vital information for phylogenetics and phylogeography; however, 2ISPs are bound to arise due to the genome's complexity (in comparison to chloroplast and mitochondrial genomes), biparental inheritance and where multiple fragments may be amplified from across the genome (e.g. ITS). These 2ISPs have generally been treated as uncertainties across a range of phylogenetic software (e.g. PAUP* and MrBayes). Depending on the source of 2ISPs, I have established that treating 2ISPs as informative characters can aid, rather than hinder, phylogenetic inference.

The problematic effect of polymorphism, irrespective of data source (e.g. morphology or DNA), on phylogeny reconstruction is a long-standing problem (reviewed in Wiens 1999). For nuclear sequence data, different approaches have been used to address the decline in topological resolution and support values in the presence of 2ISPs: 1) removing samples identified as hybrids based on 2ISP patterns (e.g. Aguilar & Feliner 2003, Fehrer *et al.* 2009, Whittall *et al.* 2000), 2) removing any samples with two or more 2ISPs (e.g. Hanna *et al.* 2007), 3) removing any sites that contain 2ISPs prior to analysis (e.g. Scherson *et al.* 2008), 4) treating all DNA bases (including 2ISPs) as unordered characters (e.g. Campbell *et al.* 1997, Fama *et al.* 2000), or 5) cloning samples containing 2ISPs and including clones rather than the original sequence in the phylogenetic analyses (suggested standard practice by Feliner & Rosselló 2007). In an attempt to address analytical problems created by 2ISPs, statistical haplotype

inference methods have even been used to infer variants within ITS sequences based on 2ISP patterns (e.g. Lorenz-Lemke *et al.* 2005), despite such methods being reserved for low-copy regions. Statistical haplotype inference assumes that the gene region in question is diploid and, at most, two alleles are present (Clark 1990, Stephens *et al.* 2001) - this assumption is violated with complex multi-gene families such as ITS. All of these approaches result in the loss of potentially informative data as either potentially useful information is removed or the association between variants and samples is lost (but see Göker & Grimm 2008, Grimm *et al.* 2007, Joly & Bruneau 2006 for examples on how to preserve the information of 2ISPs represented in cloned variants). As shown here, the primary cause for the loss of topological resolution and decrease of branch support when 2ISPs are present in a dataset is due to their treatment as uncertain characters or missing data in commonly used phylogenetic software.

Comparative assessment of simulated and experimental data sets shows that treating 2ISPs as informative characters provides an opportunity to incorporate additional information in phylogenetic inference in a straightforward and cost-sensitive manner. The performance of treating 2ISPs as informative is considerably better than the standard approach for both the simulated and real-world data (Figures 3.5, 3.6, 3.7, 3.8, 3.10, 3.11, and 3.13). Moreover, the presence of hybrid samples appears to have equally negative effect on ambiguous and informative approaches (except for MP; Figures 3.2, 3.3, and 3.4). Thus, the positive results from analyses of real-world datasets, which demonstrate significant improvements under the informative approach (Figure 3.8), suggests that the 2ISP variation provides additional information regarding the phylogenetic relationships between samples. This is most likely because the 2ISPs represent inherited variation. Also, this suggests that phylogenetic inference using the 2ISP-informative approach can be done without knowing the sources of intra-individual polymorphism, the processes that induce intra- and inter-array variation, which, in many cases, may be hard to identify at all. The 2ISP-informative approach is especially more powerful when datasets have a negative P index (2ISPs outweigh standard parsimony informative characters); however, even under positive P index values an improvement in tree topology and support is evident (Figure 3.8).

Intra-individual site polymorphisms can arise through many different mechanisms, some of which maintain evolutionary signal (e.g. incomplete concerted evolution where mutation rates outstrip homogenisation processes) and others that blur it (e.g. hybridisation). Our hybrid simulation results (Figures 3.2, 3.3, and 3.4) and the

Hieracium case study (Figure 3.10) demonstrate that the 2ISP-informative approach does not solve the general problem of conflicting signal within data induced by hybrids, or reticulation in general, in phylogenetic reconstruction. However, the MP-I approach does outperform the standard MP-A approach in resolving at least some of the underlying topology with support, primarily because the number of conflicting equally parsimonious trees is far greater under MP-A when 2ISPs are present. Hybridisation causes problems because a hybrid may be equally closely related to both its progenitors. Thus, the placement of hybrids within a tree is unpredictable - at best they group with one of the progenitors, but the support will typically be low. Hybrid samples have traditionally been removed because of the conflicting signal they induce when it comes to phylogenetic tree-building (Vriesendorp & Bakker 2005). When removing the *Hieracium* hybrids centrally located in the splits graph (Figure 3.9), the basic division of the genus into two major clades observed by Fehrer *et al.* (2009) is retrieved by both 2ISP-informative and 2ISP-ambiguous treatments by most methods (Figure 3.11). In general, the 2ISP-informative treatment outperforms that of 2ISP-ambiguous in providing greater supported topological resolution within the two clades (Figures 3.8 and 3.11). However, as highlighted by Vriesendorp & Bakker (2005, p. 598), 'the common practice of leaving suspected [hybrid] taxa out of the analysis to avoid confounding effects on phylogenetic reconstruction will not stimulate further progress'. The incapacity of dichotomous trees to deal with reticulate signal can be overcome to some degree by using phylogenetic networks based on collections of trees or distance matrices (Huson & Bryant 2006, Lockhart 2006, Vriesendorp & Bakker 2005), but most implementations of these network methods also do not treat 2ISPs as informative characters (e.g. statistical parsimony networks, Clement *et al.* 2000)(NeighbourNet, Bryant & Moulton 2004 – but see the 'Average' option for calculated uncorrected p-distance in SplitsTree). Intra-individual site polymorphisms can be included into networks by either using the 2ISP-informative approach to construct the tree collection or base the distance matrices on the polymorphism p-distance (e.g. Figures 3.9 and 3.12).

When 2ISPs are derived from population-level processes (or independent evolutionary histories in general) that maintain the underlying evolutionary signal, then the 2ISP-informative approach not only offers a dramatic improvement over the standard approach, but also greater resolution towards the species or population tree. The independent evolution of different variants represents unique gene histories (Maddison

1997), which in combination may culminate in an improved prediction of the actual species or population tree (e.g. see Fig. 5 in Göker & Grimm 2008). The simplistic independent-variant simulations used here demonstrate that the combined variants $A_{1\&2}$ datasets have significantly more supported nodes and are generally closer to the 'true' tree used to simulate those data than the independent variant datasets (A_1 and A_2; Figures 3.5 and 3.6). Thus, intra-individual polymorphisms may offer a consensus signal of the underlying coalescent gene trees of a number of variants.

An example of multiple variants representing population level processes is shown in the case study of *Nymania capensis*. The 2ISPs in this dataset are most likely not due to hybridisation between species (*Nymania* is a monotypic genus), but can instead be attributed to a lack of concerted evolution, divergence of ITS copies, and inheritance of polymorphisms. The clusters found in the splits graph (Figure 3.12) and clades resolved in the NJ-I, MP-I and ML-I trees (Figure 3.13) provide valuable phylogeographic information, and suggest that populations have been isolated in different drainage basins. This is supported by chloroplast data (Chapter 4). The high proportion of 2ISPs in the *Nymania* dataset would normally preclude it from being used for any form of phylogenetic reconstruction when they are treated as ambiguous since some clades may be unsupported (e.g. NJ-A; Figure 3.13.A), or misleading relationships between samples may be inferred (e.g. ML-A; Figure 3.13.C) or so many equally parsimonious trees exist that the strict consensus tree is star-like (MP-A; Figure 3.13.B). Thus, where population-level heterogeneity exists, treating 2ISPs as informative provides both a means for dealing with this heterogeneity which includes provision for polymorphisms in hypothetical ancestral taxa and an estimate of the 'population' tree from the summarised information from the different variants.

Unfortunately, the process of concerted evolution can greatly reduce the variant heterogeneity and thus remove much of the evolutionary signal detectable by direct sequencing (although rare variants can persist in the genome even through multiple speciation events, Mahelka & Kopeck 2010). Therefore, the studies that will benefit the most from the 2ISP-informative approach are those where the target taxa have not recently experienced significant concerted evolution (as is the case with *Nymania capensis*). These will most likely be phylogeographic or intra-generic studies (e.g. Fama *et al.* 2000, Feliner *et al.* 2004, Rosselló *et al.* 2007). The probabilities of complete or near-complete concerted evolution increases as the species involved become more distantly related through evolutionary time.

The ML-I approach offers an alternative to the fixed step matrix used under NJ-I and MP-I. It allows an estimation of a transition matrix based on the shifts between all characters, including 2ISPs, and this may offer a more accurate realisation of 2ISP evolution. However, an accurate estimate of the transition matrix under the ML-I approach may require far more signal within any given dataset than the ML-A approach; under ML-A, the transition matrix is only estimated between four character states, whereas anywhere up to 15 different character states may be in the ML-I transition matrix. An increasingly complex transition matrix is less likely to be stable and also less decisive regarding the underlying data if there are insufficient data for transition rate estimation. This would decrease support and increase topological ambiguity. This most likely explains why clade 1 and Sundays samples form an unsupported basal grades in the reduced-*Hieracium* and *Nymania* datasets, respectively, under the ML-I approach (Figures 3.11.F and 4.7.F), whereas all other methods reconstruct these samples into a clade, although not always with support. Despite this potential problem when estimating the transition matrix, the branch support for ML-I trees is greater than ML-A trees for the majority of real-world datasets (Figure 3.8) and ML-I outperforms ML-A for the *N. capensis* case study (Figure 3.13). Also, the ML-I approach dramatically outperforms the ML-A approach in the 2ISP-dominated $A_{1\&2}$ datasets of independent variation (Figures 3.5, 3.6 and 3.7). However, this may, in part, be due to the current manner in which the ML-A transition matrix is estimated in RAxML; 2ISPs are incorporated into tree searching as outlined by Felsenstein (Felsenstein 2004, p. 253-256), however the estimation of the transition matrix prior to tree searching ignores 2ISPs (A. Stamatakis, personal communication). This failure to include 2ISPs when estimating the transition matrix may explain the poor performance of ML-A when analysing the $A_{1\&2}$ datasets of independent variation as the transition matrices have unrealistically low transition estimates (near zero) especially under the datasets with lower signal. This may also explain why the ML-A method marginally outperforms ML-I in the hybrid simulations (Figures 3.2, 3.3 and 3.4), as transition matrices remain fairly constant for ML-A analyses because the 2ISPs in the hybrid samples are ignored while the transition matrices for ML-I fluctuate dramatically due to the presence of 2ISPs in a single sample which increases the numbers of characters in the transition matrix without providing enough information to estimate the transitions (e.g. estimates of the alpha parameter vary by three orders of magnitude between HP datasets under ML-I). Thus, the ML-A approach may still be a viable means of including 2ISPs as informative characters but

is currently underperforming as 2ISPs are not included when estimating the transition matrix. The success of the ML-I approach will largely depend on whether there is sufficient information in a dataset to obtain realistic estimates of transition rates. This would need to be judged on a case by case basis.

3.5.1. Identifying 2ISPs

Intra-individual site polymorphism observed in trace files obtained from direct sequencing of PCR products have been treated as informative characters for a range of applications other than phylogeny reconstruction, such as hybrid identification (Aguilar & Feliner 2003, Fehrer *et al.* 2007), determining geographic structure (Feliner *et al.* 2004), and statistically inferring haplotypes from low copy nuclear genes (e.g. PHASE, Stephens *et al.* 2001). However, the relationship between 2ISPs identified through direct sequencing and the underlying population of variants is fairly unexplored, although it is clear that not every polymorphic site observed across clones is detected through direct-PCR sequencing (Tables 3.5 and 3.6). The undetected clone variation may be due to rare copies that are detected via cloning, but do not have the numerical dominancy to be detected by direct sequencing, or possibly biased amplification of copies in the PCR. Nonetheless, the 2ISPs identified from direct sequencing detect more than 75% of the underlying variation observed in clones (Table 3.6). In a simple sequencing experiment, Rauscher *et al.* (2002) demonstrated that the heights of multiple peaks observed in trace files were directly proportional to the underlying and controlled frequencies of clone copies used in the sequencing reaction. Thus, any variation observed in rare copies is unlikely to be detected via direct sequencing and even cloning (Rauscher *et al.* 2002).

Although not all 2ISPs in direct sequences were resolved in clones and not all variable sites in clones were detected by direct sequencing (Table 3.6), the degree of resolution should be sufficient for further phylogenetic and phylogeographic analyses. Detecting 2ISPs in sequences is critically important to the multitude of studies using direct-PCR sequences from multi-copy genes, as the same factors that affect 2ISP identification can affect base calling and create misleading information in the data (e.g. biased PCR amplification between variants or failure to code polymorphisms). While the relationship between direct-PCR sequences and the underlying ITS copy population is positive, it still needs to be explored further. Until this relationship

between direct sequencing and the underlying variant population has been established, I suggest the following steps to identify 2ISPs prior to phylogenetic analyses:

1. Use high fidelity (proofreading) polymerases for PCRs. This reduces the background noise from misannealed nucleotides in trace files and results in more reliable sequences (Arezi *et al.* 2003, Cline *et al.* 1996).
2. Sequence in both directions and ensure that 2ISPs are present in both sequences (Baldwin *et al.* 1995). If in doubt, then re-sequence.
3. Clone a subset of samples and compare the 2ISPs identified in direct sequences with those found in the clone alignment in order to independently verify the presence and absence of 2ISPs. If the variable sites observed in clones have a high degree of concordance with 2ISPs detected from direct sequencing, then direct sequencing is likely to provide a reliable signal across closely related species or multiple populations of the same species of the underlying ITS copy population (e.g. Table 3.6).

Although direct sequencing may work for the majority of groups, it may not work under a number of conditions: i) when numerous variants of different lengths are present within an individual, and ii) when there are two or more very different active ITS regions, such as those observed in grasses (e.g. Peng *et al.* 2010). The presence of indels results in copies that are out of phase and this causes nonhomologous sites to be overlaid; usually such direct sequences are discarded as unusable. However, there are algorithmic approaches to separating such super-imposed sequences (e.g. Dixon 2010, Dmitriev & Rakitov 2008). Although promising, these methods must be applied with caution as they assume diploidy and thus may not be applicable to multi-copy regions such as ITS. Furthermore, it remains uncertain as to how effective such methods are at detecting 2ISPs.

The strict usage of IUPAC ambiguity codes is extremely important for the 2ISP-informative approach. Previously, it used to make no difference to the majority of analyses as to whether IUPAC ambiguity codes were used to indicate a poor signal in the trace file or the presence of two bases in a single site (2ISP). With the informative treatment of 2ISPs used here, this distinction becomes extremely important. I strongly suggest that poor signal be represented by uncertainty symbols ('?') and that IUPAC ambiguity codes be reserved for 2ISPs only.

3.5.2. Widespread use of the 2ISP-informative approach

The 2ISP-informative approach can be used for any DNA sequences from high or low copy genes wherever 2ISPs are found, as demonstrated in this study. Furthermore, the step matrix used to implement the 2ISP-informative approach is an extension of the allele sharing distance (Bowcock 1994) widely used to calculate distance matrices for unlinked single nucleotide polymorphisms (SNP) data. This extension will enable the calculation of distance matrices for complex SNPs (i.e. those that have greater variability than simply two bases and the polymorphism of those bases). It will also enable the implementation of MP for tree reconstruction of SNP data, a method that has been lacking in the majority of studies using unlinked SNPs.

Phylogenetic analyses of combined data can be problematic when 2ISPs are present. Feliner & Rosselló (2007) suggest cloning any heterozygous sequences prior to analysis. However, if researchers wish to combine datasets under the 'total evidence' or 'conditional combination' approaches (reviewed in Huelsenbeck *et al.* 1996), then it is difficult to combine multiple cloned sequences and those sequences from other regions or genomes all from the same individual. A commonly used option is to select a single clone per individual per cloned gene region to be combined with other data. For example, Steele *et al.* (2010) selected single clone variants from two different nuclear gene regions to concatenate with chloroplast sequences; this leads to a loss of information and runs the risk of filtering bias. The 2ISP-informative approach will enable researchers to combine multiple regions, and account for 2ISPs, whether they are identified through direct sequencing or consensus sequences from clones.

3.5.3. Conclusions

In a phylogeographic study of *Vaccinium uliginosum*, Eidesen *et al.* (2007) state that 'as a consequence [of intra-individual site polymorphisms], unless all ITS PCR products are cloned, any phylogenetic signal useful for inferring relatively recent phylogeographical patterns is effectively concealed by the polymorphisms caused by the two [or more] paralogous ITS repeats'. This frustration has been shared by many researchers. Here it has been demonstrated that treating once problematic 2ISPs as informative characters can dramatically improve the resolution of phylogeny reconstruction. I envisage that this method should greatly aid phylogenetic inference

at the intra-generic or intra-specific level, including phylogeography studies.

4. Phylogeographic congruence between genetic patterns, river drainage basins and palaeoclimatic niche models of *Nymania capensis* (Meliaceae)

4.1. Abstract

Landscape features play an important role in the distribution of species and gene flow between populations, and the *evolutionarily discrete drainage basin* hypothesis (EDDB) suggests that this is particularly applicable to the topographically complex coastal lowlands of South Africa. Furthermore, the climatic changes during the Pleistocene glacial cycles profoundly affected species distributions, gene flow between populations and demography. The *glacial refugia hypothesis* predicts that the Albany Subtropical Thicket (AST), a coastal lowland vegetation, retracted into fragmented refugia during glacial cycles. I test these two hypotheses using multigene phylogeography and the projected Last Glacial Maximum (LGM) distribution of a dominant AST plant species, *Nymania capensis* (Meliaceae). The patterns of genetic diversity from non-coding chloroplast DNA, a high-copy nuclear region (ITS) and a low copy nuclear region, support the EDDB hypothesis as: 1) different drainage basins contain genetically distinct lineages, 2) limited genetic structuring was detected within basins whilst high structuring was detected between basins, and 3) within drainage basin populations display a high degree of genealogical lineage sorting. Both molecular dating and climate envelope modelling support the glacial refugia hypothesis as: a) the timing of chloroplast lineage diversification is restricted to the Pleistocene in a landscape that has been relatively unchanged since the Pliocene, and b) the projected LGM distribution of suitable climate for *N. capensis* suggest fragmentation into refugia that correspond to the current phylogeographic populations. This suggests that the AST was fragmented and isolated along the lowland basins of Southern Africa during the LGM and prior glacial periods.

4.2. Introduction

Both historical and current variation in topography and climate may influence the spatial patterns of genetic diversity within a species. Topographic complexity may act as a barrier to gene flow, with examples ranging from mountain-top species that cannot cross valleys (e.g. Ehrich *et al.* 2007) to lowland species that cannot disperse across uplands (e.g. Zhang *et al.* 2011). Climatically-induced shifts in the geographical distribution of species driven by variations in the Earth's orbit, termed 'orbitally forced range dynamics', will change the degree to which topography influences the connectivity between populations (Jansson & Dynesius 2002). Drainage basins and their corresponding watersheds are dominant topographic features in the terrestrial landscape that are acknowledged barriers to dispersal and gene flow; however, their effect on genetic diversity has rarely been tested beyond aquatic organisms. In this chapter I investigate the role of drainage basins and past climatic fluctuations on the structuring of genetic diversity in a plant species, *Nymania capensis* (Meliaceae). This species is a dominant and widespread component of the Albany Subtropical Thicket (AST, Figure 4.1) biome which is restricted to the coastal lowlands of South Africa.

The vegetation of the AST is characterised as dense, woody, semi-succulent and thorny, with an average height of 2-3 m that is relatively impenetrable in a pristine condition (Acocks 1953, Mucina & Rutherford 2006). It spans a number of primary drainage basins that occur between the coast and the Great Escarpment. On the basis of a floristic assessment of the AST, Vlok *et al.* (2003) suggested that these drainage basins are discrete biogeographical units. Here this proposal is termed the 'evolutionarily discreet drainage basin hypothesis' (EDDB). Based on this hypothesis, drainage basins have been treated as unique entities in large-scale conservation planning that aims to ensure the persistence of evolutionary processes for the AST biota (Rouget *et al.* 2006). This hypothesis is supported by phylogeographic research on freshwater redfins (Swartz *et al.* 2009) and terrestrial cicadas (Price *et al.* 2010), but there is limited data on whether it applies to terrestrial plants. *Nymania capensis* is a widespread and abundant shrub species in the AST, and thus an ideal candidate to test this hypothesis. I test the EDDB hypothesis using genetic data from both the chloroplast and nuclear genomes, given the associated problems with single genome studies (reviewed in Avise 2000, Schaal *et al.* 1998). A number of predictions can be made if drainage basins and their associated watersheds are important landscape

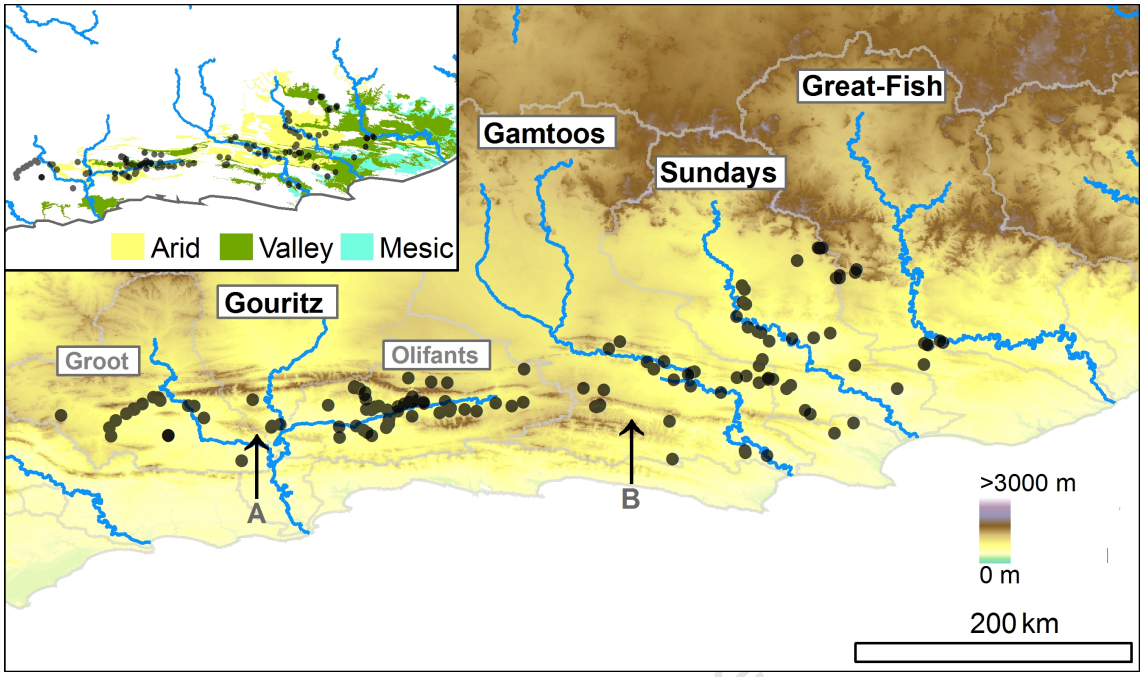

Figure 4.1. The distribution of *Nymania capensis* sampling localities along the coastal lowlands of the Albany Subtropical Thicket (AST). The three inland AST vegetation subtypes are shown in the inset. The watersheds separating drainage basins are shown in grey. Two landscape features mentioned in the text are highlighted: (A) the Rooiberg inselberg, and (B) the Baviaanskloof valley.

features responsible for structuring genetic diversity (e.g. Price *et al.* 2010):

1. Different drainage basins should contain genetically distinct lineages,
2. There should be limited genetic structuring within drainage basins and a high degree of genetic structuring between basins, and
3. If watersheds have been evolutionarily long term barriers to gene flow, then isolated drainage basin populations should contain evidence of genealogical sorting, which would eventually result in monophyly.

Climate change is also a major driving force that determines both the distribution and population dynamics of species (Comes & Kadereit 1998, Hewitt 2004, Jansson & Dynesius 2002). The effect of Pleistocene climatic changes on the distribution of plant species within the AST, and surrounding regions, remains largely unknown as there are insufficient reliable proxy records to provide a model of palaeoclimate or

palaeovegetation during glacial periods (Chase & Meadows 2007, Lewis 2008). Limited paleodata, however, suggests that the thicket was highly fragmented during glacial periods (Scholtz 1986), most likely due to lower temperatures and the frost sensitivity observed in many of the AST species. Thus, Cowling *et al.* (2005) suggest that the AST most likely retracted into refugia during glacial periods. Species distribution modelling (also known as ecological niche modelling) has been applied to construct habitat suitability models for taxa during the Last Glacial Maximum (Hugall *et al.* 2002, Richards *et al.* 2007). These models have proven to be a valuable approach to explore historical distributions, especially in the absence of extensive palaeontological data, as they often coincide with hypothesised refugia (Waltari *et al.* 2007). In order to test the glacial refugia hypothesis I follow this relatively novel approach of extrapolating species distribution models (SDM) onto statistically downscaled global climate model simulations. If AST species were adversely affected by glacial climate through the Pleistocene then the following would be predicted:

4. The timing of population subdivision between the catchments should coincide with the onset of glacial-interglacial cycles as these drainage basins are largely fixed and stable landscape features through this period (Cowling *et al.* 2009), and
5. The SDMs will demonstrate range contraction and fragmentation in *N. capensis*, consistent with phylogeographic patterns.

I use both molecular data and palaeoclimate distribution modelling of *N. capensis* to unravel the impacts of drainage basin topography and the Pleistocene climate oscillations on the genetic structure of *Nymania capensis*. The objectives are to test the EDDB and glacial refugia hypotheses. In order to do this I will: (i) assess the levels of genetic diversity and structure in *N. capensis* across its distribution in the AST, (ii) calculate the level of correlation between genetic and spatial distance between and within drainage basins to determine the level of isolation, (iii) determine if genealogical sorting has occurred within drainage basin populations, (iv) identify potential range changes and refugial areas during past climate scenarios, and (v) determine the timing of population splits using a molecular clock approach and a wide range of substitution rates.

4.3. Methods

4.3.1. Study system

The study system is outlined in Chapter 1. In brief, the Albany Subtropical Thicket is restricted to the year-round rainfall zone along the coastal lowlands of South Africa. A winter rainfall zone is found to the west and a summer rainfall zone to the east. The coastal lowlands form a series of short but deeply incised drainage basins separated from an unusually elevated interior plateau (Lithgow-Bertelloni & Silver 1998, Moore *et al.* 2009) by the Great Escarpment. The coastal lowland landscape has been topographically stable and relatively unchanged since the end of the Pliocene (∼2.6 Ma; Cowling *et al.* 2009).

In this study I investigate the role of primary drainage basins on genetic structuring. The distribution of *N. capensis* spans three drainage basins, specifically the Gouritz, Gamtoos and Sundays drainage basins. However, the Gouritz basin has added topographic complexity as two parallel mountain ranges associated with the Cape Fold Belt run across it creating an intermontane basin, known as the Little Karoo. The Rooiberg Mountain is an inselberg within the Little Karoo and splits this region along a west-east orientation (see Figure 4.1). This mountain range has been found to be a barrier to gene flow in another terrestrial plant species (A. J. Potts, unpublished data), so the secondary drainage basins west and east of the Rooiberg (Groot and Olifants, respectively) were sampled as extensively as primary drainage basins.

4.3.2. Study species

Nymania is a monotypic genus restricted to southern Africa. *Nymania capensis* is a large shrub to small tree with a maximum height of six metres, but usually no more than three metres (Figure 4.2.A). Solitary flowers are borne on leaf axils and are generally pollinated by insects (personal observations). The fruit is an inflated and deeply lobed capsule with papery thin membranes that contains numerous small seeds (Figure 4.2.C). The seeds are carried away form the parent plant within the light inflated capsules that are easily blown along the ground by wind. The inflated capsule is not known anywhere else in the Meliaceae (Pennington & Styles 1975).

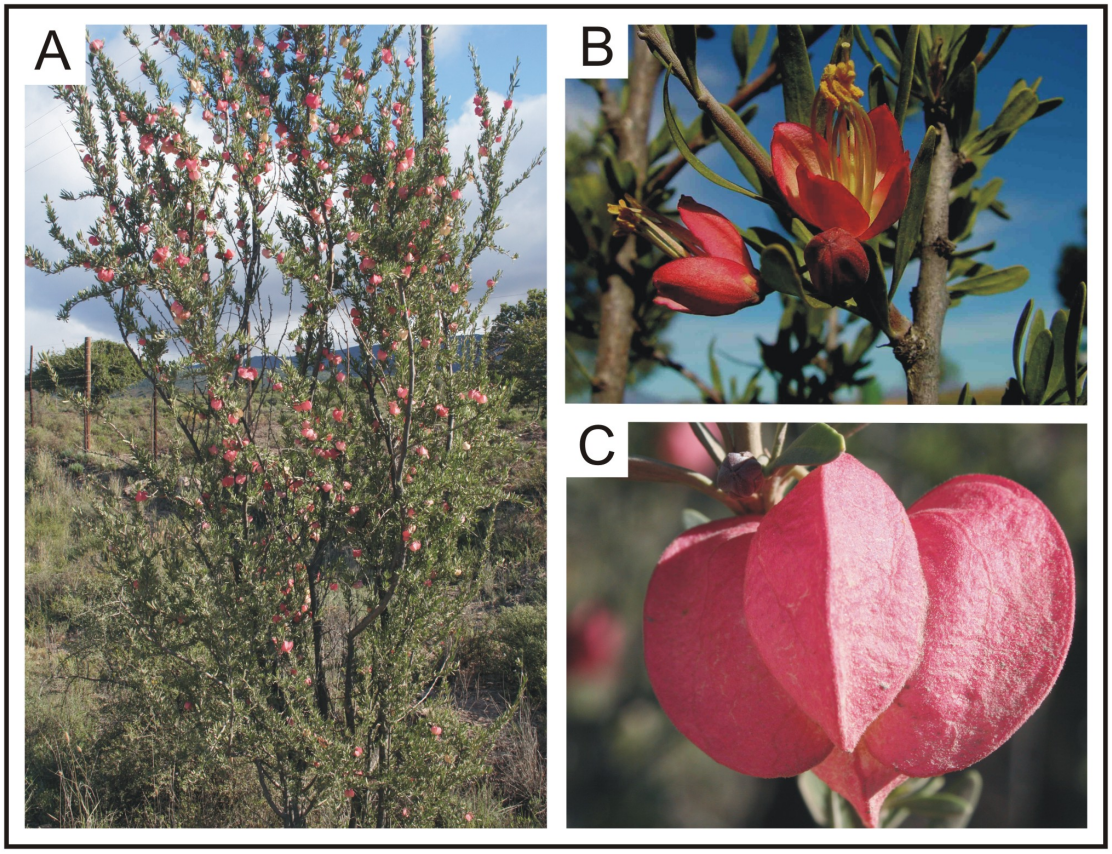

Figure 4.2. *Nymania capensis* (A) growth form, (B) flower, and (C) inflated fruit capsule. Photo credit: (B) J.H.J. Vlok

Nymania capensis is a dominant component in the arid and valley thicket subtypes of the Albany Subtropical Thicket (Vlok *et al.* 2003). This species occurs across a wide range of environmental conditions and is split into two disjunct distributions, one in the southern semi-arid region (~300–600 mm annual rainfall) of the AST and the other in the northern arid region (~100–300 mm annual rainfall) of South Africa where it also grows in Orange River Broken Veld (Acocks 1953) thicket-mosaic vegetation. Preliminary morphological (A.E. van Wyk, personal communication) and genetic data (this study) suggest that the northern distribution, which extends from the Nama Karoo into Namibia, and southern distribution restricted to the AST are different species. The northern and southern distributions of *N. capensis* are separated in distance by over 300 km as well as the Great Escarpment. This species was selected because (i) the southern distribution is characteristic of the AST biome, (ii) for its ease

of location and identification in the landscape, and (iii) the population demographics appear not to have been adversely affected by human activities such as livestock farming.

4.3.3. Sample collection and DNA extraction

Populations were identified at 133 different localities within the Albany Subtropical Thicket (Figure 4.1). Leaf material was obtained from 118 of these localities. The material for phylogenetic and phylogeographic analyses consisted of 78 individuals selected from a subset of these localities, with one individual per locality (Appendix Table A.1, Pg. 238). The sampling strategy focussed on maximising the number of sites sampled in order to explore the broad scale regional patterns rather than intra-population differences. This scattered sampling strategy is not affected by local and rapid coalescence events (Städler *et al.* 2009) and thus gives an unbiased view of population structuring and demographic history. Such sampling is also necessary when populations cannot be defined *a priori* (Harwood 2009), as is the case for *N. capensis*. Thus, each drainage basin was represented by between 14 to 26 individuals, with most individuals selected from localities that were over 10 km apart. Two samples of *N. capensis* from the northern distribution stored in the Bolus Herbarium were used as outgroup samples (BOL48535 and BOL60966).

Genomic DNA was extracted from silica-dried leaf material using a modified version of the method specified by Gawel & Jarret (1991) where reaction volumes were downscaled to 1.5 ml and polyvinylpyrrolidone-40 (PVP) was added when grinding the leaf material in liquid nitrogen using a mortar and pestle.

4.3.4. Chloroplast and nuclear sequencing

Chloroplast DNA sequence data from two separate regions, $trnQ\text{-}5' - rps16$ and $atpI - atpH$ were generated using PCR and direct sequencing. Nuclear sequence data were generated from two independent loci: 1) the 5.8S cistron and the flanking internal transcribed spacers, ITS-1 and ITS-2, hereafter referred to as ITS, and 2) the chloroplast-expressed glutamine synthetase gene, hereafter referred to as ncpGS. The ITS region is a high copy nuclear gene whereas ncpGS is a low copy nuclear gene. All primer sequences used to amplify each region are listed in Table 4.1. The ITS sequence

data were generated using PCR with high fidelity Taq (KAPA HiFi DNA polymerase, KapaBiosystems, Cape Town), to ensure accurate base calling, and direct sequencing. Chloroplast and ncpGS sequence data were generated using PCR with standard Taq (KAPATaq ReadyMix DNA polymerase, KapaBiosystems, Cape Town) and direct sequencing as the high fidelity Taq proved to be problematic for ncpGS amplification. The PCR conditions and protocols varied for each gene region. For the chloroplast and ncpGS regions, the PCRs were performed in volumes of 36 μl containing 1.2 μl of template DNA, 3.6 μl of 10X KAPATaq polymerase reaction buffer (Kapa Biosystems, Boston, Massachusetts, United States), 0.72 μl $MgCl_2$ (50mM), 1.2 μl of each primer (10 μM), 1.44 μl of dNTPs (10 mM), 0.24 μl Taq polymerase and sterile H_2O up to 36 μl. The chloroplast protocol consisted of an initial 2 minutes (min) denaturing step at 94°C; 28 cycles, each comprising 94°C for 1 min, 50°C for 30 seconds, 72°C for 1 min; and a final 6 min extension step at 72°C. The ncpGS protocol consisted of an initial 5 min denaturing step at 95°C; 30 cycles, each comprising 95°C for 1 min, 50°C for 1 min, 72°C for 2 min; and a final 7 min extension step at 72°C. The PCR conditions and protocols for the ITS regions using KAPAHiFi DNA polymerase are given in Chapter 3 (Pg. 70). All PCRs were performed on a GeneAmp 2700 PCR System (Applied Biosystems, USA), and were directly sequenced using both forward and reverse primers. All sequencing was performed by either Macrogen (Korea) or the University of Stellenbosch Sequencing Facility. Many nuclear PCR products generated sequences that exhibited polymorphic sites or length heterogeneity. A subset of these samples was cloned using the pGEM-T Easy Vector cloning kit (Promega, Madison, USA), and for each cloned PCR, five to eight colonies were sequenced to identify alleles. Cloning of ITS samples involved an extra A-tailing step as the high-fidelity Taq produced blunt-ended products. The A-tailing procedure followed the instructions in the cloning kit manual. Point mutations appearing in a single clone that did not correspond to a mutation detected in direct sequences nor observed in any other clones were not considered in the analysis since they are likely to be the result Taq polymerase errors (Eyre-Walker *et al.* 1998).

Table 4.1. The primers used for PCR amplification of chloroplast and nuclear regions of *Nymania capensis*

Genome	Region	Primer name	Primer Sequence	Reference
Chloroplast	$trnQ\text{-}5' - rps16$	$trnQ^{UUG}$	GCG TGG CCA AGY GGT AAG GC	(Shaw *et al.* 2007)
		$rpS16x1$	GTT GCT TTY TAC CAC ATC GTT T	
	$atpI - atpH$	$atpI$	TAT TTA CAA GYG GTA TTC AAG CT	
		$atpH$	CCA AYC CAG CAG CAA TAA C	
Nuclear	ITS1–5.8S–ITS2	$ITS4$	TCC TCC GCT TAT TGA TAT GC	(White *et al.* 1990)
		$ITS5m$	GGA AGG AGA AGT CGT AAC AAG G	(Sang *et al.* 1995)
	ncpGS	$GScp687f$	GAT GCT CAC TAC AAG GCT TG	(Emshwiller & Doyle 1999)
		$GScp994r$	AAT GTG CTC TTT GTG GCG AAG	

4.3.5. Sequence assembly, alignment and characterisation

Both chloroplast and nuclear DNA sequences generated in this study were assembled with CODONCODE ALIGNER version 3.7 (Codon Code Corp, http://www.codoncode.com) and automatically aligned using the softwares in-built alignment algorithm. The following steps were followed in order to identify polymorphic sites across and within sequences: (1) each base-call within every sequence was assigned a quality score using the automated base-calling program PHRED (Ewing *et al.* 1998), (2) sites that contained secondary peaks that were greater than 20% of the primary peaks were scored as polymorphic, and (3) all polymorphic sites were verified by eye. The boundaries of the ITS sequences and ribosomal coding regions were determined by comparison with an annotated sequence of *Nymania capensis* from Genbank (DQ861633). Numerous paralogous ribotypes were identified from cloned samples. Intra-individual site polymorphisms (2ISPs) were observed in both nuclear regions; these have been observed in many other plant species (Bailey *et al.* 2003), and are unsurprising given the complexity of the nuclear regions, especially ITS (reviewed in Chapter 3). The presence of 2ISPs indicates that there is more than one DNA variant present in the genome.

Summary diversity statistics were generated for each DNA marker and drainage basin using the PEGAS library version 0.3.4 (Paradis 2010) in R version 2.13 (R Development Core Team 2011); specifically, Nei's (Nei 1987) nucleotide diversity (π), haplotype diversity (h), number of variable sites and parsimony informative sites.

4.3.6. Phylogenetic networks and trees

The two chloroplast regions were merged for analyses as they are inherited in tandem. The phylogenetic relationships between chloroplast ncpGS haplotypes and all ITS ribotypes were performed using Statistical Parsimony (SP) or NeighbourNet (NN) networks and Neighbour Joining (NJ), Maximum Parsimony (MP), and Maximum Likelihood (ML) phylogenies. The ITS dataset was not reduced for analysis due to the high number of ribotypes in the dataset, thus all analyses were carried out at the level of individuals. The SP network was implemented in TCS version 1.21 (Clement *et al.* 2000) using uncorrected p-distances (cpDNA) or polymorphism p-distances (ncpGS). A NN network, implemented in SPLITSTREE version 4.8 (Huson & Bryant 2006) using

polymorphism p-distances, was constructed for the ITS dataset due to the high number of unique ribotypes and the complex relationships among them.

As both ITS and ncpGS datasets were rich in 2ISPs they were analysed using the 2ISP-informative approach for NJ, MP and ML (Chapter 3). Neighbour Joining was performed using uncorrected p-distances (cpDNA) or polymorphism p-distances (ITS and ncpGS) in the APE library version 2.7.1 (Paradis 2010) in R. The MP analyses were implemented in PAUP* version 4.0b10 (Swofford 2002) and involved a heuristic search strategy with 1000 replicates of random addition sequences, and the default branch swapping and character optimization options (TBR and ACCTRAN, respectively). RAxML version 7.2.6 was used to estimate trees and perform bootstrap analyses under ML (Stamatakis *et al.* 2008) using the GTR-Γ model of sequence evolution and multi-state characters. Branch support was evaluated using bootstrapping for both NJ and MP with 10,000 replicates; the MP bootstrapping followed the suggestions of Müller (2005), as each replicate comprised a single random sequence replicate and TBR branch swapping. Branch support for ML was evaluated using 1000 rapid bootstraps.

4.3.7. Isolation by distance and genetic variation

Pairwise genetic and geographic distances among sampling locations within and between drainage basins were used to test patterns of isolation by distance using a Mantel test (Mantel 1967). Genetic distance for cpDNA was calculated using uncorrected p-distances, whereas ITS and ncpGS genetic distances were calculated using polymorphism p-distances. Both genetic and geographic distances were calculated using the APE library in R, and the probability and significance of the correlation coefficient (Spearman's R) was estimated after 10,000 permutations.

4.3.8. Genealogical tests of population divergence

The level of genealogical divergence was assessed in the chloroplast and nuclear gene trees for lineages falling within the different drainage basins using the genealogical sorting index (*gsi*, Cummings *et al.* 2008). The *gsi* statistic is a standardised measure of the extent to which predefined groups in a gene tree exhibit exclusive ancestry; the *gsi* statistic ranges from 0 (a complete lack of genealogical divergence with respect to other groups) to 1 (monophyly). The significance of the *gsi* statistic is obtained

through randomisation of the group labels across the tips in a gene tree. Thus, lineages within groups can be tested against a null hypothesis of no divergence. The *gsi* was calculated on 100 MP trees sampled from the bootstrap using the same MP settings specified above, except full datasets (i.e. not haplotype datasets) for each DNA region were used. Maximum Parsimony analyses were used as this could be applied to all DNA regions, and the MP 2ISP-informative approach estimated trees with greater resolution and support for ITS than the equivalent ML approach. Thus, 100 trees were sampled from the 10,000 MP bootstrap replicates. These 100 individual *gsi* measurements were then used to calculate an ensemble *gsi* statistic *(gsi_T)* for each DNA region. The gsi_T statistic represents a summary of genealogical exclusivity that incorporates tree topology uncertainty. All analyses were performed using the GENEALOGICALSORTING library version 0.91 (available from http://www.genealogicalsorting.org/) implemented in R.

4.3.9. Molecular dating

Assigning a timescale to phylogenies typically involves using dated ingroup fossils or applying a rate of molecular evolution. No fossil or pollen data specific to *N. capensis* is available, so I test the hypothesis of Pleistocene lineage evolution using a highly conservative approach by using the 'extremes' of substitution rates found in the literature for non-coding chloroplast DNA, specifically 1.0×10^{-9} (Richardson *et al.* 2001) and 31×10^{-9} substitutions per site per year (Fu & Allaby 2010).

The dating of lineage divergence was carried out using BEAST version 1.4.8 (Bayesian Evolutionary Analysis Sampling Trees, Drummond & Rambaut 2007) which estimates the tree structure and the date of nodes simultaneously, using a Bayesian Markov Chain Monte Carlo (MCMC) procedure. This also provides a Bayesian Inference (BI) of the cpDNA tree – no BI methods are currently available that utilise a 2ISP-informative approach for nDNA. The BEAST analysis was performed using all cpDNA samples as this sampling is important for coalescent estimation of lineage divergence. The chloroplast dataset were analysed using the K3Puf substitution model, as selected using the Akaike Information Criterion in JMODELTEST version 0.1.1 (Posada 2008), and the Dollo model for indels. The hypothesis of rate constancy between samples was rejected using the relative rate test (Tajima 1993) as implemented within the APE library in R using an outgroup sample (BOL48535). Therefore, an

uncorrelated lognormal clock was used for all BEAST analyses. All other priors were left to the defaults. The BEAST analysis were replicated (N = 3) to verify convergence to stationarity and the estimation of effective sample size (ESS); this was confirmed by inspection of each analysis using TRACER version 1.4 (http://beast.bio.ed.ac.uk/Tracer). Each analysis was run for 10^7 generations, sampling every 10^4 steps. After discarding the first 2×10^6 samples as burnin, the parameter and tree estimates from the three runs were combined. Tree files from the separate runs were combined using LOGCOMBINER version 1.4.8 and TREEANNOTATOR version 1.6.1 (from the BEAST package) to yield a consensus maximum clade credibility using a posterior probability limit set to 0.5 and summarizing median node heights and the 95% higher posterior densities (HPD) of age estimates. These were visualized using the PHYLOCH library version 1.4.49 (available from http://www.christophheibl.de/Rpackages.html) in R.

4.3.10. Species distribution modelling

Distribution information on the southern thicket distribution of *N. capensis* was based on localities where leaf material was collected and georeferenced localities from herbarium material, resulting in a dataset of n = 133 (GPS: 118; georeferenced: 15). However, there was a definite sampling bias in this dataset, as most localities were identified near roads and/or concentrated in certain regions. This can mislead SDM algorithms (Phillips *et al.* 2006). I used the Clark-Evans nearest neighbour ratio (R) to define this bias along with the cumulative distribution function for edge corrections (Clark & Evans 1954); an R value significantly less than 1 indicates clumped samples, an R value equal to 1 indicates randomly dispersed samples and an R value significantly greater than one indicates evenly dispersed samples. The R value indicated that the total locality dataset was severely clumped ($R = 0.52, p \leq 0.01$). In order to remove this bias, samples were selected at random with a 10 km buffer, and any subsequently selected samples were removed if they fell within the buffer from a sample already selected. Five pseudo-replicate sub-sampled datasets were created in this manner, ranging from 55 to 59 localities, and used for the subsequent species distribution modelling. The R values of these subsets were all close to 1 ($R = 1.01$ to $1.05, p \geq 0.10$), indicating that the locality data were randomly dispersed.

The species distribution model (SDM) used in this study exclusively focus on areas of climatic suitability as the present and LGM distribution of *N. capensis* was

modelled only using environmental variables that capture biologically meaningful aspects of climate variation (e.g. Hugall *et al.* 2002). The SDM was based on the bioclimatic variables (which follow ANUCLIM; Australian National University: http://cres.anu.edu.au/outputs/anuclim.php) derived from the WorldClim data set (http://www.worldclim.org/, Hijmans *et al.* 2005) sampled at a resolution of 2.5 arc-minutes ($\sim 4 \times 4$ km resolution). These variables characterise the dimensions of climate considered pertinent in determining species distributions and represent summaries of annual trends for temperature and precipitation, aspects of seasonality, and extreme or potentially limiting environmental factors. The 19 bioclimatic variables for the LGM climate layers were derived from simulations from two global climate models (GCM): the Community Climate System Model (CCSM, Collins *et al.* 2004) and the Model for Interdisciplinary Research on Climate (MIROC, version 3.2, Hasumi & Emori 2004). These climate estimates were statistically downscaled by R. J. Hijmans using the WorldClim data set (Hijmans *et al.* 2005) and data provided by the Paleoclimate Modelling Intercomparison Project II (PMIP2) and made available through the WorldClim website (www.worldclim.org). Many of the bioclimatic variables are correlated in the study region (Figure 2.2, Pg. 34); to reduce this redundancy all but one variable found in a correlation cluster were removed leaving 11 variables (Table 2.1, Pg. 31).

The present and past SDMs for the species was generated using the maximum entropy machine learning algorithm MAXENT version 3.3.3c (Phillips *et al.* 2006), a method appropriate for presence-only data which performs well in comparison with other methods (Elith *et al.* 2006) and can be used for predicting past and future distributions (Hijmans & Graham 2006). The default settings in MAXENT were used as these have been optimised across a wide range of data sets and automatically select suitable regularization values and functions of environmental variables (Phillips & Dudík 2008). Model evaluation was performed using 5-fold (K-fold) cross-validation on each of the five occurrence subsets for each species. The K-fold evaluation method randomly partitions the data into K subsamples, each of which is used in turn as test data, while the remaining K-1 subsamples are used for training data. The model performance was evaluated using the standard statistical measure of predictive ability, the area under the receiver operating characteristic curve (AUC). The AUC statistic ranges from 0.5 (model prediction is no better than random) to 1.0 (perfect model prediction of presence versus absence). The models were calibrated using the survey

data and the past potential distributions were estimated by projecting the predicted present-day SDMs onto the LGM bioclimatic layers. MAXENT outputs a grid map with each cell having a continuous probability value of environmental suitability that ranges from 0 (not suitable) to 100 (most suitable). In order to compare and graphically represent the results between the five locality datasets, each of the folds was converted into binary present-absent values using the maximum test sensitivity (true positive rate) plus specificity (false positive rate) criterion; this criterion optimises the correct discrimination of presences and pseudoabsences in the test data and has performed well in comparison with other threshold criteria (Liu *et al.* 2005). These binary presence/absence layers were then summed and overlaid onto present and past niche models on the map of the study regions to examine visually if drainage basins maintained areas of suitable environmental conditions.

4.4. Results

4.4.1. Genetic data characteristics

The final chloroplast dataset is comprised of 1948 bp ($trnQ$-$5' - rps16$: 791 bp; $atpI - atpH$: 1157 bp). The ITS and ncpGS datasets were comprised of 666 bp and 1089 bp, respectively. All datasets aligned readily and reliable gaps corresponding to insertion or deletion events were included as informative characters; indels that were not associated with homopolymer repeats were coded as binary characters and used in all analyses (multiple base indels were treated as single characters). Sequence characteristics for each dataset are summarised in Table 4.2. Haplotype and nucleotide diversity were high in the ITS dataset, moderate in the chloroplast dataset, and low in the ncpGS dataset. A summary of variable sites across the chloroplast haplotypes, a subset of ITS samples and ncpGS haplotypes are shown in Tables 4.3, 4.4 and 4.5, respectively. Many 2ISPs were detected in both ITS and ncpGS sequences; 2ISPs that comprised an indel and a base were detected at five sites within the ITS dataset. Both ITS and ncpGS cloned samples demonstrated that these regions contained multiple copies of each region per sample. This rules out statistically inferring phased haplotypes from nuclear genes as these methods assume a maximum of two copies per individual (e.g. PHASE, Stephens *et al.* 2001) and cloning all samples was prohibitively expensive. Thus, the 2ISP-informative approach (Chapter 3) was used for all analyses of nuclear

data.

Table 4.2. Summary statistics, genealogical sorting indices (gsi_T) and mantel test (MT_R) results for *Nymania capensis* within and across basins. Summary statistics include nucleotide diversity (π), haplotype diversity (h), and number of segregating sites (s).

Genome	Drainage Basin	n	h	s	π	MT_R	gsi_T
cpDNA	Overall	78	20	28	0	0.4510 ***	
	Groot	14	4	7	0	-0.0517	0.290 (0.108) **
	Olifants	26	8	11	0	0.0638	0.508 (0.164) **
	Gamtoos	18	5	11	0	-0.0250	0.797 (0.276) **
	Sundays	18	7	8	0	0.1077	0.728 (0.183) **
	Gro+Oli	40	10	12	0	0.0885 *	0.840 (0.237) **
	Oli+Gam					0.3930 ***	0.372 (0.115) **
	Sun+Gam	38	11	16	0	0.3460 ***	0.859 (0.140) **
ITS	Overall	75	68	54	0.03	0.6234 ***	
	Groot	14	14	20	0.01	0.1889	0.250 (0.025) **
	Olifants	24	22	28	0.01	0.1078	0.474 (0.039) **
	Gamtoos	15	13	23	0.01	0.3433 *	0.996 (0.018) **
	Sundays	22	21	21	0.01	0.1619 *	0.994 (0.034) **
	Gro+Oli	38	34	29	0.01	0.0829 *	0.945 (0.007) **
	Oli+Gam					0.6448***	0.456 (0.083) **
	Sun+Gam	37	34	35	0.02	0.5622 ***	0.990 (0.082) **
ncpGS	Overall	69	13	11	0	0.5717 ***	
	Groot	13	3	4	0	-0.1778	0.251 (0.098) **
	Olifants	22	5	3	0	-0.1435 *	0.393 (0.132) **
	Gamtoos	14	5	5	0	0.2020	0.187 (0.060) **
	Sundays	20	2	1	0	-0.1236	0.546 (0.174) **
	Gro+Oli	35	7	7	0	0.0372	0.842 (0.284) **
	Oli+Gam					0.5454 ***	0.000 (0.000) **
	Sun+Gam	34	6	6	0	0.0875	0.848 (0.267) **

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 4.3. Variable sites across the chloroplast DNA haplotypes from two gene regions of *Nymania capensis* accessions. All sequences are compared to the reference haplotype A.

	trnQ$^{(UUG)}$*-5'rps16*													*atpI-atpH*																				
Haplotypes	71	126	246	293	422	461	517	542	693	723	752	174	176	327	362	469	501	518	528	616	623	624	642	692	907	960	974	999	1061	1101
A	A	A	C	A	A	A	T	A	A	A	1[a]	C	C	1[b]	C	T	C	A	T	A	C	T	A	T	C	A	A	C	G	G
B	.	.	.	.	.	.	.	.	.	A	.	.	C	.	.	.	C	.	.	.	.	.	.	.	.	.	.	T	.	.
C	.	C	.	.	.	.	.	.	.	A	.	.	C	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.
D	.	.	.	.	.	.	.	.	.	A	.	.	C	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.
E	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
F	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.
G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.
H	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	T	.	.	.	.	.	.	.	.	.
I	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.
J	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	G	.	.	.	.	.	.
K	.	.	.	.	.	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
L	.	.	T	.	.	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
M	.	.	.	.	G	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.
N	.	.	.	.	G	.	.	.	C	.	0	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.
O	.	.	.	.	G	.	.	.	C	.	0	T	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.
P	.	.	.	.	G	.	.	.	C	.	0	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	C	.	.	.
Q	C	.	.	.	G	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	A
R	.	.	.	.	.	.	.	.	.	.	0	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	T	.
S	.	.	.	G	.	.	.	.	.	.	0	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	T	.
T	.	.	.	.	.	.	G	G	.	.	0	.	.	.	G	.	.	.	.	C	.	C	.	.	.	.	.	.	T	.

Number '0/1' in the sequences indicate the absence/presence of length polymorphisms whereby the superscripts identify corresponding character states. Note that poly-T stretches were excluded from analyses.

a, TAAGA; b, AAT

Table 4.4. Variable sites in direct-PCR ribosomal ITS sequences from a subset of *Nymania capensis* accessions. Five samples were selected at random from each ITS cluster to demonstrate the variability within and between the clusters. All sequences are compared to the reference consensus sequence. Intra-individual site polymorphisms (2ISPs) are coded using IUPAC nomenclature. 2ISPs involving a base and an indel are coded using 'X'.

		ITS1																										5.8S	ITS2								
	Sample	25	51	54	61	63	64	72	74	75	79	83	86	89	92	98	114	119	130	133	145	175	207	213	225	235	252	260	436	472	474	515	519	520	521	527	528	537
Consensus		C	A	G	T	G	A	G	T	C	G	C	C	G	T	A	G	C	C	C	G	C	C	C	G	C	C	C	C	G	G	C	A	T	T	C	C	T
Cluster 1	AJP0074	M	W	·	W	.	M	R	·	·	R	·	·	K	·	M	K	·	Y	.	K	·	·	X	·	·	·	·	·	·	·	.	R	·	Y	.	·	Y
	AJP0075	M	W	·	A	K	·	A	·	·	R	·	·	K	·	M	K	·	Y	Y	K	·	·	X	·	·	·	·	·	·	·	M	R	·	Y	M	·	·
	AJP0077	.	T	·	A	·	·	A	·	·	A	·	·	T	·	C	T	·	T	·	K	·	·	X	·	·	·	·	·	·	·	.	·	·	·	·	·	·
	AJP0083	.	T	·	A	·	·	A	Y	·	A	·	·	T	·	C	T	·	T	·	·	·	·	.	·	·	·	·	·	·	·	M	·	·	·	·	·	·
	AJP0789	M	W	K	A	·	·	A	Y	·	A	·	·	T	·	C	K	·	Y	·	·	M	·	X	·	·	·	·	·	·	·	M	R	·	Y	·	·	·
Cluster2	AJP0269	H	·	·	·	·	M	·	·	·	·	·	X	·	·	·	·	·	·	·	·	·	Y	·	·	Y	·	.	·	·	·	·	R	W	Y	·	·	C
	AJP0270	A	·	·	·	·	C	·	·	·	·	·	.	·	·	·	·	·	·	·	·	·	.	·	·	.	·	M	·	·	·	·	R	W	.	·	·	C
	AJP0545	M	·	·	·	·	M	·	·	·	·	·	X	K	·	M	·	·	·	·	·	·	Y	·	R	Y	·	M	·	·	·	·	R	W	Y	·	·	Y
	AJP0551	M	·	·	·	·	M	·	·	·	·	·	X	·	·	M	·	·	·	·	·	·	Y	·	R	Y	·	·	·	·	·	·	R	W	Y	·	·	C
	AJP0586	A	·	·	·	·	C	·	·	Y	·	·	.	·	·	.	·	·	·	·	·	·	.	·	.	X	·	·	·	·	·	·	G	A	C	·	·	C
Cluster3	AJP0444	M	·	·	·	·	M	·	·	·	·	·	·	K	·	M	K	·	·	·	·	·	·	·	·	·	M	·	X	.	R	·	·	·	·	·	.	·
	AJP0466	·	·	·	·	·	·	·	·	·	·	·	·	K	·	M	K	·	·	·	·	·	·	·	·	·	M	·	X	R	R	·	·	·	·	·	Y	·
	AJP0490	·	·	·	·	·	·	·	·	·	·	·	·	.	Y	·	·	·	·	·	·	·	·	·	·	·	A	·	X	·	R	·	·	·	·	·	Y	·
	AJP0726	M	·	·	·	·	·	·	·	·	·	·	·	K	Y	·	·	·	·	·	·	·	·	·	·	·	A	·	X	·	R	·	·	·	·	·	Y	·
	AJP0729	M	·	·	·	·	·	·	·	·	·	Y	·	K	.	·	K	·	·	·	·	·	·	·	·	·	M	·	X	·	R	·	·	·	·	·	Y	·
No cluster	AJP0810	M	.	.	.	.	M	.	.	.	.	.	.	.	.	.	.	M	.	.	.	.	.	.	.	.	M	·	X	.	R	.	R	W	Y	·	·	Y

Table 4.5. Variable sites in direct-PCR ncpGS sequences from *Nymania capensis* accessions. All sequences are compared to the consensus sequence. Sequence polymorphisms follow IUPAC codes. Nucleotide positions refer to the aligned sequences in the dataset.

	Positions											
Haplotype	5	36	54	71	216	283	336	417	421	584	707	1013
Consensus	C	C	C	A	A	C	C	T	T	T	T	C
a	.	.	.	.	.	.	M	.	.	.	.	.
b	.	.	.	.	.	.	.	.	.	.	Y	.
c	.	.	.	.	.	.	.	.	.	.	C	.
d	.	Y	.	.	.	.	.	.	.	.	.	.
e	.	.	S	.	.	.	.	.	.	.	.	.
f	G	.	.	.	.	.	.	.	.	.	.	.
g	.	.	.	.	.	.	.	Y	.	.	.	M
h	.	.	.	.	.	.	.	Y	K	.	.	M
i	.	.	.	.	.	Y	.	C	.	.	.	M
j	.	.	.	M	.	Y	.	C	.	.	.	M
k	.	.	.	.	.	.	.	C	.	.	.	A
l	.	.	.	.	.	.	.	C	.	W	.	A
m	.	.	.	.	G	.	.	C	.	.	.	A

Pseudogenes may mislead phylogenetic and phylogeographic inferences, and this is a known problem in the high copy number ITS region (Feliner & Rosselló 2007). In order to determine if pseudogenes were present in ITS, I 1) compared sequences for length variation, 2) confirmed the presence of conserved angiosperm motifs (Harpke & Peterson 2008, Liu & Schardl 1994), and 3) compared the number of mutations from the relatively stable 5.8S region with those in the ITS regions. The ITS dataset was determined to be free of pseudogene sequences as length variation was minimal, all conserved motifs were present and the number of mutations in the ITS regions greatly outranked those found in the 5.8S region.

Two individuals found in the Gamtoos drainage basin had chloroplast and ITS haplotypes associated with the Sundays drainage basin lineage, however these were very close (≤5 km) to the watershed suggesting that the presence of these haplotypes beyond their drainage basin borders was due to recent migration rather than incomplete lineage sorting. Thus, these samples were grouped with the Sundays basin samples for all analyses.

4.4.2. Phylogeographic analyses

Different drainage basins should contain genetically distinct lineages (Prediction 1). The drainage basin association was strong with chloroplast haplotypes, with all but one haplotype restricted to one of the drainage basins or sub-basins (Figure 4.3). Haplotype F spanned both Groot and Olifants sub-basins, but was restricted to the Gouritz basin. A strong association was also evident between primary drainage basins and ITS clusters (Figure 4.4) with two exceptions (indicated with arrows, discussed further below). The rare ncpGS unphased haplotypes were restricted to drainage basins, but two haplotypes were widespread either in the Gouritz or across the Gamtoos and Sundays basins (*a* and *k*, respectively; Figure 4.4). Two AST samples contained anomalous ncpGS haplotypes (*g* and *h*); these haplotypes are found in an intermediate position between the western and eastern haplotype clades. These haplotypes may represent inherited ancestral copy diversity or recent gene flow between samples from the western and eastern clade. However, given the slow rate of mutation in this region, evident by the low genetic diversity, and that haplotype *g* is also shared with the outgroup sample from the northern distribution, it is likely that these haplotypes represent ancestral and unsorted ncpGS copies.

Samples AJP0537 and AJP0810 (indicated with black and grey arrows, respectively, in Figures 4.3, 4.4, 4.5) lie within the Gamtoos basin, but close to samples with a Sundays basin genetic signature found on the watershed boundary between the basins. These samples have Gamtoos chloroplast haplotypes (haplotype R), and for ncpGS, a rare Gamtoos haplotype (AJP0810, haplotype *j*) and a common and widespread haplotype (AJP05537, haplotype *k*). However, AJP0537 has an ITS signal that nests it within the Sundays ITS cluster 3, while AJP0810 displays a recombinant signal that lies between the Gamtoos cluster 2 and cluster 3 (Figure 4.4). The contrasting basin associations between these two samples suggest that they represent a contact zone between the Sundays and Gamtoos lineages, possibly caused by pollen flow between plants in this zone. As these samples are anomalous to the overall patterns of association, and are geographically restricted, they were removed for all subsequent analyses.

Isolation by distance using the Mantel Test (MT) was used to determine if there was genetic structuring within and between basins (Prediction 2). In general, non-significant MT_R values were observed within the basins or sub-basins (Table 4.2) across

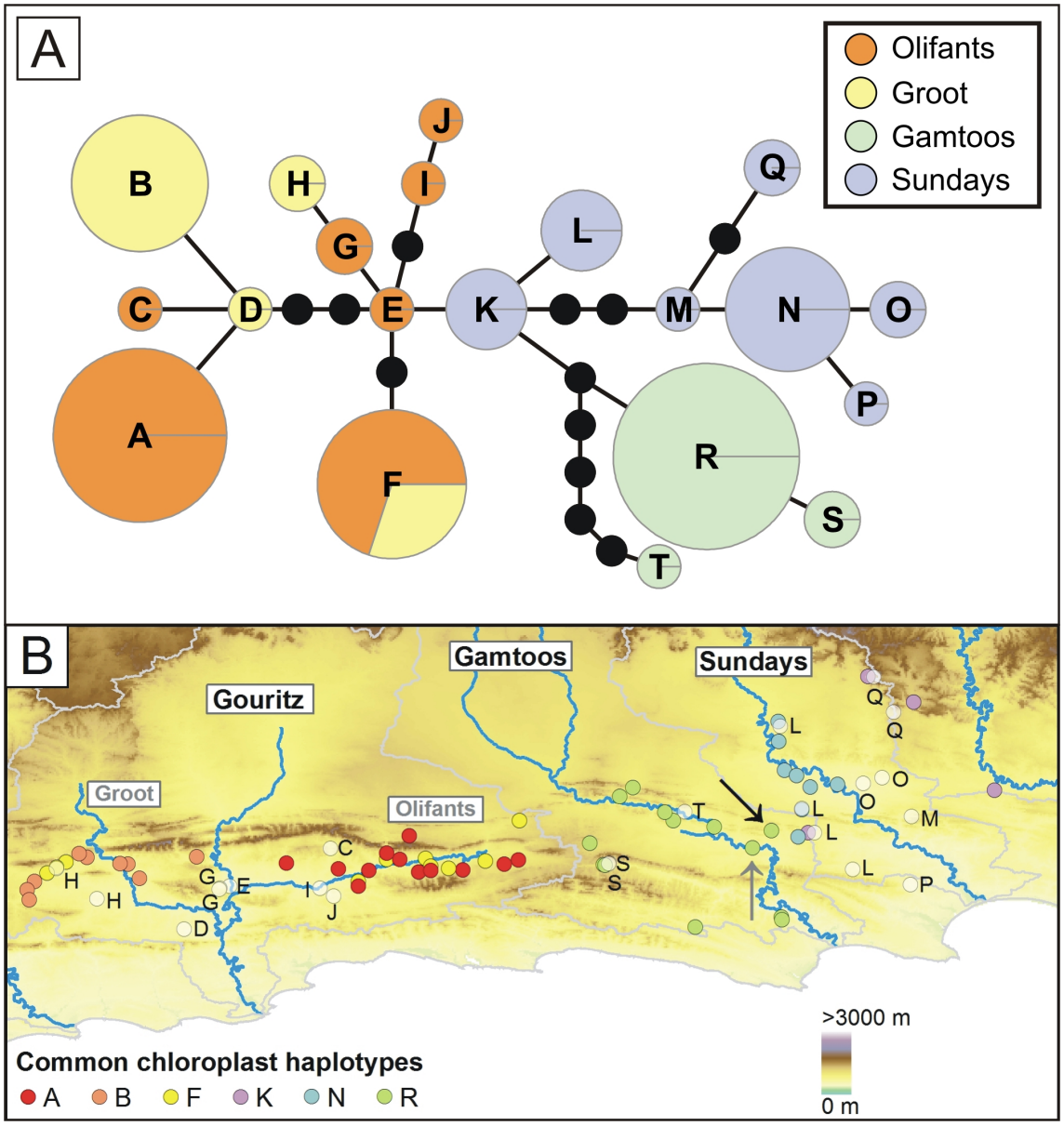

Figure 4.3. Chloroplast haplotype network and haplotype distribution of *Nymania capensis*. (A) The statistical parsimony haplotype network is based on the two combined chloroplast regions. Lineages are identified drainage basin are shown. (B) The distribution of lineages within drainage basins (watershed boundaries are shown in grey). Outgroup samples formed an unconnected network. Black and grey arrows represent samples AJP0537 and AJP0810, respectively, which are discussed in the text.

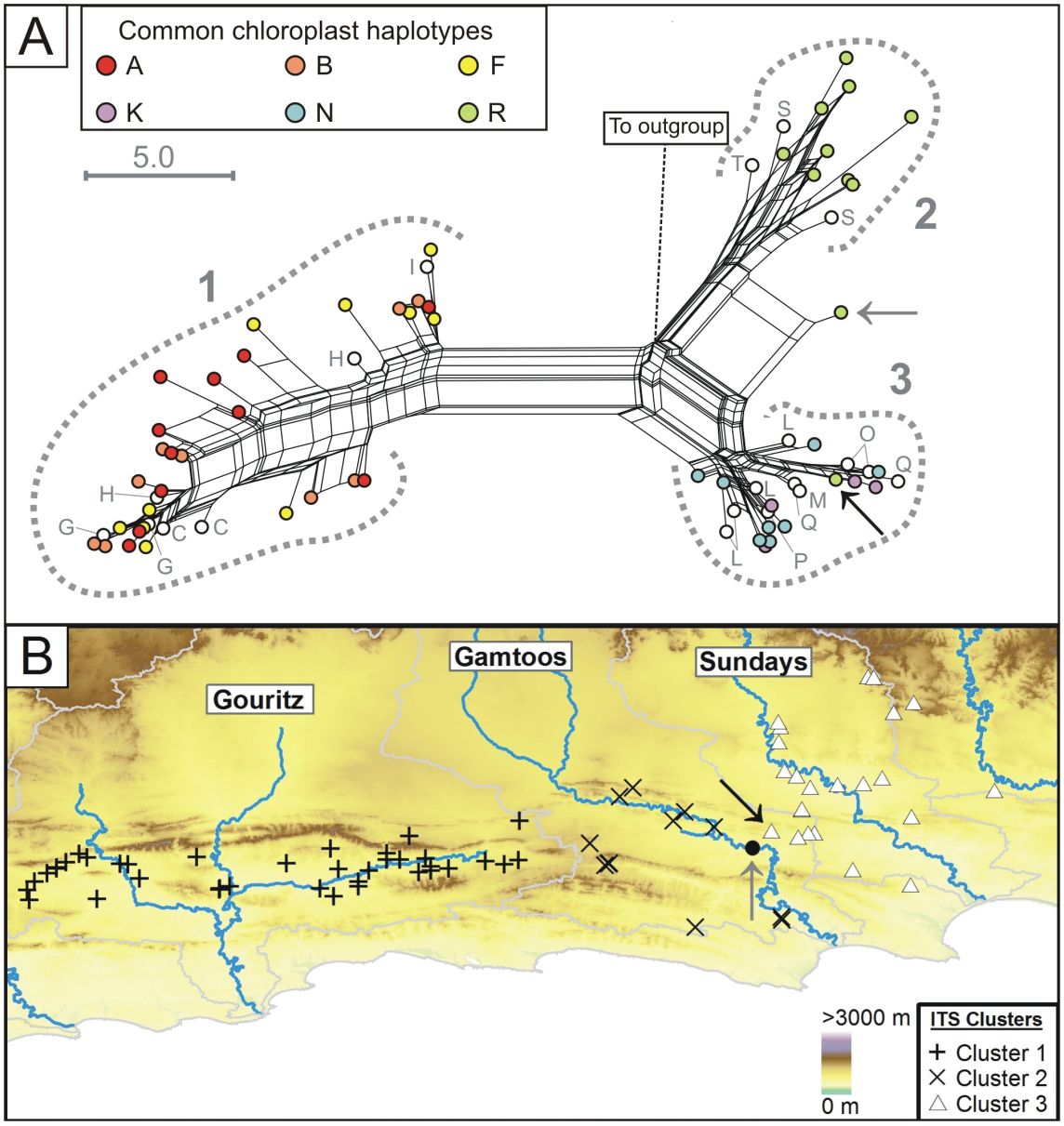

Figure 4.4. Network, clusters and cluster distribution of ITS sequences from *Nymania capensis*. (A) The NeighbourNet splits phylogenetic network is based on polymorphism p-distances. The outgroup samples are from the disjunct northern distribution (BOL48535 and BOL60966). (B) The distribution of clusters identified in the network across the major drainage basins (watershed boundaries are shown in grey). Black and grey arrows represent samples AJP0537 and AJP0810, respectively, which are discussed in the text.

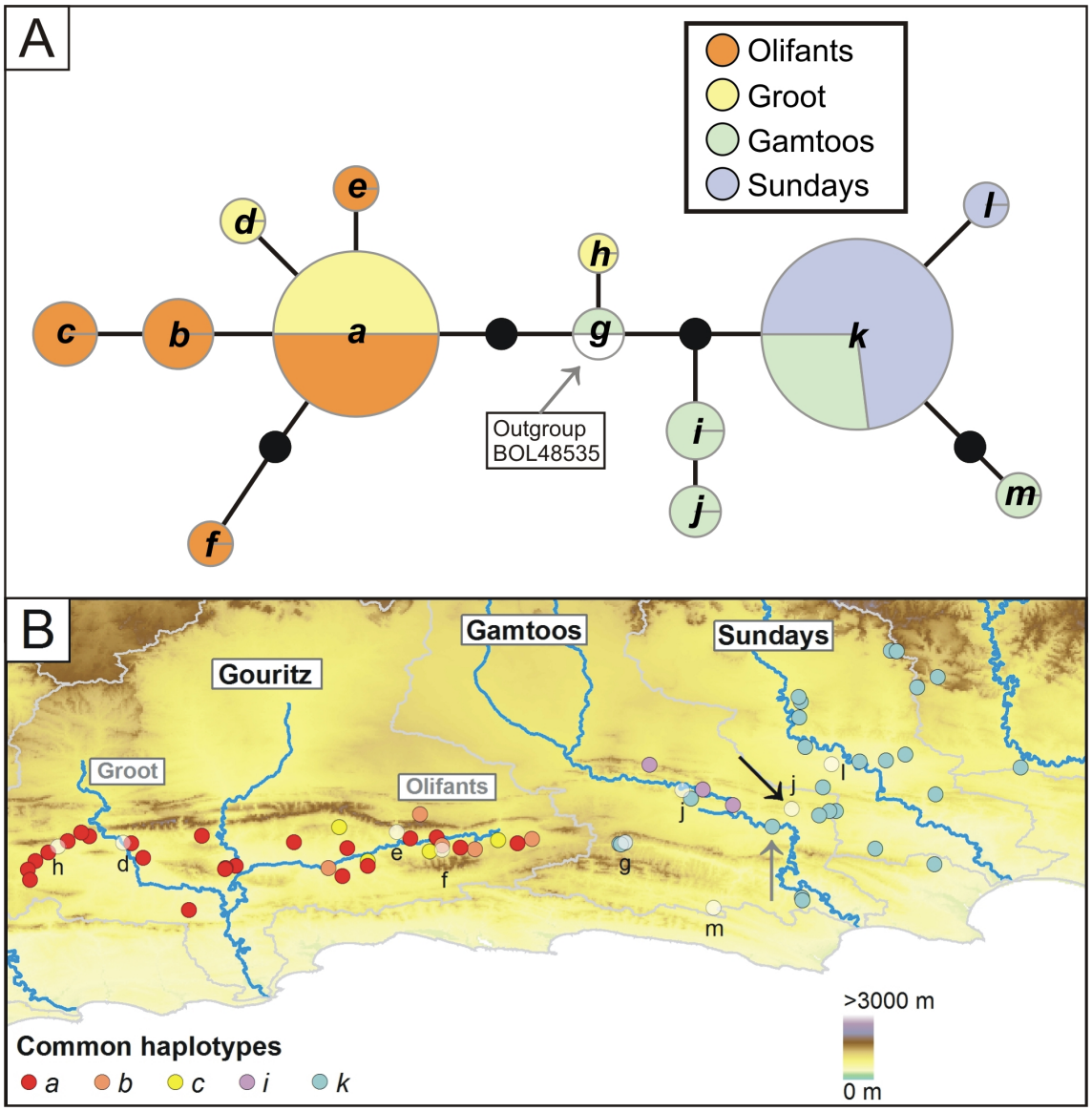

Figure 4.5. Network, clusters and cluster distribution of ncpGS haplotypes from *Nymania capensis*. (A) The NeighbourNet splits phylogenetic network is based on polymorphism *p*-distances. The outgroup sample is from the disjunct northern distribution (BOL48535). (B) The distribution of clusters identified in the network across major drainage basins (watershed boundaries are shown in grey). Black and grey arrows represent samples AJP0537 and AJP0810, respectively, which are discussed in the text.

the different DNA regions; however, there were a few instances where significant but low MT_R (ITS in Sundays; ncpGS in Olifants) and significant but high MT_R values (in comparison to the overall value; ITS in Gamtoos) were observed. There were significant and high MT_R values between the Olifants and Gamtoos basins across all DNA regions. Significant MT_R values between the Groot and Olifants and the Gamtoos and Sundays were only observed in the cpDNA and ITS; however, the MT_R values were low for the former pair of basins and high for the latter pair.

If watersheds have been evolutionarily long term barriers to gene flow then monophyly of lineages within basins is expected (Prediction 3). Only two monophyletic chloroplast lineages (containing haplotypes A-D and M-Q) were detected with moderate ($\geq$ 70%) to high ($\geq$ 90%) support consistently across the NJ, MP and ML trees (Figure 4.6); these lineages were restricted to the Gouritz and Sundays basins, respectively. No monophyletic lineages in the cpDNA were found that are restricted to a single basin or sub-basin that also contain all the haplotypes found in that basin. In contrast, the ITS phylogenetic trees find moderate to high support for Gouritz and Gamtoos lineages (Figure 4.7). The samples from the Sundays basin form an unsupported clade in NJ and MP, and a basal grade in the ML. The reason for lack of support for a Sundays clade is that these samples are the least differentiated and divergent group; they cluster in the NN splits graph (Figure 4.4.A), connected by bundles of short edges to the centre of the graph, which can be considered to represent the putative root or common ancestor. Thus, although these individuals are more closely related to one another than to either of the other two clusters, which would be indicative of a common origin, there is simply not enough signal in the data to support a Sundays clade.

For ncpGS phylogenetic trees, no monophyletic lineages are restricted to basins or sub-basins, but this is due to the two widespread ncpGS haplotypes. However, the divergence between the western Gouritz and eastern Gamtoos and Sundays basins is demonstrated as two monophyletic lineages are detected with moderate to high support (although one lineage has low support in the MP analyses) that are restricted to these basins. The AST samples with the anomalous ncpGS haplotypes *h* and *g* do not form part of either clade and are closely related to the outgroup sample.

If watersheds have been barriers to gene flow for a moderate period of evolutionary time (i.e. post fragmentation but not long enough to form monophyletic lineages) then

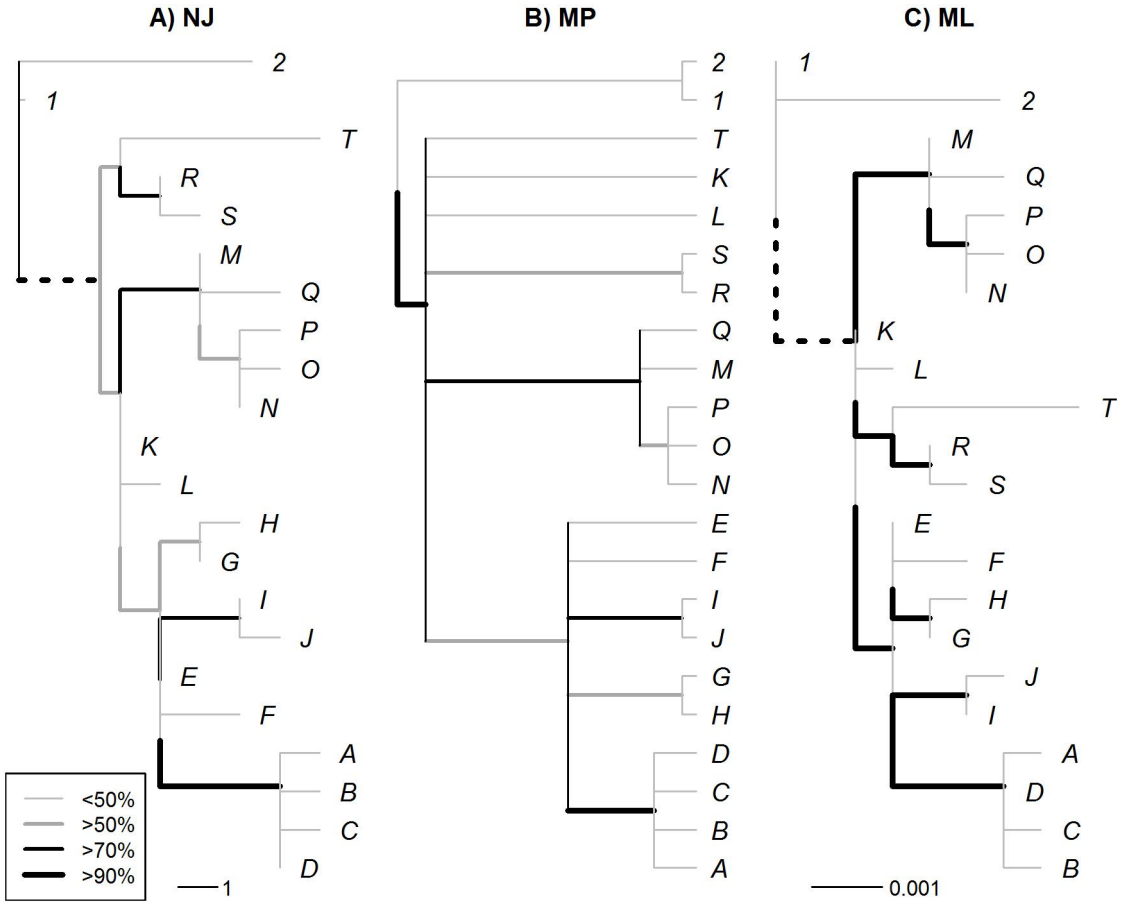

Figure 4.6. Phylogeny reconstructions of *Nymania capensis* chloroplast haplotypes. Methods used for reconstruction are (A) Neighbour-Joining, (B) Maximum Parsimony, and (C) Maximum Likelihood. Branch support is shown using colour and thickness of branches. Dashed branches have been reduced in length by a factor of 10. Haplotypes 1 and 2 are outgroup samples from Namibia (BOL48535 and BOL60966, respectively).

isolated basin populations should contain evidence of genealogical sorting (Prediction 3). The genealogical sorting index was significant for all sample groups explored (Table 4.2). However, the *gsi* values (indicating the degree of sorting) varied considerably between basins and basin groups. The Gouritz, Gamtoos and Sundays basins had very high *gsi* values (≥ 0.700) for both cpDNA and ITS trees. The ncpGS trees had high gsi values for the Gouritz basin, moderate (≥ 0.500) values for the Sundays basin, and low (< 0.500) values for the Gamtoos basin. The western and eastern paired basins (i.e. Groot and Olifants, Gamtoos and Sundays) had high *gsi* values across all DNA regions, whereas the centrally paired basins (Olifants and Gamtoos) had low

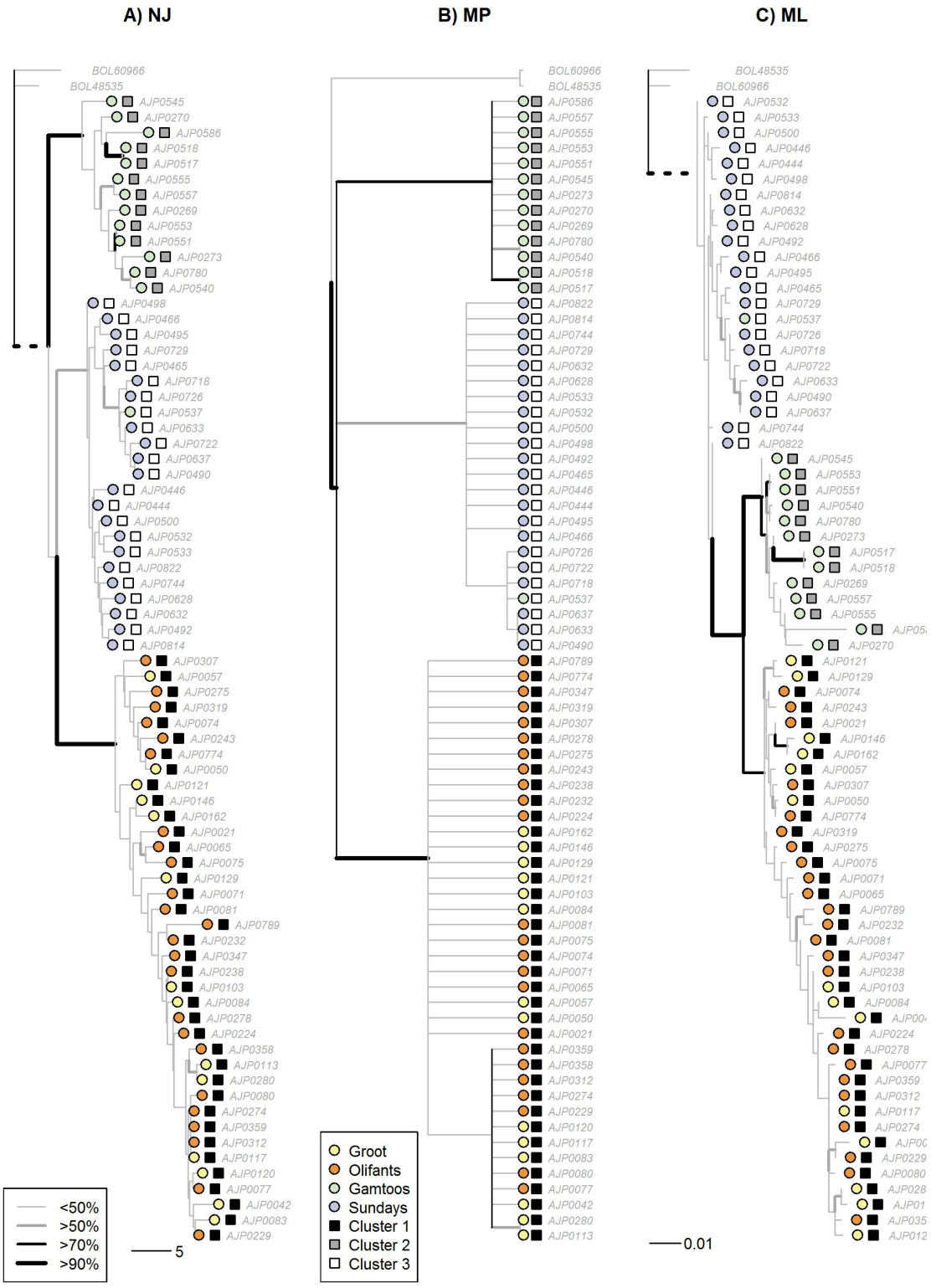

Figure 4.7. Phylogeny reconstructions of ITS *Nymania capensis* sequences. Methods used for reconstructions are Neighbour Joining (NJ), Maximum Parsimony (MP) and Maximum Likelihood (ML) with intra-individual site polymorphisms treated as informative characters. Branch support is shown using a combination of line thickness and colour. Dotted branches are reduced by a factor of 20.

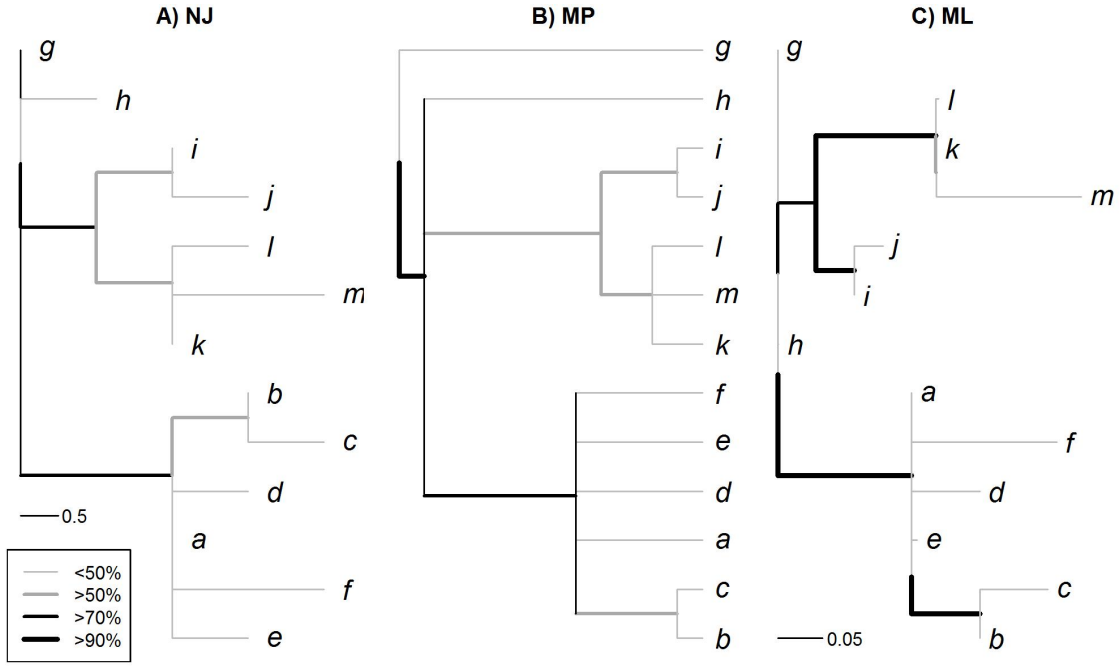

Figure 4.8. Phylogeny reconstructions of ncpGS *Nymania capensis* sequences. Methods used for reconstructions are Neighbour Joining (NJ), Maximum Parsimony (MP) and Maximum Likelihood (ML) with intra-individual site polymorphisms treated as informative characters. Branch support is shown using a combination of line thickness and colour.

gsi values. This reflects the deep divergence found between the western and eastern drainage basins.

4.4.3. Molecular clock and species distribution modelling analyses

A Beast analysis was used to date the diversification of chloroplast lineages in order to determine whether the Pleistocene climate cycles have also played a role in population isolation in drainage basin (Prediction 4). The timing of all of the AST lineages fall firmly within Quaternary whether a 'fast' or 'slow' rate of chloroplast mutation is used (Figure 4.9).

Species distribution modelling was used to further explore whether *N. capensis* experienced a range contraction and fragmentation during the Last Glacial Maximum (Prediction 5). The combined distribution models derived from the 25 locality subset datasets with thresholds for Maxent matched the fine-scale geographic mapping of

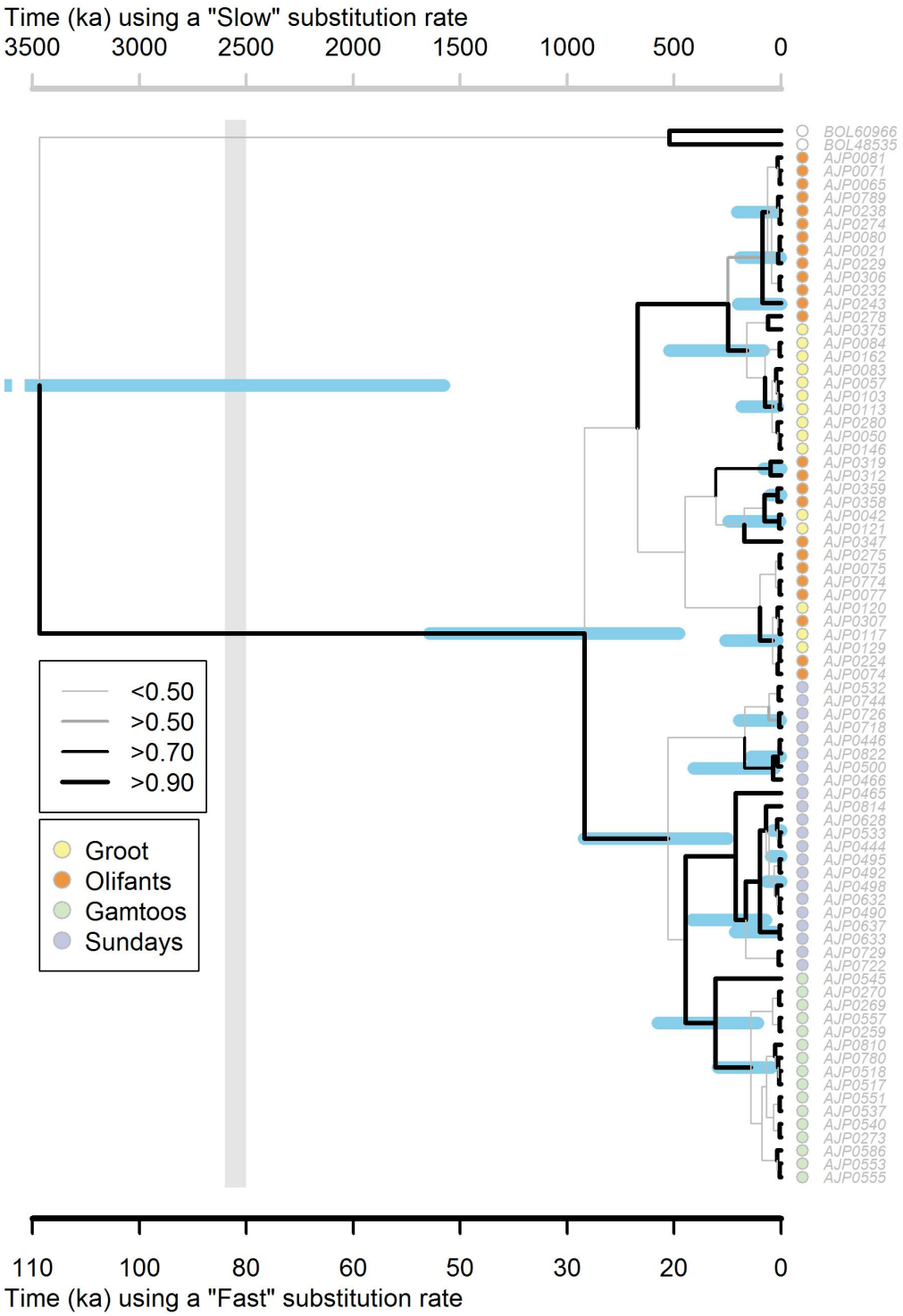

Figure 4.9. Molecular dating of *Nymania capensis* chloroplast sequences. The majority rule Bayesian chronogram generated in BEAST using the chloroplast data with the posterior probabilities of branches shown using a combination of branch width and colour. Nodes are centred on the mean TMRCA with blue shaded bars indicating the distribution of the 95% HPD for each estimate. The timing of divergence estimates are show using a 'slow' and 'fast' substitution rate (see text for details). The vertical grey bar indicates the Pliocene-Pleistocene boundary (~2.6 Ma) under the 'slow' substitution rate.

AST subtypes (arid and valley thicket) where *N. capensis* is a dominant component of the vegetation (Figure 4.10.A). The models also predicted suitable conditions outside of the mapped distribution of the thicket subtypes, primarily in the Gouritz basins. All MAXENT models were accurate in the target region, with AUC values higher than null expectations ($p \leq 0.01$, AUC$= 0.8877 \pm 0.0345$). The high AUC values and a projected distribution that coincides well with sampling localities and vegetation types suggest a strong fit between model and data. The projected models show a greatly reduced and fragmented distribution during the LGM for both global climate models (Figures 4.10.B and 4.10.C). Refugia are predicted by both GCMs in the Groot, Olifants and Sundays basins. A refugium in the Gamtoos basin is only predicted by the CCSM global climate model.

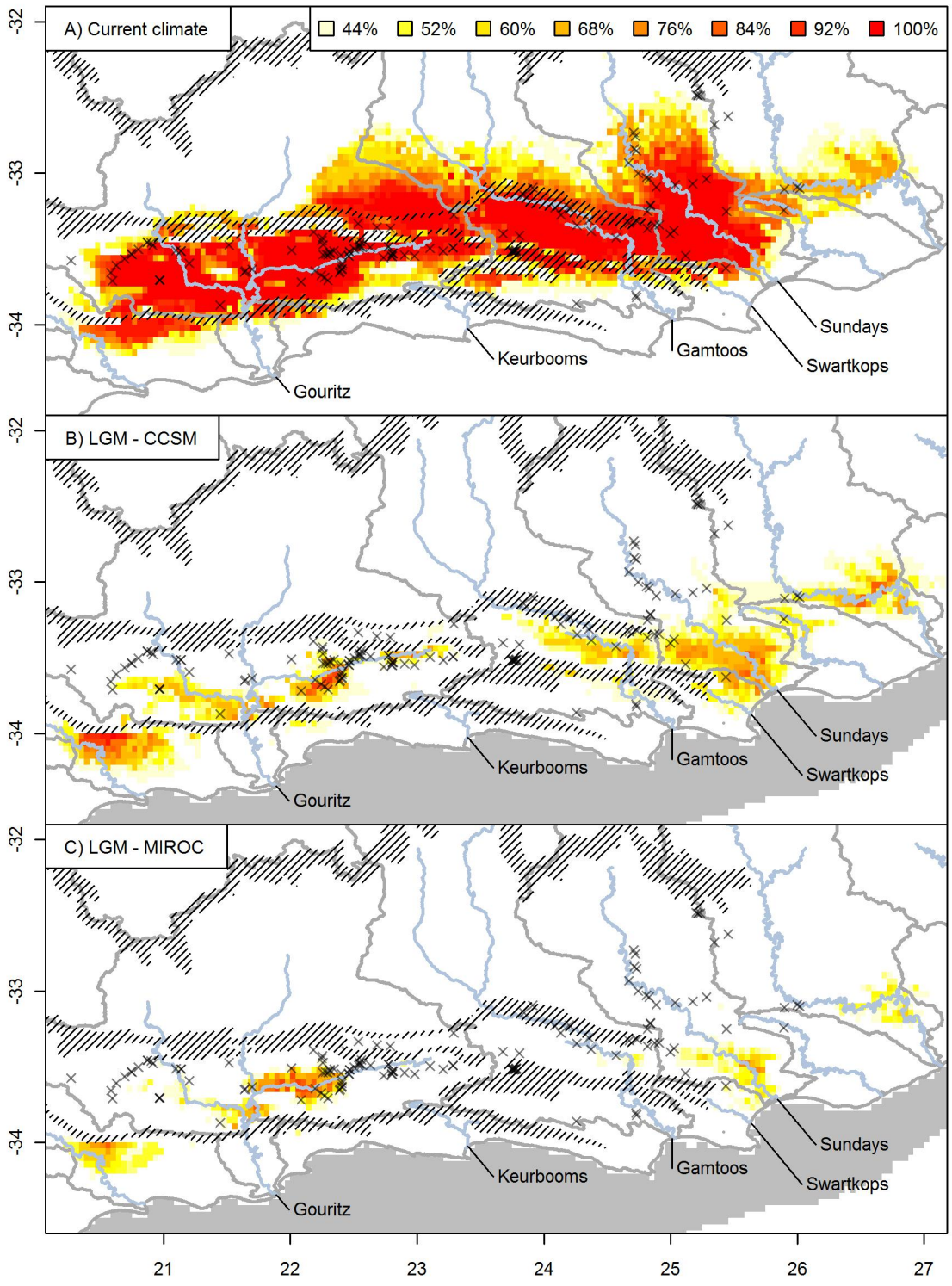

Figure 4.10. The modelled present and Last Glacial Maximum areas of suitable climate for *Nymania capensis*. The summary of modelled distributions from five pseudo-replicate datasets and k-folding subsets (see text) representing the climatically suitable areas in the Albany Thicket Biome under (A) present conditions, (B) the CCSM Last Glacial Maximum (LGM; 21,000 BP) simulation, and (C) the MIROC LGM simulation. Black crosses are localities used in the analysis. Hashed areas indicate major mountain ranges.

4.5. Discussion

I use an integrative approach that employs both genetic and geospatial data in order to test two hypotheses regarding the effects of topography and Pleistocene climate on the genetic diversity within the AST: the EDDB and glacial refugia hypotheses. From these hypotheses, a number of predictions were deduced and tested.

4.5.1. Evolutionarily discrete drainage basin hypothesis: watersheds and catchments as drivers of diversification

Investigating the role that drainage basin topography plays in reducing gene flow and driving population divergence has largely been restricted to either obligate freshwater species (e.g. redfins, Swartz *et al.* 2009) or freshwater species capable of terrestrial movement (e.g. freshwater crayfish, crabs, salamanders, and spotted frogs; Cook *et al.* 2008, Funk *et al.* 2005, Giordano *et al.* 2007, Ponniah & Hughes 2006). Only recently have drainage basins and watersheds been explored as drivers of diversification in terrestrial species that do not rely on the riparian system, and these have mostly focussed on invertebrates (e.g. springtails and cicadas; Garrick *et al.* 2004, 2007, Price *et al.* 2010). The EDDB hypothesis for the AST basins is supported by two studies, one on fish (Swartz *et al.* 2009) and the other on cicadas (Price *et al.* 2010).

The phylogeographic patterns of *N. capensis* are largely consistent with the three predictions deduced from the EDDB hypothesis. Firstly, the predominant pattern is one of genetically distinct lineages restricted to single drainage basins (Prediction 1; Figures 4.3, 4.4, and 4.5), although a few lineages span neighbouring drainage basins. The number of genetically distinct lineages varies between the different markers; this is expected given their different rates of mutation and effective population sizes. Secondly, there is limited genetic structuring within drainage basins as most comparisons did not detect a within-basin signal of isolation by distance. In contrast, a large and significant isolation by distance effect is evident across all drainage basins (Prediction 2; Table 4.2). This suggests that gene flow is hindered between basins, while this is not the case within basins. Lastly, although drainage basins do not contain exclusive monophyletic lineages, there is strong evidence of genealogical sorting within each basin (Prediction 3; Table 4.2).

The strong phylogeographic break between the western Gouritz and the eastern

Gamtoos and Sundays drainage basins detected in this study corresponds to previously documented breaks observed within *Meleuphorbia* (Euphorbiaceae, Ritz *et al.* 2003) and *Platypleura plumosa* (Hemiptera: Cicadidae, Price *et al.* 2010). The Rooiberg inselberg that separates the Groot and Olifants basin has also been observed as a phylogeographic break in *Platypleura karooensis* (Hemiptera: Cicadidae, Price *et al.* 2010) and *Berkheya cuneata* (Asteraceae; A.J. Potts, unpublished data). However, only the cpDNA supports this as a barrier. This may be due to restricted seed flow but continuous pollen flow over this inselberg - in most Angiosperms, the chloroplast genome is mostly maternally-inherited and thus dispersed in seeds, whereas the nuclear genome is usually biparentally-inherited and is dispersed by both seeds and pollen (Ennos 1994).

The results of this study further supports the EDDB hypothesis as the marked genetic structuring revealed in *N. capensis* is consistent with patterns predicted on the basis of drainage basin divisions. However, although the landscape has been stable since the Late Pliocene, dramatic cycling between glacial and inter-glacial climates has occurred during the Pleistocene. These shifts may have also affected the distribution, fragmentation and divergence of AST species such as *Nymania capensis* (Cowling *et al.* 2005, Dynesius & Jansson 2000).

4.5.2. Glacial refugia hypothesis: Pleistocene climatic cycles as drivers of diversification

Climatic changes during the Pleistocene glacial cycles have induced distributional shifts in species, often resulting in fragmentation and divergence of populations (Hewitt 2004, Jansson & Dynesius 2002). Identifying refugial areas during glacial periods through the Pleistocene has been a strong focus within many phylogeographic studies (Avise 2000), with the majority of studies focussed on the previously glaciated northern hemisphere regions (Abbott *et al.* 2000, Soltis *et al.* 1997, Taberlet *et al.* 1998). Southern Africa did not experience glaciation during the Pleistocene climate cycles (Partridge 1997) and determining refugial areas on plant species distributions in areas that did not experience such glaciation is a complex process (e.g. Byrne 2007). In the Cape Floristic Region, which neighbours the AST, climatic fluctuations during the Pleistocene have been suggested to be the main driver causing fragmentation and shifts in faunal species distributions contributing to allopatric diversification (e.g.

rock agama, dwarf chameleons, and cicadas; Price *et al.* 2007, Swart *et al.* 2009, Tolley *et al.* 2006). The lower temperatures both globally (Zachos *et al.* 2001) and regionally (Holmgren *et al.* 2003, Talma & Vogel 1992) coupled with the frost sensitivity observed in many of the thicket's component species is suggested to have driven AST vegetation into fragmented refugia (Cowling *et al.* 2005). This glacial refugia hypothesis is supported by the postdicted distribution of suitable climate for the AST vegetation subtypes during the LGM (Chapter 2); these results suggest that areas of possible refugia for AST vegetation lay within the primary drainage basins during the Last Glacial Maximum. The findings of this study further support this glacial refugia hypothesis in that the results are consistent with the two predictions derived from this hypothesis.

If the Pleistocene climate cycling between glacial and interglacial was responsible for isolating populations in a landscape that has been stable since the late Pliocene (∼2.6 Ma; Cowling *et al.* 2009), then it would be expected that lineage diversification would coincide or occur after the onset of these cycles (Prediction 4). There are many potential problems with the molecular clock approach (Graur & Martin 2004, Ho 2007), and selecting an accurate rate of substitution for target taxa is of primary concern when fossil calibration is not available. Using the widest range of published substitution rates for non-coding chloroplast DNA should circumvent the lack of an accurate rate for *Nymania capensis*. Using both slow and fast substitution rates, the divergence of all AST lineages falls well within the Pleistocene (Figure 4.9), suggesting that this species has experienced fragmentation and isolated diversification during this period. Under the fast substitution rate, many of the lineages diverge after the Last Glacial Maximum. This is an unlikely scenario given that this is an exceptionally fast rate that has been derived from a genus of annual herbs (*Linum*), and the fast turnover in generations is likely to have greatly increased the substitution rate (Kay *et al.* 2006). *Nymania capensis* is a perennial plant that requires more than five years of ideal environmental conditions before flowers and seeds are produced (Jan Vlok, personal communication), thus it most likely has a rate much slower than the fast substitution rate, which would push the timing of lineage diversification backwards into the Pleistocene. The slower substitution rate also falls within the generally accepted range of 1.0 to 3.0×10^{-9} (Wolfe *et al.* 1987).

Climatic changes will drive shifts in geographic distributions of species (Jansson & Dynesius 2002) if their ecological niches cannot change radically, at least over

moderate periods of time. Evidence suggests that niche conservatism may be general and pervasive across most species over moderate periods of time, despite profound changes in climate and environmental conditions (Martínez-Meyer & Peterson 2006, Martínez-Meyer *et al.* 2004, Peterson *et al.* 1999). Here I assume that the niche of *N. capensis* has been largely conserved from the LGM to the present.

The SDM results of *N. capensis* suggest this species' range fragmented and contracted into the primary drainage basins (Prediction 5; Figure 4.10). This is largely consistent with the phylogeographic evidence that suggests this species has been isolated into at least three refugia which correspond to the delimitation of primary basins (Figures 4.3, 4.4, and 4.5). An exception is that an area of suitable LGM climate is not postdicted for the Gamtoos under the MIROC3.2 global climate model. However, given the phylogeographic evidence of a population restricted to the Gamtoos, it is likely that the modelled LGM climate in this GCM may be more extreme than the actual conditions. The CCSM results are likely to be more reliable as this mirrors the phylogeographic evidence. Thus, the retraction into drainage basins during glacial periods through the Pleistocene would have strengthened the effects of watershed barriers to gene flow. This retraction is consistent with limited palaeodata (reviewed in Cowling *et al.* 2005) and community distribution modelling of AST subtypes (Chapter 2) that suggests that AST suffered significant range constrictions during the most recent Pleistocene glacial period. Also, the absence of vertebrates endemic to the AST, which would be expected given the present-day area of the biome, is suggestive of historical reduction and fragmentation of the biomes distribution (Mucina & Rutherford 2006).

4.6. Conclusions

The EDDB hypothesis has been used as the basis for conservation planning in the AST biome in order to conserve both the biodiversity patterns and evolutionary processes of this vegetation (Rouget *et al.* 2006). Specifically, a number of conservation corridors have been identified to create a mega-conservancy network; these corridors are predominantly focussed on the conserving major environmental gradients primarily within drainage basins. This study offers the first intra-specific validation of this hypothesis. Although it must still be shown that the pattern observed in *N. capensis* is representative of other AST species, this is positive validation of the EDDB hypothesis

and its use in conservation planning.

Both phylogeographic and niche modelling results suggest that the genetic structuring of *N. capensis* has been determined by landscape topology and Pleistocene climate. Populations have been restricted to drainage basins with no obvious present-day or historical gene flow during the previous glacial period. These results are largely consistent with the predictions if landscape topography and climatic fluctuations are responsible for structuring populations across the southern African lowlands. These findings validate the decision by conservation planners to identify drainage basins as discrete and significant units important for maintaining evolutionary processes.

5. A tale of two trees: contrasting evolutionary history and reproductive ecology of *Pappea capensis* (Sapindaceae) and *Schotia afra* (Fabaceae) in the Albany Subtropical Thicket

5.1. Abstract

The Albany Subtropical Thicket (AST) vegetation of southern Africa is suggested to have experienced severe range reductions and fragmentation during the glacial periods of the Pleistocene climatic cycling. Support for this hypothesis is accumulating through phylogeographic and distribution modelling studies. Here I use a multidisciplinary approach to assess if two dominant AST tree species, *Pappea capensis* and *Schotia afra*, were restricted to isolated refugia along the coastal lowlands during the Last Glacial Maximum (LGM). First, I evaluated the phylogeographic and population expansion patterns of both species using chloroplast and nuclear sequence data. Secondly, I used species distribution modelling (SDM) to predict present and LGM distributions using two different global climate models. The phylogeographic analyses indicate that the *P. capensis* chloroplast haplotypes are highly structured across drainage basins, which has some support in the nuclear data. In contrast, the chloroplast haplotypes of *S. afra* are widespread across the coastal lowlands and this pattern is further supported by the nuclear data. These results are consistent with SDM predictions that *P. capensis* was restricted to three isolated refugia distributed across the AST during the LGM, whereas *S. afra* was isolated to a single large refugium in the eastern Albany Subtropical Thicket. I suggest that these contrasting phylogeographic patterns can be attributed to differences in reproductive ecology responsible for post-glacial expansion. *Pappea capensis* has small red fruit that are bird-dispersed, whereas *S. afra* has large Fabaceae pods that are eaten by a range of large mammals including two mega-herbivores (African Savannah Elephants and Rhino Rhinos). Long distance dispersal by birds is rare due to territoriality and short

gut retention times of seeds. I suggest that migrating mega-herbivores with poor digestion and large intestinal tracts that retain seeds for long periods are responsible for the dramatic post-glacial colonisation and lack of phylogeographic structure of *S. afra* across the Albany Subtropical Thicket.

5.2. Introduction

Investigating the phylogeographic patterns of a taxon can reveal cryptic biodiversity, which is otherwise overlooked by traditional taxonomy (Beheregaray & Caccone 2007). This cryptic divergence can reveal the evolutionary history and ecological processes responsible for the generation and maintenance of biodiversity which may provide generalities that extend to unstudied codistributed taxa (Avise 2000, Garrick *et al.* 2007, Sunnucks *et al.* 2006). Both cryptic divergence and the evolutionary history of communities can be explored using comparative phylogeography (e.g. Garrick *et al.* 2004, 2007, 2008, Sunnucks *et al.* 2006). The Albany Subtropical Thicket (hereafter abbreviated to AST) vegetation of southern Africa has only recently been distinguished as a biome based on its unique floristic, growth form and ecological characteristics (Low & Rebelo 1996, Vlok *et al.* 2003), and there is limited information regarding the evolutionary history and processes of this biome (Cowling *et al.* 2005). Here I use phylogeography of two tree species from the AST in order explore the cryptic divergence, evolutionary history and ecological processes operating within this biome.

In contrast to phylogeographic studies of animals, studies targeting plants are routinely hampered either by a lack of appropriate molecular variation, or the complexity of the nuclear genome (Chapter 1; Álvarez & Wendel 2003, Feliner & Rosselló 2007, Schaal *et al.* 1998). This has meant that plant phylogeography has lagged behind animal phylogeography (Avise 2000, Beheregaray 2008) and that animal taxa have often been used as proxies to infer the evolutionary history of vegetation (Garrick *et al.* 2004, Hugall *et al.* 2002, Moussalli *et al.* 2009). However, plants can have highly divergent life history strategies and reproductive ecologies. This complexity is unlikely to be captured by faunal proxies. Plant communities offer very interesting terrestrial models to explore the evolutionary history of biogeographical vegetation units, such as biomes, with taxa that have different life history or reproductive strategies. Comparative phylogeography of plants is now more feasible due to the availability of chloroplast regions with greater variability (Shaw *et al.* 2007) and new

methods to analyse complex nuclear multigene families (Chapter 3).

In this chapter I examine two co-distributed tree taxa from the Albany Subtropical Thicket: *Pappea capensis* (Sapindaceae) and *Schotia afra* (Fabaceae). Both of these species are from genera considered to have ancient origins as *Pappea* is a monotypic genus and the genus *Schotia*, which is restricted to southern Africa, diverges basally within the Fabaceae (Cowling *et al.* 2005, Schrire *et al.* 2005). These species have very similar life history strategies with similar morphologies as low-growing trees usually no taller than five meters (Figures 5.1 and 5.2) and a high ratio of woody conductive tissue to leaf tissue with accompanying slow growth, likely as an adaptation to periodic drought (Holmes & Cowling 1993). Within the AST, these two species are dominant components of the valley and arid thicket subtypes (Vlok *et al.* 2003) and are often found growing in close proximity to one another. However, *P. capensis* and *S. afra* have contrasting zoochorous seed dispersal syndromes. *Pappea capensis* has small fruit (1.0–1.5 cm) with a furry green capsule and a bright red aril surrounding the seeds that is exposed when ripe (Figures 5.1.B). The fruit is primarily dispersed by birds (Sigwela 2004), but may also be dispersed by a range of small to medium mammals such as Vervet Monkeys, Chacma Baboons and antelope (van Wyk 1972). *Schotia afra* has a fairly typical, but large, woody Fabaceae pod (5 - 12 cm x ∼3 cm) that contains a number of large roundish seeds (Figures 5.2.C). The pods are eaten by a range of large mammal herbivores. The majority of woody AST species are thought to be adapted for endozoochorous seed dispersal by mammals, birds, or both (Castley *et al.* 2001, Cowling 1983, Cowling *et al.* 1997, Sigwela 2004, Watson 2002). Mammals and birds are likely to have differing cues and requirements to elicit seed dispersal, and AST species may have adapted their seed morphology to mainly target of one of these guilds as dispersal agents (Howe & Smallwood 1982), although long-distance dispersal may still occur via non-standard dispersal vectors (Higgins *et al.* 2003). The species selected for this study likely lie closely to the two extremes along the bird-dispersed and mammal-dispersed continuum, with *P. capensis* primarily bird-dispersed (but also small mammals) and *S. afra* primarily dispersed by two mega-herbivores (African Savannah Elephants and Black Rhinos) and other large mammals (e.g. Greater Kudu).

These two species are used to explore the patterns of cryptic divergence and provide insights into the evolutionary history of the AST biome. Details of the AST as a research model are given in Chapter 1 (Pg. 1.3). Briefly, the coastal lowlands are a series of deeply incised drainage basins separated from the interior plateau

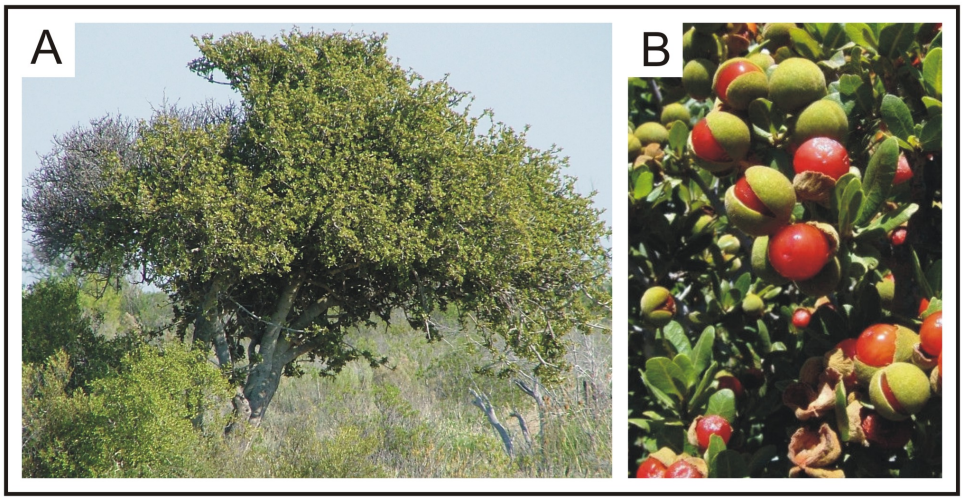

Figure 5.1. ***Pappea capensis*** **(A) growth form and (B) fruit comprised of a single seed surrounded by a fleshy red aril.** Photo credits: (A) W. Bass, (B) J.H.J. Vlok.

by the horse-shoe shaped Great Escarpment, and this topography has been fairly stable since the late Pliocene (Cowling *et al.* 2009). The AST is found along the southern coastal lowlands and is largely restricted to a zone of all-year rainfall, with a winter-rainfall zone in the west and a summer-rainfall zone in the east (Carr *et al.* 2006). It is a fire- and frost-intolerant vegetation that is bounded by the fire-driven biomes, fynbos to the west and savannah to the east, and the frost-region of the Great Escarpment and interior plateau to the north. The AST is considered an ancient and once-widespread assemblage that has become restricted due to the evolution of fire-driven biomes (Cowling *et al.* 2005). The Pleistocene fluctuations between glacial and interglacial periods are suggested to have had a dramatic impact on the distribution of the AST (Cowling *et al.* 2005), with the AST retreating into valleys during glacial periods resulting in a highly reduced and fragmented distribution (the glacial refugia hypothesis). Both the watersheds separating lowland drainage basins and the climatic oscillations through the Pleistocene have left signatures in lowland taxa such as redfins (Swartz *et al.* 2009), cicadas (Price *et al.* 2010) and a small tree species with wind-dispersed seeds (*Nymania capensis*, Chapter 4). Significant genetic structuring across basins supports the evolutionarily distinct drainage basin (EDDB) hypothesis; this hypothesis suggests that the watersheds act as barriers to gene flow for AST species, especially during glacial periods when the AST was restricted to isolated refugia in

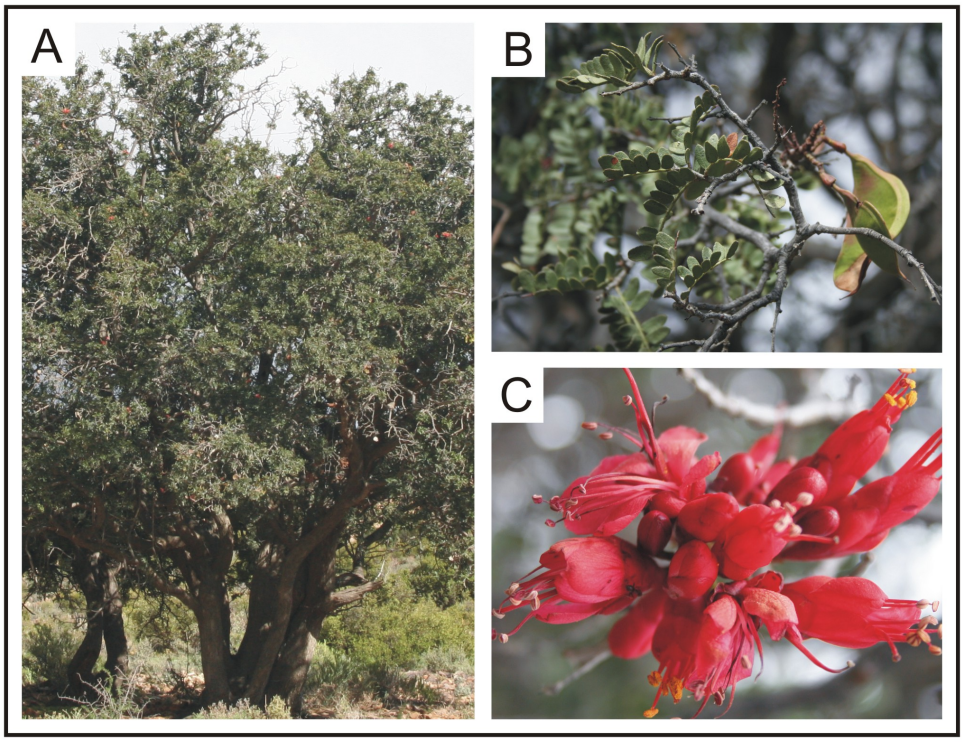

Figure 5.2. ***Schotia afra*** **(A) growth form, (B) leaves and pods, and (C) flowers.** Photo credits: (A,B,C) R. Potts.

basin valleys.

In order to investigate the roles of seed-dispersal ecology, the complex regional topography and Pleistocene climatic fluctuations, I have used a multidisciplinary approach. First, I examine the history of sequence data from both the chloroplast and nuclear genomes of *P. capensis* and *S. afra*, two tree species in the AST. Secondly, I use species distribution modelling (SDM) and its historical extension (palaeodistribution modelling) to explore potential range changes during Pleistocene glacial periods by focussing on the Last Glacial Maximum (LGM). This combination of techniques has proved to be complementary in generating and testing alternative biogeographic hypotheses (Hugall *et al.* 2002, Richards *et al.* 2007, Vega *et al.* 2010).

5.3. Methods

5.3.1. Sampling collection

In total, 75 individuals of *P. capensis* and 74 of *S. afra* from the AST were used for the phylogeographic analyses of chloroplast and nuclear sequences. Both *P. capensis* and *S. afra* have distributions that extend beyond the AST (Figures 5.3.B and 5.4.B). *Pappea capensis* and *S. afra* have a high density on a regional scale within the AST, but beyond this these species have very low regional densities although they may be locally dominant (Cowling *et al.* 2005). In order to sample the variation beyond the AST, samples from 39 individuals of *P. capensis* were included in order to place the AST samples relative to the rest of the species' distribution. *Schotia afra* belongs to a small genus, comprised of four species, that is restricted to Southern Africa. In order to place the *S. afra* AST samples relative to the rest of the species distribution and in context of the other species, 1 individual from the northern disjunct population of *S. afra* was included, as well as 21 individuals from the three other species (13 *S. latifolia*, 7 *S. brachypetala*, and 1 *S. capitata*). The majority of samples were obtained from fieldwork, but 10 samples of *P. capensis* were from herbarium collections (Pretoria National Herbarium) and 28 *Schotia* samples were from a previous study on the genus by Ramdhani *et al.* (2010). My sampling strategy focussed on maximising the number of sites sampled in order to explore the broad scale regional patterns rather than intra-population differences. This scattered sampling strategy is not affected by local and rapid coalescence events (Städler *et al.* 2009) and thus gives an unbiased view of population structuring and demographic history. Field identification was unambiguous as both *P. capensis* and *S. afra* have distinctive leaf and growth morphology. Species-level identification within *Schotia* can be hampered by a lack of flowering material; nonetheless, *S. afra* was easily identified even in its vegetative state.

5.3.2. DNA extraction, chloroplast and nuclear sequencing

Total DNAs were extracted from silica-dried leaf samples collected in the field using, with minor modifications, the protocol of Gawel & Jarret (1991). Polyvinylpyrrolidone-40 (PVP) was added when grinding the leaf material in liquid nitrogen using a mortar and pestle.

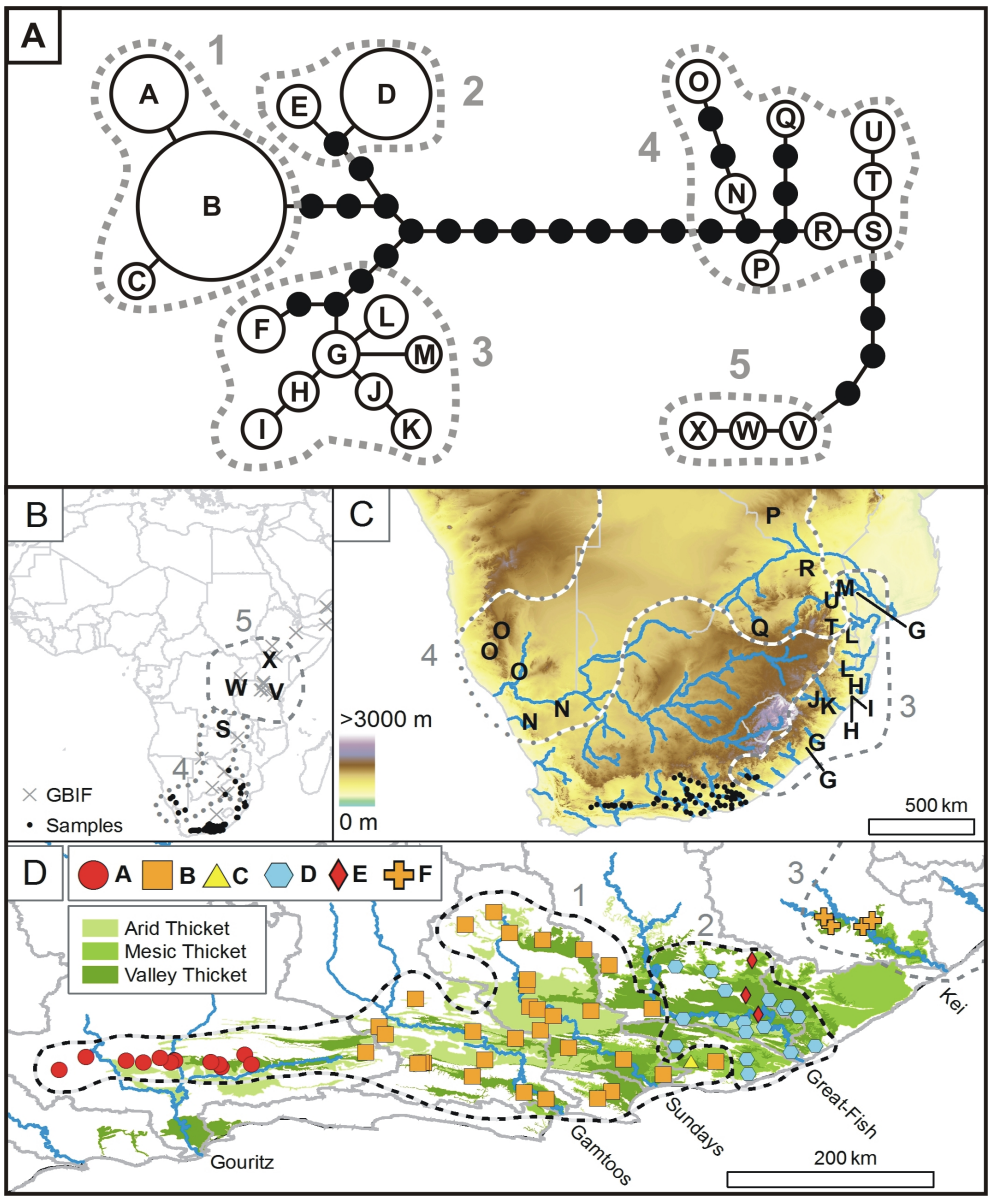

Figure 5.3. Chloroplast phylogeography and distribution of *Pappea capensis*. (A) The statistical parsimony network of 24 haplotypes found in *P. capensis* as defined on the basis of chloroplast *trnL–trnF* and *trnQ-5′–rps16* sequences. Lines represent single mutation steps; black, small circles represent unsampled or extinct haplotypes. Dotted lines encompass major chloroplast lineages and these are shown in the subsequent distribution maps. (B) The distribution of *P. capensis* and four haplotypes found in central Africa (S,V-X). (C) The distribution of haplotypes in southern Africa which occur beyond the Albany Subtropical Thicket. (D) The distribution of haplotypes within the Albany Subtropical Thicket

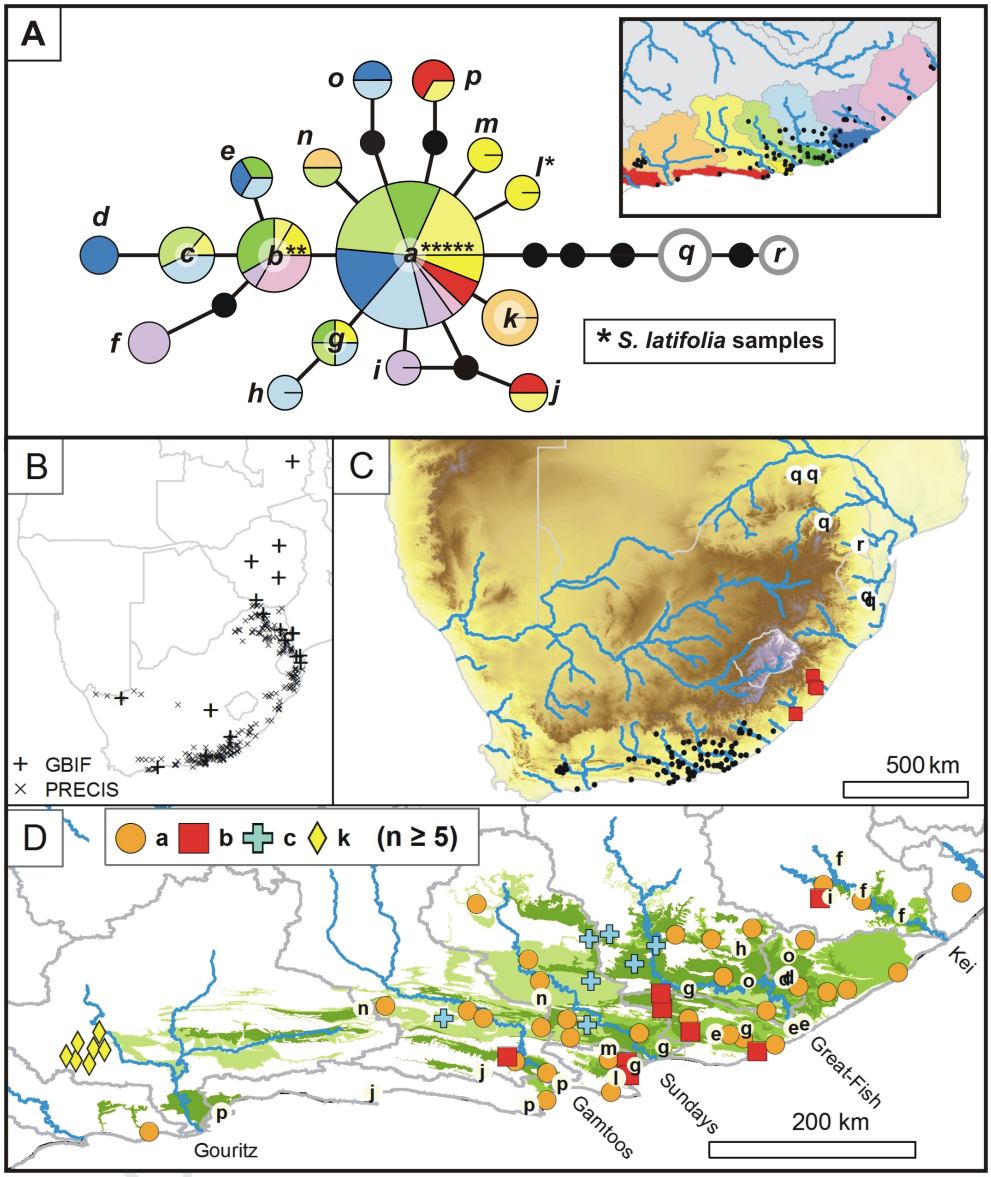

Figure 5.4. Chloroplast phylogeography and distribution of *Schotia afra* and closely related species. (A) The statistical parsimony network of 18 haplotypes found in *S. afra* and related species as defined on the basis of chloroplast *trnQ–5′-rps16* and *psbD–trnT*$^{(GGU)}$ sequences. The size of the circle representing each haplotype is proportional to its frequency and the colours indicate the presence and frequency of drainage basins in which each haplotype is found (see inset). Lines represent single mutation steps; black, small circles represent unsampled or extinct haplotypes. Samples of *S. latifolia* that share haplotypes with *S. afra* are indicated (*). Haplotypes *q* and *r* are found in samples of *S. brachypetala* and *S. capitata*. (B) The distribution of the genus *Schotia*. (C) The distribution of haplotypes in southern Africa which occur beyond the Albany Subtropical Thicket. (D) The distribution of haplotypes within the Albany Subtropical Thicket.

Two chloroplast regions were used per taxon: (1) *trnL-trnF* (Taberlet *et al.* 1991) and *trnQ-5′–rps16* intergenic spacer for *P. capensis*, and (2) *trnQ-5′–rps16* and *psbD–trnT*$^{(GGU)}$ intergenic spacers for *Schotia* (Shaw *et al.* 2007). The chloroplast target regions were PCR-amplified from gross cellular DNA extracts. PCR reactions were as given in Chapter 4 (Page 107) using KAPATaq DNA polymerase.

Nuclear variation was sampled for the ITS-1, 5.8S and ITS-2 region using the primers ITS5m (Sang *et al.* 1995) and ITS4 (White *et al.* 1990). The nuclear target regions were PCR-amplified from gross cellular DNA extracts. PCR reactions were as given in Chapter 3 (Pg. 70) using KAPAHiFi DNA polymerase. Nine *P.capensis* samples and six *S. afra* were cloned to verify the presence of intra-individual site polymorphisms (2ISPs, see Chapter 3) observed in direct sequences. Cloning was performed using the pGEM-T Easy Vector System II (Promega) following the manufacturer's instructions, but downscaled to half reactions. To facilitate cloning, Kapa HiFi PCR products were incubated at 72°C for 10 minutes with Kapa Taq polymerase to provide 5′ terminal thymidine overhangs. Eight clones were sequenced per sample. All clones were PCR amplifications of colonies. All PCR products were sequenced using BigDye technology run on an ABI 3300 sequence analyser by Macrogen, Korea (http://dna.macrogen.com). All samples were sequenced in both directions to obtain reliable sequences for all regions.

5.3.3. Sequence assembly, alignment and characterisation

All sequences were edited with reference to chromatograms and aligned using Codon Code Aligner version 3.5.7 (Codon Code Corp, http://www.codoncode.com). Each base-call within every sequence was assigned a quality score using the automated base-calling program phred (Ewing *et al.* 1998) to improve the speed and accuracy of identifying DNA variations among assembled sequences. Polymorphic sites in the nDNA dataset were identified within sequences using the following steps: (1) each base-call within every sequence was assigned a quality score (using phred), (2) sites that contained secondary peaks that were greater than 20% of the primary peaks were scored as polymorphic using the 'Call second peaks' option in Codon Code Aligner, and (3) all polymorphic sites were verified by eye. Polymorphic sites where then coded using IUPAC ambiguous codes. All scored phylogenetically informative characters were nucleotide polymorphisms or indels that were not in simple repeats. A number of

long mononucleotide and dinucleotide repeats were observed to have length variation, but these were excluded because they are difficult to score with confidence. Indels in the cpDNA dataset that were not associated with such repeats were treated as informative characters and were coded for phylogenetic and phylogeographic analyses. All positions with missing data were excluded from subsequent analyses.

5.3.4. Phylogenetic networks and trees

Haplotypes of the combined cpDNA sequences were identified using the PEGAS library version 0.3.4 (Paradis 2010) in R version 2.13 (R Development Core Team 2011). The genealogical relationships among cpDNA haplotypes and nDNA sequences were estimated using both networks and trees. The cpDNA haplotype networks were estimated using the statistical parsimony (SP) method (Templeton *et al.* 1992) in TCS version 1.13 (Clement *et al.* 2000). The ITS networks were estimated using polymorphism *p*-distances (Chapter 3) and Neighbour-Net splits graphs (Bryant & Moulton 2004) generated in SplitsTree version 4.8 (Huson & Bryant 2006).

Phylogenetic trees of the cpDNA haplotypes and nDNA sequences were estimated using Neighbour Joining (NJ), Maximum Parsimony (MP) and Maximum Likelihood (ML). Bayesian Inference (BI) was not used as suitable outgroup sequences were not available; the current software implementation of BI requires complete outgroup sequences for rooting (e.g. MrBayes). Furthermore, current implementations of BI are unable to incorporate variation represented by the intra-individual site polymorphisms (2ISPs, Chapter 3) that are present in the nDNA. Neighbour Joining analyses were implemented in the APE library version 2.7.1 (Paradis *et al.* 2004) in R using uncorrected *p*-distances. Maximum Parsimony analyses were undertaken in PAUP* version 4b10 (Swofford 2002) using heuristic searches performed with 100 random sequence addition replicates with TBR branch swapping, and also saving no more than 1000 trees of length greater or equal to one per replicate. RAxML version 7.2.6 (Stamatakis *et al.* 2008) was used for ML, with the GTR-Γ model of sequence evolution. This is the simplest model available in RAxML for sequence data because thorough topological searching has a greater impact on the final tree quality than modelling details (Stamatakis *et al.* 2008). Assessment of bootstrap support included 10,000 bootstrap replicates; for MP following the suggestions of Müller (2005) with each replicate composed of a single random sequence replicate and TBR branch

swapping, and ML using rapid-bootstrapping (Stamatakis *et al.* 2008). Due to the presence of 2ISPs, I used the 2ISP-informative approach for NJ, MP and ML analyses of nDNA sequences (Chapter 3). The outgroup for *P. capensis* analyses included *trnL-trnF* and ITS accessions from *Plagioscyphus unijugatus* and *Plagioscyphus aff. louvelii* (Buerki *et al.* 2009). Following Ramdhani *et al.* (2010), the outgroup used for *Schotia* sample analyses was *Barnebydendron riedellii* (Fougère-Danezan *et al.* 2007). Only ITS accessions were available for this species, thus mid-point rooting was used for the cpDNA analyses.

5.3.5. Genetic and population expansion analyses

Standard sequence polymorphism indices (number of haplotypes or ribotypes, polymorphic sites and parsimony informative sites) and genetic diversity values (nucleotide diversity and haplotype or ribotype diversity) were estimated using the PEGAS library in R or using adjusted functions that take 2ISPs into account.

Possible population expansion of the two species in the AST was evaluated using the chloroplast data and the following methods: 1) Tajima's *D*, which is based on the infinite-sites model without recombination and tests for selective neutrality, but significant values can also be caused by population expansion, bottlenecks or heterogeneous mutation rates (Tajima 1989), 2) Fu's F_S, also based on the infinite sites model with no recombination but including information from haplotype frequencies (Fu 1997); large negative F_S value indicates there has been demographic population expansion, 3) R_2 test of neutrality, based on the number of singletons, total number of segregating sites and average number of nucleotide differences between sequences (Ramos-Onsins & Rozas 2002), and lastly, 4) the mismatch distribution (Rogers & Harpending 1992) and the associated raggedness index (*rg*; Rogers & Harpending 1992). The mismatch distribution uses the number of pairwise differences between haplotypes to compare the population demography with expectations of a sudden expansion model. The *rg* index measures the smoothness of the observed distribution and provides a means to statistically validate the estimated model of expansion; statistical significance was calculated using a parametric bootstrap approach that sums the square deviations (SSD) between the observed and expected mismatch (Schneider & Excoffier 1999) across 10,000 bootstraps. Tajima's D, Fu's F_S and the mismatch indices were calculated in ARLEQUIN version 3.5.1.2 (Excoffier *et al.* 2005), while the

R_2 test was calculated in the PEGAS library in R with 10,000 bootstrap replicates. The nuclear data was not used to explore possible population expansion as these indices currently are unable to incorporate intra-individual site polymorphisms.

5.3.6. Molecular dating

Although the absence of suitable fossil data precludes direct calibration of divergence rates, application of molecular clocks can still be useful and provide rough estimates of divergence times, even when certain assumptions of such methods are violated (see Bromham & Penny 2003). Rate constancy of chloroplast haplotypes from the AST and eastern lowlands was evaluated using the relative rate test (Tajima 1993) implemented in the APE library in R. The minimum age of divergence between the chloroplast haplotypes is estimated using a molecular clock approach (Sarich & Wilson 1973). A highly conservative approach is used, based on maximum and minimum substitution rates for non-coding chloroplast regions found in the literature, specifically 1.0×10^{-9} (Richardson *et al.* 2001) and 31×10^{-9} substitutions per site per year (Fu & Allaby 2010).

5.3.7. Spatial analyses

To establish whether either of the species have experienced isolation-by-distance, matrices of genetic distance (uncorrected *p*-distances for cpDNA and polymorphism *p*-distances for ITS) and the logarithms of geographical distance data (log km) between all individuals within the AST were constructed. The degree of correlation between these matrices, and its significance, was estimated by the Mantel test statistic (MR) as implemented in the R PEGAS library using 10 000 randomisations (Mantel 1967). Directional autocorrelation between genetic distance and geographic distance was examined over 14 classes of 50 km each using a Moran's I correlogram. The Moran's I test statistic (M_I) was calculated for each geographic distance-class matrix versus the genetic distance matrix. Significance levels of individual R-values were tested against the null hypothesis of no spatial arrangement by a resampling procedure (1000 permutations). The overall significance of all correlograms was assessed by the Bonferroni technique (Hommel 1988).

Spatial principle component (sPCA; Jombart *et al.* 2008) analyses were used to

explore the complex spatial patterns of genetic variability of the nDNA datasets. This is a spatially explicit multivariate method: spatial autocorrelation of allele frequencies is incorporated into the traditional principal component analysis framework that summarises the spatial patterns of genetic structure using synthetic components that optimise the product of the variance in the data and Moran's I (Jombart *et al.* 2008). The synthetic components can be separated into global (positive eigenvalues) and local (negative eigenvalues) structures. A global structure would indicate the presence of genetically distinguishable spatial clusters or a cline. Local structures would be found when there are stronger genetic differences among neighbours than among random pairs of entities (e.g. repulsion of individuals from the same genetic pool). Spatial PCA has two features making it expressly suitable for analysing the nDNA datasets: 1) it does not assume any genetic model that requires linkage equilibrium to exist between loci or populations to meet the Hardy-Weinberg equilibrium, and 2) the method can identify clusters and allele frequency gradients or clines. These features are not available in other methods such as those based on Bayesian clustering of individuals (Chen *et al.* 2007, Pritchard *et al.* 2000). The ability to detect gradients or clines is important because deviations from random mating not caused by hard barriers (such as isolation by distance or spatial autocorrelation) can mislead clustering algorithms (Frantz *et al.* 2009). Implementing sPCA requires that 'neighbouring sites' are defined (in order to calculate Moran's I) using a spatial connection network. There are several algorithms available to build this connection network (Legendre & Legendre 1998, pp 752-756); here I use the distance-based and the inverse-distance neighbouring graphs as these are appropriate for the irregular and aggregated distribution of our samples. Under the inverse-distance neighbouring graph, all individuals were connected and spatial weights were proportional to the inverse of the distance between the sites. Under the distance-based neighbouring graph, all individuals that were within 0 km and 100 km of one another were connected. I also used the global and local structuring tests to detect whether significant global or local sPCA eigenvalues are present (Jombart *et al.* 2008).

5.3.8. Species distribution modelling

Species distribution modelling is a burgeoning field with many uncertainties and assumptions (Dormann 2007; also discussed in Chapter 2). A primary source of uncertainty is methodological, where different modelling techniques produce different

results and that these may fundamentally change the predicted areas of suitability (Buisson *et al.* 2010, Elith *et al.* 2006, Thuiller 2004). These uncertainties can be compounded when projections of current species distributions are projected onto different areas or time periods (e.g. Diniz-Filho *et al.* 2009; known as the 'transferability problem'). Ensemble modelling has been suggested as a potential solution to this problem (Araújo & New 2007, Diniz-Filho *et al.* 2009), where results are averaged across predictions made from numerous different algorithms. However, not all methods perform equally well (e.g. Elith *et al.* 2006) and ensemble modelling may mislead the results if methods that perform poorly outweigh those that perform well. A synthesis that explores all methods and teases apart the causes for differences in prediction under a range of scenarios is still lacking. Thus, selecting methods for ensemble modelling is still highly subjective. Instead of using the ensemble approach, I use a single method that has been extensively tested (in comparison to other methods): MAXENT (Phillips *et al.* 2006). This method has performed well in terms of predictive performance and success in comparison to other methods (Elith *et al.* 2006), with estimates of the probability of occurrence correlating with local abundance (VanDerWal *et al.* 2009*b*), and it has performed well in comparison with mechanistic models under current climate and transferred onto past and future climate conditions (Hijmans & Graham 2006). Very importantly it has consistent results when 'transferred' (i.e. SDM results projected onto a different time) onto different global climate models (Diniz-Filho *et al.* 2009).

MAXENT was used to model the current areas of climatic suitability and project this onto the LGM climate layers. MAXENT is a machine-learning method based on the principle of maximum entropy (Phillips *et al.* 2006). It fits a probability distribution to the environmental conditions at the locations where a species has been observed; this is extrapolated over climate surfaces to create distribution maps with a probability of occurrence in each cell (Phillips *et al.* 2006). I used the default settings in MAXENT 3.3.3c (www.cs.princeton.edu/ schapire/maxent/) controlled via the DISMO library version 0.6.10 (Hijmans *et al.* 2011) in R. The properties of the MAXENT method means that it can cope with correlations and interactions among the climatic variables (Farber & Kadmon 2003, Phillips *et al.* 2006).

The primary source of occurrence data were GPS localities visited while collecting material for genetic analyses. A second source of occurrence data was georeferenced locality descriptions from the National Herbarium database (PRECIS). Only georef-

erenced localities that were considered to have a precision of five kilometres (the approximate size of the raster cells) or less were included. In total there were 231 records for *P. capensis* (201 GPS, 30 georeferenced) and 166 records for *S. afra* (124 GPS, 42 georeferenced). The occurrence data displayed some clustering, because i) the national road network was used to survey the region, and ii) not all sub-regions were subject to the same sampling effort. To correct for this sampling bias, I create five random subsets from the original dataset by limiting the minimum distance between samples to 15 km. This procedure yielded five subsets with 85 to 89 unique occurrences for *P. capensis* and 74 to 78 unique occurrences for *S. afra*.

Model evaluation was performed using 5-fold (K-fold) cross-validation on each of the five occurrence subsets for each species. The K-fold evaluation method randomly partitions the data into K subsamples, each of which is used in turn as test data, while the remaining K-1 subsamples are used for training data. In order to compare and graphically represent the results between the locality datasets, each of the folds was converted into binary present-absent values using the maximum test sensitivity (true positive rate) plus specificity (false positive rate) criterion (MSS); this criterion optimises the correct discrimination of presences and pseudoabsences in the test data and has performed well in comparison with other threshold criteria (Liu *et al.* 2005). Background data from 5000 cells without species observations were sampled from a mask extending from 20.0° to 29.0° east and 31.5° to 35.0° south (Appendix Figure A.1, Pg. 221); this was used for model evaluation and calculating the MSS threshold.

The climate data used to for the present-day and LGM SDM is explained in detail in Chapter 2 (Pg. 30). In brief, the bioclimatic variables used are from the WorldClim dataset (spatial resolution of 2.5 arc-minutes, www.worldclim.org) which are derived from temperature and precipitation values gathered from weather stations around the world from 1950-2000 which have been statistically interpolated to climatic surfaces (Hijmans *et al.* 2005). Many of the bioclimatic variables are correlated in the study region (see Chapter 2 for details); to reduce this redundancy all but one variable found in a correlation cluster were removed leaving 11 variables (Figure 2.2, Pg. 34; Table 2.1, Pg. 31). The same variables with the same spatial resolution were used for the LGM; these variables were derived from two statistically downscaled global circulation models (www.worldclim.org): CCSM3 (Collins *et al.* 2004) and MIROC3.2 (Hasumi & Emori 2004).

5.4. Results

5.4.1. Genetic data characteristics and phylogenetic reconstructions

All samples used in phylogeographic analyses including their respective chloroplast haplotype and ITS cluster are given in Appendix Tables A.2 (Pg. 240) and A.3 (Pg. 243). All datasets aligned readily and gaps corresponding to insertion or deletion events were included as informative characters; indels that were not associated with homopolymer repeats were coded as binary characters and used in all analyses (multiple base indels were treated as single characters).

The chloroplast DNA alignment was 1939 bp in *P. capensis* (915 bp from *trnL–trnF* and 1024 bp from *trnQ-5′–rps16*) and 1495 bp in Schotia (473 bp from *trnQ-5′–rps16* and 1022 bp from *psbD–trnT*$^{(GGU)}$). Of the 114 *P. capensis* and 90 *Schotia* samples used for genetic analyses, 11 and six failed amplification or sequencing for one or both of the chloroplast regions, respectively. These samples were dropped for subsequent chloroplast analyses. Sequence characteristics were consistent with typical chloroplast intergenic spacer regions with relatively low variability, and nucleotide and haplotype diversity (Table 5.1). A summary of variable sites across the chloroplast haplotypes for each species is shown Tables 5.2 and 5.3.

Table 5.1. Summary statistics for chloroplast and ITS datasets of *Pappea capensis* and *Schotia*. Summary statistics are shown for the complete dataset and samples restricted to the Albany Subtropical Thicket (AST); only *Schotia afra* samples from the AST are summarised. The mean and standard deviation (in brackets) is given for nucleotide and haplotype or ribotype diversity.

		Pappea capensis		*Schotia*	
		Complete	AST	Complete	*S. afra* in AST
cpDNA[1]	N	103	72	90	71
	Bases	1939	1939	1495	1495
	Var. Sites	49	14	23	16
	Indels	12	2	1	1
	Pars. Inf.	42	14	20	13
	Haplotypes	24	6	18	15
	Nuc. Div.	0.0045	0.002	0.0014	0.0012
		(0.0010)	(0.0004)	(0.0003)	(0.0003)
	Hap. Div	0.8471	0.6933	0.834	0.8262
		(0.0259)	(0.0378)	(0.0319)	(0.0373)
ITS	N	94	59	87	68
	Bases	628	628	638	638
	Var. Sites	76	32	144	111
	Indels	4	1	9	7
	Pars. Inf.	33	17	78	64
	Ribotypes	84	42	84	65
	Nuc. Div.	0.0169	0.0097	0.0350	0.0246
		(0.0620)	(0.0472)	(0.0894)	(0.0752)
	Rib. Div	0.9899	0.9766	0.9989	0.9982
		(0.0042)	(0.0096)	(0.0021)	(0.0031)

[1] Two chloroplast regions were analysed for each taxa: *trnL–trnF* and *trnQ-5′–rps16* for *P. capensis* and *trnQ-5′–rps16* and *psbD–trnT*$^{(GGU)}$ for *Schotia*.

Table 5.2. Variable sites across the chloroplast DNA haplotypes from two gene regions of *Pappea capensis* accessions. All sequences are compared to the reference haplotype A.

Nucleotide positions: *trnL--trnF*

Haplotype	N	59	152	182	187	189	246	292	392	405	587	604	611	710	737	826	887	902	909
A	13	T	T	T	1[a]	C	0	C	A	G	1[c]	0	T	C	A	A	A	T	G
B	34	.	.	.	.	.	.	.	.	.	.	0	.	.	.	.	.	.	.
C	1	.	.	.	.	.	.	.	.	.	.	0	.	.	.	.	.	G	.
D	17	G	.	.	0	0	.	T	.	.	.	0	.	.	.	.	.	.	.
E	3	G	.	.	0	0	.	T	.	.	.	0	.	.	.	.	.	.	.
F	4	.	.	.	.	.	.	T	G	.	.	0	.	.	.	C	.	.	.
G	4	.	C	.	.	.	.	T	G	.	.	0	.	.	.	C	.	.	.
H	2	.	C	.	.	.	.	T	G	.	.	0	.	.	.	C	.	.	.
I	2	.	C	.	.	T	.	T	G	.	.	0	.	.	.	C	.	.	.
J	1	.	C	.	.	.	.	T	G	.	.	0	.	.	.	C	.	.	.
K	2	.	C	.	.	.	.	T	G	.	.	0	.	.	.	C	.	.	.
L	3	.	C	.	.	.	.	T	G	.	.	0	.	.	.	C	.	.	A
M	1	.	C	.	.	.	.	T	G	.	.	1[d]	.	.	.	C	.	.	.
N	2	.	.	.	.	.	.	T	.	.	0	0	.	T	.	.	.	.	.
O	3	.	.	G	.	.	.	T	.	.	0	0	.	T	.	.	.	.	.
P	1	.	.	.	.	.	.	T	.	.	0	0	.	T	.	.	G	.	.
Q	1	.	.	.	.	.	.	T	.	.	0	0	.	T	.	.	.	.	.
R	1	.	.	.	.	.	.	T	.	.	0	0	.	T	.	.	.	.	.
S	1	.	.	.	.	.	.	T	.	.	0	0	.	T	G	.	.	.	.
T	1	.	.	.	.	.	.	T	.	.	0	0	.	T	A	.	.	.	.
U	3	.	.	.	.	.	.	T	.	A	0	0	.	T	A	.	.	.	.
V	1	.	.	.	.	.	.	T	.	.	0	0	.	T	G	.	.	.	.
W	1	.	.	.	.	.	1[b]	T	.	.	0	0	.	T	G	.	.	.	.
X	1	.	.	.	.	.	.	T	.	.	0	0	G	T	G	.	.	.	.

Nucleotide positions: *trnQ-5'--rps16*

Haplotype	4	5	22	30	150	180	250	261	288	347	364	419	430	455	519	546	555	561	586	598	623	640	829	859	872	917	942	976	986
A	T	0	C	A	0	C	A	0	1[h]	A	G	G	G	G	T	G	G	A	G	T	C	T	T	0	A	0	T	G	C
B	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	0	.	.	.	.	.
C	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	0	.	.	.	.	.
D	.	.	C	.	.	.	.	.	.	.	T	.	.	.	.	.	T	G	T	.	.	.	.	0	.	.	.	.	.
E	.	.	C	.	.	.	.	.	.	.	T	.	.	.	G	.	.	G	T	.	.	.	.	0	.	.	.	.	.
F	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	G	T	.	.	C	.	1[i]	.	.	.	.	.
G	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	G	T	.	.	.	.	0	.	.	.	.	.
H	.	.	.	.	.	.	.	.	.	.	T	.	A	.	.	.	.	G	T	.	.	.	.	0	.	.	.	.	.
I	.	.	.	.	.	.	.	.	.	.	T	.	A	.	.	.	.	G	T	.	.	.	.	0	.	.	.	.	.
J	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	G	T	.	.	.	.	0	.	.	.	.	.
K	.	.	.	.	1[f]	.	.	.	.	.	T	.	.	.	.	.	.	G	T	.	.	.	.	0	.	.	.	.	.
L	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	G	T	.	.	.	.	0	.	.	.	.	.
M	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	G	T	.	.	.	.	0	.	.	.	.	.
N	.	1[e]	.	.	.	.	.	.	0	.	T	-	.	A	.	.	.	G	T	C	.	.	G	0	.	1[j]	G	.	.
O	.	1[e]	.	.	.	.	.	.	0	G	T	-	.	A	.	C	.	G	T	C	.	.	G	0	.	1[j]	G	.	.
P	.	1[e]	.	.	.	.	.	.	0	.	T	-	.	A	.	.	.	G	T	C	.	.	.	0	.	1[j]	G	C	.
Q	.	1[e]	.	.	.	T	.	1[g]	0	.	T	-	.	A	.	.	.	G	T	C	.	.	.	0	.	1[j]	G	C	T
R	.	1[e]	.	.	.	.	.	.	0	.	T	-	.	A	.	.	.	G	T	C	.	.	.	0	C	1[j]	G	C	.
S	.	1[e]	.	.	.	.	.	.	0	.	T	-	.	A	.	.	.	G	T	C	.	.	.	0	C	1[j]	G	C	.
T	C	1[e]	.	.	.	.	.	.	0	.	T	-	.	A	.	.	.	G	T	C	.	.	.	0	C	1[j]	G	C	.
U	C	1[e]	.	.	.	.	.	.	0	.	T	-	.	A	.	.	.	G	T	C	.	.	.	0	C	1[j]	G	C	.
V	.	1[e]	.	C	.	.	C	.	0	.	T	-	.	A	.	.	T	G	T	C	T	.	.	0	.	1[j]	G	C	.
W	.	.	.	C	.	.	C	.	0	.	T	-	.	A	.	.	T	G	T	C	T	.	.	0	.	1[j]	G	C	.
X	.	.	.	C	.	.	C	.	0	.	T	-	.	A	.	.	T	G	T	C	T	.	.	0	.	1[j]	G	C	.

Number '0/1' in the sequences indicate the absence/presence of length polymorphisms whereby the superscripts identify corresponding character states. Note that poly-T stretches were excluded from analyses.

a, ATCGAATATCGAAACTCCATAAAGGATGAAGGTTAAGTCTATATACACTATGGATAATG; b, AATG; c, CCCACCCT; d, CTTTTTTT;

e, ATATATTAGA; f, AAATAATACGA; g, GAC; h, ATAGTAAAG; i, AAATAAAAA; j, AATTCTT.

Table 5.3. Variable sites across the chloroplast DNA haplotypes from two gene regions of *Schotia* accessions. All sequences are compared to the reference haplotype *a*.

	Nucleotide positions																																		
	psbD-trnT$^{(GGU)}$																					trnQ-5'-rps16														
Haplotype	46	161	184	392	431	436	469	558	560	600	615	636	658	702	827	828	900		28	127	133	153	169	173	176	192	197	216	218							
a	G	A	T	C	A	G	C	C	T	G	A	T	1[a]	C	T	A	A		C	T	G	G	A	T	T	A	A	A	A							
b	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.	.	.							
c	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.		.	.	.	A	.	.	.	.	.	.	.							
d	.	.	.	A	.	.	.	.	.	.	.	.	.	T	.	.	.		.	.	.	A	.	.	.	.	.	.	.							
e	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.		.	.	C	.	.	.	.	.	.	.	.							
f	A	.	.	A	.	.	.	.	.	.	C	.	.	.	A	.	.		.	.	.	.	.	.	.	.	.	.	.							
g	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	C	.	.							
h	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	C	.	.							
i	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.	.	.							
j	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.		.	.	.	.	.	.	.	C	.	.	.							
k	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.	.	.							
l	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.	.	.							
m	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	A	.	.	.	C	.							
n	.	.	.	.	.	.	.	.	.	.	.	.	0	.	.	.	.		.	.	.	.	.	.	.	.	.	.	.							
o	.	.	.	.	.	.	.	T	.	T	.	.	.	.	.	.	.		.	.	.	.	.	A	.	.	.	.	.							
p	.	.	.	.	.	.	.	.	G	.	.	G	.	.	.	T	.		.	.	.	.	T	.	.	.	.	.	.							
q	.	.	.	.	C	.	.	.	.	.	.	.	.	.	A	.	C		A	A	.	.	.	.	G	.	.	.	.							
r	.	.	.	.	C	A	.	.	.	.	A	.	.	.	A	.	C		A	A	.	.	.	.	G	.	.	.	C							

Number '0/1' in the sequences indicate the absence/presence of length polymorphisms whereby the superscripts identify corresponding character states. Note that poly-T and poly-AT stretches were excluded from analyses.

a, CAAA

The ITS DNA alignment was 628 bp in *P. capensis* and 638 bp in *Schotia*. Of the 114 *P. capensis* and 90 *Schotia* samples used for genetic analyses, 20 and nine failed amplification or sequencing of the ITS region, respectively. Six of the nine *Schotia* samples that amplified but failed to sequence were cloned. Consensus clone sequences were included in the *Schotia* ITS dataset for each of these samples. A single *Schotia* sample (AJP0742) contained two indels 80 bp apart the intervening bases were inferred using the 'split heterozygous indels' algorithm in CodonCode Aligner. These bases were compared with the other *Schotia* sequences and considered to be reliable. A summary of variable sites across the ITS ribotypes for each species is shown Tables 5.4 and 5.5. Multiple variants were found within cloned sequences from both species (Appendix Tables A.4 [Pg. 245] and A.5 [Pg. 247]). There is a high detection rate of 2ISPs in direct sequences that correspond to variability observed in clone sequences (94% across nine *P. capensis* samples and 97% across six *Schotia* samples; Table 5.6).

These rates are equivalent to those observed in others studies, and I deemed the direct sequences to be reliable indicators of the underlying variability in the ITS datasets for both taxa. Sequence characteristics were consistent with typical ITS with relatively high variability, and nucleotide and ribotype diversity (Table 5.1).

Table 5.4. Variable sites in direct-PCR ribosomal ITS sequences from a subset of *Pappea capensis* accessions. Eight samples were selected at random from each ITS cluster (identified in Figure 5.6) to demonstrate the variability within and between the clusters. All sequences are compared to the consensus sequence. Intra-individual site polymorphisms (2ISPs) are coded using IUPAC nomenclature with capital letters indicating complete overlap of bases in the trace file and small letters indicating one base dominant over another in the trace file. 2ISPs involving a base and an indel are coded using 'X'.

		ITS-1																														5.8S		ITS-2																	
Cluster	Sample	57	78	84	89	94	95	103	107	119	127	129	131	134	145	152	166	173	189	204	222	223	240	247	250	251	253	256	257	262	268	299	448	457	465	469	481	485	491	527	528	529	595	616	620	622	637	642	656	665	672
	Consensus	C	T	T	C	G	G	A	C	C	C	C	T	T	A	C	G	A	G	G	T	T	G	G	C	G	G	G	G	A	T	C	C	A	C	C	C	G	T	G	C	G	A	C	C	C	G	C	G	G	A
1	AJP0004	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.
	AJP0022	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	A	.	.	.	.	y	.	.	.	.	.	Y	s	.	.	.	A	.	.	.	.	.	.	.
	AJP0023	y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	R	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.
	AJP0400	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.
	AJP0443	.	.	.	.	.	X	.	.	.	.	.	y	.	.	.	.	.	.	.	Y	.	.	.	.	.	.	R	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	m	s	.	.	.	.	.	.
	AJP0630	.	.	.	.	.	X	.	.	.	.	.	Y	.	.	.	.	.	.	.	Y	.	.	.	.	.	.	R	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	M	s	.	s	.	.	.	.
	AJP0672	.	.	.	.	.	X	.	.	.	.	.	y	.	.	.	.	.	.	.	y	.	.	.	.	.	.	r	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	m	s	.	.	.	.	.	.
	AJP0728	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Y	.	.	.	.	.	.	R	.	.	.	.	.	.	.	y	.	.	y	.	.	.	.	M	.	.	.	.	.	.	.
2	AJP0293	.	.	.	.	.	-	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	AJP0294	.	.	.	.	.	X	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	r	.	.	.	.	.	.	.	.
	AJP0461	.	.	.	.	.	X	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	.	r	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	AJP0690	.	.	.	.	.	X	.	.	.	.	.	y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	y	.	.	.	.	.	.	r	.	.	.	.	.	.	.	.
	AJP0705	.	.	.	.	.	X	.	.	.	.	.	y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	r	.	.	.	.	.	.	.	.
	AJP0719	.	.	.	.	.	X	.	.	.	.	.	y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	r	.	s	.	.	.	.	.	.
	AJP0839	.	.	.	.	.	X	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	y	.	.	m	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	AJP0862	.	.	.	.	s	X	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3	AJP0436	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	AJP0437	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	AJP0439	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Y	.	X	.	.	R
	AJP0451	.	Y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	AJP0462	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	AJP0980	.	.	.	.	.	.	.	.	.	.	Y	Y	.	.	.	.	R	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.
	AJP0983	.	.	.	.	.	.	.	.	.	y	y	C	.	.	.	.	r	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.
	AJP0990	.	.	.	.	.	.	.	.	.	.	.	y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	AJP0036	.	.	Y	M	.	-	.	T	Y	.	.	.	Y	.	T	.	.	.	.	.	.	.	.	Y	.	.	.	.	R	.	.	.	.	Y	.	.	.	.	.	Y	S	G	.	.	.	.	.	T	.	.
	AJP0447	.	.	Y	M	.	-	.	T	.	.	.	.	Y	R	T	.	.	.	.	.	.	.	.	y	r	.	.	.	r	.	.	.	.	y	.	.	.	.	.	y	s	r	.	.	.	.	.	T	.	.
	PRE563739	.	.	.	.	.	-	.	T	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	T	r	.	.	.	.	.	.	.	.	y	.	.	.	.	.	y	s	G	.	.	.	.	.	T	.	.
	PRE561643	.	.	.	.	.	-	.	T	.	.	.	.	.	.	T	.	.	.	.	.	.	.	k	y	A	.	.	.	.	.	.	.	.	y	.	.	.	.	.	Y	S	G	.	.	.	.	.	.	.	.
	CM3187	.	.	.	.	.	-	.	T	.	.	.	.	.	.	T	S	.	.	K	.	C	.	.	T	.	.	.	.	.	.	.	.	.	.	.	y	.	.	.	.	.	G	.	.	.	.	.	.	.	.
	CM3183	.	.	.	.	.	-	.	T	.	.	.	.	.	.	T	.	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	K	.
	AJP0734	.	.	.	.	.	-	.	T	.	.	.	.	.	.	T	.	.	.	.	.	C	.	.	T	.	R	.	.	.	.	M	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.
	AJP0984	.	.	.	.	.	-	r	T	.	.	.	.	.	.	T	.	.	.	.	.	Y	.	.	Y	R	.	.	.	R	.	.	.	.	Y	.	.	.	.	.	y	s	G	.	.	.	.	.	T	.	.

Table 5.5. Variable sites in direct-PCR or clone consensus ribosomal ITS sequences from *Schotia* accessions. Samples are ordered according to the cluster in which they were identified in the ITS splits graph (Figure 5.8). Cluster *I* is represented by fifteen samples selected at random. The species sampled were: *S. afra* (*S. afr*), *S. brachypetala* (*S. bra*), *S. capitata* (*S. cap*), and *S. latifolia* (*S. lat*). A single sample considered to be a hybrid between *S. afra* and *S. latifolia* by Ramdhani *et al.* (2010) is included (*S. lat**). All sequences are compared to the consensus sequence. 2ISPs involving a base and an indel are coded using 'X'.

			ITS-1																																																											
Cluster	Species	Sample	8	14	16	20	22	24	30	33	35	40	41	43	44	48	50	51	59	66	67	70	75	76	85	89	91	92	96	97	101	102	103	104	115	118	122	123	124	127	151	153	167	168	169	171	172	175	183	185	194	199	204	208	210	215	220	222	223	224	227	228	229	234	241	243
		Consensus	G	C	C	C	-	C	T	T	C	T	A	-	A	G	C	G	C	C	T	T	C	C	C	T	G	A	G	C	C	G	-	C	C	C	C	C	T	T	C	G	A	C	A	A	T	C	G	G	C	C	G	G	G	C	C	A	G	C	C	G	C	T	C	A
I	*S. afr*	AJP0689	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	A	G	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	T	.	.	.	.	.	G	.	.	.	G	.	.	.	.	.	.	.	.
		AJP0692	.	.	.	.	.	.	Y	.	.	Y	.	.	.	.	.	R	.	Y	Y	.	.	.	.	C	.	.	M	G	M	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	T	.	.	.	.	.	G	.	.	.	G	.	M	.	.	.	.	.	.
		AJP0702	.	.	.	.	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	W	.	.	.	.	C	.	.	R	G	M	R	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	T	.	.	.	.	.	.	.	.	Y	G	.	.	Y	.	.	.	.	M
		AJP0742	.	.	Y	.	X	.	.	.	M	.	.	.	.	.	.	.	.	Y	Y	.	.	.	.	C	.	.	A	G	A	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	-	T	.	.	.	.	.	.	.	.	.	G	.	A	.	.	.	.	.	.
		AJP0788	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Y	C	.	.	A	G	A	.	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	-	T	.	R	.	.	K	.	.	Y	Y	G	.	.	Y	.	.	.	.	.
		AJP0795	.	.	.	.	.	Y	.	.	.	C	.	.	.	.	.	.	.	.	A	.	.	.	.	C	.	.	.	G	C	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	T	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.
		AJP0808	.	.	Y	.	X	.	.	.	M	Y	.	.	.	.	.	.	.	Y	Y	.	.	.	.	Y	.	.	A	G	M	.	.	.	.	.	.	.	.	.	.	R	.	.	.	.	-	T	.	.	Y	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.
		AJP0809	.	.	.	.	.	.	.	.	M	Y	.	X	.	.	.	.	.	T	Y	.	.	.	.	.	.	.	A	G	C	.	.	.	.	.	.	.	.	.	.	R	.	.	.	.	-	T	.	.	Y	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.
		AJP0858	.	.	.	.	.	.	.	.	M	C	.	.	.	.	.	.	.	Y	Y	.	.	.	.	C	.	.	R	G	C	.	.	.	.	.	.	.	Y	.	.	.	.	.	.	.	-	T	.	.	.	.	.	.	S	.	.	G	.	.	.	.	.	.	.	.
		AJP0876	.	.	.	.	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	.	.	Y	.	.	C	.	.	R	G	M	.	.	.	.	.	.	.	.	.	.	.	.	.	.	R	-	T	.	.	.	.	.	.	.	.	K	G	.	.	Y	.	.	.	.	.
		AJP0881	.	.	.	.	.	.	.	.	.	Y	.	.	.	.	.	.	.	Y	Y	.	.	.	.	C	.	.	A	G	M	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	T	.	.	.	.	.	R	S	.	Y	G	.	.	Y	.	.	.	.	.
		AJP0902	.	.	.	.	.	.	Y	.	.	Y	M	.	.	.	.	.	.	.	.	.	.	.	Y	C	.	.	M	G	M	.	.	.	.	.	.	.	.	.	.	.	.	M	.	.	-	T	.	.	.	.	.	.	.	.	.	G	.	M	.	.	.	W	.	.
		AJP0945	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	G	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	T	.	.	.	.	.	.	.	S	M	R	R	.	.	.	.	.	.	.
		YvW4407	R	.	.	.	.	.	.	.	.	C	.	.	.	R	.	.	.	Y	Y	.	.	.	.	C	.	.	M	G	C	C	.	.	.	.	.	.	.	.	Y	.	.	.	.	.	-	T	.	.	.	.	.	.	.	Y	M	R	.	M	.	.	.	.	.	.
	S. lat	NB1973	.	.	Y	.	X	.	.	.	.	Y	.	.	.	.	.	.	.	Y	Y	.	.	.	.	C	.	.	A	G	M	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	T	.	.	.	.	.	.	.	.	.	G	.	M	.	.	.	.	.	.
II	*S. bra*	BP16	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	A	.	.	.	.	.	.	.	.	G	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
		BP35	.	.	.	.	.	.	.	.	.	.	.	.	R	.	.	.	.	.	C	.	.	.	.	.	A	.	.	.	.	.	.	.	.	S	T	.	.	.	.	.	.	.	.	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	.	.	.	Y	.	.	.
		BP36	.	.	.	.	.	.	Y	.	.	.	.	.	R	.	.	.	.	.	C	.	.	.	.	.	A	.	.	.	.	.	.	.	Y	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Y	.	.	.
		SR799	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	T	.	.	A	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
		SR812	.	.	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	S. cap	SR789	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	A	.	.	.	.	.	.	.	.	G	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	S. lat	ML1144	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	C	.	.	.	.	.	A	.	.	.	G	.	.	.	.	.	T	.	.	.	.	.	M	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
III	*S. afr*	VetterX	.	Y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	S. bra	SR785	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	R	R	.	.	.	.	.	Y	.	.	K	.	.	Y	.	.	.	Y	.	.	.	.	R	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	S. lat	AJP0756	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
		DS3144	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Y	.	.	.	.	.
		NB2018	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	R	G	.	.	.	.	.	Y	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	.	.	.	.	.	.
		NB2020	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	R	G	.	.	.	.	.	T	.	.	G	.	.	.	.	.	.	Y	.	.	.	.	R	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
		TD4485	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	G	.	.	.	.	.	T	.	.	G	.	.	.	.	.	.	T	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	*S. lat**	SR824	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
IV	*S. lat*	AJP0933	.	.	.	Y	.	.	T	.	.	C	G	.	.	.	A	.	.	T	.	Y	.	.	.	.	A	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	T	.	.	.	.	R	.	.	A	.
		AJP1000	.	.	.	.	.	.	T	.	.	C	G	.	.	.	A	.	.	T	.	C	.	.	.	.	A	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	T	.	.	.	.	.	.	.	A	.
		SR816	.	.	.	.	.	.	T	.	.	C	G	.	.	.	A	.	.	T	.	C	.	.	.	.	A	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	T	.	.	.	.	.	.	.	A	.

Table 5.5. Continued.

Cluster	Species	Sample	5.8S				ITS-2																																		
			297	303	336	394	419	427	432	436	437	439	447	448	458	464	475	481	482	497	515	516	520	522	532	541	551	563	575	576	585	588	594	596	602	605	613	614	620	622	628
		Consensus	G	C	C	C	A	C	G	C	G	0	C	C	G	G	G	C	T	C	G	T	G	G	G	A	-	A	T	C	C	C	C	G	C	C	G	C	C	C	T
I	*S. afr*	AJP0689	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.
		AJP0692	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	Y	.	.	.
		AJP0702	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	R	.	.	.	.	.	.	S	.	.	.	.	.	.	.	.	.	.	.
		AJP0742	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.
		AJP0788	.	.	Y	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Y	.	.	.	.	.
		AJP0795	.	.	.	.	.	.	.	.	.	.	G	Y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	T	.	.	.	.	.	.	.	.
		AJP0808	.	.	.	Y	.	.	.	.	.	.	G	.	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	.	.	.
		AJP0809	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	.	.	.
		AJP0858	R	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
		AJP0876	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
		AJP0881	.	.	.	.	.	.	.	.	R	.	G	.	.	.	.	.	.	.	.	.	R	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
		AJP0902	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	Y	Y	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	Y	.	.	.	.	.
		AJP0945	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	S	.	.	.	.	Y	M	.	.	.	.	.
		YvW4407	.	.	.	.	.	.	R	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	M	.	.	Y	.	.
	S. lat	NB1973	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	Y	.
II	*S. bra*	BP16	.	.	.	.	.	.	.	.	.	1[a]	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	M	.	.	.	.	.	.	.	.	.
		BP35	.	Y	.	.	.	Y	.	.	.	1[a]	Y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
		BP36	.	.	.	.	.	T	.	.	.	1[a]	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	S	.	.	.	.	.	.	.	.	.	.
		SR799	.	.	.	.	.	.	.	.	.	1[a]	.	.	.	.	A	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
		SR812	.	.	.	.	.	.	.	.	.	1[a]	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	W	.	.	.	.	.	.	.	.	.	.	.	.
	S. cap	SR789	.	.	.	.	.	.	.	.	.	1[a]	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	S. lat	ML1144	.	.	.	.	M	.	.	.	.	1[a]	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	M	.	.	.	.	.	.	.	Y	.	.	.	.	.
III																																									
	S. afr	VetterX	.	.	.	.	.	.	.	.	.	1[b]	.	.	.	.	.	.	.	.	.	Y	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	S. bra	SR785	.	.	.	.	.	.	.	.	.	1[b]	.	.	.	.	.	.	.	.	.	Y	.	.	R	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	S. lat	AJP0756	.	.	.	.	.	.	.	.	.	1[b]	.	.	.	.	.	.	.	.	.	C	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
		DS3144	.	.	.	.	.	.	.	.	.	1[b]	.	.	.	.	.	.	.	.	.	C	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
		NB2018	.	.	.	.	.	.	.	.	.	1[b]	.	.	.	.	.	.	.	Y	.	Y	.	.	R	G	.	.	.	.	.	.	.	.	.	.	.	Y	.	.	.
		NB2020	.	.	.	.	.	.	.	.	.	1[b]	.	.	.	.	.	.	.	.	.	.	.	R	R	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
		TD4485	.	.	.	.	.	.	.	.	.	1[b]	.	.	.	.	.	.	.	.	.	.	.	.	R	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	*S. lat**	SR824	.	.	.	.	.	.	.	.	.	1[b]	.	.	.	R	.	.	.	.	S	C	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
IV	*S. lat*	AJP0933	.	.	.	.	.	.	.	.	.	1[b]	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	R	.	.	R	.	.	.	.
		AJP1000	.	.	.	.	.	.	.	.	.	1[b]	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	R	.	.	.	.	.	.	.
		SR816	.	.	.	.	.	.	.	.	.	1[b]	.	.	R	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.

Number ‘0/1’ in the sequences indicate the absence/presence of length polymorphisms whereby the superscripts identify corresponding character states.

a, GCGC; b, ACGC.

Table 5.6. Comparison of intra-individual site polymorphisms detected from direct sequencing versus cloning across multiple sequences from *Pappea capensis* and *Schotia* and three other plant taxa. The number of samples and number of clones per sample are shown in brackets.

	Direct sequences									
	Pappea capensis (9; 8-10)		*Schotia*[1] (6; 7-9)		*Buxus balearica*[2] (5; 3-11)		*Asarum* sect. *Asiasarum*[3] (30; 13-31)		*Nymania capensis*[4] (8; 8)	
Clone sequences	Present	Absent	Present	Absent	Present	Absent	Present	Absent	Present	Absent
Present	48	15	36	2	29	4	177	54	69	20
Absent	3	-	1	-	3	-	1	-	4	-

[1] The presence or absence of variability across all clones is compared with that from all direct sequences of *Schotia afra* and *S. latifolia* as direct sequences were not available for the cloned samples due to an excess of ITS variants with indels; the presence of numerous variants of different lengths prevents clear sequencing across the entire region. Comparison of all clones with the entire dataset biases the estimates towards detection in both clones and direct sequences.

[2] Yamaji *et al.* (2007)

[3] Rosselló *et al.* (2007)

[4] Chapter 3

The SP network of *P. capensis* chloroplast data is shown in Figure 5.3. Five nested clades are identified which received moderate (>70%) to high (>90%) bootstrap support across the NJ, MP and ML trees (Figure 5.5). A major divergence is revealed in both the networks and trees as two well-supported clades: one consisting of the samples restricted to the coastal lowlands and the other consisting of the samples from the rest of Africa (Figures 5.3 and 5.5). Clade 1 (haplotypes A-C) is restricted to the western AST drainage basins, with the Gouritz containing a unique haplotype, while haplotype B has the widest geographical distribution of all the AST haplotypes (Figure 5.3.D). Clade 2 (haplotypes D-E) is largely restricted to the Great Fish drainage basin. Clade 3 (haplotypes F-M) occurs across the central and eastern coastal lowlands. Clade 4 (haplotypes N-U) occurs in the rest of southern Africa and clade 5 (haplotypes V-X), which is nested in clade 4, is found in central Africa.

The majority of ITS sequences were unique in both *P. capensis* and *Schotia* datasets, therefore they were not concatenated into ribotypes. The nDNA NN phylogenetic network of the *P. capensis* ITS also displays a split between the coastal lowland samples and those from the rest of Africa (Figure 5.6) supporting the splits found in the cpDNA trees and SP network. These clusters, or any other clades, did not receive any consistent support across the different phylogeny-reconstruction methods (Appendix Figure A.15, Pg. 248]). The coastal lowlands samples form three clusters, which show a degree of association with the chloroplast haplotype clades and geography. The ITS clusters 1 and 2 are predominantly composed of samples from chloroplast clades 1 and 2, respectively, although they do contain haplotypes from other lowland clades. Cluster 3 is solely comprised of samples from chloroplast clade 3. The three coastal clusters also show geographic affinities, with cluster 1 restricted to the west, cluster 3 in the east, and cluster 2 found between these two and overlapping somewhat. Clusters 4 and 5 are congruent with the chloroplast clades 4 and 5. Both the chloroplast haplotypes and clades and ITS clusters demonstrate a strong geographic association between these haplotypes and primary drainage basins within the AST (Figures 5.3 and 5.6). This pattern may extend to the eastern lowland samples, but inference is limited due to the small sample size.

The chloroplast SP network of *Schotia afra* presented a star-like pattern with the most central haplotype (*a*) being the most abundant and very widespread (Figure 5.4). Haplotype *b* was also widespread extending from the AST eastward along the coastal lowlands. The only dominant haplotype ($n \geq 5$) that was restricted to a

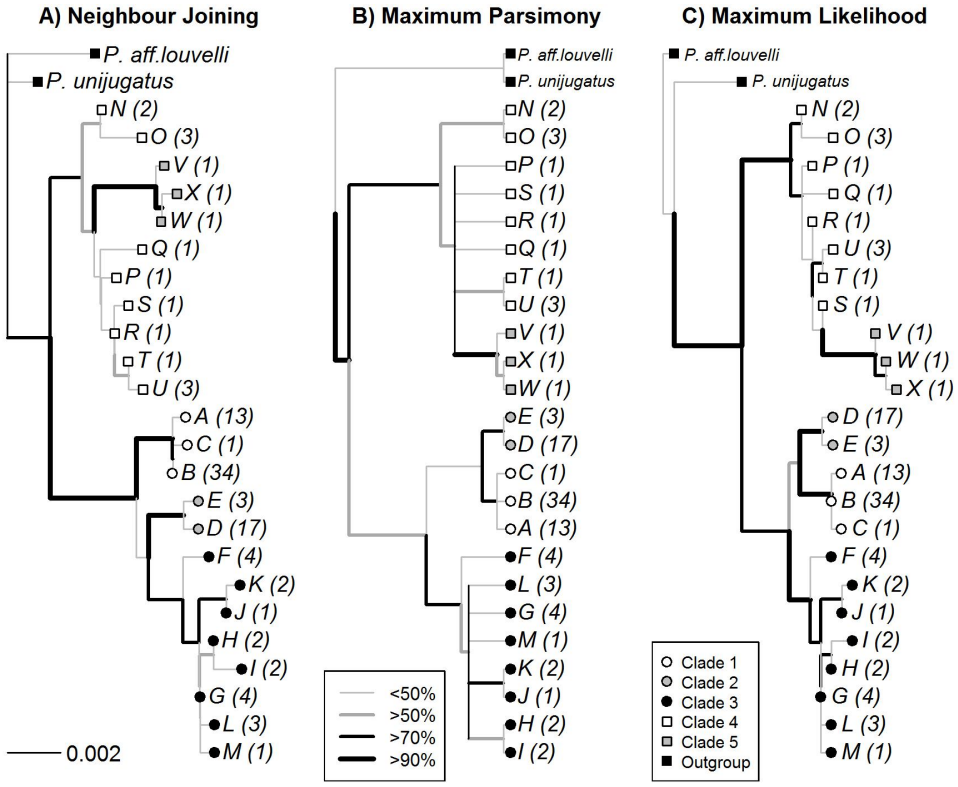

Figure 5.5. Phylogenies and clade definitions of of 24 haplotypes found in *Pappea capensis* as defined on the basis of chloroplast *trnL–trnF* and *trnQ-5′–rps16* sequences using (A) Neighbour Joining, (B) Maximum Parsimony, and (C) Maximum Likelihood. The number of samples per haplotype is shown in brackets. Outgroup samples are from the genus *Plagioscyphus*. Branch support is shown using a combination of line thickness and colour.

specific drainage basin was haplotype k in the Gouritz basin. The haplotypes (q and r) were restricted to the eastern lowlands and were three mutational steps away from the western lowland samples. These two haplotypes come from *S. brachypetala*, *S. capitata*, and an isolated and morphologically distinct population of *S. latifolia* (M. Lotter, personal communication). These eastern haplotypes were identified as a separate divergent clade by NJ and MP midpoint rooting and were treated as outgroup samples for bootstrapping analyses. The ingroup haplotypes (i.e. predominantly from the AST) formed a well-supported clade (Figure 5.7); however, there was very little resolution or support for the relationships between the ingroup haplotypes.

The ITS NeighbourNet phylogenetic network of the *Schotia* samples revealed

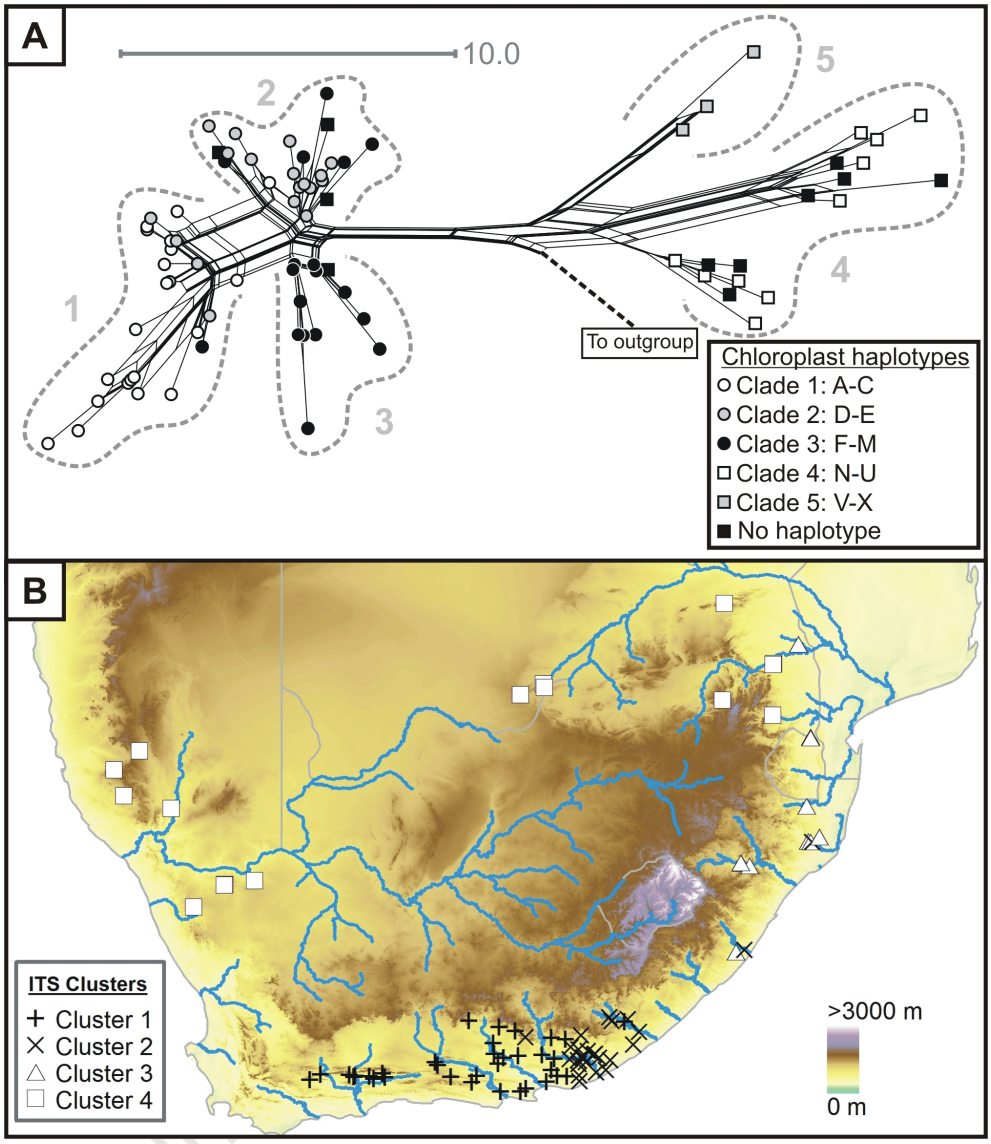

Figure 5.6. Phylogeography of nuclear DNA (ITS-1–5.8S–ITS-2) sequences from *Pappea capensis* with the (A) NeighbourNet splits graph and identified clusters, and (B) the distribution of clusters. The clades identified in the chloroplast data are represented using tip symbols in the splits graph. The outgroup accession are from *Plagioscyphus unijugatus* and *P.* aff. *louvelii*. The length of the branch connecting the outgroup to the ingroup is a polymorphism *p*-distance of 37.5.

four divergent clusters, with long branch lengths between cluster *I* and the remaining clusters. These clusters generally received moderate to high support in the ITS phylogeny reconstructions across a range of methods (Appendix Figure A.16, Pg.

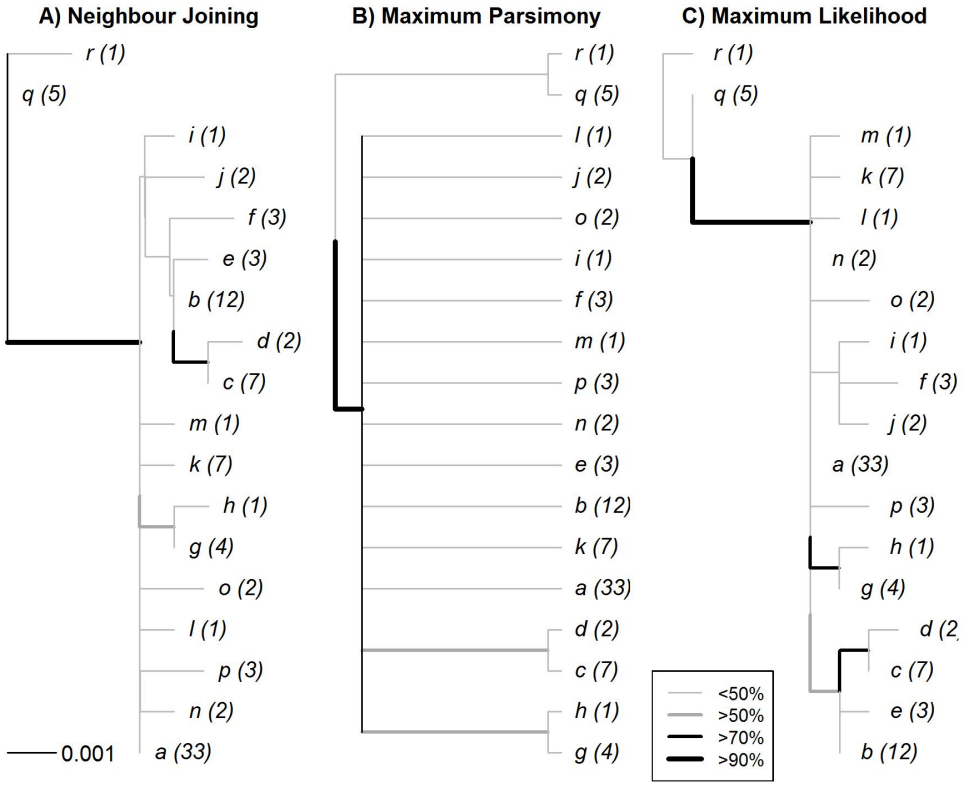

Figure 5.7. Phylogenies and clade definitions of 18 haplotypes found in *Schotia* as defined on the basis of chloroplast *trnQ–5′-rps16* and *psbD–trnT*$^{(GGU)}$ sequences using (A) Neighbour Joining, (B) Maximum Parsimony, and (C) Maximum Likelihood. The number of samples per haplotype is shown in brackets. The trees were rooted using midpoint-rooting. Branch support is shown using a combination of line thickness and colour.

249; however, clades within cluster *I* did not receive consistent support. As clade *I* was comprised primarily of the target species, *S. afra*, with no obvious phylogenetic divisions, the phylogeographic patterns were further explored using sPCA (reported below). The only notable congruence between network clusters and chloroplast haplotypes are found in the eastern *Schotia* samples, specifically haplotypes *q* and *r* fall within cluster *II*. Cluster *I* was only comprised of *S. afra* individuals, with the exception of one sample identified as *S. latifolia* (NB1973) by Ramdhani *et al.* (2010). Cluster *II* comprised the eastern samples of *S. brachypetala*, *S. capitata*, and the anomalous eastern *S. latifolia*. Cluster *III* is exceptionally widespread as it occurs in the AST and eastward along the coastal lowlands, but also one sample is from Namibia. This cluster is predominantly comprised of *S. latifolia*, but also contains a

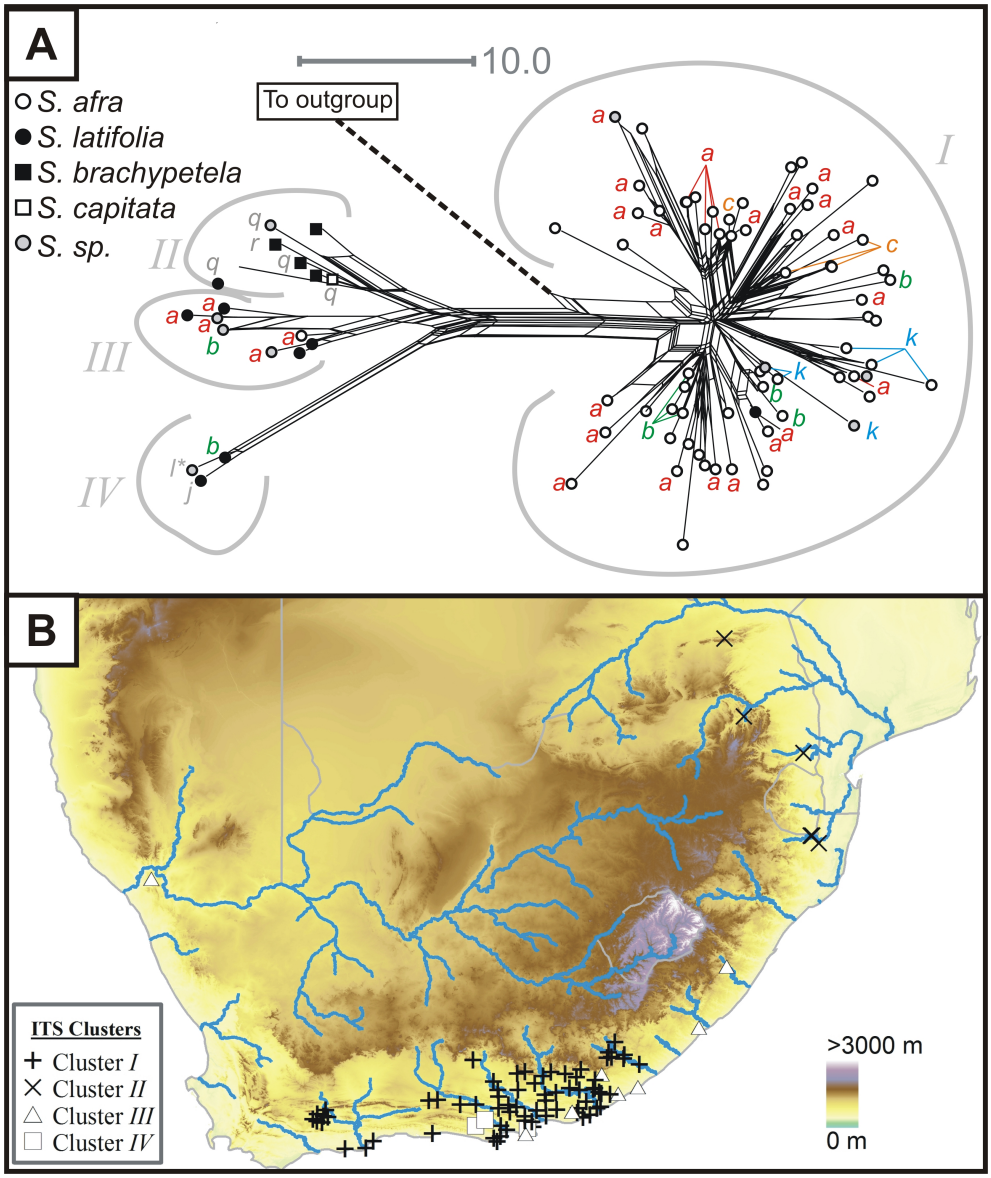

Figure 5.8. Phylogeography of nuclear DNA (ITS-1–5.8S–ITS-2) sequences from *Schotia afra* and closely related species: (A) NeighbourNet splits graph and (B) the distribution of clusters. Clusters are identified using roman numerals. The common chloroplast haplotypes are shown using letters near the tips of the respective sample in the splits graph. The outgroup accession is from *Barnebydendron riedellii*. The length of the branch connecting the outgroup to the ingroup is a polymorphism p-distance of 98.9.

sample of *S. afra* (Vetter sn) from its north-western distribution in Namibia. This can be considered an anomalous population of *S. afra*, which is already recognised as a sub-species (*S. afra* var. *angustifolia* [E. Meyer] Harvey). Cluster *IV* is comprised of individuals of *S. latifolia* in a very restricted area of the AST which includes the topographically complex Baviaanskloof range.

5.4.2. Population expansion, molecular clock and spatial statistical analyses

The mismatch distribution of the *P. capensis* AST chloroplast haplotypes was bimodal, consistent with pairwise differences between sequences belonging to divergent lineages (Figure 5.9). The mismatch distribution of the *S. afra* AST chloroplast haplotypes showed a unimodel distribution that, visually, fitted almost perfectly the expect values of a model of population expansion (Figure 5.9). The goodness of fit test showed no significant differences between the observed and expected values under a sudden expansion model for the *P. capensis* or *S. afra* AST samples (*P. capensis*: SSD = 0.0790, $p > 0.05$; *S. afra*: SSD = 0.0012, $p > 0.05$). Positive and non-significant values of Tajima's D (D = 0.7966, $p > 0.05$) and Fu's F_S (F_S = 2.9372, $p > 0.05$) statistics for *P. capensis* do not support the hypothesis of a sudden population expansion, and neither does the R_2 test for neutrality (R_2 = 0.1134, $p > 0.05$). In contrast, negative and significant Tajima's D (D = -1.5084, $p < 0.05$) and Fu's F_S (F_S = -8.0692, $p < 0.01$) values of *S. afra* samples are consistent with a sudden population expansion. This is also supported by the R_2 test for neutrality ($R_2 = 0.0473$, $p < 0.05$).

The relative rate test did not find significant deviation from the neutral molecular clock between any of the haplotype pairs of *P. capensis* from AST and eastern coastal lowlands, or between any haplotype pairs of *S. afra* (all $p > 0.05$). Utilising a clock-based approach with a wide range of biologically plausible mutation rates for cpDNA, the divergence within and between the haplotype clades all fall within the Pleistocene (10 ka to 2.6 Ma) for both *P. capensis* and *S. afra* (Table 5.7). Given the slow growth of both of these species, it is likely that the slow substitution rate (i.e. older dates) is more reliable than the fast rate and the timing of divergence is likely to fall within the early to mid-Pleistocene.

The Mantel test of AST *P. capensis* samples detected strong and significant patterns of isolation by distance for both cpDNA ($M_R = 0.5554$, $p < 0.01$) and nDNA

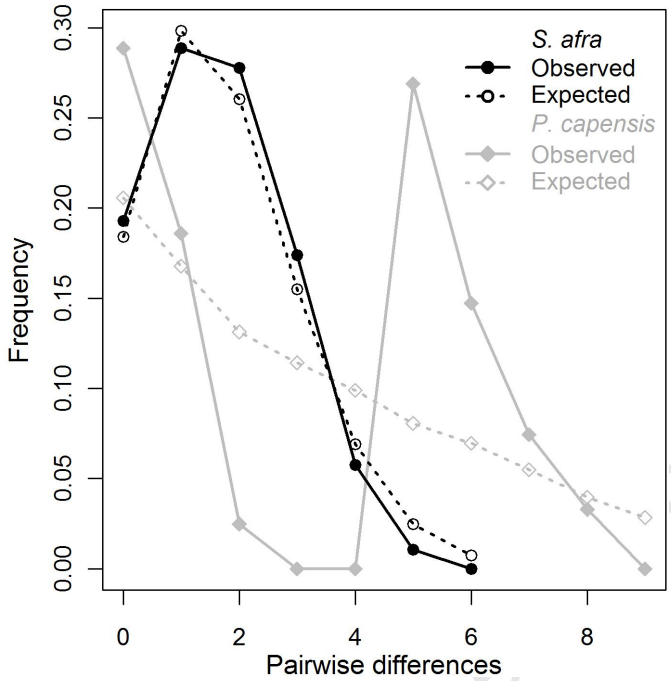

Figure 5.9. Mismatch distribution for observed and expected pairwise comparison under a sudden population expansion model among chloroplast sequences of *Pappea capensis* and *Schotia afra* samples from the Albany Subtropical Thicket. The bimodal observed distribution of *P. capensis* is comprised of a first peak of pairwise comparisons between closely related individuals (intra-clade), while the second peak corresponds to pairwise comparisons among distantly related individuals (inter-clade). The observed mismatch distribution among sequences from *S. afra* has a unimodal distribution that closely mirrors the expected mismatch distribution; this genetic signature corresponds to the expected distribution for sudden population expansion.

($M_R = 0.4815$, $p < 0.01$) datasets. The mantel test of the AST *S. afra* samples detected significant but negligible patterns of isolation by distance for cpDNA ($M_R = 0.0244$, $p < 0.05$) and nDNA ($M_R = 0.2648$, $p < 0.01$). The spatial autocorrelation analysis of cpDNA data of *P. capensis* shows that a strong effect of decreasing relatedness with distance operates most strongly over distances of about 150 km (Figure 5.10). This pattern is mirrored in the ITS dataset of *P. capensis*, but is non-significant. A significant negative autocorrelation between the distance bins of 350 km to 450 km suggests genetic divergence between samples greater than 350 km apart. Examination of all pairs of samples found in the 350 to 450 km classes showed that this result was based on sample pairs between cluster 1 and cluster 2. The negative but non-significant

Moran's I values greater than 450 km is likely due to the decline in the number of sample pairs at these distance bins. Both the spatial autocorrelation analyses of cpDNA and ITS of *S. afra* reveal non-significant Moran's I values in the correlogram with fluctuations close to 0, which indicates a lack of any spatial structure. This is congruent with the low M_R values.

The nDNA datasets contain complex patterns in AST samples for both *P. capensis* and *S. afra* (Figures 5.6 and 5.8). Spatial PCA analyses were used to provide a visual summary of the genetic variation observed in ITS in the two plant species. The sPCA analyses were focussed on all AST samples of *P. capensis*, and the nDNA cluster primarily comprised of *S. afra* (cluster *I*). The global test confirmed the presence of global structure in both *P. capensis* (max(t) = 0.0508, $p < 0.01$) and *S. afra* (max(t) = 0.0427, $p < 0.01$), while the local tests found no evidence for local structures for either *P. capensis* (max(t) = 0.0315, $p > 0.10$) or *S. afra* (max(t) = 0.0310, $p > 0.10$). The first sPCA eigenvalue was strikingly large compared to the others in *P. capensis*, and thus only the first eigenvalue and corresponding scores were retained (Figure 5.11.A; Appendix Figure A.17, Pg. 250). No such clear distinction of sPCA is found in *S. afra* (Figure 5.11.B), but the first three eigenvalues ($\lambda 1$, $\lambda 2$, and $\lambda 3$) are distinct from the successive values in the screeplot of the spatial and variance components (Appendix Figure A.18, Pg. 250). Thus, the first three eigenvalues were retained for sPCA exploration in this species. The retained eigenvalues for each species were

Table 5.7. Dating chloroplast haplotype divergences of *Pappea capensis* and *Schotia afra* using the molecular clock approach. The mean (minimum; maximum) ages estimated from two substitution rates are given. As no substitution rates were available for the target taxa, two biologically plausible substitution rates are used that represent the fast and slow extremes found in the literature. The rates used for fast and slow are 31×10^{-9} and 1.0×10^{-9} substitutions per site per year (Fu & Allaby 2010, Richardson *et al.* 2001)

Species	Haplotypes	Fast rate	Slow rate
P. capensis	A-E	36,438 (9,096; 54,675)	1,129,578 (281,968; 1,694,918)
	F-M	17,242 (8,615; 34,519)	534,501 (267,058; 1,070,093)
	A-E vs F-M	67,250 (54,675; 82,121)	2,084,754 (1,694,918; 2,545,741)
S. afra	*a-p*	30,410 (10,889; 54,549)	942,721 (337,553; 1,691,021)
	a-p vs *q-r*[1]	71605 (43592; 98287)	2,219,742 (1,351,354; 3,046,906)

[1]Note that the geographic sampling between the haplotypes *a-p* and *q-r* is very poor. This may affect the minimum and mean molecular date estimates.

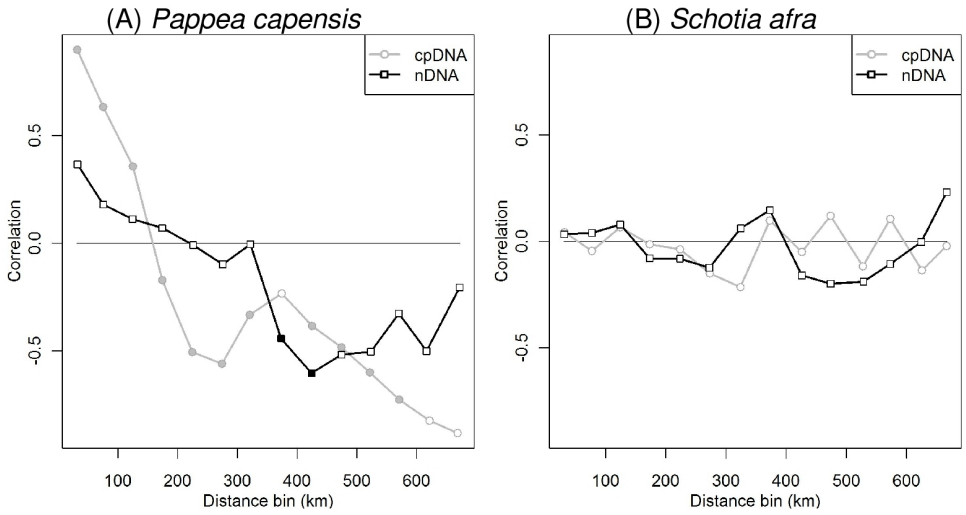

Figure 5.10. Correlograms of Moran's I per distance class of chloroplast or nuclear DNA datasets for (A) *Pappea capensis* and (B) *Schotia afra*. Filled symbols represent Bonferroni-corrected Moran's I p-values significantly different from 0 at $\alpha = 0.05$.

represented together using a colour plot (Figures 5.11.C and 5.11.D). The split between ITS clusters 1 and 2 in the phylogenetic network in *P. capensis* is detected in the sPCA colour plot. The sPCA also detects a fine scale structure in nDNA cluster 1, dividing this into subdivisions that mirror chloroplast haplotype clade 1. Each of these clusters was tested for an isolation-by-distance pattern using a Mantel test; no significant patterns were detected (clusters from west to east: M_R = -0.1812, $p > 0.10$, M_R = 0.07067, $p > 0.10$, M_R = 0.1815, $p > 0.10$). The sPCA analysis of *S. afra* nDNA cluster 1 shows a clinal pattern of variation with strong sPCA clustering of samples in the western Gouritz and the eastern Kei drainage basins. Samples in the intervening drainage basins display a high level of heterogeneity (many colours) with subtle turnover between samples sharing sPCA space (representing samples that occur in the same sPCA space) and no strong association with drainage basins.

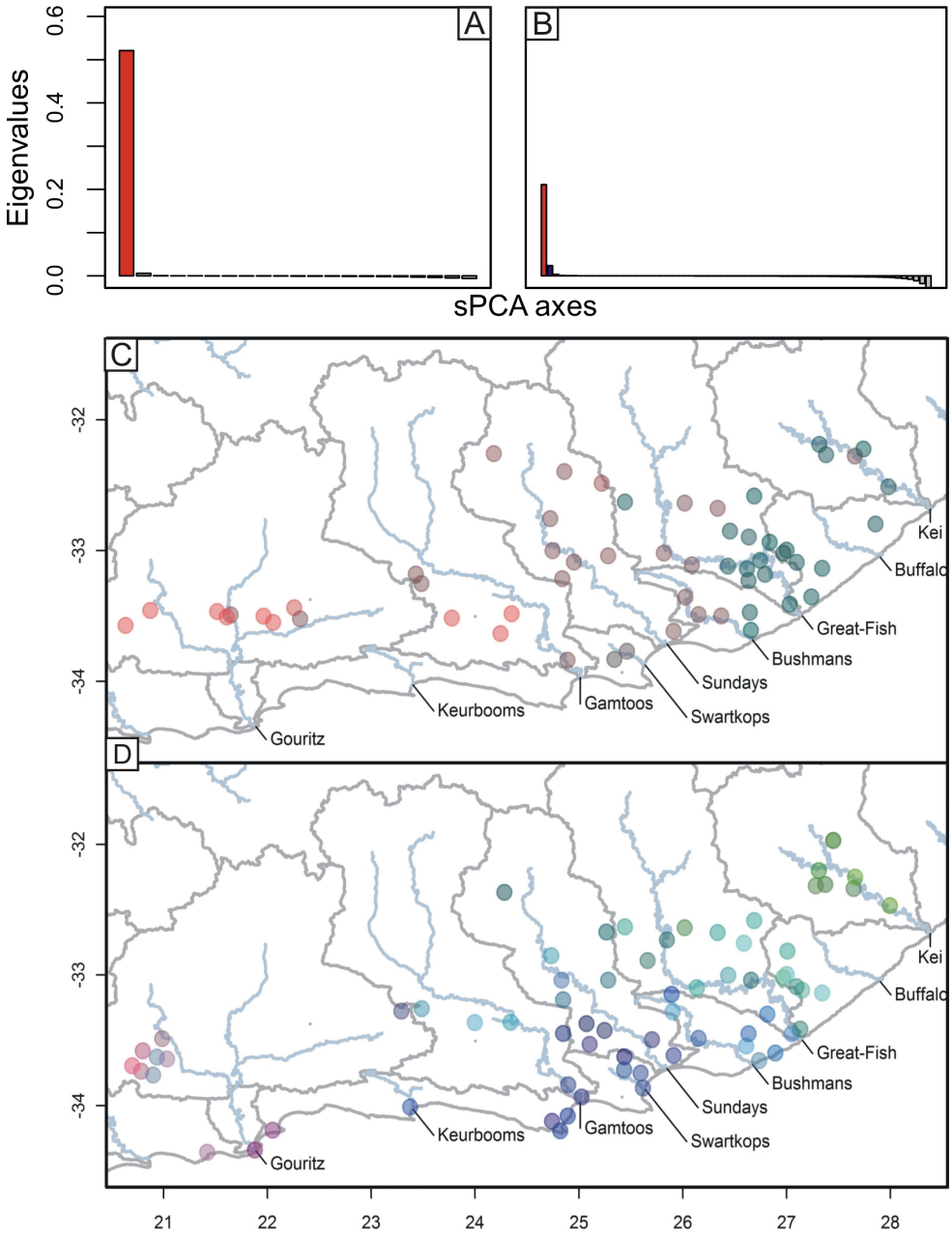

Figure 5.11. Analyses of ITS sequences from *Pappea capensis* and *Schotia afra* using spatial principle component analysis (sPCA). The screeplots of Eigenvalues for (A) *Pappea capensis*, and (B) *Schotia afra* are shown. The colour plots represent (C) one (*P. capensis*) or (D) three (*S. afra*) principle components of the sPCA. Each circle corresponds to a sampled individual and each principle component is recoded as intensities of a given colour channel of the RGB system: red (first PC), green (second PC) and blue (third PC). These channels are mixed to form colours representing the genetic similarity of individuals in sPCA space.

5.4.3. Species distribution modelling

The combined species distribution models of the 25 locality datasets with thresholds for MAXENT matched the fine-scale geographic mapping of thicket subtypes where both *P. capensis* and *S. afra* are dominant components of the vegetation. The models also predicted suitable conditions outside of the mapped distribution of the thicket subtypes, primarily in the upper reaches of the Sundays and Gouritz basins. All MAXENT models were accurate in the target region, with AUC values higher than null expectations for both *P. capensis* ($p < 0.01$, AUC $= 0.8531 \pm 0.0229$) and *S. afra* ($p < 0.01$, AUC $= 0.8644 \pm 0.0285$) datasets.

There was significant variability in postdicted suitable LGM conditions between species and GCMs. Nonetheless, both GCMs show a dramatic decline in the current distribution of the two species during the LGM with the distribution under MIROC3.2 being far more restricted than, but nested within, the distribution under the CCSM GCM (Figures 5.12 and 5.13). Refugia of suitable conditions for *P. capensis* predicted by both GCMs are found in the Gouritz, in the Little Karoo (a valley defined by east-west mountain ranges), Sundays and Great-Fish drainage basins. Under the larger distribution postdicted under CCSM, the Baviaanskloof, the intermontane valley in the Gamtoos drainage basin, is also predicted as an area with suitable climate. The LGM postdicted area of suitable climate for *S. afra* found in both GCMs is restricted to a single refugium found in the eastern AST which extends onto the exposed continental shelf. However, under the CCSM, an extensive and continuous area of suitable climate is postdicted that also includes an inland component which extends along the present-day coastline from the Sundays up to the Kei drainage basins.

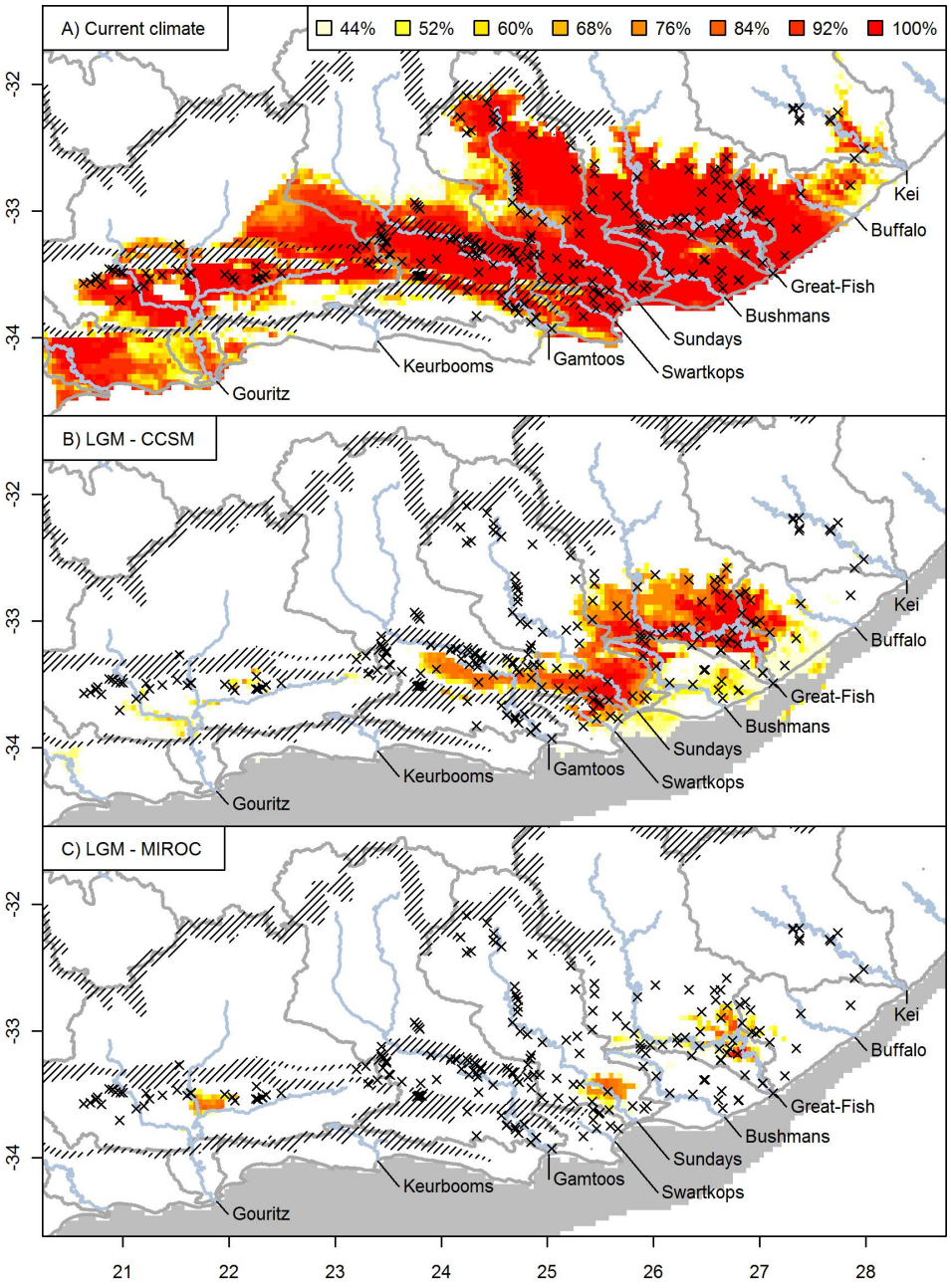

Figure 5.12. The modelled present and Last Glacial Maximum areas of suitable climate for *Pappea capensis* within the Albany Subtropical Thicket (AST). The summary of modelled distributions from five pseudo-replicate datasets and K-folding subsets (see text) representing the climatically suitable areas in the AST under (A) present conditions, (B) the CCSM Last Glacial Maximum (LGM; 21,000 BP) simulation, and (C) the MIROC LGM simulation. Black crosses are localities used in the analysis. Hashed areas indicate major mountain ranges. The continental shelf exposed due to lower sea-levels during the LGM is shaded in grey.

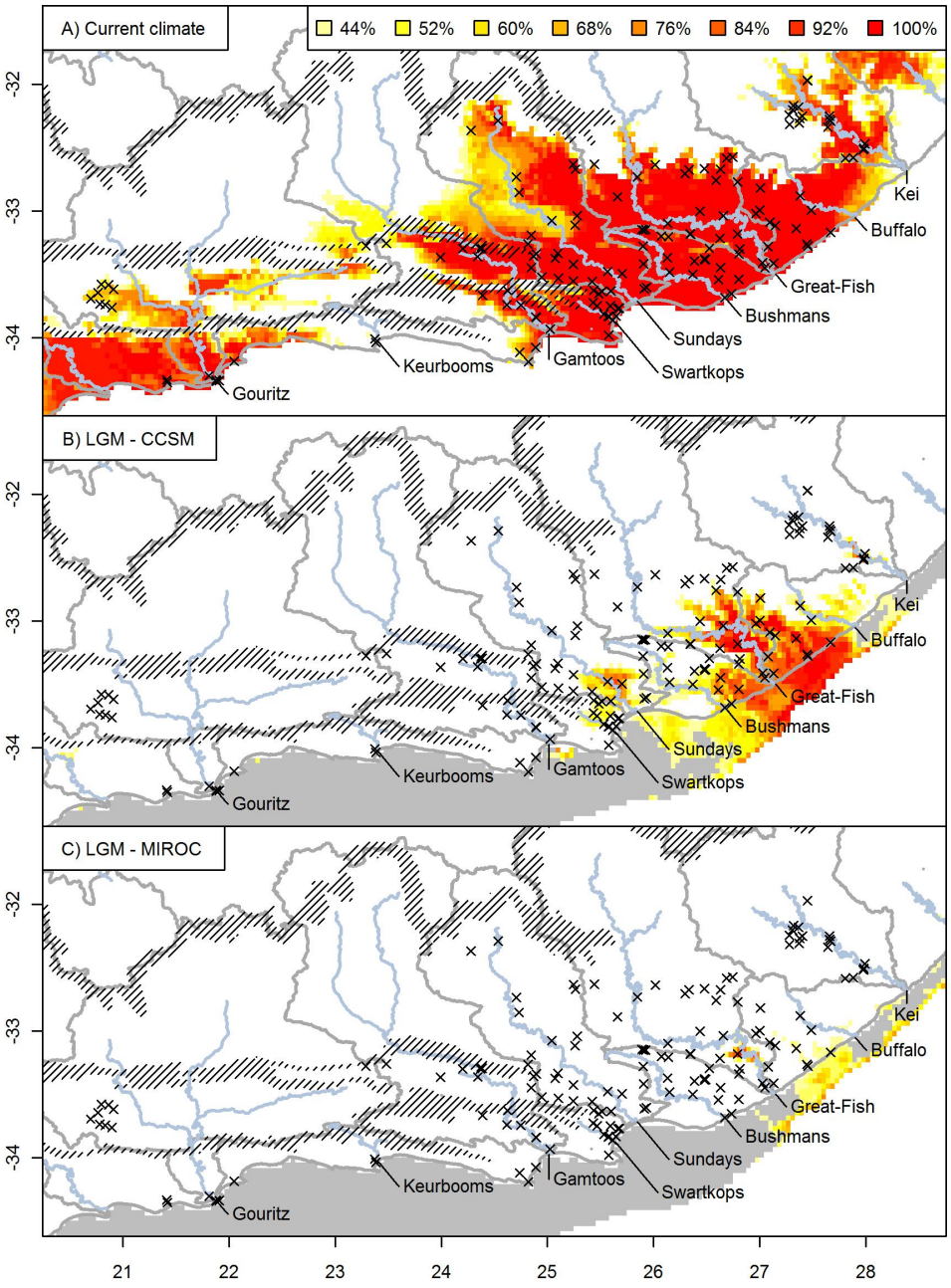

Figure 5.13. The modelled present and Last Glacial Maximum areas of suitable climate for *Schotia afra* within the Albany Subtropical Thicket (AST). The summary of modelled distributions from five pseudo-replicate datasets and K-folding subsets (see text) representing the climatically suitable areas in the AST under (A) present conditions, (B) the CCSM Last Glacial Maximum (LGM; 21,000 BP) simulation, and (C) the MIROC LGM simulation. Black crosses are localities used in the analysis. Hashed areas indicate major mountain ranges. The continental shelf exposed due to lower sea-levels during the LGM is shaded in grey.

5.5. Discussion

5.5.1. Contrasting history and phylogeographic patterns

Pappea capensis and *S. afra* are dominant tree species in AST vegetation (Mucina & Rutherford 2006, Vlok *et al.* 2003) that share a largely overlapping distribution within this biome. The species also have similar life history strategies related to their growth forms, with a slow growth strategy (Holmes & Cowling 1993). As a result, they are both middle to late colonisers in thicket succession (e.g. Jordaan 2010). Despite these similarities, these two species display very contrasting phylogeographic patterns, demographic histories and glacial period distributions through the Pleistocene.

Although the EDDB hypothesis postulates that watersheds separating drainage basins are significant barriers to gene flow, only a handful of studies have investigated and detected such patterns in plant species (e.g. Durka 1999, Gugger *et al.* 2008, Zhang *et al.* 2011). Previous phylogeographic studies have demonstrated a strong correlation between genetic lineages and these lowland drainage basins in South Africa, ranging from aquatic organisms (Daniels *et al.* 2006, Swartz *et al.* 2007, 2009) to terrestrial invertebrates (Daniels *et al.* 2009, Price *et al.* 2010) and a tree species (Chapter 4). The chloroplast phylogeography of *P. capensis* adds further evidence for the EDDB hypothesis as the strong correlation between haplotypes and drainage basins suggest that the watersheds along the coastal lowlands have acted as long-term barriers to gene flow (Figure 5.3). The phylogenetic patterns in the nDNA do not display the same deep divergences between AST basins as found in the cpDNA where lineages are restricted to specific drainage basins. Nonetheless, the results do indicate a level of phylogeographic structuring (Figure 5.6), which is similar to the cpDNA findings when the spatial aspect of genetic variability is taken into account (i.e. sPCA, Figure 5.11.C). These phylogeographic analyses provide evidence that *P. capensis* is comprised of distinct and isolated populations along the AST coastal lowlands.

The variance between the cpDNA and nDNA in *P. capensis* results may have arisen from differences in either the coalescent time between these two genomes or in the dispersal rates and distances between seeds and pollen. As the rate of coalescence is tied to genetic drift which is determined by population size, coalescent theory predicts that a larger effective population size will result in a far longer period for clear phylogenetic or phylogeographic patterns to emerge after speciation or population

fragmentation events (Avise 2000, Avise & Wollenberg 1997, Palumbi *et al.* 2001). A diploid nuclear genome is generally considered to have an effective population size four times larger than that of a haploid and uniparentally-inherited genome (Ballard & Whitlock 2004), such as the chloroplast genome. Thus, lineage sorting may occur much faster in the chloroplast genome than the nuclear genome; the difference between the rates of lineage sorting between the two genomes increases with a decline in effective population size. In addition, in the case of ITS which is comprised of numerous repeats and a lack of concerted evolution which homogenises repeats, the effective population size of ITS (nDNA) may be far greater than low-copy nuclear regions (discussed in Chapter 3). In angiosperms, seeds carry both chloroplast and nuclear DNA (Birky 1995) whereas pollen exclusively carries nuclear DNA (although there are many exceptions, e.g. Ellis *et al.* 2008, Reboud & Zeyl 1994, Testolin & Cipriani 1996). Thus, pollen flow may increase the dispersal rate and distance of the nuclear genome over the chloroplast genome. Discerning the contribution that these processes have had towards the discrepancies observed in the phylogeographic patterns between the two genomes is, unfortunately, beyond the scope of the currently available data.

No significant genetic divergence (i.e. monophyly) between *Schotia* species is observed in the chloroplast dataset as several haplotypes are shared between species (Figure 5.4), similar to the findings of Ramdhani *et al.* (2010). However, unlike the earlier study, samples of *S. afra* are separated into a highly divergent and well supported cluster in the nDNA (Figure 5.8). The discrepancy between the ITS results presented here and those from the earlier study are discussed further below. Although coalescent times are far shorter for the chloroplast genome compared to the nuclear genome (as discussed above), if the ancestral population was genetically depauperate - possibly due to a population bottleneck - then genetic divergence may be driven by mutation rather than lineage sorting. A genetically depauperate ancestor would restrict the degree to which lineage sorting results in genetically differentiated species. Thus, the lack of differentiation between *Schotia* species may simply be due to the slow accumulation of mutations in the chloroplast genome (Schaal *et al.* 1998). The ITS region has a much faster rate of mutation (Kay *et al.* 2006), and this may explain why genetic divergences among species of *Schotia* is detected in this region and not in the cpDNA. Based on the well-supported nDNA cluster of *S. afra*, this study was restricted to phylogeographic analyses and discussion of this species within the Albany Subtropical Thicket.

Schotia afra does not display any strong drainage basin affiliations in the cpDNA dataset, with the exception of the inland valley known as the Little Karoo in the Gouritz drainage basin where only haplotype *k* is found (Figure 5.4). These samples are also found close together in cluster *I* of the nDNA phylogenetic network (Figure 5.8). Beyond this basin, there is a lack of geographic structuring in both cpDNA and nDNA across the Albany Subtropical Thicket. This is supported by the low Mantel R values (Table 5.1) and the lack of spatial autocorrelation (Figure 5.10.B). The sPCA suggests a clinal pattern of genetic variation in the ITS data (Figure 5.11.D), with the western Gouritz and eastern Kei basin samples sharing sPCA space; however there are no clear association with remaining drainage basins and samples in sPCA space (inferred by samples with similar colouring) in comparison to what is found in *P. capensis*. Therefore, unlike other lowland species, the watersheds separating primary drainage basins have not played a significant role in isolating populations of *S. afra*.

The cyclic Pleistocene climate changes are suggested to have driven shifts in species distributions globally (Jansson & Dynesius 2002), as well as genetic divergence in the coastal lowlands of South Africa (*Nymania capensis*, Chapter 4; Price *et al.* 2010, Swart *et al.* 2009). The molecular dating of the chloroplast haplotypes under an extremely wide, but plausible, range of substitution rates suggest that the genetic variability arose or diverged during the Pleistocene in both *P. capensis* and *S. afra*. This suggests that the cycles of glacial and interglacial periods have played an important role in the demographic history of both species. However, these two species have contrasting demographic histories with only *S. afra* displaying a signature of an expanding population (Figure 5.9), whereas *P. capensis* has a low and isolated haplotype diversity suggestive of recent bottlenecks in isolated refugia (Figure 5.3). The SDMs provide further insights into this demographic history through the postdicted distributions of these species during the Last Glacial Maximum.

There is strong concordance between the phylogeographic results and the predicted LGM distributions using Maxent and the two global climate models. For *P. capensis* the three major refugial areas postdicted in both GCMs correspond to the distribution of the three largely allopatric chloroplast lineages (Figures 5.3.D and 5.12). Thus, *P. capensis* likely retreated into a number of small refugia within the drainage basins during glacial periods. Coupled with the demographic results and low genetic diversity, this suggests that *P. capensis* probably also experienced population bottlenecks within these isolated refugia which reduced the genetic diversity within the

populations. These findings are in agreement with the predictions made in Chapter 2 where arid and valley AST subtypes experienced severe range reductions with an overall decline in altitude and concomitant fragmentation into primary drainage basins. The LGM distribution of *S. afra* is suggested to be restricted to a single refugium along the eastern AST which extends from the Great-Fish basin onto the exposed continental shelf (Figure 5.13). This suggests that this species survived in a single and connected refugium during the Last Glacial Maximum. Given the high haplotype and ribotype diversity, it is likely that this single refugium supported a large population that could maintain both a high level of genetic diversity and a signal of Pleistocene population expansion. However, genetic diversity can also be generated when a population rapidly expands (Aris-Brosou & Excoffier 1996, Edmonds *et al.* 2004). The signal of demographic expansion sits firmly within the Quaternary under a wide range of plausible mutation rates (Table 5.7). This expansion most likely occurred during the Pleistocene and is unlikely to be post-LGM - a possibility under the 'fast' mutation rate - as *S. afra* is a slow growing tree with a long generation time. The *S. afra* refugium on the exposed coastal shelf is not postdicted by the climate envelope modelling of the vegetation subtypes (arid and valley thicket in Chapter 2). This provides a valuable demonstration that modelling of vegetation subtypes may not extend to all species within those subtypes and that such results should be treated with caution (Midgley *et al.* 2002).

The common chloroplast haplotypes of *S. afra* generally show a dispersed pattern across the Albany Subtropical Thicket. However, there are unique haplotypes restricted to the intermontane valley (Little Karoo) within the Gouritz basin. Furthermore, no area of suitable climate is postdicted for the Little Karoo during the LGM (Figure 5.13). These unique chloroplast haplotypes (Figure 5.4) and the close association of these samples in the *S. afra* cluster (cluster *I*) in the ITS phylogenetic network (Figure 5.8) suggest that this population either survived in an isolated refugium in this basin, or formed from a post-LGM founder event with no subsequent gene flow into this population. The Little Karoo basin is topographically very complex and suitable climatic conditions may have existed at a local scale to support a population. Alternatively, the Little Karoo is bordered by high mountains (these watersheds are much higher than in the eastern AST) which may have restricted seed flow into this population. There is no way to differentiate between these hypotheses given the current data. Nonetheless, the lack of any correspondence between genetic

diversity and geography across the remainder of the AST, coupled with the postdicted eastern single refugium, suggests that this species has easily dispersed over watersheds. Why then does *S. afra* have no geographic boundaries whilst the AST watersheds seem to structure genetic diversity in *P. capensis*?

A likely explanation for this difference is that, although these species share similar life history strategies and endozoochorous dispersal, they have very different dispersal vectors. The majority of woody AST species are thought to be adapted for endozoochorous seed dispersal by mammals, birds, or both (Castley *et al.* 2001, Cowling 1983, Cowling *et al.* 1997, Sigwela 2004, Watson 2002). These two species lie close to the extremes along the bird-dispersal and mammal-dispersal continuum, as the small red fruits of *P. capensis* are primarily bird-dispersed (Sigwela 2004), whereas the large Fabaceae pods of *S. afra* are likely to only be dispersed by larger mammals, including megaherbivores (Savannah Elephants and Black Rhino). Long-distance dispersal by birds is generally very limited (Jordano *et al.* 2007) because of territoriality that limits bird movement and birds have a small, very efficient, intestinal tract which means that seeds have a short residence time in the gut (Kays *et al.* in press, Westcott *et al.* 2005). In contrast, long-distance dispersal by large herbivores is likely to be fairly unrestricted across the Albany Subtropical Thicket. Savannah Elephants are likely to be prominent vectors for long-distance dispersal; their large volume of forage, limited mastication, long residency time in the gut coupled with relatively poor digestion, large home-ranges (especially along the coast lowlands, Boshoff *et al.* 2001) and periodic migrations makes them ideal vectors for seed dispersal (Campos-Arceiz & Blake in press). Furthermore, large-scale mammal migrations (including mega-herbivores) along the coastal lowlands and exposed continental shelf during glacial periods are postulated as large herbivores are thought to have tracked the seasonal rainfall between the western winter-rainfall and eastern summer-rainfall regimes (Compton 2011, Marean 2010). The general lack of geographic structuring in *S. afra* may be the first evidence for long-distance migrations along the coastal lowlands.

5.5.2. *Pappea capensis* in the AST in relation to the rest of the species' distributions

Pappea capensis has a widespread distribution throughout Africa (Figure 5.3.B), but is usually only locally dominant (e.g. Siebert & Eckhardt 2008, Werger & Coetzee 1977) or restricted to thicket clumps on termite mounds (Bloesch 2008). It is only in the AST that *P. capensis* reaches high regional densities and is a dominant component of the vegetation (Vlok *et al.* 2003). The AST is comprised of three very divergent cpDNA lineages. Sampling along the eastern drainage basins lacks the thoroughness of that performed for the AST; nonetheless, such deep divergences are not evident across the eastern lowlands and an ancestral haplotype (haplotype G) spans a wide range of drainage basins. The split between haplotype F and the remainder of clade 3 is likely a sampling issue rather than a significant divergence. Further sampling is required to establish if a drainage basin pattern of genetic diversity exists along the eastern lowlands; the current, although limited, distribution of haplotypes suggests that this may be the case. The deep divergence observed in cpDNA and nDNA between the coastal lowland clades and the samples from the rest of Africa suggest a long-term isolation between them. The close geographic distance between clades (see haplotypes M, L, U and T in Figure 5.3.C) suggests that this is not an 'isolation by distance' phenomenon, but rather long-term and most likely allopatric divergence with subsequent expansion. With the evidence from the current sampling, the deep split between the southern (Clades 1 and 2) and northern clades (Clades 3 and 4) would justify splitting this monotypic genus into at least two species based on the phylogenetic and evolutionary species concepts (Mishler & Theriot 1997, Wiley 1978). Further sampling in the possible contact zone may detect interbreeding between these southern and northern *P. capensis* clades. If a contact zone is established and no interbreeding is detected, then the splitting would be further justified under the biological species concept (Mayr 1940).

5.5.3. *Schotia afra* in relation to the rest of the *Schotia* species

The cpDNA and nDNA results generally support the findings of Ramdhani *et al.* (2010), specifically that there is limited congruence between cpDNA haplotypes and nDNA cluster (Figure 5.8), and this congruence occurs only between samples found in the eastern distribution of the genus. Ramdhani *et al.* (2010) cite hybridisation and

incomplete lineage sorting as an explanation for the lack of genetic resolution between the four *Schotia* species. In this study, the sampling of the *Schotia* species, other than *S. afra*, is fairly poor. Nonetheless, further insights can be garnered. Ramdhani *et al.* (2010) state that 'no double peaks or ambiguous base calls were found in [their ITS] electropherograms'. However, the ITS dataset in this study is rich in 2ISPs (Table 5.5), and the presence of 2ISPs was confirmed via cloning (Table A.5). This failure to detect overlapping peaks in chromatograms is possibly because neither high fidelity taq polymerase nor cloning was used in the earlier study (see Chapter 3 for the importance of using these two techniques with ITS). Detecting 2ISPs has provided a more resolved and accurate reflection of the ITS relationships between samples, and this may be the reason why supported monophyly in *S. afra* is now detectable in the nDNA under NJ and MP (Appendix Figure A.16, Pg. 249). Furthermore, given the high level of 2ISPs present in the nDNA dataset, it is likely that concerted evolution (the process where ITS variants are homogenised to a single sequence) is either not in effect or is very weak. Thus, any recent and possibly also fairly old hybridisation events should be evident in the chromatograms (e.g. Carine *et al.* 2007). Given the extensive sampling of *S. afra* in the AST, if this species and *S. latifolia* were hybridising freely and prolifically, then we should observe *S. afra* hybrid samples that share mutations between these clusters which would then fall in intermediary positions in the network between cluster *I* and the remaining clusters (Huson & Bryant 2006). The lack of any significant genetic divergence between *S. afra* and *S. latifolia* in the cpDNA dataset is most likely attributable to slow substitution rates, a genetically depauperate ancestor, incomplete lineage sorting, or a combination of these, rather than hybridisation or chloroplast capture. Hybridisation is probably restricted to a few isolated instances and areas (Ramdhani *et al.* 2010, Ross 1977).

5.5.4. Conclusions

I have investigated the phylogeography of two AST tree species, *P. capensis* and *S. afra*, which has revealed cryptic divergence that has not been detected by traditional taxonomy. This cryptic divergence has revealed that these two species with similar life history strategies and distribution in the AST have had very contrasting evolutionary histories during the Pleistocene climate cycling. Furthermore, the contrasting phylogeographic patterns are attributed to the differences in seed dispersal vectors, where drainage basin watersheds are significant barriers to the bird-dispersed *P.*

capensis while no major geographical barriers exist for the mega-herbivore dispersed *S. afra*. If these results represent generalities for these two syndromes in the AST, then they provide a new understanding of the evolutionary history and ecology of this biome's flora.

6. Synthesis

The Albany Subtropical Thicket (AST) is an ancient, widespread and once dominant vegetation which has suffered severe range reductions after the evolution of the fire-driven biomes, such as the fynbos and savannah (Cowling *et al.* 2005). The fate of AST vegetation during the global Pleistocene glacial-interglacial climate cycles within the regional setting of a complex landscape has remained largely speculative. In this thesis, I have used a combination of techniques to explore the history of AST vegetation, including detailed phylogeographic and distribution modelling analyses of three selected taxa. The results provide support for the glacial refugia and the evolutionarily distinct drainage basin (EDDB) hypotheses. Comparative phylogeography, coupled with a consideration of seed dispersal modes, provides novel insights into important ecological processes within the AST, particularly the effects of different vectors on long distance dispersal and ultimately the metapopulation dynamics of a tree species. These evolutionary and ecological insights also provide support for established AST conservation plans. I discuss each of these points in further detail below.

6.1. The glacial refugia hypothesis

The glacial refugia hypothesis postulates that AST vegetation experienced significant range retractions during the glacial periods of the Pleistocene (Cowling *et al.* 2005). Support for this hypothesis has, up until now, been in the form of field observations and fossil evidence from a single locality. Many AST plant species, especially succulents, have been observed to be intolerant to frost (Richard Cowling, personal communication). This, coupled with lower global temperatures during glacial periods (Zachos *et al.* 2001) that would have increased the range and severity of frost events suggests that AST vegetation may have experienced range reductions. The fossil sequence from the Boomplaas Cave, located in the Little Karoo, also reveals a reduction in thicket, as grazers dominated during the Last Glacial Maximum (LGM) in

an area that is now monopolised by browsers that are associated with AST vegetation (Faith in press, Klein 1980). Thus, the community distribution modelling (CDM) of the AST subtypes offers the first regional overview of potential range retractions during the most recent, and extreme, glacial period (Chapter 2). The CDM postdicts dramatic reduction of the area of suitable climate for all AST subtypes during the LGM, with concomitant fragmentation into drainage basins for the arid and valley subtypes. The CDM also provides possible locations of LGM refugia. However, the exact locations of these refugia of suitable climate must be considered with caution. Some cornerstone assumptions such as community coherence and relatively static community and species niches may very well be violated. If so, any environmental changes would result in an unpredictable shuffling and shifting of species. Furthermore, issues surrounding the modelling algorithms or global climate models used may further obscure the exact whereabouts of glacial refugia (Buisson *et al.* 2010, Heikkinen *et al.* 2006, Pearson *et al.* 2006). Unfortunately, there are insufficient palaeoarchives to provide an independent test of the CDM postdictions (e.g. Carnaval & Moritz 2008). However, limited support is found in the species distribution modelling (SDM) and molecular results presented in the subsequent chapters.

One means of testing the generalities of CDM projections is to compare these with projections derived from component species (Midgley *et al.* 2002). Unfortunately, most AST species, as well as those from many other South African biomes, lack sufficient and accurate locality information required for distribution modelling. The locality information collected for the three target species of this thesis are the only readily available datasets with sufficient distribution of localities to perform distribution modelling. Nonetheless, the SDMs of suitable climate under LGM conditions derived from *Nymania capensis* and *Pappea capensis* (Chapters 4 and 5) offer support for the CDM postdictions. The LGM SDMs for *Nymania capensis* and *P. capensis*, which are common in both arid and valley AST subtypes (Vlok & Euston-Brown 2002, Vlok *et al.* 2003), are qualitatively very similar to the joint CDM distributions for the arid and valley AST subtypes (compare Figures .2.4 [Pg. 40], 4.10 [Pg. 130] and 5.12 [Pg. 172]). The only main discrepancy is that for both species there is a much greater area of suitable climate postdicted in the Little Karoo (Gouritz basin) than suggested by CDM results. The SDM of *Schotia afra*, which is also common in arid and valley AST subtypes, does display some overlap between the CDM valley thicket subtype for the LGM; however, it also shows a significant extension onto the exposed coastal shelf that

was not postdicted by the CDM of these subtypes. The *S. afra* projection provides a valuable cautionary demonstration that the CDM climate envelope may not be representative of the community's component species. The generality of CDM results should be further investigated using wide range of AST species (e.g. Loarie *et al.* 2008), which should be focussed on narrowly distributed endemics that are at greatest risk of their climate envelope not being incorporated into that of the community distribution model (Midgley *et al.* 2002).

Phylogeography also provides some support for the refugia postdicted by the community distribution modelling. Three expectations were derived from the LGM climate refugia suggested by CDM regarding the distribution of genetic diversity, specifically that: 1) the highest genetic diversity should reside in the refugial areas, 2) clade or population subdivision should be evident in species that occur in arid and valley AST subtypes, and 3) there should be signatures of post-glacial population expansion. The slow rate of mutation in chloroplast DNA and increased coalescent times for nuclear DNA, coupled with limited sampling from these genomes in my datasets, means that there is insufficient resolution to fully explore expectations one and three. Nonetheless, the second expectation is met by *N. capensis* and *P. capensis* as each have a pattern of phylogeographic fragmentation across drainage basins that coincides with postdicted CDM refugia, despite the present-day continuous swathes of AST vegetation that span multiple drainage basins and watersheds.

The glacial refugia hypothesis is supported by community and species distribution modelling, as well as phylogeography. The CDM results offer the first regional hypothesis for the location of these refugia. This has received positive, if still limited, support. The hypothesised locations of refugia can be tested using further studies of phylogeography, palaeoarchives and fine-scale distribution mapping of AST species, especially endemics. Composite maps of species distributions may also identify areas of glacial refugia, as these should contain greater number of AST endemics and higher species richness in comparison to areas of post-glacial colonisation if species have been dispersal limited.

6.2. The evolutionarily distinct drainage basin hypothesis and long distance dispersal

The watersheds along the coastal lowlands are hypothesised to constitute dispersal barriers for a number of faunal species (Price *et al.* 2010, Swartz *et al.* 2007, 2009). Both CDM of AST subtypes and SDM of the three target species suggest that these barriers would have been strengthened during Pleistocene glacial periods as areas of suitable climate contracted into refugial areas within drainage basins (Figures 2.4 [Pg. 40], 4.10 [Pg. 130] and 5.12 [Pg. 172]). The results from both *N. capensis* and *P. capensis* provide the first phylogeographic evidence that the EDDB hypothesis extends to AST flora. However, the phylogeographic patterns of *S. afra* again provides a valuable, and cautionary, contrast demonstrating that watersheds are not dispersal barriers for all AST species. I propose that the primary reason for these contrasting patterns is because of the variance in long-distance dispersal (LDD) capabilities of different vectors.

Migrating megaherbivores are likely to be the dominant vectors for LDD of seeds in topographically complex landscapes (Nathan *et al.* 2008). Long-distance dispersal is important for the establishment of a plant species in new sites and, depending on its frequency, the subsequent maintenance of metapopulation dynamics and ultimately genetic integrity. The phylogeographic results from the three largely co-distributed tree species within the AST biome, each with a unique and specific dispersal vector, provide the first regional exploration for the genetic effects of dispersal vectors. There are no barriers to gene flow across this biome in *S. afra* which has seeds dispersed by megaherbivores (Chapter 5). In contrast, *P. capensis* and *N. capensis*, which have seeds primarily dispersed by birds or wind, respectively, have been isolated in drainage basins and unable to disperse seeds across the watersheds (Chapters 4 and 5).

The use of genetic methods to study long-distance seed dispersal is well established (Cain *et al.* 2000). However, comparative studies contrasting seed dispersal syndrome and the effect this has on the genetic diversity of co-distributed plants are lacking. The results from Chapters 4 and 5 provide extremely promising examples demonstrating that phylogeography can be use to elucidate not only historical aspects of AST taxa (Cowling *et al.* 2005) but also ecological aspects such as long distance dispersal. Further studies using this comparative method should finally allow us to get an evolutionary handle on all the aspects of seed dispersal, from dispersal syndromes,

dispersal vectors, and to population dynamics of AST flora.

The *S. afra* phylogeography gives us the first indication of the importance of megaherbivores for LDD of seeds across the African landscape, especially in the dissected landscape of the coastal lowlands. The importance of seed dispersal by megaherbivores is an old, repeated, and untested hypothesis (e.g. Burtt & Salisbury 1929, Dudley 2000). Furthermore, in the light of anticipated climate change, the present restricted movement of large mammals and predicted range shifts in plant species, this highlights the importance of the free movement of megaherbivores for seed dispersal for the maintenance of ecological and evolutionary processes (Cowling & Pressey 2001).

6.3. Phylogeography and conservation

The long-term isolation within drainage basins observed in *N. capensis* and *P. capensis* and the importance of mega-herbivores for LDD of *S. afra* adds valuable support to conservation efforts within the Albany Subtropical Thicket. The EDDB hypothesis is the basis of the hierarchical classification used by Vlok *et al.* (2003) to characterise and delimit the AST vegetation. These authors based the importance of drainage basins on the biogeographical distribution of endemics. However, the environmental drivers of these speciation events that have led to the current distribution of endemics may be ancient, and highly disparate from those of the current epoch. Thus, the population-level fragmentation across drainage basins offers the first intra-specific, and thus evolutionarily recent, support for the EDDB hypothesis. This suggests that environmental, ecological and evolutionary drivers have remained relatively stable through recent geological history and adds additional support for the acknowledgement of drainage basins as unique floristic entities.

The importance of expert-knowledge is also highlighted by the phylogeographic support of the EDDB hypothesis. Primary drainage basins were treated as discrete entities not only by the vegetation classification and mapping performed by Vlok *et al.* (2003) but also the conservation planning of a mega-conservancy network for the AST biome (Rouget *et al.* 2006). The mega-conservancy network aims to capture both ecological and evolutionary processes within the Albany Subtropical Thicket. In order to do this, within-basin environmental gradients were deemed the most important.

There was limited scientific support for including drainage basins in such a manner and the decision to use this approach was primarily based on expert-knowledge (Rouget *et al.* 2006, Vlok *et al.* 2003). Both the vegetation mapping and mega-conservancy network planning were part of the Subtropical Thicket Ecosystem Planning Project, a four year initiative (July 2000 - June 2004) funded by the Global Environment Facility. The detailed mapping and conservation planning are impressive given the short time period and that the AST has been a largely neglected biome in terms of research, primarily due to the historical uncertainty regarding its floral affinities. Thus, the phylogeographic support, which is limited, costly and time consuming, demonstrates the value of expert knowledge in kick-starting major conservation efforts that are scientifically-sound.

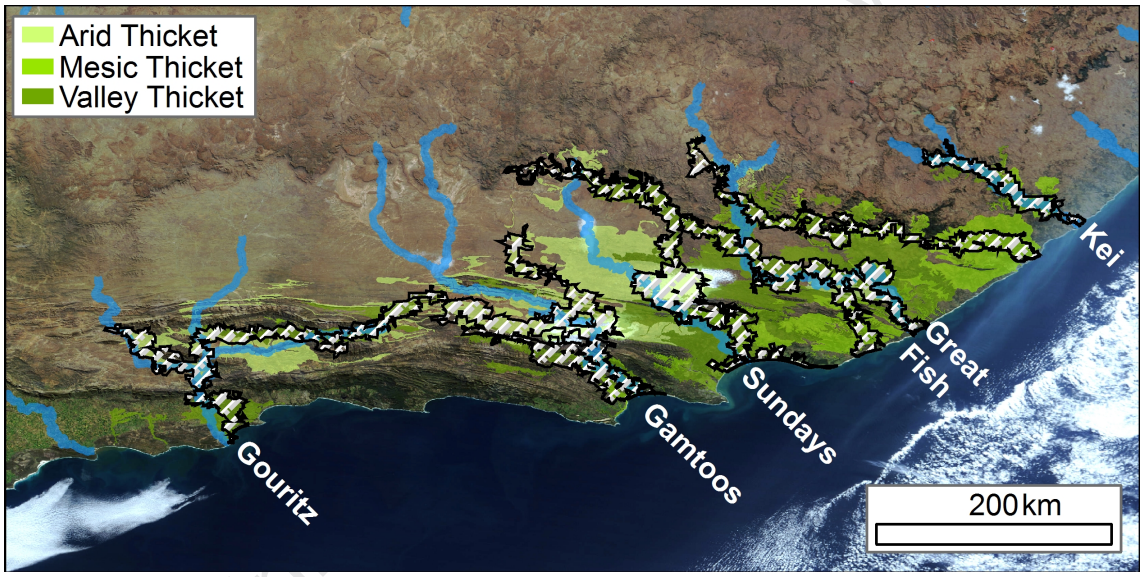

Figure 6.1. The megaconservancy network for the Albany Subtropical Thicket proposed by Rouget *et al.* (2006). The major coastal lowland rivers are shown.

Megaherbivores, primarily elephants, have been identified as keystone species with regards to disturbance in the Albany Subtropical Thicket (Kerley *et al.* 1995, 2002). Consequently, elephant suitability was one of the four critical components driving the systematic conservation planning (Rouget *et al.* 2006). However, the proposed mega-conservancy network does not presently incorporate migratory movements across the drainage basins (Figure 6.1). This is quite likely because the historical role that mega-herbivores have played in extensive LDD of seeds and the consequence this had for

metapopulation dynamics could not have been predicted, and was a surprising and unexpected finding of this thesis. If the results found for *S. afra* are mirrored in other plant species that target mega-herbivores as dispersal vectors, then corridors that span between the established networks that allow mega-herbivore migration may be an important feature for any updates or extensions to the current conservation plan.

6.4. Final thoughts and personal reflections

The phylogeographic results offered in this thesis are the tip of a very interesting iceberg. They provide a largely narrative understanding of the evolutionary and ecological history of the unique AST vegetation, which we have only started to explore relatively recently. Each species examined offers only a single descriptive glimpse into the history of the Albany Subtropical Thicket. A more comprehensive and statistically-robust overview may be obtained using more species (i.e. comparative phylogeography, Avise 2000) and model-based methods (reviewed in Bloomquist *et al.* 2010). Model-based methods that make use of coalescence theory offer even more refined inferences into the history of a species (e.g. Hey & Nielsen 2004, Lemey *et al.* 2009). However, accurate parameter estimates using such methods require signal richness in datasets that is far beyond what has been detected in this study (e.g. Won & Hey 2005). Nonetheless, the molecular era is marching on and advances in new techniques that allow for whole genome sequencing should provide the required data for non-model organisms that is not prohibitively expensive (Mardis 2008, Shendure & Ji 2008). Thus, in order to further explore the history of the AST in a statistical framework, the sample size, in terms of species and DNA data examined, will have to be increased. With this end in mind, I hope that this thesis will act as a catalyst for further phylogeographic studies of the AST flora. I envisage that this would provide a valuable understanding of, *inter alia*, the following: 1) the role of topography in driving population and species divergence along the coastal lowlands, 2) the genetic and evolutionary ramifications of targeting different dispersal vectors by AST plants, and 3) 'biome integrity' or the degree to which AST species share an evolutionary past and thus likely to share an evolutionary future.

The advent of the molecular era offers exciting opportunities to explore the natural world in a level of detail that was previously unattainable. As methodological hurdles are overcome, there is a growing shift away from simply describing molecular

patterns towards a holistic view that incorporates approaches and data from many wide-ranging disciplines. However, biological science has experienced an unfortunate shift away from natural history, a trend driven by the perceived superiority of the hypothetico-deductive scientific model over the descriptive mode of natural history (Beehler 2010). As we delve deeper into the world of DNA molecules, our need for a comprehensive understanding of natural history becomes more and more important (e.g. Avise 2004, Avise *et al.* 2002). An example of this is readily at hand in this thesis. The support for my ideas surrounding the efficiencies of dispersal vectors for the three target species was not based on any available literature, simply because there is none. Simple field observations of natural history, such as which animals eat the fruit of this or that plant species, are rarely deemed of sufficient scientific merit to be worthy of publication. Without the knowledge of natural history provided by Richard Cowling, Jan Vlok, Ayanda Sigwela and various farmers – which is based on the simple combination of observant minds and time spent in the field – I would not have been able to explore this aspect of the molecular data. Without a comprehensive understanding of natural history, the increasing resolution of molecular data available for phylogeography becomes more and more meaningless. I hope that the advances in the field of molecular ecology will rejuvenate the study of natural history and restore its rightful place at the pinnacle of biological science. Here's to spending more time in nature and away from the laboratory and my computer.

Bibliography

Abbott, R., Smith, L., Milne, R., Crawford, R., Wolff, K. & Balfour, J. (2000). Molecular analysis of plant migration and refugia in the Arctic. *Science* **289**, 1343–1346.

Acocks, J. (1953). *Veld Types of South Africa*, vol. 57 of *Memoirs of the Botanical Survey of South Africa*. Department of Agriculture, Pretoria.

Adams, K. & Wendel, J. (2005). Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology* **8**, 135–141.

Aguilar, J. & Feliner, G. (2003). Additive polymorphisms and reticulation in an ITS phylogeny of thrifts (Armeria, Plumbaginaceae). *Molecular Phylogenetics and Evolution* **28**, 430–447.

Allen, G., Soltis, D. & Soltis, P. (2003). Phylogeny and biogeography of *Erythronium* (Liliaceae) inferred from chloroplast matK and nuclear rDNA ITS sequences. *Systematic Botany* **28**, 512–523.

Alsos, I., Alm, T., Normand, S. & Brochmann, C. (2009). Past and future range shifts and loss of diversity in dwarf willow (Salix herbacea L.) inferred from genetics, fossils and modelling. *Global Ecology and Biogeography* **18**, 223–239.

Álvarez, I. & Wendel, J. (2003). Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution* **29**, 417–434.

Álvarez, I. & Wendel, J. (2006). Cryptic interspecific introgression and genetic differentiation within *Gossypium aridum* (Malvaceae) and its relatives. *Evolution* **60**, 505–517.

Araújo, M. & New, M. (2007). Ensemble forecasting of species distributions. *Trends in Ecology and Evolution* **22**, 42–47.

Araújo, M. & Rahbek, C. (2006). How does climate change affect biodiversity? *Science* **313**, 1396–1397.

Araújo, M., Cabeza, M., Thuiller, W., Hannah, L. & Williams, P. (2004). Would climate change drive species out of reserves? An assessment of existing reserve-selection methods. *Global Change Biology* **10**, 1618–1626.

Arezi, B., Xing, W., Sorge, J. & Hogrefe, H. (2003). Amplification efficiency of thermostable DNA polymerases. *Analytical Biochemistry* **321**, 226–235.

Aris-Brosou, S. & Excoffier, L. (1996). The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. *Molecular Biology and Evolution* **13**, 494–504.

Arnheim, N., Krystal, M., Schmickel, R., Wilson, G., Ryder, O. & Zimmer, E. (1980). Molecular evidence for genetic exchanges among ribosomal genes on nonhomologous chromosomes in man and apes. *Proceedings of the National Academy of Sciences, U.S.A.* **77**, 7323–7327.

Avise, J. (1998). The history and purview of phylogeography: a personal reflection. *Molecular Ecology* **7**, 371–379.

Avise, J. (2000). *Phylogeography: the History and Formation of Species*. Harvard University Press, Cambridge, Massachusetts.

Avise, J. (2004). *Molecular markers, natural history, and evolution*. Sinaueer Associates, Inc, Sunderland.

Avise, J. & Wollenberg, K. (1997). Phylogenetics and the origin of species. *Proceedings of the National Academy of Sciences, U.S.A.* **94**, 7748–7755.

Avise, J., Arnold, J., Ball, R., Bermingham, E., Lamb, T., Neigel, J., Reeb, C. & Saunders, N. (1987). Intraspecific phylogeography – the mitochondrial-DNA bridge between population-genetics and systematics. *Annual Review of Ecology and Systematics* **18**, 489–522.

Avise, J., Jones, A., Walker, D., DeWoody, J., Dakin, B., Fiumera, A., Fletcher, D., Mackiewicz, M., Pearse, D., Porter, B. *et al.* (2002). Genetic mating systems and natural history of fishes: lessons for ecology and evolution. *Annual Review of Genetics* **36**, 19–45.

Bailey, C., Carr, T., Harris, S. & Hughes, C. (2003). Characterization of angiosperm nrDNA polymorphism, paralogy, and pseudogenes. *Molecular Phylogenetics and Evolution* **29**, 435–455.

Baldwin, B., Sanderson, M., Porter, J., Wojciechowski, M., Campbell, C. & Donoghue, M. (1995). The ITS region of nuclear ribosomal DNA: a valuable source of evidence on Angiosperm phylogeny. *Annals of the Missouri Botanical Garden* **82**, 247–277.

Ballard, J. & Whitlock, M. (2004). The incomplete natural history of mitochondria. *Molecular Ecology* **13**, 729–744.

Barr, C., Neiman, M. & Taylor, D. (2005). Inheritance and recombination of mitochondrial genomes in plants, fungi and animals. *New Phytologist* **168**, 39–50.

Beaumont, L., Hughes, L. & Pitman, A. (2008). Why is the choice of future climate scenarios for species distribution modelling important? *Ecology Letters* **11**, 1135–1146.

Beehler, B. (2010). The forgotten science: a role for natural history in the twenty-first century? *Journal of Field Ornithology* **81**, 1–4.

Beheregaray, L. (2008). Twenty years of phylogeography: the state of the field and the challenges for the Southern Hemisphere. *Molecular Ecology* **17**, 3754–3774.

Beheregaray, L. & Caccone, A. (2007). Cryptic biodiversity in a changing world. *Journal of Biology* **6**, 9.

Belahbib, N., Pemonge, M., Ouassou, A., Sbay, H., Kremer, A. & Petit, R. (2001). Frequent cytoplasmic exchanges between oak species that are not closely related: *Quercus suber* and *Q.ilex* in Morocco. *Molecular Ecology* **10**, 2003–2012.

Bergh, N., Hedderson, T., Linder, H. & Bond, W. (2007). Palaeoclimate-induced range shifts may explain current patterns of spatial genetic variation in renosterbos (*Elytropappus rhinocerotis*, Asteraceae). *Taxon* **56**, 393–408.

Birky, C. (1995). Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. *Proceedings of the National Academy of Sciences, U.S.A.* **92**, 11331–11338.

Bloesch, U. (2008). Thicket clumps: a characteristic feature of the Kagera savanna landscape, East Africa. *Journal of Vegetation Science* **19**, 31–44.

Bloomquist, E., Lemey, P. & Suchard, M. (2010). Three roads diverged? Routes to phylogeographic inference. *Trends in Ecology and Evolution* **25**, 626–632.

Bond, W. (2008). What limits trees in C_4 grasslands and savannas? *Annual Review of Ecology, Evolution, and Systematics* **39**, 641–659.

Bond, W., Midgley, G. & Woodward, F. (2003). The importance of low atmospheric CO_2 and fire in promoting the spread of grasslands and savannas. *Global Change Biology* **9**, 973–982.

Born, J., Linder, H. & Desmet, P. (2007). The Greater Cape Floristic Region. *Journal of Biogeography* **34**, 147–162.

Boshoff, A., Kerley, G. & Cowling, R. (2001). A pragmatic approach to estimating the distributions and spatial requirements of the medium- to large-sized mammals in the Cape Floristic Region, South Africa. *Diversity and Distributions* **7**, 29–43.

Bowcock, A. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**, 455.

Braconnot, P., Otto-Bliesner, B., Harrison, S., Joussaume, S., Peterschmitt, J., Abe-Ouchi, A., Crucifix, M., Driesschaert, E., Fichefet, T., Hewitt, C. *et al.* (2007). Results of PMIP2 coupled simulations of the Mid-Holocene and Last Glacial Maximum – Part 1: experiments and large-scale features. *Climate of the Past,* **3**, 261–277.

Broennimann, O., Thuiller, W., Hughes, G., Midgley, G., Alkemade, J. & Guisan, A. (2006). Do geographic distribution, niche property and life form explain plants' vulnerability to global change? *Global Change Biology* **12**, 1079–1093.

Bromham, L. & Penny, D. (2003). The modern molecular clock. *Nature Reviews Genetics* **4**, 216–224.

Brown, J. & Lomolino, M. (1998). *Biogeography*. Sinauer Associates, Sunderland, MA, 2nd edn.

Brunsfeld, S., Sullivan, J., Soltis, D. & Soltis, P. (2001). Comparative phylogeography of northwestern North America: A synthesis. In Silvertown, J. & Antonovics, J., eds., *Integrating ecological and evolutionary processes in a spatial context*, pp. 319–339. Blackwell Science, Oxford.

Bryant, D. & Moulton, V. (2004). Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* **21**, 255–265.

Buckler, E., Ippolito, A. & Holtsford, T. (1997). The evolution of ribosomal DNA divergent paralogues and phylogenetic implications. *Genetics* **145**, 821–832.

Buerki, S., Forest, F., Acevedo-Rodríguez, P., Callmander, M., Nylander, J., Harrington, M., Sanmartín, I., Küpfer, P. & Alvarez, N. (2009). Plastid and nuclear DNA markers reveal intricate relationships at subfamilial and tribal levels in the soapberry family (Sapindaceae). *Molecular Phylogenetics and Evolution* **51**, 238–258.

Buisson, L., Thuiller, W., Casajus, N., Lek, S. & Grenouillet, G. (2010). Uncertainty in ensemble forecasting of species distribution. *Global Change Biology* **16**, 1145–1157.

Burtt, B. & Salisbury, E. (1929). A record of fruits and seeds dispersed by mammals and birds from the Singida District of Tanganyika Territory. *Journal of Ecology* **17**, 351–355.

Busby, J. (1991). BIOCLIM: a bioclimatic analysis and prediction system. In Margules, C. & Austin, M., eds., *Nature conservation: cost effective biological surveys and data analysis*, pp. 64–68. CSIRO, Melbourne.

Byrne, M. (2007). Phylogeography provides an evolutionary context for the conservation of a diverse and ancient flora. *Australian Journal of Botany* **55**, 316–325.

Byrne, M., Yeates, D., Joseph, L., Kearney, M., Bowler, J., Williams, M., Cooper, S., Donnellan, S., Keogh, J., Leys, R. *et al.* (2008). Birth of a biome: insights into the assembly and maintenance of the Australian arid zone biota. *Molecular Ecology* **17**, 4398–4417.

Cain, M., Milligan, B. & Strand, A. (2000). Long-distance seed dispersal in plant populations. *American Journal of Botany* **87**, 1217–1227.

Campbell, C., Wojciechowski, M., Baldwin, B., Alice, L. & Donoghue, M. (1997). Persistent nuclear ribosomal DNA sequence polymorphism in the *Amelanchier* agamic complex (Rosaceae). *Molecular Biology and Evolution* **14**, 81–90.

Campos-Arceiz, A. & Blake, S. (in press). Megagardeners of the forest – the role of elephants in seed dispersal. *Acta Oecologica* **doi:10.1016/j.actao.2011.01.014**.

Carine, M., Robba, L., Little, R., Russell, S. & Guerra, A. (2007). Molecular and morphological evidence for hybridization between endemic Canary Island Convolvulus. *Botanical Journal of the Linnean Society* **154**, 187204.

Carnaval, A. & Moritz, C. (2008). Historical climate modelling predicts patterns of current biodiversity in the Brazilian Atlantic forest. *Journal of Biogeography* **35**, 1187–1201.

Carr, A., Thomas, D., Bateman, M., Meadows, M. & Chase, B. (2006). Late Quaternary palaeoenvironments of the winter-rainfall zone of southern Africa: Palynological and sedimentological evidence from the Agulhas Plain. *Palaeogeography, Palaeoclimatology, Palaeoecology* **239**, 147–165.

Castley, J., Bruton, J., Kerley, G. & McLachlan, A. (2001). The importance of seed dispersal in the Alexandria Coastal Dunefield, South Africa. *Journal of Coastal Conservation* **7**, 57–70.

Chase, B. & Meadows, M. (2007). Late Quaternary dynamics of southern Africa's winter rainfall zone. *Earth-Science Reviews* **84**, 103–138.

Chen, C., Durand, E., Forbes, F., Fran & Ois, O. (2007). Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Notes* **7**, 747–756.

Chen, Z. & Pikaard, C. (1997). Transcriptional analysis of nucleolar dominance in polyploid plants: Biased expression/silencing of progenitor rRNA genes is developmentally regulated in *Brassica*. *Proceedings of the National Academy of Sciences, U.S.A.* **94**, 3442–3447.

Chiang, T. & Schaal, B. (1999). Phylogeography of North American populations of the moss species *Hylocomium sylendens* based on the nucleotide sequence of internal transcribed spacer 2 of nuclear ribosomal DNA. *Molecular Ecology* **8**, 1037–1042.

Clark, A. (1990). Inference of haplotypes from PCR-amplified samples of diploid populations. *Molecular Biology and Evolution* **7**, 111–122.

Clark, P. & Evans, F. (1954). Distance to nearest neighbour as a measure of spatial relationships in populations. *Ecology* **35**, 445–453.

Clement, M., Posada, D. & Crandall, K. (2000). TCS: a computer program to estimate gene genealogies. *Molecular Ecology* **9**, 1657–1660.

Cline, J., Braman, J. & Hogrefe, H. (1996). PCR Fidelity of Pfu DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Research* **24**, 3546–3551.

Coleman, A. (2003). ITS2 is a double-edged tool for eukaryote evolutionary comparisons. *Trends in Genetics* **19**, 370–375.

Collins, W., Bitz, C., Blackmon, M., Bonan, G., Bretherton, C., Carton, J., Chang, P., Doney, S., Hack, J., Henderson, T. *et al.* (2004). The Community Climate System Model Version 3 (CCSM3). *Journal of Climate* **19**, 2122–2143.

Comes, H. & Kadereit, J. (1998). The effect of Quaternary climatic changes on plant distribution and evolution. *Trends in Plant Science* **3**, 432–438.

Compton, J. (2011). Pleistocene sea-level fluctuations and human evolution on the southern coastal plain of South Africa. *Quaternary Science Reviews* **30**, 506–527.

Cook, B., Pringle, C. & Hughes, J. (2008). Phylogeography of an island endemic, the Puerto Rican freshwater crab (*Epilobocera sinuatifrons*). *Journal of Heredity* **99**, 157–164.

Copenhaver, G. & Pikaard, C. (1996). Two-dimensional RFLP analyses reveal megabase-sized clusters of rRNA gene variants in *Arabidopsis thaliana*, suggesting local spreading of variants as the mode for gene homogenization during concerted evolution. *The Plant Journal* **9**, 273–282.

Cowling, R. (1983). Phytochorology and vegetation history in the south-eastern Cape, South Africa. *Journal of Biogeography* **10**, 393–419.

Cowling, R. & Pressey, R. (2001). Rapid plant diversification: planning for an evolutionary future. *Proceedings of the National Academy of Sciences, U.S.A.* **98**, 54525457.

Cowling, R., Kirkwood, D., Midgley, J. & Pierce, S. (1997). Invasion and persistence of bird-dispersed, subtropical thicket and forest species in fire-prone coastal Fynbos. *Journal of Vegetation Science* **8**, 475–488.

Cowling, R., Prochés, S. & Vlok, J. (2005). On the origin of southern African subtropical thicket vegetation. *South African Journal of Botany* **71**, 1–23.

Cowling, R., Prochés, S. & Partridge, T. (2009). Explaining the uniqueness of the Cape flora: incorporating geomorphic evolution as a factor for explaining its diversification. *Molecular Phylogenetics and Evolution* **51**, 64–74.

Crisp, M., Arroyo, M., Cook, L., Gandolfo, M., Jordan, G., McGlone, M., Weston, P., Westoby, M., Wilf, P. & Linder, H. (2009). Phylogenetic biome conservatism on a global scale. *Nature* **458**, 754–756.

Cronn, R., Cedroni, M., Haselkorn, T., Grover, C. & Wendel, J. (2002). PCR-mediated recombination in amplification products derived from polyploid cotton. *Theoretical and Applied Genetics* **104**, 482–489.

Cummings, M., Neel, M., Shaw, K. & Otto, S. (2008). A genealogical approach to quantifying lineage divergence. *Evolution* **62**, 2411–2422.

Daniels, S., Gouws, G. & Crandall, K. (2006). Phylogeographic patterning in a freshwater crab species (Decapoda: Potamonautidae: *Potamonautes*) reveals the signature of historical climatic oscillations. *Journal of Biogeography* **33**, 1538–1549.

Daniels, S., Picker, M., Cowlin, R. & Hamer, M. (2009). Unravelling evolutionary lineages among South African velvet worms (Onychophora: *Peripatopsis*) provides evidence for widespread cryptic speciation. *Biological Journal of the Linnean Society* **97**, 200–216.

Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. J. Murray, London.

de Sousa Queiroz, C., de Carvalho Batista, F. & de Oliveira, L. (2011). Evolution of the 5.8S nrDNA gene and internal transcribed spacers in *Carapichea ipecacuanha* (Rubiaceae) within a phylogeographic context. *Molecular Phylogenetics and Evolution* **59**, 293–302.

Deacon, H., Deacon, J., Scholtz, A., Thackeray, J. & Brink, J. (1984). Correlation of palaeoenvironmental data from the Late Pleistocene and Holocene deposits at Boomplaas Cave, southern Cape. In Vogel, J., ed., *Late Cainozoic Palaeoclimates of the Southern Hemisphere*, pp. 339–360. Balkema, Rotterdam.

Diniz-Filho, J., Mauricio Bini, L., Fernando Rangel, T., Loyola, R., Hof, C., Nogués-Bravo, D. & Araújo, M. (2009). Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. *Ecography* **32**, 897–906.

Dixon, C. (2010). OLFinder–a program which disentangles DNA sequences containing heterozygous indels. *Molecular Ecology Resources* **10**, 335–340.

Dmitriev, D. & Rakitov, R. (2008). Decoding of superimposed traces produced by direct sequencing of heterozygous indels. *PLoS Computational Biology* **4**, e1000113.

Dormann, C. (2007). Promising the future? Global change projections of species distributions. *Basic and Applied Ecology* **8**, 387–397.

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

Dubcovsky, J. & Dvorak, J. (1995). Ribosomal RNA multigene loci: nomads of the Triticeae genomes. *Genetics* **140**, 1367–1377.

Dudley, J. (2000). Seed dispersal by elephants in semiarid woodland habitats of Hwange National Park, Zimbabwe. *Biotropica* **32**, 556–561.

Durka, W. (1999). Genetic diversity in peripheral and subcentral populations of *Corrigiola litoralis* L. (Illecebraceae). *Heredity* **83**, 476–484.

Dynesius, M. & Jansson, R. (2000). Evolutionary consequences of changes in species' geographical distributions driven by Milankovitch climate oscillations. *Proceedings of the National Academy of Sciences, U.S.A.* **97**, 9115–9120.

Edmonds, C., Lillie, A. & Cavalli-Sforza, L. (2004). Mutations arising in the wave front of an expanding population. *Proceedings of the National Academy of Sciences, U.S.A.* **101**, 975–979.

Edwards, S. & Beerli, P. (2000). Perspective: Gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution* **54**, 1839–1854.

Eeley, H., Lawes, M. & Piper, S. (1999). The influence of climate change on the distribution of indigenous forest in KwaZulu-Natal, South Africa. *Journal of Biogeography* **26**, 595–617.

Ehleringer, J., Cerling, T. & Helliker, B. (1997). C_4 photosynthesis, atmospheric CO_2, and climate. *Oecologia* **112**, 285–299.

Ehrich, D., Gaudeul, M., Assefa, A., Koch, M., Mummenhoff, K., Nemomissa, S., Consortium, I. & Brochmann, C. (2007). Genetic consequences of Pleistocene range shifts: contrast between the Arctic, the Alps and the East African mountains. *Molecular Ecology* **16**, 2542–2559.

Eidesen, P., Alsos, I., Popp, M., Stensrud, O., Suda, J. & Brochmann, C. (2007). Nuclear vs. plastid data: complex Pleistocene history of a circumpolar key species. *Molecular Ecology* **16**, 3902–3925.

Elder, J. & Turner, B. (1995). Concerted evolution of repetitive DNA sequences in eukaryotes. *The Quarterly Review of Biology* **70**, 297–320.

Elith, J., Graham, C., Anderson, R., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R., Huettmann, F., Leathwick, J., Lehmann, A. *et al.* (2006). Novel methods improve prediction of species distributions from occurrence data. *Ecography* **29**, 129–151.

Elith, J., Kearney, M. & Phillips, S. (2010). The art of modelling range-shifting species. *Methods in Ecology and Evolution* **1**, 330–342.

Ellis, J., Bentley, K. & McCauley, D. (2008). Detection of rare paternal chloroplast inheritance in controlled crosses of the endangered sunflower *Helianthus verticillatus*. *Heredity* **100**, 574–580.

Emshwiller, E. & Doyle, J. (1999). Chloroplast-Expressed Glutamine Synthetase (ncpGS): Potential utility for phylogenetic studies with an example from *Oxalis* (Oxalidaceae). *Molecular Phylogenetics and Evolution* **12**, 310–319.

Ennos, R. (1994). Estimating the relative rates of pollen and seed migration among plant populations. *Heredity* **72**, 250–259.

Eriksson, T. & Donoghue, M. (1997). Phylogenetic relationships of *Sambucus* and *Adoxa* (Adoxoideae, Adoxaceae) based on nuclear ribosomal ITS sequences and preliminary morphological data. *Systematic Botany* **22**, 555–573.

Ewing, B., Hillier, L., Wendl, M. & Green, P. (1998). Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Research* **8**, 175–185.

Excoffier, L., Laval, G. & Schneider, S. (2005). Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* **1**, 47–50.

Eyre-Walker, A., Gaut, R., Hilton, H., Feldman, D. & Gaut, B. (1998). Investigation of the bottleneck leading to domestication of maize. *Proceedings of the National Academy of Sciences, U.S.A.* **95**, 4441–4446.

Faith, D. (1994). Phylogenetic pattern and the quantification of organismal biodiversity. *Philosophical Transactions of the Royal Society of London B Biological Sciences* **345**, 45–58.

Faith, J. (in press). The last 18,000 years at Boomplaas Cave (South Africa) and the future of the Cape mountain zebra (*Equus zebra zebra*). *Diversity and Distributions* .

Fama, P., Olsen, J., Stam, W. & Procaccini, G. (2000). High levels of intra- and inter-individual polymorphism in the rDNA ITS1 of *Caulerpa racemosa* (Chlorophyta). *European Journal of Phycology* **35**, 349–356.

Farber, O. & Kadmon, R. (2003). Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. *Ecological Modelling* **160**, 115–130.

Fehrer, J., Gemeinholzer, B., Chrtek Jr, J. & Brutigam, S. (2007). Incongruent plastid and nuclear DNA phylogenies reveal ancient intergeneric hybridization in Pilosella hawkweeds (*Hieracium*, Cichorieae, Asteraceae). *Molecular Phylogenetics and Evolution* **42**, 347–361.

Fehrer, J., Krak, K. & Chrtek, J. (2009). Intra-individual polymorphism in diploid and apomictic polyploid hawkweeds (*Hieracium*, Lactuceae, Asteraceae): disentangling phylogenetic signal, reticulation, and noise. *BMC Evolutionary Biology* **9**, 239.

Feliner, G. & Rosselló, J. (2007). Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. *Molecular Phylogenetics and Evolution* **44**, 911–919.

Feliner, G., Larena, B. & Aguilar, J. (2004). Fine-scale geographical structure, intra-individual polymorphism and recombination in nuclear ribosomal internal transcribed spacers in *Armeria* (Plumbaginaceae). *Annals of Botany* **93**, 189–200.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts.

Feng, Y., Oh, S.-H. & Manos, P. (2005). Phylogeny and historical biogeography of the genus *Platanus* as inferred from nuclear and chloroplast DNA. *Systematic Botany* **30**, 786–799.

Flato, G., Boer, G., Lee, W., McFarlane, N., Ramsden, D., Reader, M. & Weaver, A. (2000). The Canadian centre for climate modelling and analysis global coupled model and its climate. *Climate Dynamics* **16**, 451–467.

Foden, W., Midgley, G., Hughes, G., Bond, W., Thuiller, W., Hoffman, M., Kaleme, P., Underhill, L., Rebelo, A. & Hannah, L. (2007). A changing climate is eroding the geographical range of the Namib Desert tree *Aloe* through population declines and dispersal lags. *Diversity and Distributions* **13**, 645–653.

Fougère-Danezan, M., Maumont, S. & Bruneau, A. (2007). Relationships among resin-producing Detarieae s.l. (Leguminosae) as inferred by molecular data. *Systematic Botany* **32**, 748–761.

Frantz, A., Cellina, S., Krier, A., Schley, L. & Burke, T. (2009). Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance? *Journal of Applied Ecology* **46**, 493–505.

Fu, Y. (1997). Statistical test of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**, 915–925.

Fu, Y.-B. & Allaby, R. (2010). Phylogenetic network of *Linum* species as revealed by non-coding chloroplast DNA sequences. *Genetic Resources and Crop Evolution* **57**, 667–677.

Funk, W., Blouin, M., Corn, P., Maxell, B., Pilliod, D., Amish, S. & Allendorf, F. (2005). Population structure of Columbia spotted frogs (*Rana luteiventris*) is strongly affected by the landscape. *Molecular Ecology* **14**, 483–496.

Garrick, R., Sands, C., Rowell, D., Tait, N., Greenslade, P. & Sunnucks, P. (2004). Phylogeography recapitulates topography: very fine-scale local endemism of a saproxylic 'giant' springtail at Tallaganda in the Great Dividing Range of south-east Australia. *Molecular Ecology* **13**, 3329–3344.

Garrick, R., Sands, C., Rowell, D., Hillis, D. & Sunnucks, P. (2007). Catchments catch all: long-term population history of a giant springtail from the southeast Australian highlands – a multigene approach. *Molecular Ecology* **16**, 1865–1882.

Garrick, R., Rowell, D., Simmons, C., Hillis, D., Sunnucks, P. & Brown, J. (2008). Fine-scale phylogeographic congruence despite demographic incongruence in two low-mobility saproxylic springtails. *Evolution* **62**, 1103–1118.

Gawel, N. & Jarret, R. (1991). A modified CTAB DNA extraction procedure for *Musa* and *Ipomoea*. *Plant Molecular Biology Reporter* **9**, 262–266.

Giordano, A., Ridenhour, B. & Storfer, A. (2007). The influence of altitude and topography on genetic structure in the long-toed salamander (*Ambystoma macrodactulym*). *Molecular Ecology* **16**, 1625–1637.

Göker, M. & Grimm, G. (2008). General functions to transform associate data to host data, and their use in phylogenetic inference from sequences with intra-individual variability. *BMC Evolutionary Biology* **8**, 86.

Gordon, C., Cooper, C., Senior, C., Banks, H., Gregory, J., Johns, T., Mitchell, J. & Wood, R. (2000). The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. *Climate Dynamics* **16**, 147–168.

Govender, N., Trollope, W. & van Wilgen, B. (2006). The effect of fire season, fire frequency, rainfall and management on fire intensity in savanna vegetation in South Africa. *Journal of Applied Ecology* **43**, 748–758.

Graur, D. & Martin, W. (2004). Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends in Genetics* **20**, 80–86.

Grimm, G., Denk, T. & Hemleben, V. (2007). Coding of intraspecific nucleotide polymorphisms: a tool to resolve reticulate evolutionary relationships in the ITS of beech trees (Fagus L., Fagaceae). *Systematics and Biodiversity* **5**, 291–309.

Gugger, P., McLachlan, J., Manos, P. & Clark, J. (2008). Inferring long-distance dispersal and topographic barriers during post-glacial colonization from the genetic structure of red maple (*Acer rubrum* L.) in New England. *Journal of Biogeography* **35**, 1665–1673.

Hamilton, A. & Taylor, D. (1991). History of climate and forests in tropical Africa during the last 8 million years. *Climate Change* **19**, 65–78.

Hanna, J., Klopfenstein, N., Kim, M., McDonald, G. & Moore, J. (2007). Phylogeographic patterns of *Armillaria ostoyae* in the western United States. *Forest Pathology* **37**, 192–216.

Hannah, L., Midgley, G., Hughes, G. & Bomhard, B. (2005). The view from the Cape: extinction risk, protected areas, and climate change. *BioScience* **55**, 231–242.

Harpke, D. & Peterson, A. (2008). 5.8S motifs for the identification of pseudogenic ITS regions. *Botany* **86**, 300–305.

Harte, J., Ostling, A., Green, J. & Kinzig, A. (2004). Biodiversity conservation: climate change and extinction risk. *Nature* **430**, 6995.

Harwood, T. (2009). The circular definition of populations and its implications for biological sampling. *Molecular Ecology* **18**, 765–768.

Hasumi, H. & Emori, S. (2004). *K-1 coupled GCM (MIROC) description*. Center for Climate System Research, University of Tokyo, Tokyo.

Heikkinen, R., Luoto, M., Araújo, M., Virkkala, R., Thuiller, W. & Sykes, M. (2006). Methods and uncertainties in bioclimatic envelope modelling under climate change. *Progress in Physical Geography* **30**, 751–777.

Hemleben, V., Ganal, M., Gerstner, J., Schiebel, K. & Torres, R. (1988). Organization and length heterogeneity of plant ribosomal RNA genes. In Kahl, G., ed., *Architecture of eukaryotic genes*, pp. 371–383. VCH Verlagsgesellschaft mbH, Weinheim.

Hewitt, G. (2000). The genetic legacy of the Quaternary ice ages. *Nature* **405**, 907–913.

Hewitt, G. (2004). Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society of London Series B* **359**, 183–195.

Hey, J. & Nielsen, R. (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167**, 747–760.

Higgins, S., Nathan, R. & Cain, M. (2003). Are long-distance dispersal events in plants usually caused by nonstandard means of dispersal? *Ecology* **84**, 1945–1956.

Hijmans, R. & Graham, C. (2006). The ability of climate envelope models to predict the effect of climate change on species distributions. *Global Change Biology* **12**, 2272–2281.

Hijmans, R., Cameron, S., Parra, J., Jones, P. & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* **25**, 1965–1978.

Hijmans, R., Phillips, S., Leathwick, J. & Elith, J. (2011). dismo: Species distribution modeling.

Hilbert, D., Graham, A. & Hopkins, M. (2007). Glacial and interglacial refugia within a long-term rainforest refugium: the wet tropics bioregion of NE Queensland, Australia. *Palaeogeography, Palaeoclimatology, Palaeoecology* **251**, 104–118.

Hirst, A., Gordon, H. & O'Farrell, S. (1996). Global warming in a coupled climate model including oceanic eddy-induced advection. *Geophysical Research Letters* **23**, 3361–3364.

Ho, S. (2007). Calibrating molecular estimates of substitution rates and divergence times in birds. *Journal of Avian Biology* **38**, 409–414.

Hoare, D. & Frost, P. (2004). Phenological description of natural vegetation in southern Africa using remotely-sensed vegetation data. *Applied Vegetation Science* **7**, 19–28.

Holmes, P. & Cowling, R. (1993). Effects of shade on seedling growth, morphology and leaf photosynthesis in six subtropical thicket species from the eastern Cape, South Africa. *Forest Ecology and Management* **61**, 199–220.

Holmgren, K., Lee-Thorp, J., Cooper, G., Lundblad, K., Partridge, T., Scott, L., Sithaldeen, R., Talma, A. & Tyson, P. (2003). Persistent millennial-scale climatic variability over the past 25,000 years in southern Africa. *Quaternary Science Reviews* **22**, 2311–2326.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75**, 383–386.

Hooker, J. (1867*a*). Insular floras. *Gardeners' Chronicle* pp. 6–7; 27; 50–51; 75–76.

Hooker, J. (1867*b*). On insular floras: a lecture. *Journal of Botany* **5**, 23–31.

Howe, H. & Smallwood, J. (1982). Ecology of seed dispersal. *Annals of Review of Ecology and Systematics* **13**, 201–228.

Huelsenbeck, J., Bull, J. & Cunningham, C. (1996). Combining data in phylogenetic analysis. *Trends in Ecology and Evolution* **11**, 152–158.

Hugall, A., Moritz, C., Moussalli, A. & Stanisic, J. (2002). Reconciling paleodistribution models and comparative phylogeography in the Wet Tropics rainforest land snail *Gnarosophia bellendenkerensis* (Brazier 1875). *Proceedings of the National Academy of Sciences, U.S.A.* **99**, 6112–6117.

Hughes, M., Möller, M., Bellstedt, D., Edwards, T. & Villiers, M. (2005). Refugia, dispersal and divergence in a forest archipelago: a study of *Streptocarpus* in eastern South Africa. *Molecular Ecology* **14**, 4415–4426.

Huson, D. & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* **23**, 254–267.

Hutchinson, M. (1995). Interpolating mean rainfall using thin plate smoothing splines. *International Journal of Geographical Information Systems* **9**, 385–403.

Jabaily, R. & Sytsma, K. (2010). Phylogenetics of Puya (Bromeliaceae): Placement, major lineages, and evolution of Chilean species. *American Journal of Botany* **97**, 337–356.

Jansson, R. & Dynesius, M. (2002). The fate of clades in a world of recurrent climatic change: Milankovitch oscillations and evolution. *Annual Review of Ecology and Systematics* **33**, 741–777.

Joly, S. & Bruneau, A. (2006). Incorporating allelic variation for reconstructing the evolutionary history of organisms from multiple genes: an example from *Rosa* in North America. *Systematic Biology* **55**, 623–636.

Jombart, T., Devillard, S., Dufour, A. & Pontier, D. (2008). Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity* **101**, 92–103.

Jordaan, J. (2010). The proposed colonisation sequence of woody species in the Sourish Mixed Bushveld of the Limpopo province, South Africa. *African Journal of Range and Forage Science* **27**, 105–108.

Jordano, P., García, C., Godoy, J. & García-Castaño, J. (2007). Differential contribution of frugivores to complex seed dispersal patterns. *Proceedings of the National Academy of Sciences, U.S.A.* **104**, 3278–3282.

Kay, K., Whittall, J. & Hodges, S. (2006). A survey of nuclear ribosomal internal transcribed spacer substitution rates across angiosperms: an approximate molecular clock with life history effects. *BMC Evolutionary Biology* **6**, 36.

Kays, R., Jansen, P., Knecht, E., Vohwinkel, R. & Wikelski, M. (in press). The effect of feeding time on dispersal of *Virola* seeds by toucans determined from GPS tracking and accelerometers. *Acta Oecologica* .

Kembel, S., Cowan, P., Helmus, M., Cornwell, W., Morlon, H., Ackerly, D., Blomberg, S. & Webb, C. (2010). Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**, 1463–1464.

Kerley, G., Knight, M. & De Kock, M. (1995). Desertification of subtropical thicket in the Eastern Cape, South Africa: Are there alternatives? *Environmental Monitoring and Assessment* **37**, 211–230.

Kerley, G., Wilson, S. & Massey, A. (2002). *Elephant conservation and management in the Eastern Cape. Workshop proceedings. Report No. 35*. Terrestrial Ecology Research Unit, University of Port Elizabeth, Port Elizabeth, South Africa.

Kershaw, A. & Nix, H. (1988). Quantitative palaeoclimatic estimates from pollen data using bioclimatic profiles of extant taxa. *Journal of Biogeography* **15**, 589–602.

Kgope, B., Bond, W. & Midgley, G. (2010). Growth responses of African savanna trees implicate atmospheric [CO_2] as a driver of past and current changes in savanna tree cover. *Austral Ecology* **35**, 451–463.

Kim, K.-J. & Jansen, R. (1994). Comparisons of phylogenetic hypotheses among different data sets in dwarf dandelions (*Krigia*, Asteraceae): additional information from internal transcribed spacer sequences of nuclear ribosomal DNA. *Plant systematics and Evolution* **190**, 157–185.

King, M. & Roalson, E. (2008). Exploring evolutionary dynamics of nrDNA in *Carex* subgenus *Vignea* (Cyperaceae). *Systematic Botany* **33**, 514–524.

King, R. & Ferris, C. (2000). Chloroplast DNA and nuclear DNA variation in the sympatric alder species, *Alnus cordata* (Lois.) Duby and *A. glutinosa* (L.) Gaertn. *Biological Journal of the Linnean Society* **70**, 147–160.

Klein, R. (1980). Environmental and ecological implications of large mammals from Upper Pleistocene and Holocene sites in southern Africa. *Annals of the South African Museum* **81**, 223–283.

Komarova, N., Grabe, T., Huigen, D., Hemleben, V. & Volkov, R. (2004). Organization, differential expression and methylation of rDNA in artificial *Solanum* allopolyploids. *Plant Molecular Biology* **56**, 439–463.

Lechmere-Oertel, R., Cowling, R. & Kerley, G. (2005*a*). Landscape dysfunction and reduced spatial heterogeneity in soil resources and fertility in semi-arid succulent thicket, South Africa. *Austral Ecology* **30**, 615–624.

Lechmere-Oertel, R., Kerley, G. & Cowling, R. (2005*b*). Patterns and implications of transformation in semi-arid succulent thicket, South Africa. *Journal of Arid Environments* **62**, 459–474.

Legendre, P. & Legendre, L. (1998). *Numerical Ecology*. Elsevier Science, Amsterdam, 2nd edn.

Lemey, P., Rambaut, A., Drummond, A. & Suchard, M. (2009). Bayesian phylogeography finds its roots. *PLoS Computational Biology* **5**, e1000520.

Levin, D. (1983). Polyploidy and novelty in flowering plants. *American Naturalist* **122**, 1–25.

Lewis, C. (2008). Late Quaternary climatic changes, and associated human responses, during the last ∼45000 yr in the Eastern and adjoining Western Cape, South Africa. *Earth-Science Reviews* **88**, 167–187.

Linder, C., Goertzen, L., Heuvel, B., Francisco-Ortega, J. & Jansen, R. (2000). The complete external transcribed spacer of 18S-26S rDNA: amplification and phylogenetic utility at low taxonomic levels in Asteraceae and closely allied families. *Molecular Phylogenetics and Evolution* **14**, 285–303.

Lithgow-Bertelloni, C. & Silver, P. (1998). Dynamic topography, plate driving forces and the African superswell. *Nature* **395**, 269–272.

Liu, C., Berry, P., Dawson, T. & Pearson, R. (2005). Selecting thresholds of occurrence in the prediction of species distributions. *Ecography* **28**, 385–393.

Liu, J. & Schardl, C. (1994). A conserved sequence in internal transcribed spacer 1 of plant nuclear rRNA genes. *Plant Molecular Biology* **26**, 775–778.

Loarie, S., Carter, B., Hayhoe, K., McMahon, S., Moe, R., Knight, C. & Ackerly, D. (2008). Climate change and the future of California's endemic flora. *PloS ONE* **3**, e2502.

Lobo, J., Jiménez-Valverde, A. & Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography* **17**, 145–151.

Lockhart, K. (2006). Reconstructing reticulate evolutionary histories of plants. *Trends in Plant Science* **11**, 398–404.

Lorenz-Lemke, A., Muschner, V., Bonatto, S., Cervi, A., Salzano, F. & Freitas, L. (2005). Phylogeographic inferences concerning evolution of Brazilian *Passiflora actinia* and *P. elegans* (Passifloraceae) based on ITS (nrDNA) variation. *Annals of Botany* **95**, 799–806.

Low, A. & Rebelo, A. (1996). *Vegetation of South Africa, Lesotho and Swaziland*. Department of Environmental Affairs and Tourism, Pretoria.

Maddison, W. (1997). Gene trees in species trees. *Systematic Biology* **46**, 523–536.

Maddison, W. & Maddison, D. (2010). Mesquite: a modular system for evolutionary analysis. Version 2.73.

Mahelka, V. & Kopeck, D. (2010). Gene capture from across the grass gamily in the allohexaploid *Elymus repens* (L.) Gould (Poaceae, Triticeae) as evidenced by ITS, GBSSI, and molecular cytogenetics. *Molecular Biology and Evolution* **27**, 1370–1390.

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

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research* **27**, 209–220.

Mardis, E. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics* **24**, 133–141.

Marean, C. (2010). Pinnacle Point Cave 13B (Western Cape Province, South Africa) in context: The Cape Floral kingdom, shellfish, and modern human origins. *Journal of Human Evolution* **59**, 425–443.

Martínez-Meyer, E. & Peterson, A. (2006). Conservatism of ecological niche characteristics in North American plant species over the Pleistocene-to-Recent transition. *Journal of Biogeography* **33**, 1779–1789.

Martínez-Meyer, E., Peterson, A. & Hargrove, W. (2004). Ecological niches as stable distributional constraints on mammal species, with implications for Pleistocene extinctions and climate change projections for biodiversity. *Global Ecology and Biogeography* **13**, 305–314.

Mather, J. & Yoshioka, G. (1968). The role of climate in the distribution of vegetation. *Annals of the Association of American Geographers* **58**, 29–41.

Mayr, E. (1940). Speciation phenomena in birds. *American Naturalist (Sheffield)* **74**, 249–278.

Médail, F. & Diadema, K. (2009). Glacial refugia influence plant diversity patterns in the Mediterranean Basin. *Journal of Biogeography* **36**, 1333–1345.

Meerow, A., Francisco-Ortega, J., Kuhn, D. & Schnell, R. (2006). Phylogenetic relationships and biogeography within the Eurasian clade of Amaryllidaceae based on plastid ndhF and nrDNA ITS sequences: lineage sorting in a reticulate area? *Systematic Biology* **31**, 42–60.

Midgley, G. & Roberts, R. (2003). Past climate change and the generation and persistence of species richness in a biodiversity hotspot, the Cape flora of South Africa. In Visconti, G., Beniston, M., Iannorelli, E. & Barba, D., eds., *Global Change and Protected Areas*, pp. 393–402. Kluwer Academic Publishers, Boston.

Midgley, G., Hannah, L., Millar, D., Rutherford, M. & Powrie, L. (2002). Assessing the vulnerability of species richness to anthropogenic climate change in a biodiversity hotspot. *Global Ecology and Biogeography* **11**, 445–451.

Mills, A. & Cowling, R. (2006). Rate of carbon sequestration at two thicket restoration sites in the Eastern Cape, South Africa. *Restoration Ecology* **14**, 38–49.

Mills, A. & Fey, M. (2004). Transformation of thicket to savanna reduces soil quality in the Eastern Cape, South Africa. *Plant and Soil* **265**, 153–163.

Mills, A., Cowling, R., Fey, M., Kerley, G., Donaldson, J., Lechmere-Oertel, R., Sigwela, A., Skowno, A. & Rundel, P. (2005). Effects of goat pastoralism on ecosystem carbon storage in semiarid thicket, Eastern Cape, South Africa. *Austral Ecology* **30**, 797–804.

Mishler, B. & Theriot, E. (1997). Monophyly, apomorphy, and phylogenetic species concepts. In Wheeler, Q. & Meier, R., eds., *Species Concepts and Phylogenetic Theory: A Debate*. Columbia University Press, New York.

Moffett, R. & Deacon, H. (1977). The flora and vegetation in the surrounds of Boomplaas Cave: Cango Valley. *The South African Archaeological Bulletin* **32**, 127–145.

Moore, A., Blenkinsop, T. & Cotterill, F. (2009). Southern African topography and erosion history: plumes or plate tectonics? *Terra Nova* **21**, 310–315.

Moore, M., Tye, A. & Jansen, R. (2006). Patterns of long-distance dispersal in *Tiquilia* subg. *Tiquilia* (Boraginaceae): implications for the origins of amphitropical disjuncts and Galapagos Islands endemics. *American Journal of Botany* **93**, 1163–1177.

Morgan, D., Korn, R.-L. & Mugleston, S. (2009). Insights into reticulate evolution in Machaerantherinae (Asteraceae: Astereae): 5S ribosomal RNA spacer variation, estimating support for incongruence, and constructing reticulate phylogenies. *American Journal of Botany* **96**, 920–932.

Moussalli, A., Moritz, C., Williams, S. & Carnaval, A. (2009). Variable responses of skinks to a common history of rainforest fluctuation: concordance between phylogeography and palaeo-distribution models. *Molecular Ecology* **18**, 483–499.

Moyle, L., Olson, M. & Tiffin, P. (2004). Patterns of reproductive isolation in three Angiosperm genera. *Evolution* **58**, 1195–1208.

Mucina, L. & Rutherford, M. (2006). *Vegetation map of South Africa, Lesotho and Swaziland*, vol. 19. South African National Biodiversity Institute, Pretoria.

Muller, H. (1925). Why polyploidy is rarer in animals than in plants. *The American Naturalist* **59**, 346–353.

Müller, K. (2005). The efficiency of different search strategies in estimating parsimony jackknife, bootstrap, and Bremer support. *BMC Evolutionary Biology* **5**, 58.

Musil, C., Van Heerden, P., Cilliers, C. & Schmiedel, U. (2009). Mild experimental climate warming induces metabolic impairment and massive mortalities in southern African quartz field succulents. *Environmental and Experimental Botany* **66**, 79–87.

Myers, N., Mittermeier, R., Mittermeier, C., da Fonseca, G. & Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature* **403**, 853–858.

Nakicenovic, N. & Swart, R. (2000). *IPCC Special Report on Emissions Scenarios*. Cambridge University Press, New York, USA.

Nathan, R., Schurr, F., Spiegel, O., Steinitz, O., Trakhtenbrot, A. & Tsoar, A. (2008). Mechanisms of long-distance seed dispersal. *Trends in Ecology and Evolution* **23**, 638–647.

Nei, M. (1987). *Molecular evolutionary genetics*. Columbia University Press, New York.

Nei, M. & Rooney, A. (2005). Concerted and birth-and-death evolution of multigene families. *Annual Review of Genetics* **39**, 121–152.

Newton, A., Allnutt, T., Gillies, A., Lowe, A. & Ennos, R. (1999). Molecular phylogeography, intraspecific variation and the conservation of tree species. *Trends in Ecology and Evolution* **14**, 140–145.

Otto, S. (2003). In polyploids, one plus one does not equal two. *Trends in Ecology and Evolution* **18**, 431–433.

Palmer, J. (1992). Mitochondrial DNA in plant systematics: applications and limitations. In Soltis, P., Soltis, D. & Doyle, J., eds., *Molecular Systematics of Plants*, pp. 36–49. Chapman and Hall, New York.

Palumbi, S., Cipriano, F. & Hare, M. (2001). Predicting nuclear gene coalescence from mitochondrial data: the three-times rule. *Evolution* **55**, 859–868.

Paradis, E. (2010). PEGAS: an R package for population genetics with an integrated-modular approach. *Bioinformatics* **26**, 419–420.

Paradis, E., Claude, J. & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290.

Parmesan, C. (2006). Ecological and evolutionary responses to recent climate change. *Annual Review of Ecology, Evolution and Systematics* **37**, 637–639.

Partridge, T. (1997). Cainozoic environmental change in southern Africa, with special emphasis on the last 200 000 years. *Progress in Physical Geography* **21**, 3–22.

Partridge, T. & Maud, R. (1987). Geomorphic evolution of southern Africa since the Mesozoic. *South African Journal of Geology* **90**, 179–208.

Partridge, T. & Maud, R. (2000). Macro-scale geomorphic evolution of southern Africa. In Partridge, T. C. & Maud, R. R., eds., *The Cenozoic of Southern Africa*, pp. 3–18. Oxford University Press, Oxford.

Partridge, T., Scott, L. & Hamilton, J. (1999). Synthetic reconstructions of southern African environments during the Last Glacial Maximum (21-18 kyr) and the Holocene Altithermal (8-6 kyr). *Quaternary International* **57-58**, 207–214.

Pearman, P., Randin, C., Broennimann, O., Vittoz, P., Knaap, W., Engler, R., Le Lay, G., Zimmermann, N. & Guisan, A. (2008). Prediction of plant species distributions across six millennia. *Ecology Letters* **11**, 357–369.

Pearson, R., Thuiller, W., Araújo, M., Martinez-Meyer, E., Brotons, L., McClean, C., Miles, L., Segurado, P., Dawson, T. & Lees, D. (2006). Model-based uncertainty in species range prediction. *Journal of Biogeography* **33**, 1704–1711.

Peng, Y.-Y., Baum, B., Ren, C.-Z., Jiang, Q.-T., Chen, G.-Y., Zheng, Y.-L. & Wei, Y.-M. (2010). The evolution pattern of rDNA ITS in *Avena* and phylogenetic relationship of the Avena species (Poaceae: Aveneae). *Hereditas* **147**, 183–204.

Pennington, T. & Styles, B. (1975). A generic monograph of the Meliaceae. *Blumea* **22**, 419–540.

Penny, D. & Hendy, M. (1985). The use of tree comparison metrics. *Systemetic Zoology* **34**, 75–82.

Peterson, A., Soberon, J. & Sanchez-Cordero, V. (1999). Conservatism of ecological niches in evolutionary time. *Science* **285**, 1265–1267.

Phillips, S. & Dudík, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* **31**, 161–175.

Phillips, S., Anderson, R. & Schapir, R. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling* **190**, 231–259.

Pierce, S. & Cowling, R. (1984). Phenology of fynbos, renosterveld and subtropical thicket in the south-eastern Cape. *South African Journal of Botany* **3**, 1–16.

Pikaard, C. (2001). Genomic change and gene silencing in polyploids. *Trends in Genetics* **17**, 675–677.

Ponniah, M. & Hughes, J. (2006). The evolution of Queensland spiny mountain crayfish of the genus *Euastacus*. II. Investigating simultaneous vicariance with intraspecific genetic data. *Marine and Freshwater Research* **57**, 349–362.

Popp, M., Gizaw, A., Nemomissa, S., Suda, J. & Brochmann, C. (2008). Colonization and diversification in the African 'sky islands' by eurasian *Lychnis* L. (Caryophyllaceae). *Journal of Biogeography* **35**, 1016–1029.

Posada, D. (2008). jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution* **25**, 1253–1256.

Prentice, C., Guiot, J., Huntley, B., Jolly, D. & Cheddadi, R. (1996). Reconstructing biomes from palaeoecological data: a general method and its application to European pollen data at 0 and 6 ka. *Climate Dynamics* **12**, 185–194.

Price, B., Barker, N. & Villet, M. (2007). Patterns and processes underlying evolutionary significant units in the *Platypleura stridula* L. species complex (Hemiptera:Cicadidae) in the Cape Floristic Region, South Africa. *Molecular Ecology* **16**, 2574–2588.

Price, B., Barker, N. & Villet, M. (2010). A watershed study on genetic diversity: Phylogenetic analysis of the *Platypleura plumosa* (Hemiptera:Cicadidae) complex reveals catchment-specific lineages. *Molecular Phylogenetics and Evolution* **54**, 617–626.

Pritchard, J., Stephens, M. & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics and Breeding* **155**, 945–959.

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

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Ramdhani, S., Cowling, R. M. & Barker, N. P. (2010). Phylogeography of *Schotia* (Fabaceae): recent evolutionary processes in an ancient thicket biome lineage. *International Journal of Plant Sciences* **171**, 626–640.

Ramos-Onsins, S. & Rozas, J. (2002). Statistical properties of new neutrality tests against population growth. *Molecular Biology and Evolution* **19**, 2092–2100.

Rauscher, J., Doyle, J. & Brown, A. (2002). Internal transcribed spacer repeat-specific primers and the analysis of hybridization in the *Glycine tomentella* (Leguminosae) polyploid complex. *Molecular Ecology* **11**, 2691–2702.

Reboud, X. & Zeyl, C. (1994). Organelle inheritance in plants. *Heredity* **72**, 132–140.

Richards, C., Carstens, B. & Lacey Knowles, L. (2007). Distribution modelling and statistical phylogeography: an integrative framework for generating and testing alternative biogeographical hypotheses. *Journal of Biogeography* **34**, 1833–1845.

Richardson, J., Pennington, R., Pennington, T. & Hollingsworth, P. (2001). Rapid diversification of a species-rich genus of neotropical rain forest trees. *Nature* **293**, 2242–2245.

Rieseberg, L. & Soltis, D. (1991). Phylogenetic consequences of cytoplasmic gene flow in plants. *Evolutionary Trends in Plants* **5**, 65–84.

Rieseberg, L., Church, S. & Morjan, C. (2003). Integration of populations and differentiation of species. *New Phytologist* **161**, 59–69.

Riordan, E. & Rundel, P. (2009). Modelling the distribution of a threatened habitat: the California sage scrub. *Journal of Biogeography* **36**, 2176–2188.

Ritz, C., Zimmermann, N. & Hellwig, F. (2003). Phylogeny of subsect. *Meleuphorbia* (A. Berger) Pax and Hoffm. (*Euphorbia* L.) reflects the climatic regime in South Africa. *Plant Systematics and Evolution* **241**, 245–259.

Robertson, M. & Palmer, A. (2002). Predicting the extent of succulent thicket under current and future climate scenarios. *African Journal of Range and Forage Science* **19**, 21–28.

Rogers, A. & Harpending, H. (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution* **9**, 552–569.

Rogers, S. & Bendich, A. (1987). Ribosomal RNA genes in plants: variability in copy number and in the intergenic spacer. *Plant Molecular Biology* **9**, 509–520.

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

Ross, J. (1977). Fabaceae, Caesalpinoideae. *Flora of Southern Africa* **16**, 1–142.

Rosselló, J., Lzaro, A., Cosín, R. & Molins, A. (2007). A phylogeographic split in *Buxus balearica* (Buxaceae) as evidenced by nuclear ribosomal markers: when ITS paralogues are welcome. *Journal of Molecular Evolution* **64**, 143–157.

Rouget, M., Cowling, R., Lombard, A., Knight, A. & Kerley, G. (2006). Designing large-scale conservation corridors for pattern and process. *Conservation Biology* **20**, 549–561.

Sang, T., Crawford, D. & Stuessy, T. (1995). Documentation of reticulate evolution in peonies (*Paeonia*) using internal transcribed spacer sequences of nuclear ribosomal DNA: implications for biogeography and concerted evolution. *Proceedings of the National Academy of Sciences, U.S.A.* **92**, 6813–6817.

Sarich, V. & Wilson, A. (1973). Generation time and genomic evolution in primates. *Science* **179**, 1144–1147.

Schaal, B., Hayworth, D., Olsen, K., Rauscher, J. & Smith, W. (1998). Phylogeographic studies in plants: problems and prospects. *Molecular Ecology* **7**, 465–474.

Scheiter, S. & Higgins, S. (2009). Impacts of climate change on the vegetation of Africa: an adaptive dynamic vegetation modelling approach. *Global Change Biology* **15**, 2224–2246.

Scherson, R., Vidal, R. & Sanderson, M. (2008). Phylogeny, biogeography, and rates of diversification of New World *Astragalus* (Leguminosae) with an emphasis on South American radiations. *American Journal of Botany* **95**, 1030–1039.

Schliep, K. (2011). PHANGORN: Phylogenetic analysis in R. *Bioinformatics* **27**, 592–593.

Schlötterer, C. & Tautz, D. (1994). Chromosomal homogeneity of *Drosophila* ribosomal DNA arrays suggests intrachromosomal exchanges drive concerted evolution. *Current Biology* **4**, 777–783.

Schneider, S. & Excoffier, L. (1999). Estimation of demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. *Genetics* **152**, 1079–1089.

Scholtz, A. (1986). *Palynological and palaeobotanical studies in the Southern Cape*. Ph.D. thesis, Stellenbosch University.

Schrire, B., Lavin, M. & Lewis, G. (2005). Global distribution patterns of the Leguminosae: insights from recent phylogenies. *Biologiese Skrifter* **55**, 375–422.

Schulze, R., Maharaj, M., Lynch, S., Howe, B. & Melvill-Thomson, B. (1997). South Africa atlas of agrohydrology and climatology. Tech. Rep. Report TT82/96, Water Research Commission. Pretoria. South Africa.

Scott, L., Holmgren, K. & Partridge, T. (2008). Reconciliation of vegetation and climatic interpretations of pollen profiles and other regional records from the last 60 thousand years in the Savanna Biome of Southern Africa. *Palaeogeography, Palaeoclimatology, Palaeoecology* **257**, 198–206.

Shaw, J., Lickey, E., Schilling, E. & Small, R. (2007). Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *American Journal of Botany* **94**, 275–288.

Shendure, J. & Ji, H. (2008). Next-generation DNA sequencing. *Nature* **26**, 1135–1145.

Siebert, F. & Eckhardt, H. (2008). The vegetation and floristics of the Nkhuhlu exclosures, Kruger National Park. *Koedoe* **50**, 126–144.

Sigwela, A. (2004). *Animal-seed Interactions in the Thicket Biome: Consequences of Faunal Replacement and Land Use on Seed Dynamics*. Ph.D. thesis, University of Port Elizabeth.

Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K., Tignor, M. & Miller, H. (2007). *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom.

Soltis, D., Gitzendanner, M., Strenge, D. & Soltis, P. (1997). Chloroplast DNA intraspecific phylogeography of plants from the Pacific Northwest of North America. *Plant Systematics and Evolution* **206**, 353–373.

Soltis, D., Morris, A., McLachlan, J., Manos, P. & Soltis, P. (2006). Comparative phylogeography of unglaciated eastern North America. *Molecular Ecology* **15**, 4261–4293.

Städler, T., Haubold, B., Merino, C., Stephan, W. & Pfaffelhuber, P. (2009). The impact of sampling schemes on the site frequency spectrum in non-equilibrium subdivided populations. *Genetics* **182**, 205–216.

Stamatakis, A. (2006). RAxML-VI-HPC: Maximum Likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690.

Stamatakis, A., Hoover, P. & Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML web servers. *Systematic Biology* **57**, 758–771.

Steele, P., Friar, L., Gilbert, L. & Jansen, R. (2010). Molecular systematics of the neotropical genus *Psiguria* (Cucurbitaceae): Implications for phylogeny and species identification. *American Journal of Botany* **97**, 156–173.

Steenkamp, Y., van Wyk, B., Victor, J., Hoare, D., Smith, G., Dold, T. & Cowling, R. (2004). Maputaland-Pondoland-Albany. In Mittermeier, R. A., Gil, P. R., Hoffman, M., Pilgrim, J., Brooks, T., Mittermeier, C. G., Lamoreux, J. & da Fonseca, G., eds., *Hotspots revisited: Earth's biologically richest and most endangered terrestrial ecoregions*, pp. 219–228. CEMEX, Mexico City.

Stephens, M., Smith, N. & Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* **68**, 978–989.

Stuart-Hill, G. (1992). Effects of elephants and goats on the Kaffrarian Succulent Thicket of the Eastern Cape, South Africa. *Journal of Applied Ecology* **29**, 699–710.

Sunnucks, P., Blacket, M., Taylor, J., Sands, C., Ciavaglia, S., Garrick, R., Tait, N., Rowell, D. & Pavlova, A. (2006). A tale of two flatties: different responses of two terrestrial flatworms to past environmental climatic fluctuations at Tallaganda in montane south-eastern Australia. *Molecular Ecology* **15**, 4513–4531.

Svenning, J.-C., Flojgaard, C., Marske, K., Nógues-Bravo, D. & Normand, S. (2011). Applications of species distribution modeling to paleobiology. *Quaternary Science Reviews* **30**, 2930–2947.

Swart, B., Tolley, K. & Matthee, C. (2009). Climate change drives speciation in the southern rock agama (Agama atra) in the Cape Floristic Region, South Africa. *Journal of Biogeography* **36**, 78–87.

Swartz, E., Skelton, P. & Bloomer, P. (2007). Sea-level changes, river capture and the evolution of populations of the Eastern Cape and fiery redfins (*Pseudobarbus afer* and *Pseudobarbus phlegethon*, Cyprinidae) across multiple river systems in South Africa. *Journal of Biogeography* **34**, 2086–2099.

Swartz, E., Skelton, P. & Bloomer, P. (2009). Phylogeny and biogeography of the genus *Pseudobarbus* (Cyprinidae): Shedding light on the drainage history of rivers associated with the Cape Floristic Region. *Molecular Phylogenetics and Evolution* **51**, 75–84.

Swofford, D. (2002). *PAUP*: Phylogenetic analysis using parsimony (*and other methods). Version 4.10b.*. Sinauer Associates, Sunderland, Massachusetts.

Taberlet, P., Gielly, L., Pautou, G. & Bouvet, J. (1991). Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology* **17**, 1105–1109.

Taberlet, P., Fumagalli, L., Wust-Saucy, A. & Cosson, J. (1998). Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Ecology* **7**, 453–464.

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

Tajima, F. (1993). Simple methods for testing molecular clock hypothesis. *Genetics* **135**, 599–607.

Talma, A. & Vogel, J. (1992). Late Quaternary paleotemperatures derived from a speleothem from Cango Caves, Cape Province, South Africa. *Quaternary Research* **37**, 203–213.

Taylor, M. & Hellberg, M. (2006). Comparative phylogeography in a genus of coral reef fishes: biogeographic and genetic concordance in the Caribbean. *Molecular Ecology* **15**, 695–707.

Templeton, A., Crandall, K. & Sing, C. (1992). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data: III. Cladogram estimation. *Genetics* **132**, 619–633.

Testolin, R. & Cipriani, G. (1996). Paternal inheritance of chloroplast DNA and maternal inheritance of mitochondrial DNA in the genus *Actinidia*. *Theoretical and Applied Genetics* **94**, 897–903.

Thomas, C., Cameron, A., Green, R., Bakkenes, M., Beaumont, L., Collingham, Y., Erasmus, B., de Siqueira, M., Grainger, A., Hannah, L. *et al.* (2004). Extinction risk from climate change. *Nature* **427**, 145–148.

Thompson, S. & Whitton, J. (2006). Patterns of recurrent evolution and geographic parthenogenesis within apomictic polyploid Easter daisies (*Townsendia hookeri*). *Molecular Ecology* **15**, 3389–3400.

Thuiller, W. (2004). Patterns and uncertainties of species range shifts under climate change. *Global Change Biology* **10**, 2020–2027.

Ting, I. & Hanscom, Z. (1977). Induction of acid metabolism in *Portulacaria afra*. *Plant Physiology* **59**, 511–514.

Tolley, K., Burger, M., Turner, A. & Matthee, C. (2006). Biogeographic patterns and phylogeography of dwarf chameleons (*Bradypodion*) in an African biodiversity hotspot. *Molecular Ecology* **15**, 781–793.

Trewick, S., Morgan-Richards, M., Russell, S., Henderson, S., Rumsey, F., Pinter, I., Barrett, J., Gibby, M. & Vogel, J. (2002). Polyploidy, phylogeography and Pleistocene refugia of the rockfern *Asplenium ceterach*: evidence from chloroplast DNA. *Molecular Ecology* **11**, 2003–2012.

Tribsch, A. & Schonswetter, P. (2003). Patterns of endemism and comparative phylogeography confirm palaeoenvironmental evidence for Pleistocene refugia in the Eastern Alps. *Taxon* **52**, 477–497.

Trollope, W. (1974). Role of fire in preventing bush encroachment in the Eastern Cape. *African Journal of Range and Forage Science* **9**, 67–72.

Tyson, P. (1986). *Climatic change and variability in southern Africa*. Oxford University Press, Cape Town.

van Wyk, P. (1972). *Trees of the Kruger National Park*. Purnell, Cape Town.

van Zinderen Bakker, E. (1976). The evolution of late Quaternary paleoclimates of southern Africa. *Palaeoecology of Africa* **9**, 160–202.

VanDerWal, J., Shoo, L. & Williams, S. (2009*a*). New approaches to understanding late Quaternary climate fluctuations and refugial dynamics in Australian wet tropical rain forests. *Journal of Biogeography* **36**, 291–301.

VanDerWal, J., Shoo, L. P., Johnson, C. N. & Williams, S. E. (2009*b*). Abundance and the environmental niche: environmental suitability estimated from niche models predicts the upper limit of local abundance. *The American Naturalist* **174**, 282–291.

Vega, R., Fljgaard, C., Lira-Noriega, A., Nakazawa, Y., Svenning, J.-C. & Searle, J. (2010). Northern glacial refugia for the pygmy shrew *Sorex minutus* in Europe revealed by phylogeographic analyses and species distribution modelling. *Ecography* **33**, 260–271.

Vlok, J. & Euston-Brown, D. (2002). Report by biological survey component for conservation planning for biodiversity in the thicket biome. Unpublished report.

Vlok, J., Euston-Brown, D. & Cowling, R. (2003). Acocks' Valley Bushveld 50 years on: new perspectives on the delimitation, characterisation and origin of subtropical thicket vegetation. *South African Journal of Botany* **69**, 27–51.

Vlok, J., Cowling, R. & Wolf, T. (2005). A vegetation map for the Little Karoo.

Volkov, R., Borisjuk, N., Panchuk, I., Schweizer, D. & Hemleben, V. (1999). Elimination and rearrangement of parental rDNA in the allotetraploid *Nicotiana tabacum*. *Molecular Biology and Evolution* **16**, 311–320.

Volkov, R., Komarova, N. & Hemleben, V. (2007). Ribosomal DNA and plant hybrids: Inheritance, rearrangement, expression. *Systematics and Biodiversity* **5**, 261–276.

Von Willert, D., Eller, B., Werger, M. & Brinckmann, E. (1990). Desert succulents and their life strategies. *Vegetatio* **90**, 133–143.

Vriesendorp, B. & Bakker, F. (2005). Reconstructing patterns of reticulate evolution in angiosperms: what can we do? *Taxon* **54**, 593–604.

Wallace, A. (1880). *Island life, or the phenomena and causes of insular faunas and floras, including a revision and attempted solution of the problem of geological climates*. Macmillan, London.

Waltari, E., Hijmans, R., Peterson, A., Nyari, A., Perkins, S. & Guralnick, R. (2007). Locating Pleistocene refugia: comparing phylogeographic and ecological niche model predictions. *PLoS ONE* **2**, e563.

Warren, D., Glor, R., Turelli, M. & Funk, D. (2008). Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution* **62**, 2868–2883.

Watson, J. (2002). *Bontveld ecosystem functioning, and rehabilitation after strip mining*. Ph.D. thesis, University of Port Elizabeth.

Webb, T., Shuman, B. & Williams, J. (2003). Climatically forced vegetation dynamics in eastern North America during the late Quaternary Period. In Gillespie, A., Porter, S. C. & Atwater, B. F., eds., *Developments in Quaternary Sciences*, vol. 1, pp. 459–478. Elsevier.

Wendel, J. (2000). Genome evolution in polyploids. *Plant Molecular Biology* **42**, 225–249.

Werger, M. & Coetzee, B. (1977). A phytosociological and phytogeographical study of Augrabies Falls National Park, Republic of South Africa. *Koedoe* **20**, 11–51.

Westcott, D., Bentrupperbumer, J., Bradford, M. & McKeown, A. (2005). Incorporating patterns of disperser behaviour into models of seed dispersal and its effects on estimated dispersal curves. *Oecologia* **146**, 57–67.

White, T., Bruns, T., Lee, S. & Taylor, J. (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In Innis, M., Gelfand, D., Sninsky, J. & White, T., eds., *PCR protocols: A guide to methods and applications*, pp. 315–322. Academic Press, San Diego.

Whittall, J., Liston, A., Gisler, S. & Meinke, R. (2000). Detecting nucleotide additivity from direct sequences is a SNAP: an example from *Sidalcea* (Malvaceae). *Plant Biology* **2**, 211–217.

Wiens, J. (1999). Polymorphism in systematics and comparative biology. *Annual Review of Ecology and Systematics* **30**, 327–362.

Wigley, B., Bond, W. & Hoffman, M. (2009). Bush encroachment under three contrasting land-use practices in a mesic South African savanna. *African Journal of Ecology* **47**, 62–70.

Wigley, B., Bond, W. & Hoffman, M. (2010). Thicket expansion in a South African savanna under divergent land use: local vs. global drivers? *Global Change Biology* **16**, 964–976.

Wiley, E. (1978). The evolutionary species concept reconsidered. *Systematic Biology* **27**, 17–26.

Williams, J., Shuman, B. & Webb, T. (2001). Dissimilarity analyses of Late-Quaternary vegetation and climate in eastern North Amercia. *Ecology* **82**, 3346–3362.

Williams, J., Shuman, B., Webb, T., Bartlein, P. & Leduc, P. (2004). Late-Quaternary vegetation dynamics in North America: scaling from taxa to biomes. *Ecological Monographs* **74**, 309–334.

Wolfe, K., Li, W.-H. & Sharp, P. (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast and nuclear DNAs. *Proceedings of the National Academy of Sciences, U.S.A.* **84**, 9054–9058.

Won, Y. & Hey, J. (2005). Divergence population genetics of chimpanzees. *Molecular Biology and Evolution* **22**, 297–307.

Woodward, F. (1987). *Climate and plant distribution*. Cambridge University Press, Cambridge.

Woodward, F., Lomas, M. & Kelly, C. (2004). Global climate and the distribution of plant biomes. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **359**, 1465–1476.

Yamaji, H., Fukuda, T., Yokoyama, J., Pak, J., Zhou, C., Yang, C., Kondo, K., Morota, T., Takeda, S., Sasaki, H. *et al.* (2007). Reticulate evolution and phylogeography in *Asarum* sect. *Asiasarum* (Aristolochiaceae) documented in internal transcribed spacer sequences (ITS) of nuclear ribosomal DNA. *Molecular Phylogenetics and Evolution* **44**, 863–884.

Zachos, J., Pagani, M., Sloan, L., Thomas, E. & Billups, K. (2001). Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* **292**, 686–693.

Zhang, T.-C., Comes, H. & Sun, H. (2011). Chloroplast phylogeography of *Terminalia franchetii* (Combretaceae) from the eastern Sino-Himalayan region and its correlation with historical river capture events. *Molecular Phylogenetics and Evolution* **60**, 1–12.

Zimmer, E., Martin, S., Beverly, S., Kan, Y. & Wilson, A. (1980). Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. *Proceedings of the National Academy of Sciences, U.S.A.* **77**, 2158–2162.

A. Appendix

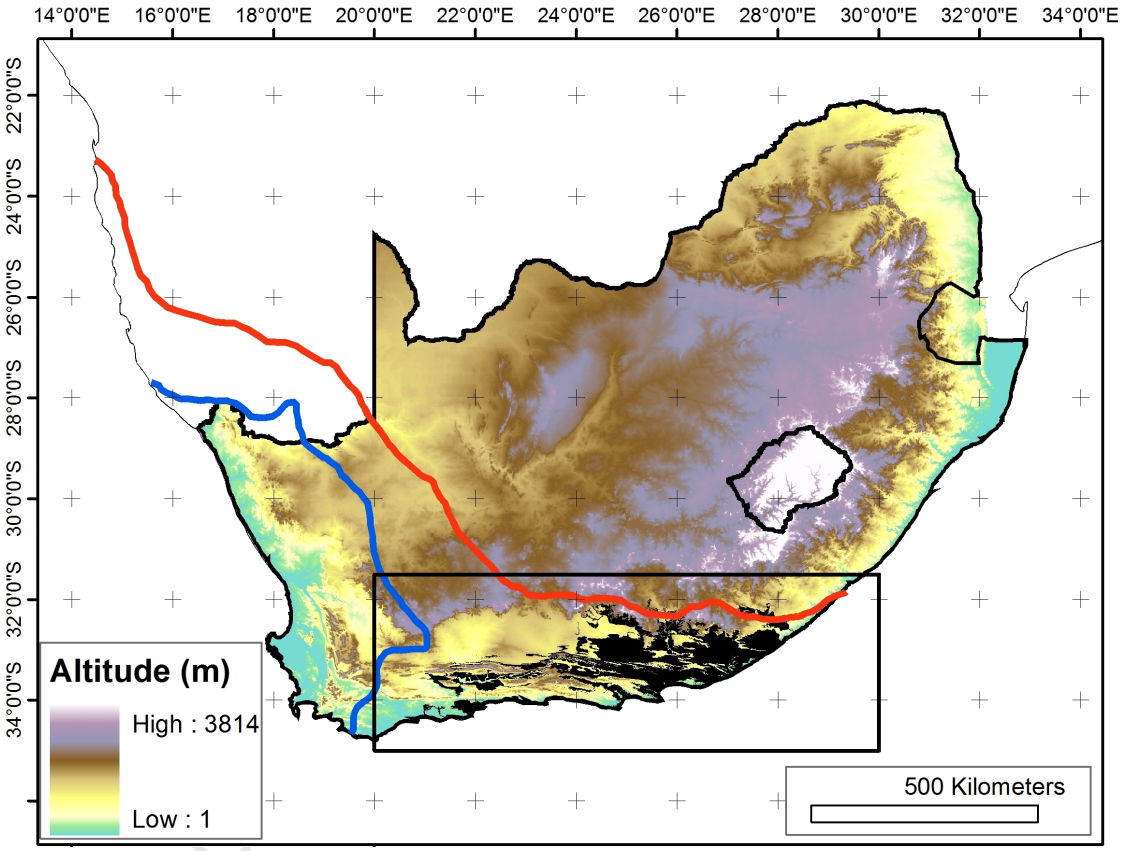

Figure A.1. The study area and mask used to sample background points for community and species distribution modelling. The distribution of the Albany Subtropical Thicket (shaded in black) and the mask (box) used to sample background points for all MAXENT analyses. The winter rainfall zone (defined as $\geq 66\%$ of annual rainfall falling within winter months; Carr *et al.* 2006) lies west of the blue line, the summer rainfall zone (defined as $\geq 66\%$ of annual rainfall falling within summer months) lies east of the red line and the annual rainfall zone falls in between.

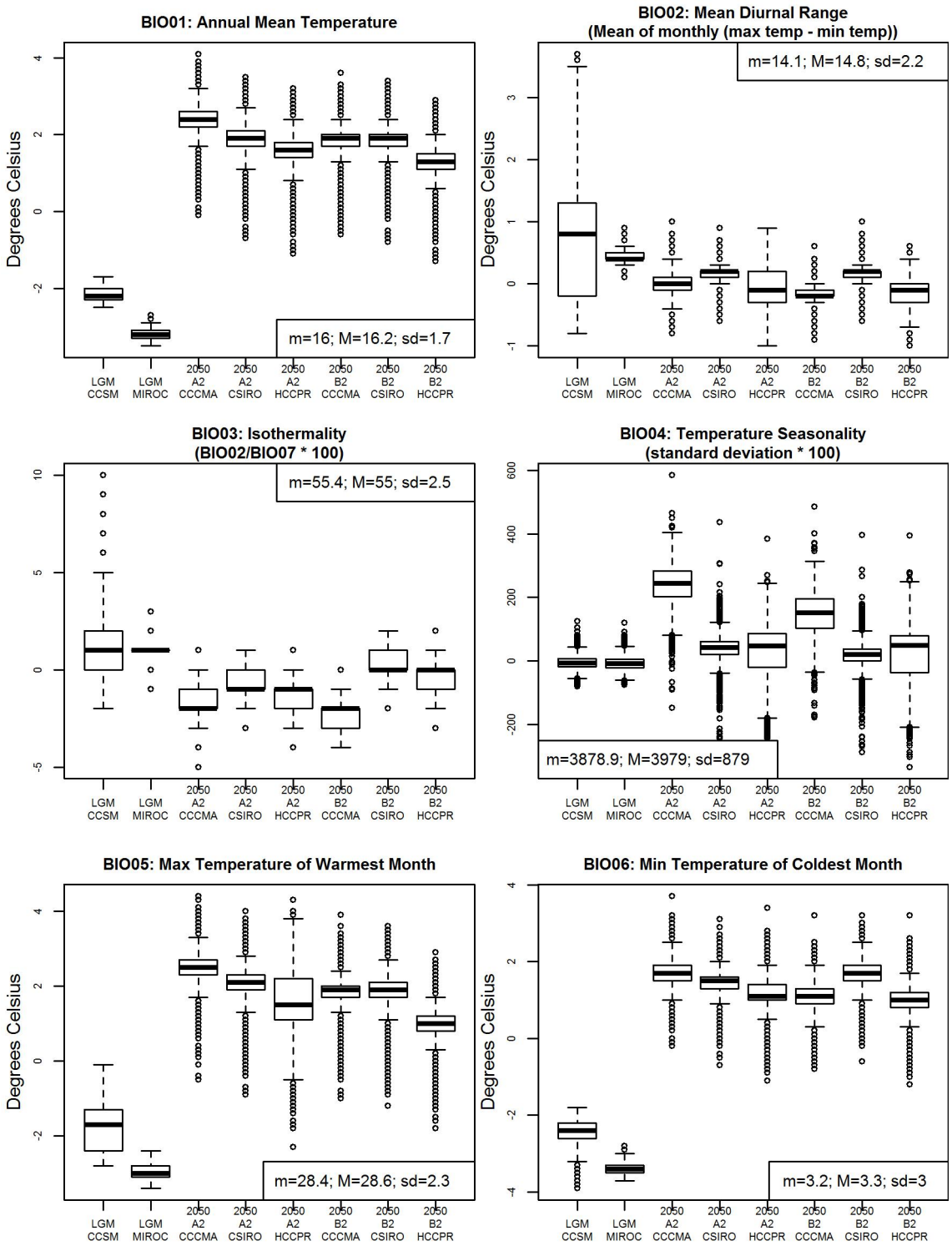

Figure A.2. The difference between altered climatic conditions projected by past and future global climate models and current climatic conditions for 19 bioclimatic variables. Comparisons are between cells sampled from the within the mask shown in Appendix Figure A.1 (Pg. 221). The mean (m), median (M) and standard deviation (sd) of the current climatic conditions for each variable are shown.

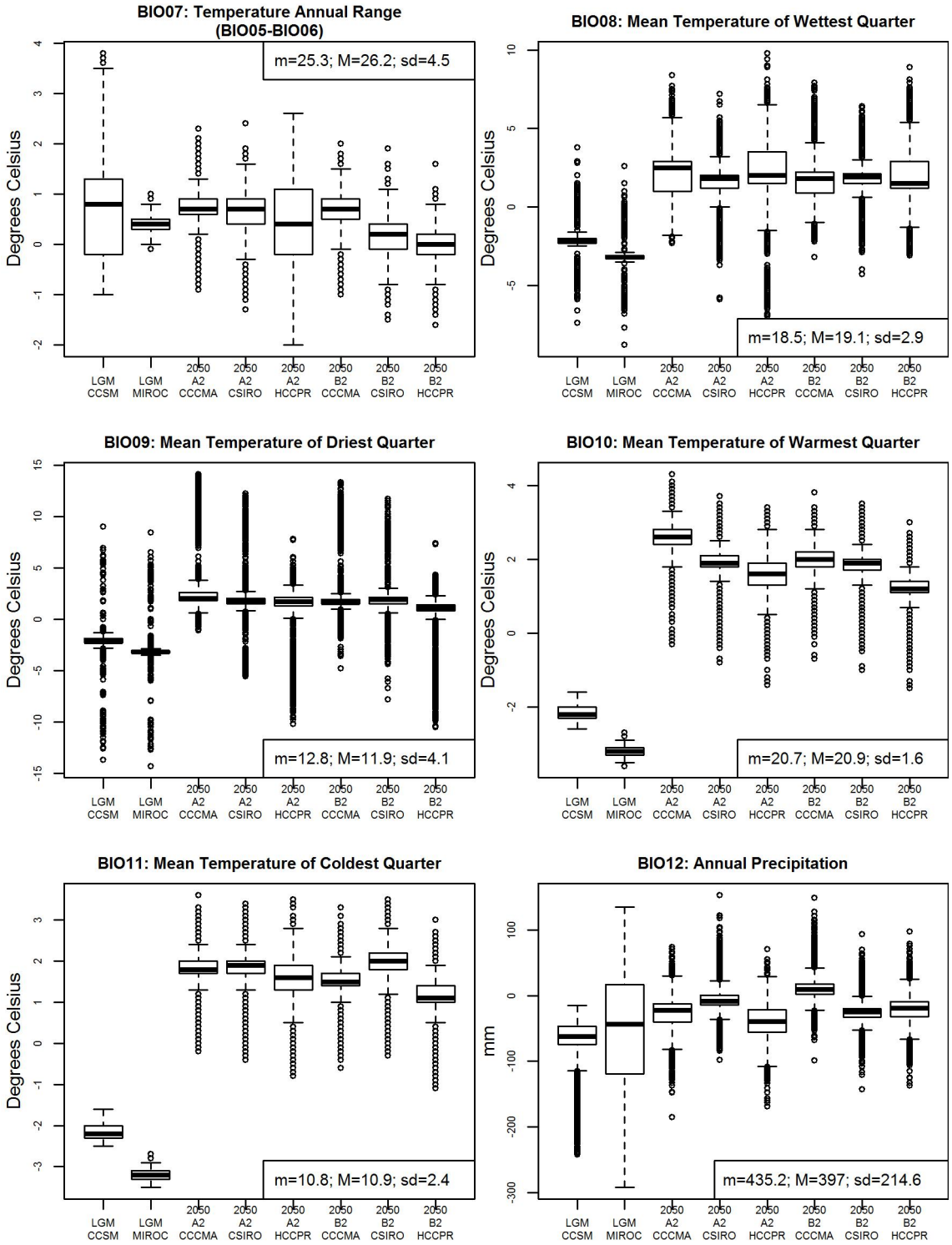

Figure A.2. Continued

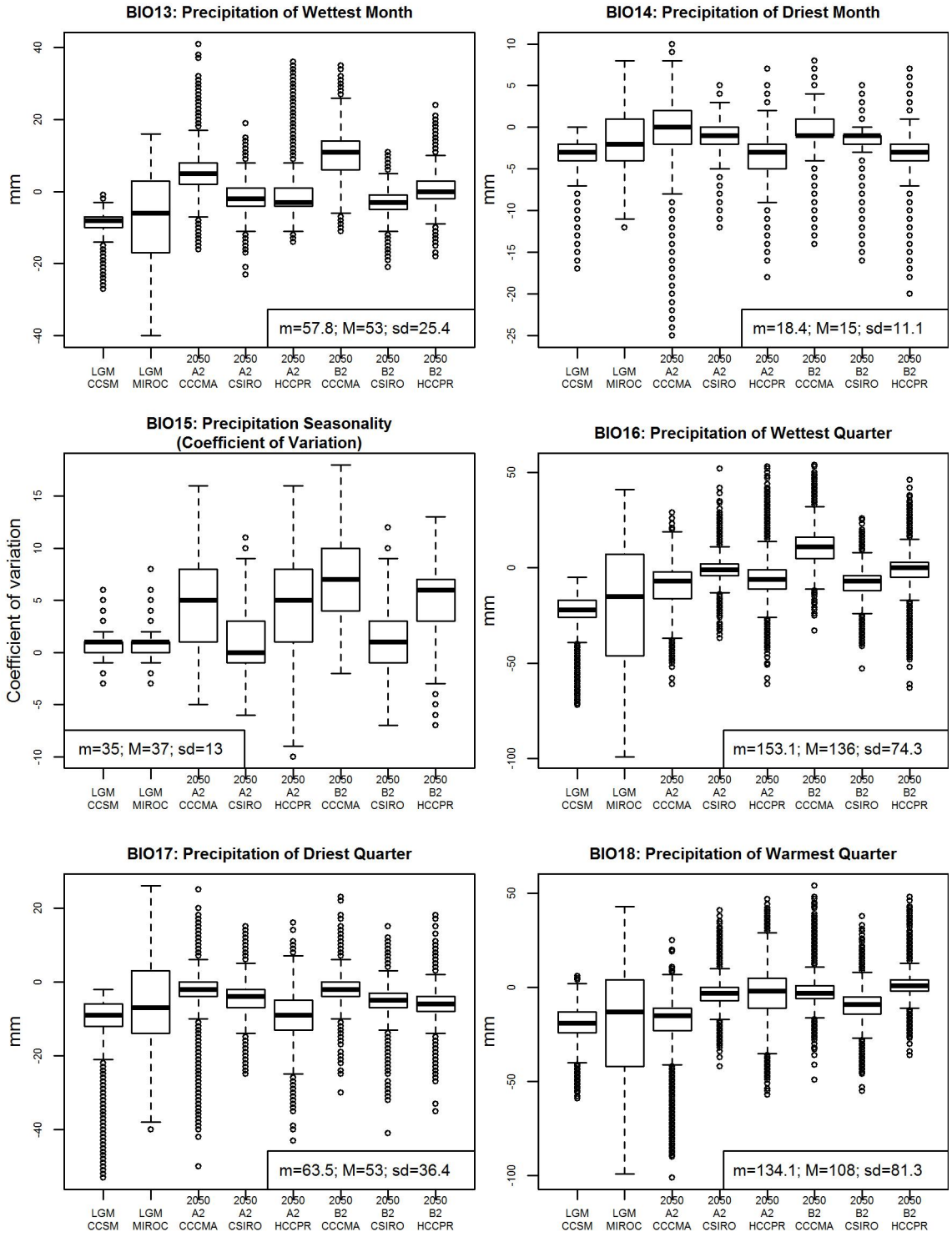

Figure A.2. Continued

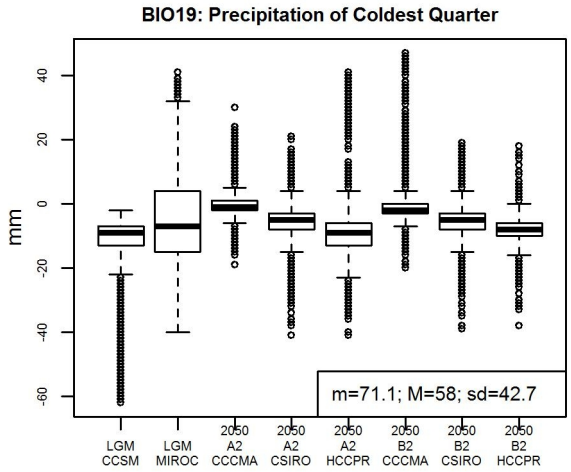

Figure A.2. Continued

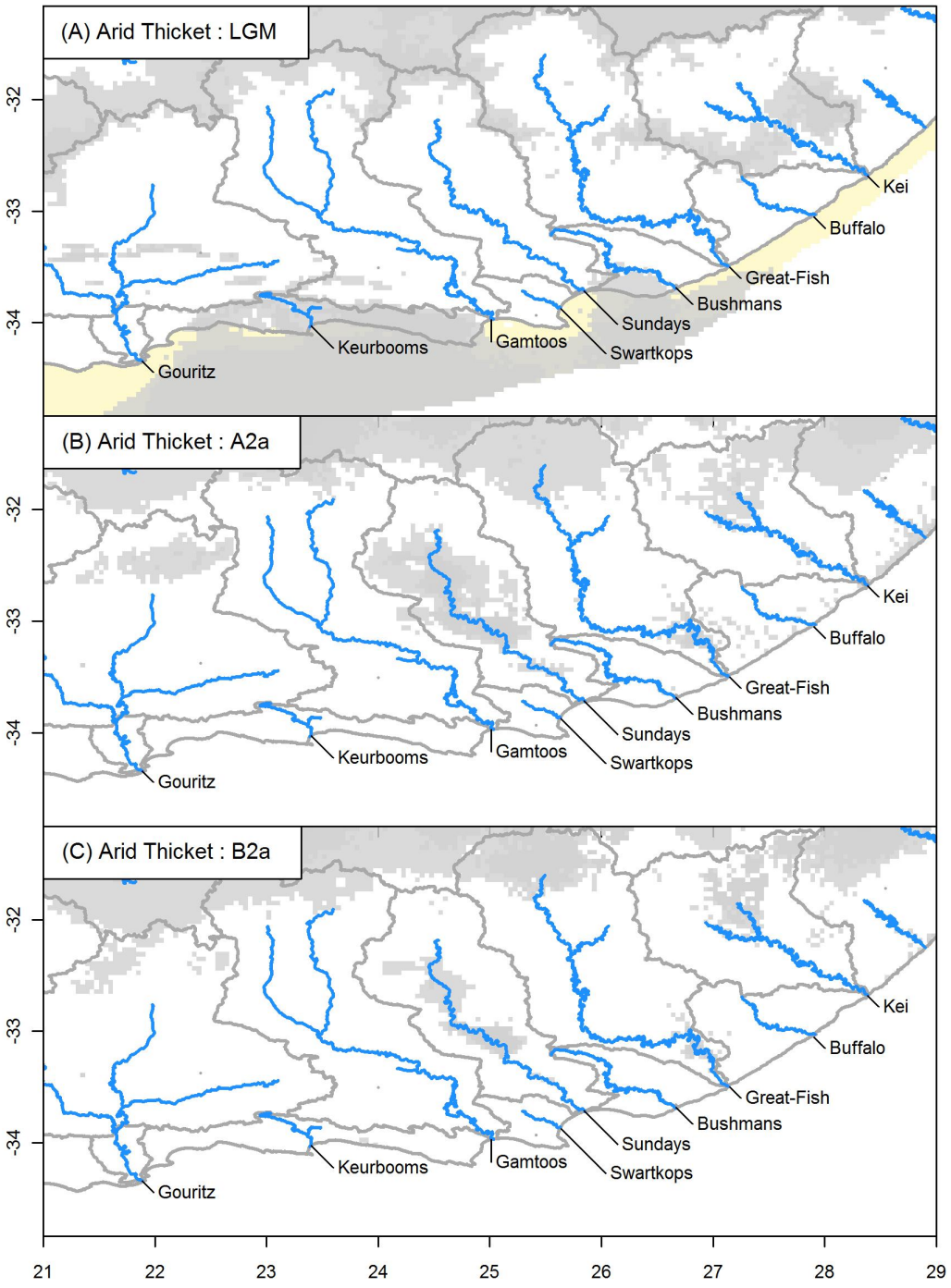

Figure A.3. The multivariate environmental similarity surface (MESS) ≤ -5 (grey) for community distribution models of the *Arid AST subtype* projected onto global climate models (GCMs) of (A) the Last Glacial Maximum (LGM), (B) 2050 scenario *A2a*, and (C) 2050 scenario *B2a*. The maps represent composites of two GCMs for the LGM and three GCMs for each of the 2050 scenarios. The MESS surface calculation represents how similar a point is to a set of reference points, which in this case are sampled from the current climate conditions. Negative MESS values indicate sites where at least one variable has a value that is outside the range of environments over the reference set, so these are novel environments.

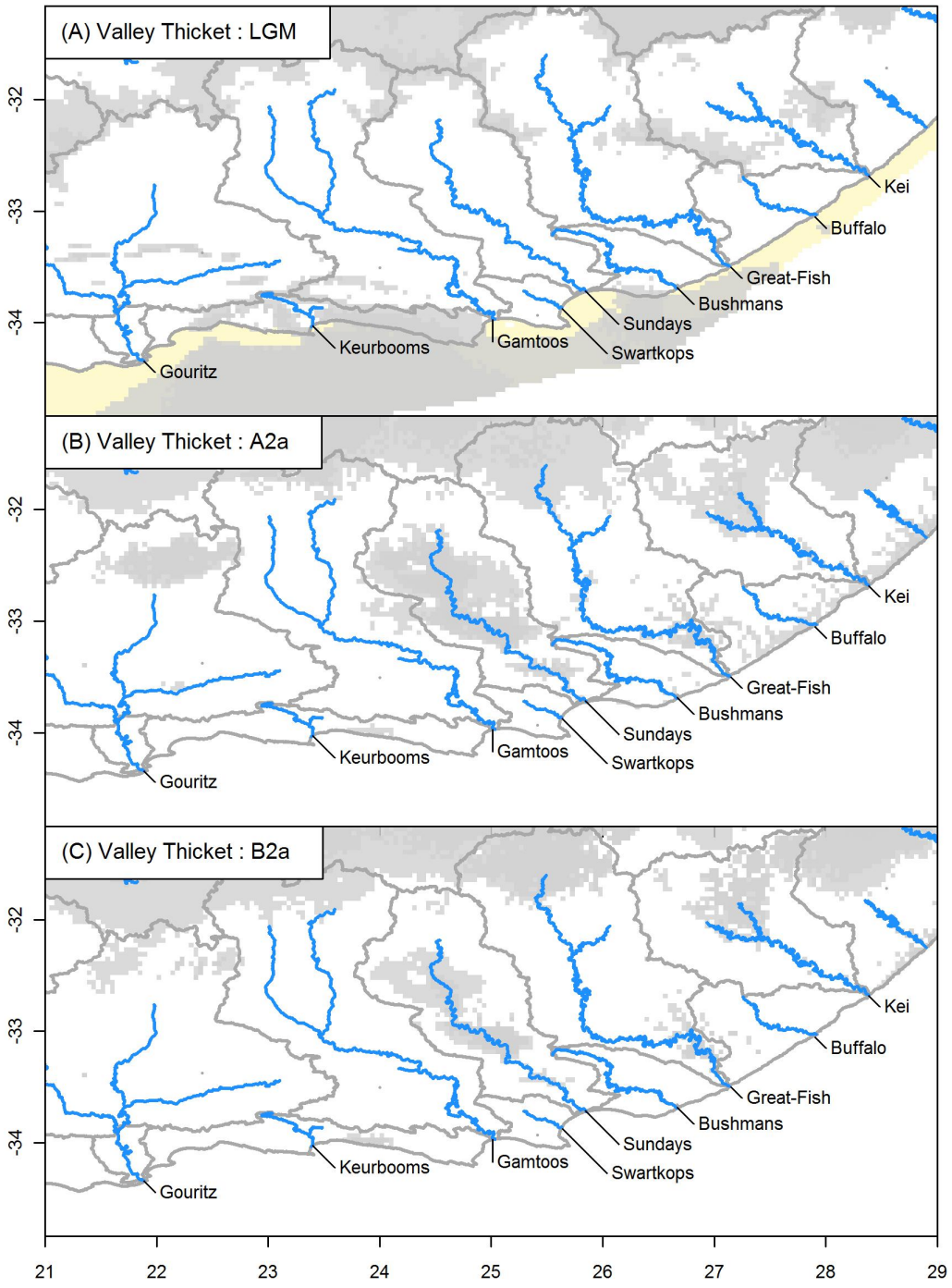

Figure A.4. The multivariate environmental similarity surface (MESS) ≤ -5 (grey) for community distribution models of the *Valley AST subtype* projected global climate models of (A) the Last Glacial Maximum (LGM), (B) 2050 scenario *A2a*, and (C) 2050 scenario *B2a*. The maps represent composites of two GCMs for the LGM and three GCMs for each of the 2050 scenarios. The MESS surface calculation represents how similar a point is to a set of reference points, which in this case are sampled from the current climate conditions. Negative MESS values indicate sites where at least one variable has a value that is outside the range of environments over the reference set, so these are novel environments.

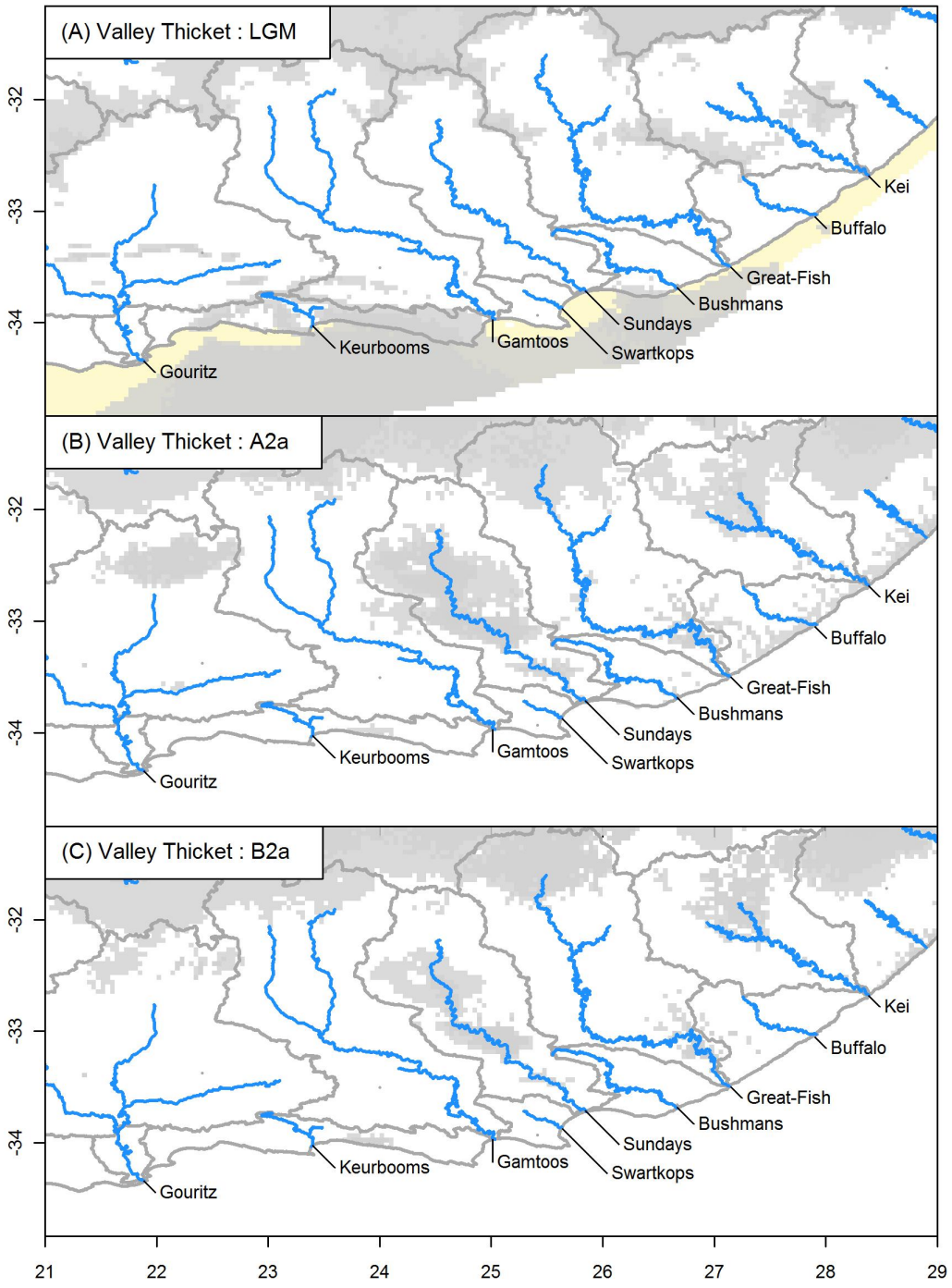

Figure A.5. The multivariate environmental similarity surface (MESS) ≤ -5 (grey) for community distribution models of the *Mesic AST subtype* projected onto global climate models of (A) the Last Glacial Maximum (LGM), (B) 2050 scenario *A2a*, and (C) 2050 scenario *B2a*. The maps represent composites of two GCMs for the LGM and three GCMs for each of the 2050 scenarios. The MESS surface calculation represents how similar a point is to a set of reference points, which in this case are sampled from the current climate conditions. Negative MESS values indicate sites where at least one variable has a value that is outside the range of environments over the reference set, so these are novel environments.

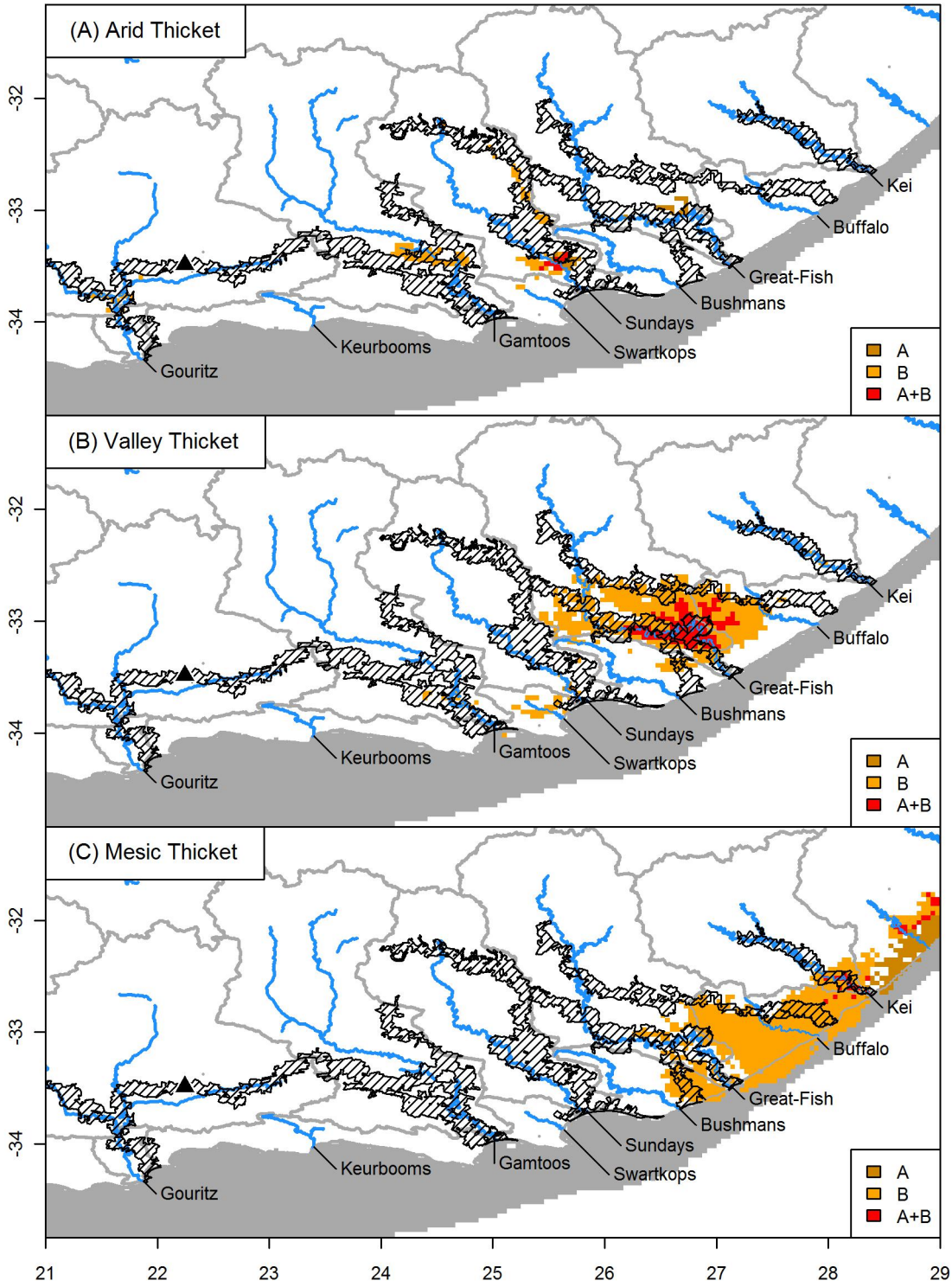

Figure A.6. The proposed mega-conservancy network and community distribution models of three Albany Subtropical Thicket subtypes projected onto Last Glacial Maximum (shaded) climatic conditions. Two Global Climate Models were used to derive the LGM distribution: CCSM (A) and MIROC (B). Grey shading shows the continental shelf exposed due to lower sea levels during the Last Glacial Maximum. The Boomplaas Cave is shown (black triangle; see text for details, Pg. 49).

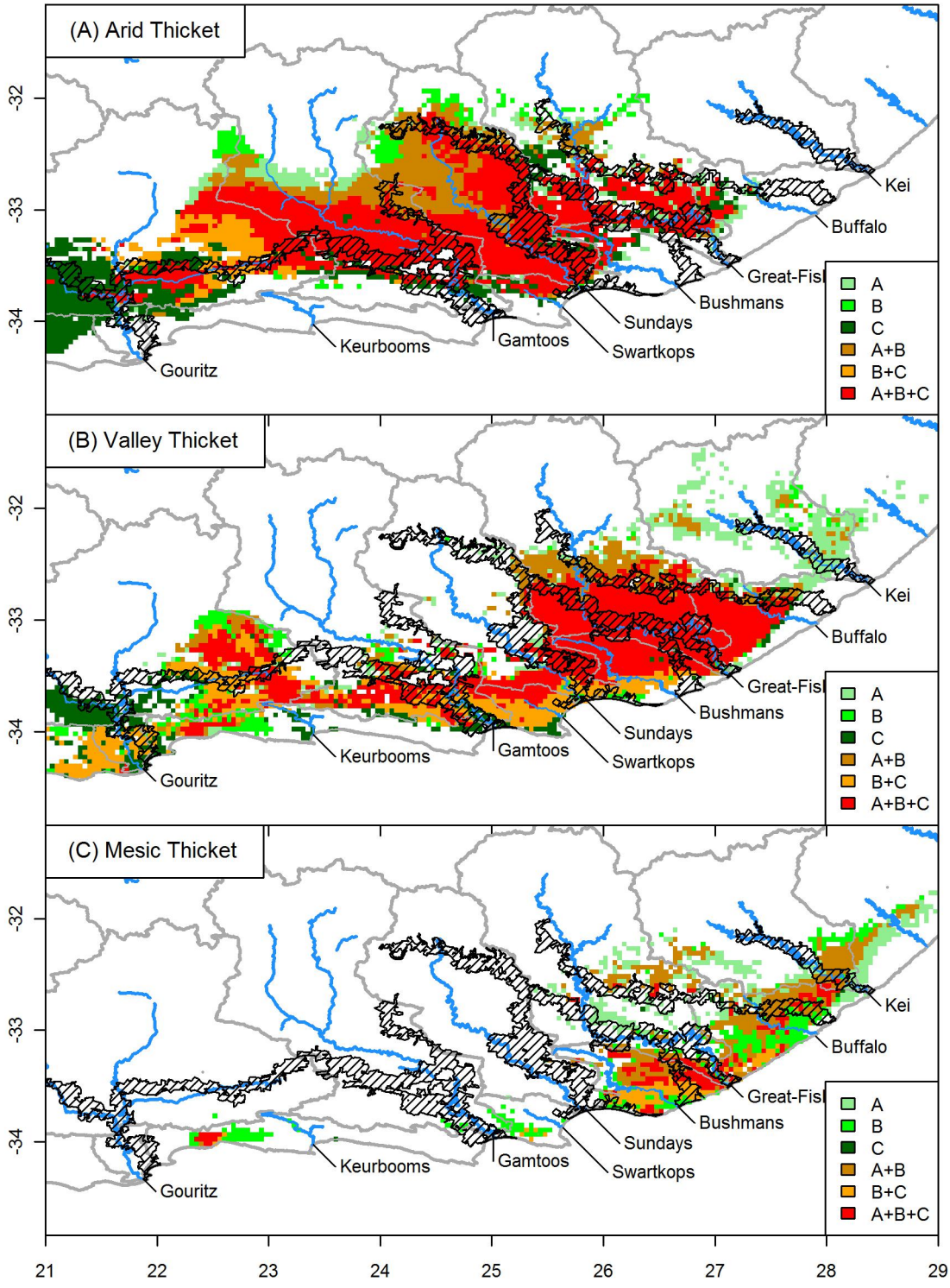

Figure A.7. The proposed megaconservancy network and the community distribution models of three Albany Subtropical Thicket subtypes projected onto 2050 scenario *A2a* climatic conditions (shaded). The GCMs used were: CCCMA (A), CSIRO (B), and HCCPR (C).

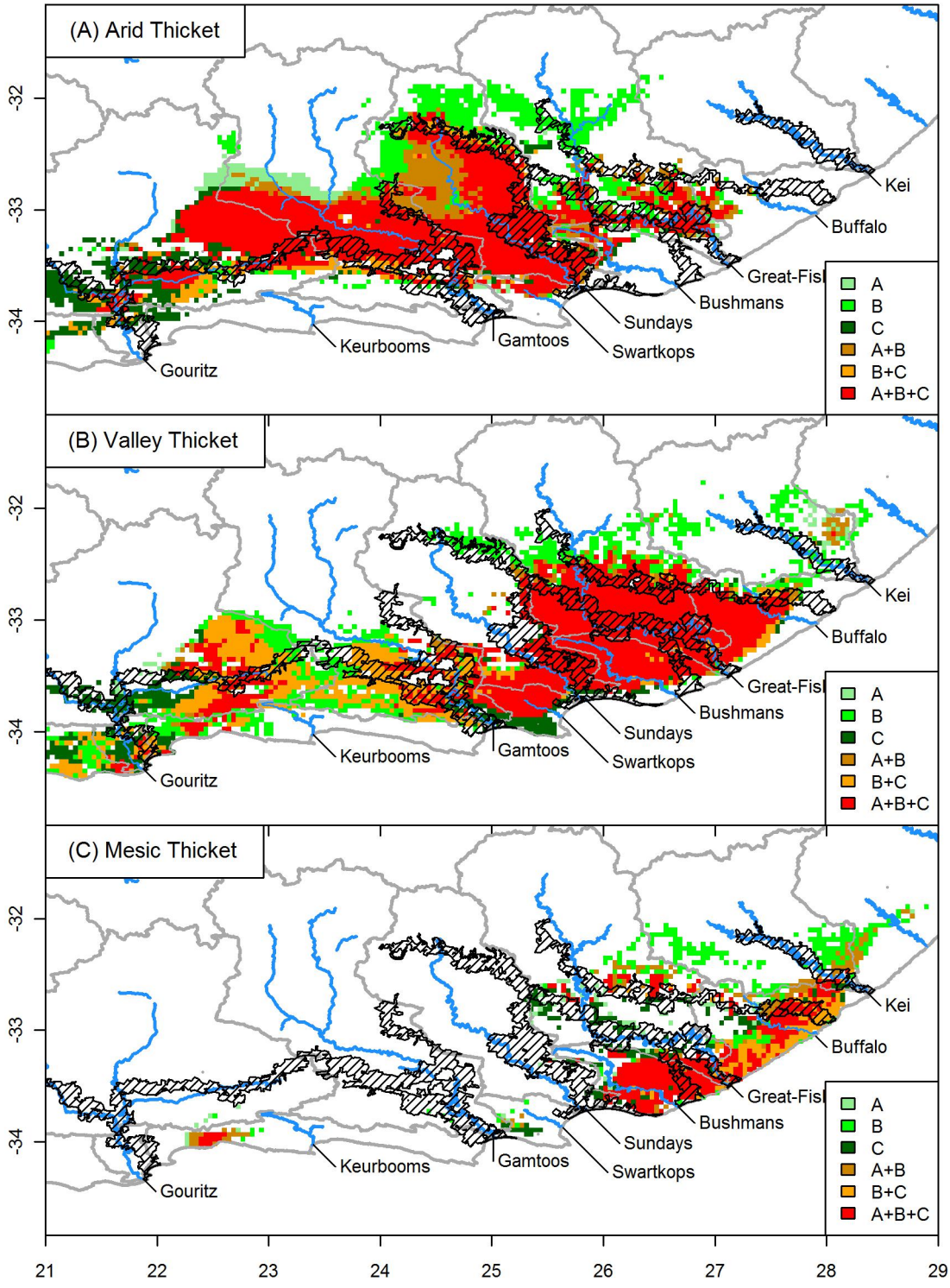

Figure A.8. The proposed megaconservancy network and the community distribution models of three Albany Subtropical Thicket subtypes under projected 2050 scenario *B2a* climatic conditions (shaded). The GCMs used were: CCCMA (A), CSIRO (B), and HCCPR (C).

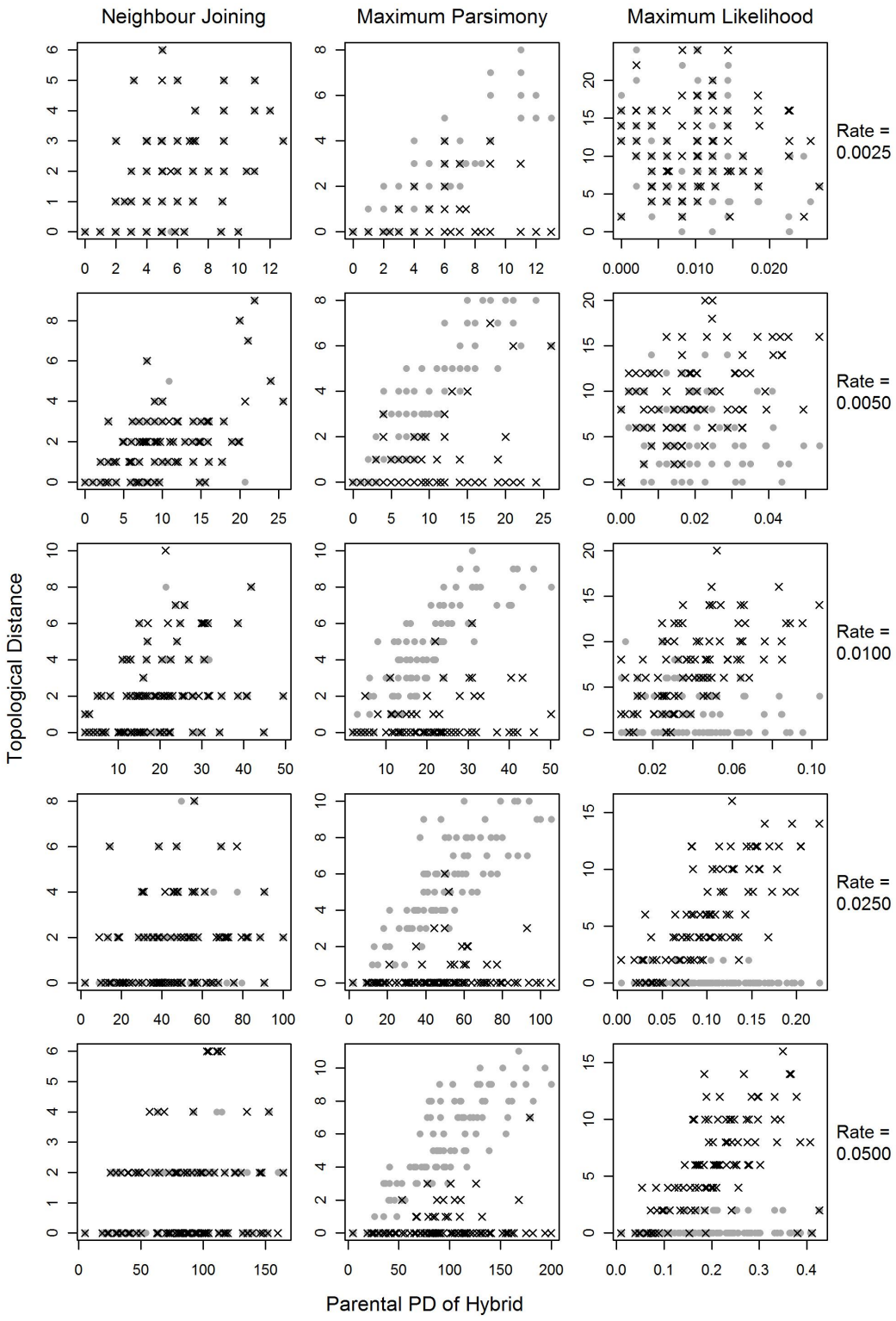

Figure A.9. Parental phylogenetic diversity (PD) of a single hybrid sample and the topological distance between trees inferred from hybrid-free (HF) or hybrid-present (HP) datasets (*simulated under a range of mutation rates*) using different phylogenetic methods with intra-individual site polymorphisms treated as ambiguous (grey circles) or informative (black crosses). Three different methods are used: (A) Neighbour Joining, (B) Maximum Parsimony, and (C) Maximum Likelihood. Different mutation rates were used to generate signal-poor to signal-rich datasets (see text for details, Pg. 62). The frequencies of the topological differences for the informative (black bars) and ambiguous (grey bars) treatments are given on the right-hand side y-axis. Note that topological distances are not directly comparable between the three methods (see text for details).

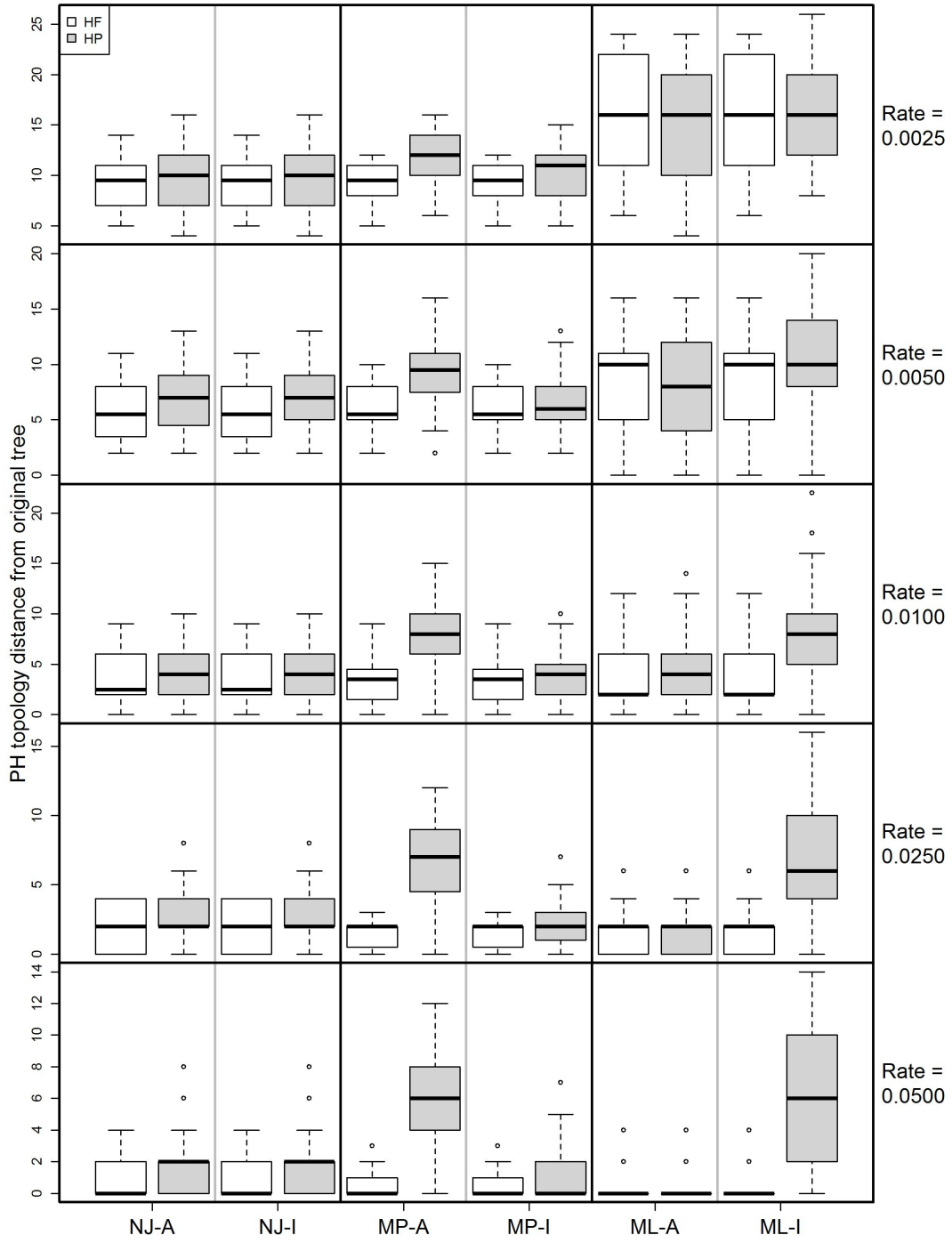

Figure A.10. The topological distance between the original tree used to simulate the data and the trees inferred from the hybrid-free (HF; white) or hybrid-present (HP; light grey) datasets (*simulated across a range of mutation rates*) using Neighbour Joining (NJ), Maximum Parsimony (MP) or Maximum Likelihood (ML) with intra-individual site polymorphisms treated as either ambiguous (-A) or informative (-I). Different mutation rates were used to generate signal-poor to signal-rich datasets (see text for details, Pg. 62). Note that the topological distance of polytomies in the MP consensus trees are not penalised as much as the incorrect relationships in the NJ and ML trees.

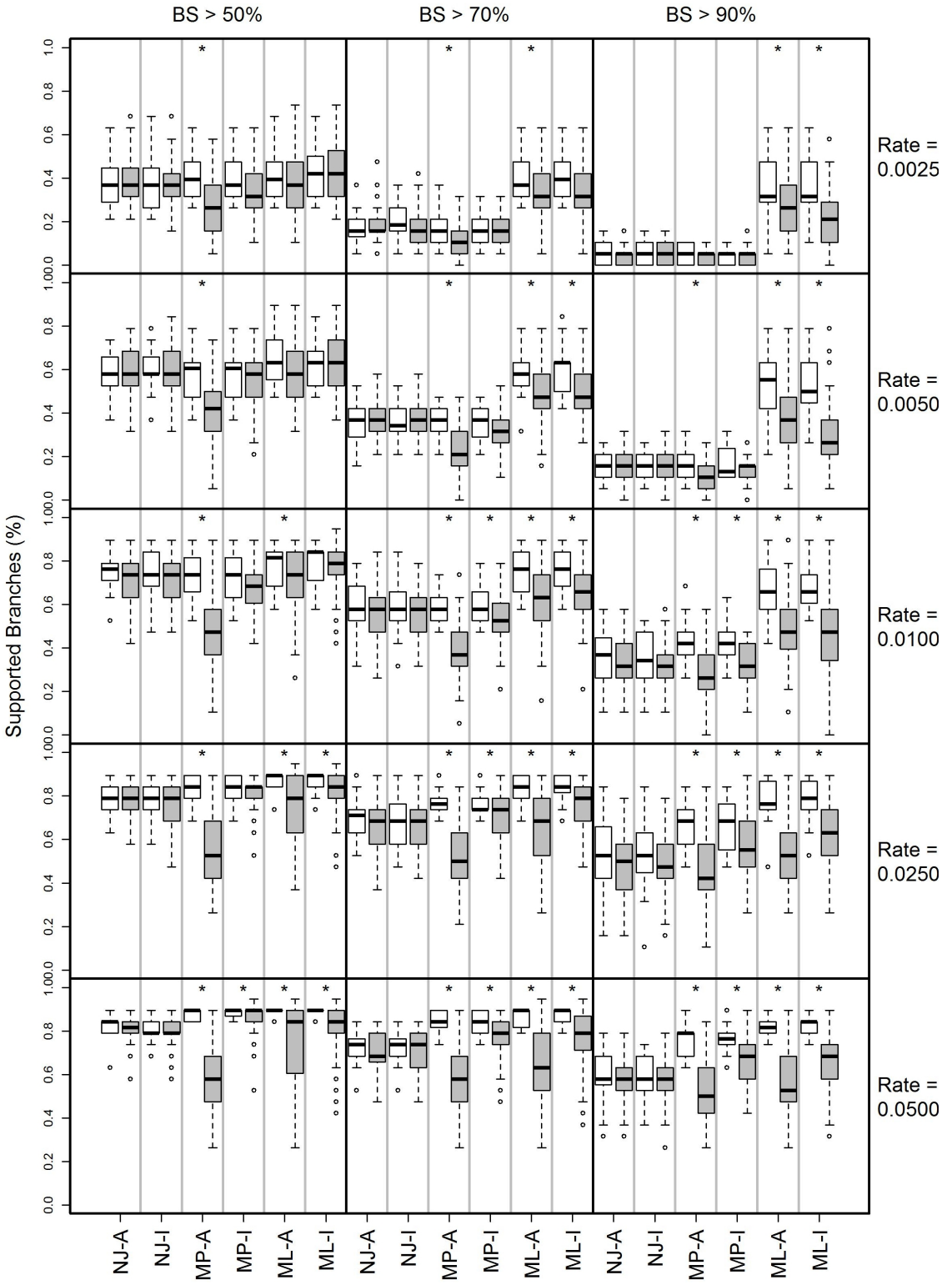

Figure A.11. The percentage of branches with bootstrap values greater than 50%, 70% and 90% from Neighbour-Joining (NJ), Maximum Parsimony (MP) and Maximum Likelihood (ML) trees estimated from hybrid-free (HF; white) or hybrid-present (HP; grey) datasets (*simulated across a range of mutation rates*) with intra-individual polymorphisms treated as ambiguous (-A) or informative (-I). Different mutation rates were used to generate signal-poor to signal-rich datasets (see text for details, Pg. 62). Significant differences, ascertained using a Student's t-test, between HF and HP datasets for a given method are indicated with stars (*).

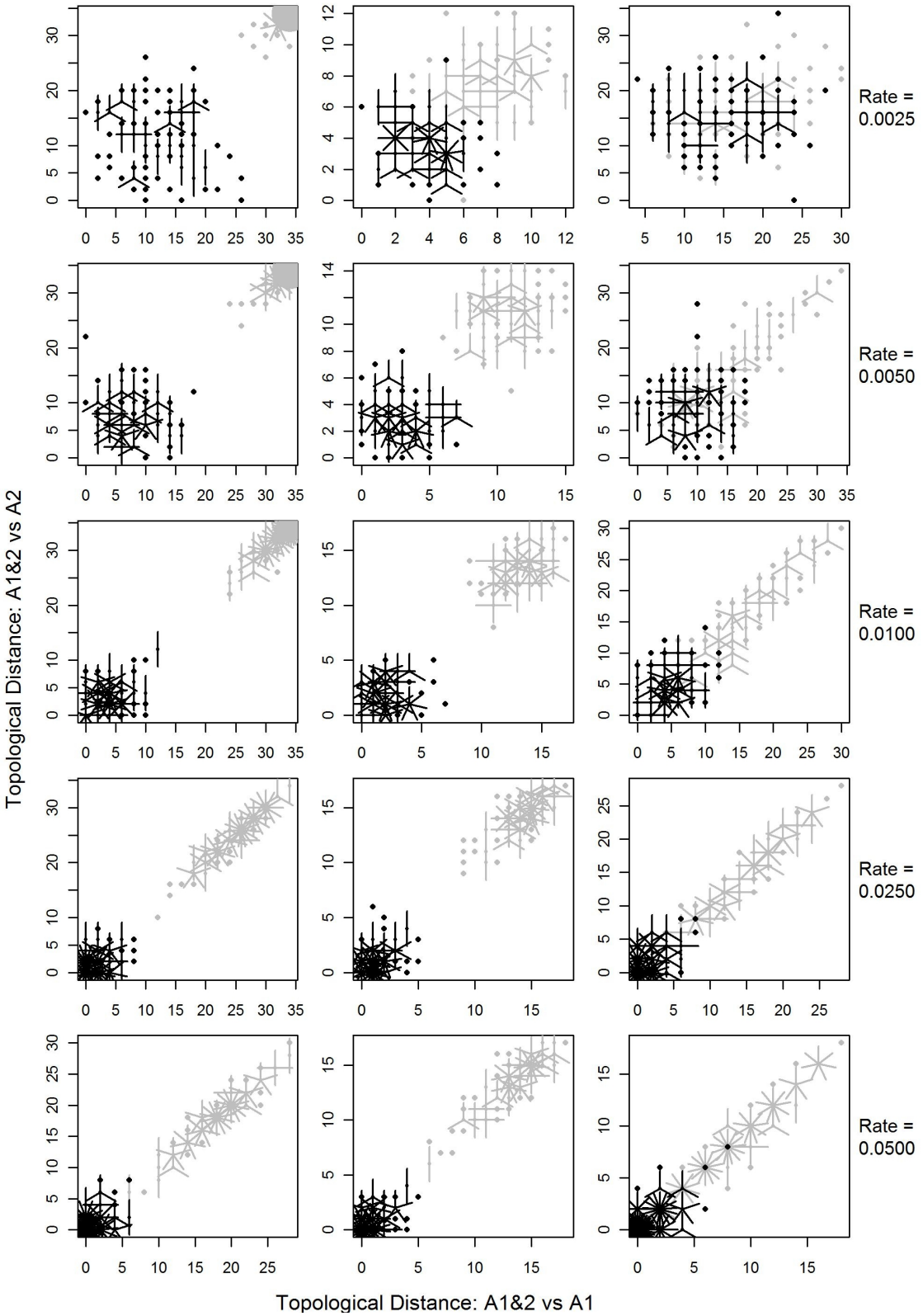

Figure A.12. Topological distance between trees inferred from combined variants ($A_{1\&2}$) and individual variant (A_1 and A_2) datasets (*simulated across a range of mutation rates*) using (A)Neighbour Joining, (B) Maximum Parsimony, or (C) Maximum Likelihood with intra-individual site polymorphisms treated as ambiguous ('grey') or informative ('black'). The number of samples that overlap on a given point correspond to the number of petals shown. The mutation rate per branch length used to simulate the data was 0.0050 (see text for details, Pg. 62).

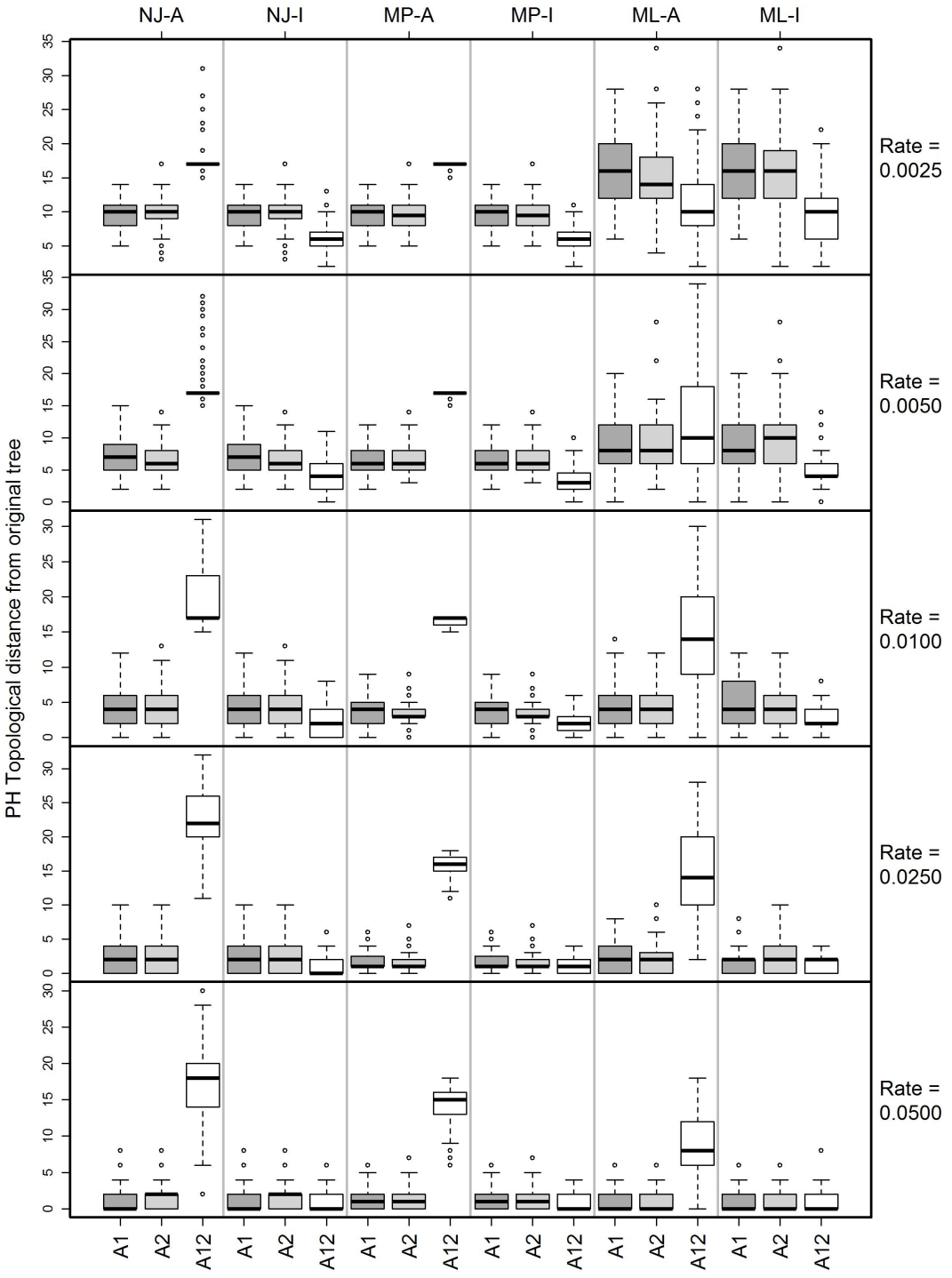

Figure A.13. The topological distance between the original tree used to simulate the data and the combined variants ($A_{1\&2}$, white) and individual variant (A_1, dark grey; A_2, light grey) datasets inferred using Neighbour Joining (NJ), Maximum Parsimony (MP) or Maximum Likelihood (ML) with intra-individual site polymorphisms treated as either ambiguous (-A) or informative (-I). Different mutation rates were used to generate signal-poor to signal-rich datasets (see text for details, Pg. 62).

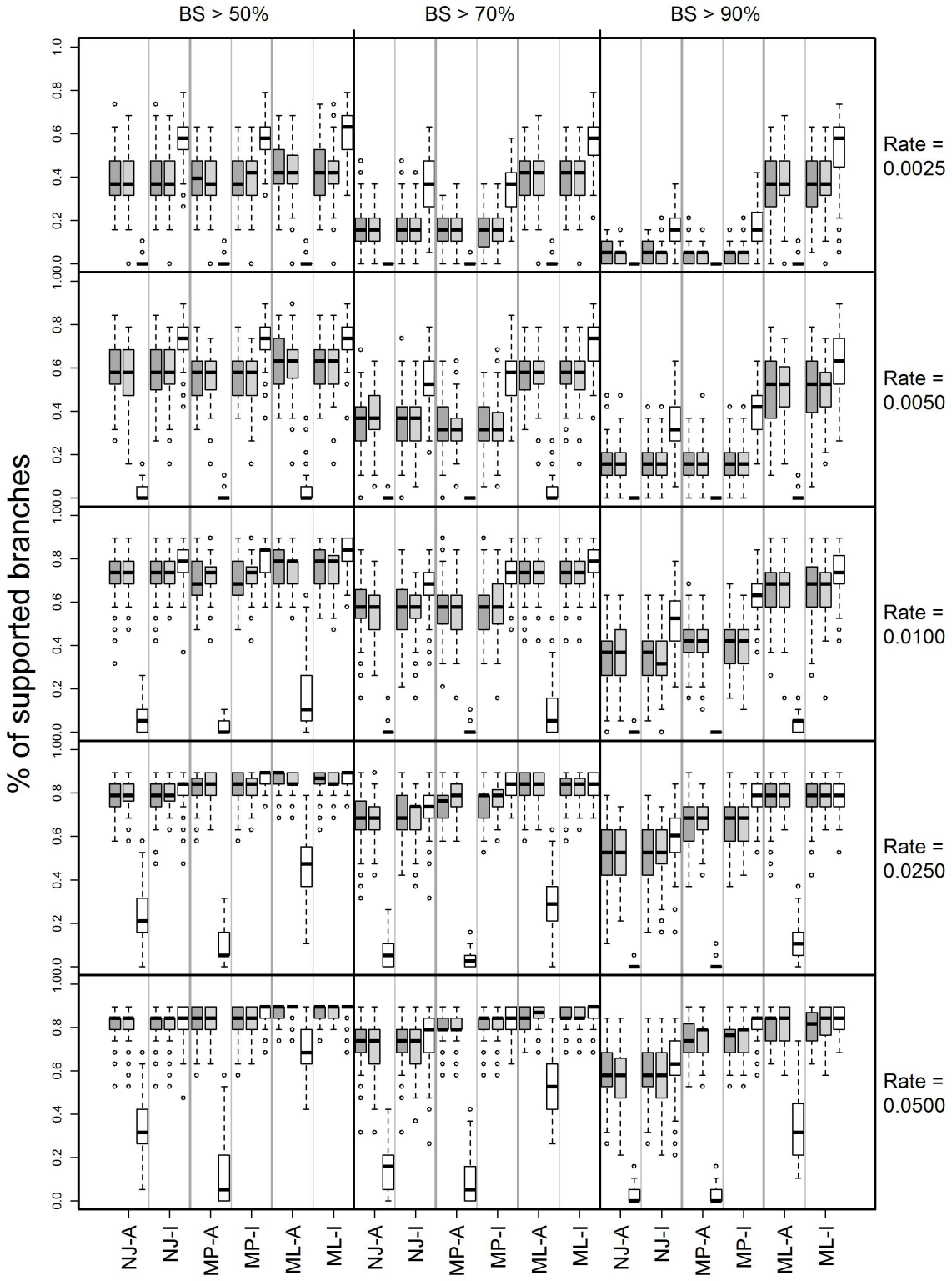

Figure A.14. The percentage of nodes with low, medium and high bootstrap support values (≥50%, ≥70% and ≥90%, respectively) from trees inferred from the combined variants and individual variant (A_1, dark grey; A_2, light grey) datasets and analysed using Neighbour Joining (NJ), Maximum Parsimony (MP) or Maximum Likelihood (ML) with intra-individual site polymorphisms treated as ambiguous (-A) or informative (-I). Different mutation rates were used to generate signal-poor to signal-rich datasets (see text for details, Pg. 62).

Table A.1. Origin of *Nymania capensis* specimens used for phylogeographic analyses. All samples were collected from South Africa. All voucher specimens are retained by the collector: AJP, A. J. Potts. Two samples are stored in the Bolus Herbarium (BOL). Chloroplast haplotypes (see Figure 4.3 [Page 121]), ITS clusters (see Figure 4.4 [Page 122]) and ncpGS haplotypes are included (see Figure 4.5 [Page 123]).

Province/Region	District/ Constituency	Voucher Specimen	Latitude	Longitude	Chloroplast hapotype	ITS Cluster	ncpGS haplotype
Eastern Cape	Hankey	AJP0517	-33.806600	24.728710	R	2	*k*
		AJP0518	-33.820320	24.729750	R	2	*k*
	Humansdorp	AJP0586	-33.860020	24.255359	R	2	*m*
	Jansenville	AJP0444	-33.215200	24.839900	N	3	*k*
		AJP0446	-32.755600	24.723500	L	3	*k*
		AJP0466	-33.210417	24.839944	L	3	-
		AJP0490	-32.731370	24.709310	N	3	*k*
		AJP0492	-32.839710	24.713340	N	3	*k*
		AJP0495	-33.000480	24.745880	N	3	*k*
		AJP0498	-33.029240	24.811260	N	3	-
		AJP0628	-33.092010	24.886390	N	3	*l*
		AJP0632	-33.077230	25.035360	N	3	*k*
	Pearston	AJP0726	-32.484020	25.203250	K	3	*k*
		AJP0729	-32.485970	25.234490	Q	3	*k*
	Somerset East	AJP0465	-33.251306	25.442722	M	3	*k*
		AJP0633	-33.070510	25.176080	O	3	*k*
		AJP0637	-33.039110	25.282500	O	3	*k*
		AJP0718	-32.624990	25.455020	K	3	*k*
		AJP0722	-32.679940	25.344410	Q	3	*k*
		AJP0744	-33.107720	25.897590	K	3	*k*
	Steytlerville	AJP0537	-33.329350	24.674060	R	3	*j*
		AJP0540	-33.309960	24.358470	R	2	*i*
		AJP0545	-33.226750	24.198500	T	2	*i*
		AJP0551	-33.276000	24.134070	R	2	*k*
		AJP0553	-33.228690	24.088070	R	2	*j*
		AJP0810	-33.426440	24.568880	R	4	*k*
	Uitenhage	AJP0500	-33.341830	24.909610	L	3	*k*
		AJP0532	-33.342660	24.873420	K	3	*k*
		AJP0533	-33.364260	24.818660	N	3	*k*
		AJP0814	-33.626520	25.436450	P	3	*k*
		AJP0822	-33.542490	25.119780	L	3	*k*
	Willowmore	AJP0259	-33.517931	23.751302	R	-	*k*
		AJP0269	-33.508563	23.782502	S	2	*g*
		AJP0270	-33.513244	23.780121	S	2	*k*
		AJP0273	-33.521486	23.760162	R	2	*k*
		AJP0555	-33.143820	23.841080	R	2	-
		AJP0557	-33.094920	23.912550	R	2	*i*
		AJP0774	-33.275740	23.288660	F	1	-
		AJP0780	-33.399970	23.675520	R	2	-

Table A.1. Continued.

Province/Region	District/ Constituency	Voucher Specimen	Latitude	Longitude	Chloroplast hapotype	ITS Cluster	ncpGS haplotype
Northern Cape	Namakwaland	BOL48535	-28.256006	17.241669	U	-	*g*
		BOL60966	-28.316668	17.249999	V	-	-
Western Cape	Calitzdorp	AJP0347	-33.634659	21.697334	E	1	*a*
		AJP0358	-33.646765	21.644877	G	1	*a*
		AJP0359	-33.652930	21.639690	G	1	*a*
	George	AJP0312	-33.690933	22.269231	J	1	*a*
	Ladismith	AJP0042	-33.705970	20.968903	H	1	-
		AJP0050	-33.593233	21.201433	B	1	*a*
		AJP0057	-33.474133	21.517933	B	1	*a*
		AJP0083	-33.515213	21.137982	B	1	*a*
		AJP0084	-33.512422	21.097586	B	1	*d*
		AJP0103	-33.478789	20.915313	B	1	*a*
		AJP0113	-33.457254	20.871332	B	1	*a*
		AJP0117	-33.502843	20.800455	F	1	*a*
		AJP0120	-33.531895	20.746276	F	1	*h*
		AJP0121	-33.536824	20.748489	H	1	-
		AJP0129	-33.564117	20.695078	F	1	*a*
		AJP0146	-33.608824	20.626011	B	1	*a*
	Oudtshoorn	AJP0021	-33.509060	22.010440	A	1	*a*
		AJP0065	-33.539567	22.293967	A	1	*a*
		AJP0071	-33.548433	22.463717	A	1	-
		AJP0074	-33.487867	22.561767	F	1	-
		AJP0081	-33.490200	22.631383	A	1	*a*
		AJP0224	-33.610121	22.405158	F	1	*c*
		AJP0229	-33.635754	22.403977	A	1	*a*
		AJP0232	-33.454246	22.560153	A	1	*e*
		AJP0278	-33.428425	22.252120	C	1	*c*
		AJP0319	-33.646264	22.194532	I	1	*b*
	Prince Albert	AJP0789	-33.358260	22.683330	A	1	*b*
	Riversdal	AJP0375	-33.871438	21.447890	D	-	*a*
	Swellendam	AJP0162	-33.657569	20.585944	B	1	*a*
		AJP0280	-33.711470	20.596740	B	1	*a*
	Uniondale	AJP0075	-33.484733	22.774033	F	1	*a*
		AJP0077	-33.527900	22.803783	F	1	*b*
		AJP0080	-33.559400	22.736617	A	1	*c*
		AJP0238	-33.549492	22.802571	A	1	*f*
		AJP0243	-33.491221	23.282956	A	1	*b*
		AJP0274	-33.515851	23.206252	A	1	*a*
		AJP0275	-33.496553	23.102923	F	1	*c*
		AJP0306	-33.548124	22.979498	A	-	*b*
		AJP0307	-33.538869	22.901916	F	1	*a*

Table A.2. Origin of *Pappea capensis* specimens used for phylogeographic analyses. Voucher specimens are retained by the specified collectors: AJP, A. J. Potts; AMM, A. M. Muasya; CM, C. Mannheimer; MR, M. Reekmans. Voucher specimens stored at the South African National Herbarium are preceded with PRE.

Country	Province/Region	District/ Constituency	Collection Number	Latitude	Longitude	Chloroplast hapotype	ITS Cluster
Botswana	South-East	Gaborone	PRE561642	-24.74180	25.847200		4
			PRE561643	-24.82480	25.872700		4
	Southern	Ngwaketse North	PRE563739	-24.980590	25.338640		4
Burundi	Kirundo	Bugabira	RM448	-2.381389	30.036667	W	4
DRC	Katanga	Lubumbashi	Malaisse #13092	-11.533333	27.433332	S	
Kenya	Coast	Taita Taveta	AMM966	-3.405083	38.256528	V	4
	Eastern	Machakos	AMM924	-1.532133	37.309550	X	4
Namibia	Karas	Berseba	AJP0455	-26.233300	16.833300	O	4
			CM3187	-27.526000	17.544000	O	4
		Luderitz	PRE589541	-27.241667	16.487222		4
			CM3183	-26.659722	16.262222	O	4
South Africa	Eastern Cape	Aberdeen	AJP0561	-32.932910	23.748920	B	
		Adelaide	AJP0708	-32.675130	26.334200	D	1
		Albany	AJP0289	-33.139789	26.619172	D	2
			AJP0294	-33.074578	26.744167	E	2
			AJP0296	-33.228997	26.633619	D	2
			AJP0678	-33.488950	26.152600	C	1
			AJP0683	-33.500110	26.370640	B	1
			AJP0690	-33.472710	26.644310	D	2
			AJP0699	-33.119940	26.435460	D	2
			AJP0748	-33.108290	26.086610	D	1
			AJP0835	-33.181700	26.791240	D	2
		Alexandria	AJP0672	-33.356460	26.024520	D	1
			AJP0838	-33.615050	25.910660	B	1
		Bathurst	AJP0291	-33.608861	26.654525	D	2
			AJP0295	-33.410450	27.028994	D	2
		Bedford	AJP0710	-32.636510	26.018550	D	1
		Butterworth	AJP0851	-32.511060	27.980040		2
		Cathcart	AJP0861	-32.266750	27.378410	F	2
			AJP0862	-32.187130	27.313890	F	2
		Fort Beaufort	AJP0292	-32.897367	26.636269	E	2
			AJP0705	-32.850110	26.455340	D	2
		Graaff-Reinet	AJP0467	-32.258000	24.180833	B	1
			AJP0485	-32.144370	24.432850	B	
			AJP0486	-32.334430	24.569890	B	
			AJP0731	-32.394680	24.859000	B	1
		Hankey	AJP0403	-33.838722	24.890361	B	1
			AJP0515	-33.777820	24.698220	B	
		Jansenville	AJP0443	-33.215200	24.839900	B	1
			AJP0445	-32.755600	24.723500	B	1
			AJP0491	-32.808010	24.722620	B	
			AJP0496	-33.000480	24.745880	B	1
			AJP0497	-33.029240	24.811260	B	
			AJP0630	-33.087440	24.952320	B	1
		Kirkwood	AJP0524	-33.478990	25.567050	B	
		Komga	AJP0841	-32.796990	27.856580		2
		Middledrift	AJP0872	-33.019280	26.966990	D	2
			AJP0878	-32.995460	27.000530	D	2
		Mpofu	AJP0882	-32.582340	26.687900	E	2
		Pearston	AJP0728	-32.483690	25.219350	B	1
		Peddie	AJP0288	-33.353544	27.238261	D	2
		Peddie	AJP0873	-33.088760	27.095710	D	2
		Port Elizabeth	AJP0816	-33.833310	25.340870	B	1

Table A.2. Continued.

Country	Province/Region	District/ Constituency	Voucher Specimen	Latitude	Longitude	Chloroplast hapotype	ITS Cluster
		Somerset East	AJP0636	-33.039110	25.282500	B	1
			AJP0652	-33.018490	25.817440	B	1
			AJP0719	-32.627170	25.441320	B	2
		Steytlerville	AJP0401	-33.483472	24.351667	B	1
			AJP0539	-33.290460	24.621560	B	
			AJP0544	-33.227450	24.247660	B	
		Tsomo	AJP0845	-32.275510	27.656970	F	1
			AJP0850	-32.221870	27.738270	F	2
		Uitenhage	AJP0290	-33.770103	25.461956	B	1
			AJP0530	-33.373660	25.086370	B	
		Victoria East	AJP0293	-32.935181	26.837439	D	2
		Willowmore	AJP0245	-33.415275	23.311842	B	
			AJP0252	-33.179785	23.429586	B	1
			AJP0256	-33.252600	23.483687	B	1
			AJP0265	-33.516052	23.775708	B	1
			AJP0268	-33.509622	23.797300	B	
			AJP0272	-33.516326	23.767251	B	
			AJP0400	-33.633917	24.245444	B	1
			AJP0402	-33.513667	23.819833	B	
		Zwelitsha	AJP0839	-33.136580	27.343480		2
	KwaZulu-Natal	Alfred	AJP0448	-30.724500	30.150833	G	3
			AJP0452	-30.724500	30.150833	G	
		Hlabisa	AJP0438	-28.154750	32.012730	I	3
			AJP0439	-28.154750	32.012730	I	3
		Lower Umfolozi	AJP0460	-28.263394	31.821652	H	3
			AJP0461	-28.273790	31.848140	H	2
		Mhlabathini	AJP0462	-28.250200	31.764040		3
		Msinga	AJP0982	-28.790333	30.460833	K	3
		Ngotshe	AJP0990	-27.463889	31.724722	L	3
		Port Shepstone	AJP0969	-30.673772	30.339342	G	2
		Weenen	AJP0980	-28.736389	30.268611	J	3
			AJP0983	-28.734433	30.257833	K	3
	Limpopo	Phalaborwa	AJP0033	-24.294800	30.993100	U	
			AJP0035	-24.286500	30.994500	U	4
			AJP0036	-24.305083	30.979806	U	4
			AJP0436	-23.851667	31.549722	M	3
			AJP0437	-23.851667	31.549722	G	3
		Soutpansberg	PRE562948	-22.943611	29.889167	R	4
			AJP0478	-22.943333	29.938056		
	Mpumalanga	Lydenburg	PRE528625	-25.098500	29.848000		4
		Nelspruit	AJP0447	-25.443222	30.970861	T	4
		Nkomazi	AJP0450	-25.924917	31.814194	L	3
			AJP0451	-25.937472	31.813278	L	3
	North West	Brits	AJP0453	-25.720928	27.807175	Q	
	Northern Cape	Kenhardt	AJP0734	-29.133380	19.401110	N	4
		Namakwaland	PRE575662	-29.216667	18.733333		4
			AJP0475	-29.233333	18.722500		4
			AJP0735	-29.725680	18.037330	N	4
	Western Cape	Calitzdorp	AJP0001	-33.488730	21.642400		
			AJP0002	-33.488560	21.642810	A	
			AJP0003	-33.490310	21.644900	A	1
			AJP0004	-33.507180	21.607430	A	1

Table A.2. Continued.

Country	Province/Region	District/ Constituency	Voucher Specimen	Latitude	Longitude	Chloroplast hapotype	ITS Cluster
		Ladismith	AJP0005	-33.465570	21.515250	A	1
			AJP0006	-33.506560	21.371110	A	
			AJP0007	-33.490070	21.218800	A	
			AJP0107	-33.457254	20.871332	A	1
			AJP0131	-33.572710	20.633875	A	1
		Laingsburg	AJP0456	-33.266664	21.533338		
		Oudtshoorn	AJP0019	-33.550140	22.053710	A	1
			AJP0020	-33.522780	22.033960	A	
			AJP0022	-33.501720	21.962050	A	1
			AJP0023	-33.436870	22.256690	A	1
			AJP0024	-33.522420	22.314170	A	1
Zimbabwe	Matabeleland South	Matobo	AJP0984	-20.526600	28.434200	P	4

Table A.3. Origin of *Schotia* specimens used for phylogeographic analyses. Voucher specimens for AJP samples are retained by A. J. Potts. See Ramdhani *et al.* (2010) for the storage localities of voucher specimens for the remaining samples.

Country	Province/Region	District/ Constituency	Voucher Specimen	Latitude	Longitude	Chloroplast hapotype	ITS Cluster
South Africa	Eastern Cape	Adelaide	AJP0709	-32.675130	26.334200	*a*	*I*
		Albany	AJP0675	-33.371180	26.131690	*a*	-
			AJP0676	-33.483180	26.152040	*b*	*I*
			AJP0684	-33.499450	26.373160	*e*	-
			AJP0692	-33.448820	26.633450	*g*	*I*
			AJP0702	-33.001490	26.437330	*a*	*I*
			AJP0750	-33.097910	26.134490	*g*	*I*
			AJP0756	-33.509670	26.493350	*a*	*III*
			AJP0899	-33.299250	26.813060	*a*	*I*
		Alexandria	AJP0840	-33.615050	25.910660	*g*	*I*
		Bathurst	AJP0689	-33.542950	26.610130	*a*	*I*
			AJP0902	-33.448240	27.049530	*e*	*I*
			NB2003	-33.654100	26.735936	*b*	*I*
			NB1973	-33.596804	26.890985	*a*	*I*
		Bedford	AJP0712	-32.636510	26.018550	*a*	*I*
			AJP0715	-32.730030	25.847500	*c*	*I*
		Butterworth	AJP0853	-32.469120	27.993440	*f*	*I*
		Cathcart	AJP0858	-32.305810	27.371380	*i*	*I*
			AJP0868	-32.200710	27.310710	*a*	*I*
			AJP0870	-32.317450	27.281150	*b*	*I*
		Cofimvaba	SR815	-31.969833	27.448667	*f*	*I*
		East London	TD4485	-32.966347	27.960939	*a*	*III*
		Fort Beaufort	AJP0887	-32.756110	26.588730	*h*	*I*
			NB2012	-33.037955	26.656808	*o*	*I*
		Graaff-Reinet	AJP0937	-32.366010	24.280640	*a*	*I*
		Hankey	AJP0507	-33.840630	24.896720	*a*	*I*
			AJP0514	-33.741980	24.621620	*a*	-
			AJP0807	-33.933410	25.026750	*p*	*I*
			AJP1000	-33.701260	24.551421	*b*	*IV*
		Humansdorp	AJP0928	-34.078639	24.892167	*a*	*I*
			AJP0930	-34.118913	24.739846	*p*	*I*
			AJP0933	-33.829076	24.331914	*j*	*IV*
			AJP0979	-34.190307	24.820542	-	*I*
		Jansenville	AJP0622	-32.852410	24.732280	*a*	*I*
			AJP0626	-33.038500	24.832050	*a*	*I*
			AJP0912	-33.188930	24.846770	*n*	*I*
		Keiskammahoek	NB2020	-32.682520	27.153139	*a*	*III*
		Kirkwood	AJP0761	-33.494500	25.702150	*a*	*I*
			AJP0917	-33.423800	25.246510	*c*	*I*
		Mdantsane	NB2018	-33.115614	27.520451	*a*	*III*
		Middledrift	AJP0876	-33.019280	26.966990	*d*	*I*
			AJP0880	-32.995460	27.000530	*d*	*I*
			AJP0881	-32.817350	27.007980	*o*	*I*
		Mpofu	AJP0889	-32.582340	26.687900	*a*	*I*
		Pearston	AJP0724	-32.671130	25.263680	*c*	*I*
		Peddie	AJP0829	-33.088820	27.094840	*a*	*I*
			AJP0896	-33.412370	27.131810	*e*	*I*
			AJP0877	-33.116890	27.149370	-	*I*

Table A.3. Continued.

Country	Province/Region	District/ Constituency	Voucher Specimen	Latitude	Longitude	Chloroplast hapotype	ITS Cluster
South Africa	Eastern Cape	Port Elizabeth	AJP0521	-33.770330	25.665820	*g*	-
			AJP0923	-33.750950	25.592990	*b*	*I*
			SR822	-33.863889	25.615000	*b*	*I*
			SR816	-33.888056	25.496944	*l*	*IV*
			SR824A	-34.007222	25.458611	*a*	*III*
		Port St Johns	SR785	-31.633611	29.327778	*b*	*III*
		Somerset East	AJP0639	-33.039110	25.282500	*c*	*I*
			AJP0649	-32.888470	25.659840	*c*	*I*
			AJP0659	-33.279170	25.902000	*b*	*I*
			AJP0721	-32.631180	25.442230	*c*	*I*
			AJP0742	-33.146860	25.890580	*b*	*I*
		Steytlerville	AJP0548	-33.297580	24.203200	*a*	-
			AJP0788	-33.363920	23.995590	*c*	*I*
			AJP0828	-33.362010	24.341880	*a*	*I*
		Stutterheim	AJP0859	-32.338910	27.644820	*a*	*I*
		Tsomo	AJP0854	-32.250050	27.660020	*f*	*I*
		Uitenhage	AJP0808	-33.725590	25.434440	*a*	*I*
			AJP0809	-33.447000	24.847870	*a*	*I*
			AJP0811	-33.626520	25.436450	*m*	*I*
			AJP0827	-33.530670	25.102790	*a*	*I*
			AJP0916	-33.372540	25.073840	*a*	*I*
		Willowmore	AJP0795	-33.275740	23.288660	*n*	*I*
			TB01	-33.257914	23.485131	*a*	*I*
			TD4486	-32.266557	28.519085	*a*	-
		Zwelitsha	AJP0836	-33.136580	27.343480	*a*	*I*
	KwaZulu-Natal	Alfred	DS3138	-30.643441	30.011062	*b*	-
			DS3077	-30.697558	30.085946	*b*	-
		Ixopo	DS3144	-30.268286	29.946431	*b*	*III*
		Ngotshe	SR789	-27.525000	32.010833	*q*	*II*
		Simdlangentsha	BP35	-27.349768	31.866181	*q*	*II*
			BP16	-27.356500	31.842170	-	*II*
			BP36	-27.348280	31.828650	-	*II*
	Limpopo	Soutpansberg	sn snout	-22.969500	29.331250	*q*	-
			SR812A	-22.944278	29.904917	*q*	*II*
	Mpumalanga	Barberton	SR799A	-25.501111	31.663333	*r*	*II*
		Lydenburg	ML1144	-24.689100	30.332450	*q*	*II*
	Northern Cape	Namakwaland	VetterX	-28.308508	17.087418	-	*III*
	Western Cape	Knysna	AJP0950	-34.010139	23.373361	*j*	*I*
		Ladismith	AJP0600	-33.485650	20.985530	*k*	*I*
			AJP0609	-33.628140	20.936870	*k*	*I*
			AJP0611	-33.641670	21.032280	*k*	*I*
			AJP0955	-33.579778	20.799306	*k*	*I*
		Riversdal	YvW4407	-34.337987	21.878978	-	*I*
		Mossel	YvW3537	-34.186581	22.050046	*p*	*I*
		Riversdal	SR733	-34.354611	21.420250	*a*	*I*
		Swellendam	AJP0944	-33.764083	20.898000	*k*	*I*
			AJP0945	-33.735056	20.783778	*k*	*I*
			AJP0957	-33.694806	20.697417	*k*	*I*

Table A.4. Character states of ITS clones at variable sites in direct-PCR and clone sequences of *Pappea capensis*. Direct sequences are given on the same line as the sample identifier. Clones are shown below this with identical clone variants bundled together and the number of clones per sequence is indicated at the beginning of each line. Clone sequences are referenced against the direct sequences. All indels detected in clones were a single base in length. Intra-individual site polymorphisms involving indels detected in direct sequences are indicated with 'X'.

		Nucleotide position																																											
		ITS-1																											5.8S					ITS2											
Sample		57	84	89	94	95	97	100	107	110	129	131	134	145	151	152	173	189	190	197	198	199	222	250	251	256	259	263	333	367	368	392	434	466	470	486	488	492	528	529	530	596	617	621	657
AJP0295		C	T	C	G	G	G	X	C	C	C	Y	T	A	C	C	A	G	C	C	G	C	T	C	G	G	-	A	T	T	C	C	C	C	Y	G	G	T	G	C	G	R	C	S	G
	3	.	.	.	.	.	.	-	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	A	.	C	.
	2	.	.	.	.	.	.	-	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	G	.	C	.
	2	.	.	.	.	.	.	-	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	A	.	G	.
	1	.	.	.	.	.	.	G	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	A	.	C	.
AJP0447		C	Y	M	G	G	G	-	T	C	C	T	Y	R	C	T	A	G	C	C	G	C	T	Y	R	G	-	R	T	T	C	C	C	Y	C	G	G	T	G	Y	S	R	C	C	T
	4	.	T	C	.	.	.	.	.	.	.	.	T	A	.	.	.	.	.	.	.	.	.	T	A	.	-	A	.	.	.	.	.	T	.	.	.	.	.	T	C	G	.	.	.
	1	.	T	C	.	.	.	.	.	.	.	.	T	A	.	.	.	.	.	.	.	.	.	T	G	.	-	A	.	.	.	.	.	T	.	.	.	.	.	C	G	G	.	.	.
	1	.	T	C	.	.	.	G	C	.	.	.	T	A	.	C	.	.	.	.	.	.	C	C	G	A	-	A	.	.	.	.	.	C	.	.	.	.	.	C	G	A	A	.	G
	1	.	T	C	.	.	T	.	.	.	.	.	T	A	.	.	.	.	.	.	.	.	.	T	G	.	-	A	.	.	.	.	.	C	.	.	.	.	.	C	G	A	.	.	.
	1	.	C	A	.	.	.	.	.	.	.	.	C	G	.	.	.	.	.	.	.	.	.	C	G	.	-	G	.	.	.	.	.	C	.	.	.	.	.	C	G	G	.	.	.
	1	.	C	A	.	.	.	.	.	.	.	.	C	A	.	.	.	.	.	.	.	.	.	C	G	.	-	A	.	.	.	.	.	T	.	.	.	.	.	T	C	G	.	.	.
	1	.	T	C	.	.	.	.	.	.	.	.	T	A	.	.	.	.	.	.	.	.	.	C	G	.	-	A	.	.	.	.	.	T	.	.	.	.	.	C	G	G	.	.	.
AJP0461		C	T	C	G	X	G	G	C	C	C	Y	T	A	C	C	A	G	C	C	G	C	T	C	G	G	X	A	T	T	C	C	C	C	C	G	G	T	G	C	G	A	C	C	G
	4	.	.	.	.	G	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	2	.	.	.	.	-	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	2	.	.	.	.	G	.	.	.	.	.	C	.	.	.	.	G	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
AJP0467		Y	T	C	G	G	G	G	C	C	Y	Y	T	A	C	C	R	G	C	C	G	C	Y	C	G	R	-	A	T	T	C	C	C	C	C	G	G	T	G	C	G	A	M	C	G
	3	C	.	.	.	.	.	.	.	.	C	T	.	.	.	.	A	.	.	.	.	.	C	.	.	A	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.
	2	T	.	.	.	.	.	.	.	.	C	T	.	.	.	.	A	.	.	.	.	.	C	.	.	G	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.
	2	C	.	.	.	.	.	.	.	.	T	C	.	.	.	.	G	.	.	.	.	.	T	.	.	G	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.
	1	C	.	.	.	.	.	-	.	.	C	C	.	.	.	.	A	.	.	.	.	.	T	.	.	G	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.

Table A.4. continued.

Sample		ITS-1 57	84	89	94	95	97	100	107	110	129	131	134	145	151	152	173	189	190	197	198	199	222	250	251	256	259	263	5.8S 333	367	368	392	434	ITS2 466	470	486	488	492	528	529	530	596	617	621	657
AJP0728		C	T	C	G	G	G	G	C	C	C	T	T	A	C	C	A	G	C	C	G	C	Y	C	G	R	-	A	T	T	C	C	C	C	Y	G	G	Y	G	C	G	A	M	C	G
	2	.	.	.	G	.	.	.	.	C	.	T	.	.	.	.	.	.	.	.	.	.	T	.	.	G	-	.	.	.	.	.	.	.	T	.	.	T	.	.	G	.	C	.	.
	2	.	.	.	G	.	.	.	.	C	.	T	.	.	.	.	.	.	.	.	.	.	C	.	.	A	-	.	.	.	.	.	.	.	C	.	.	T	.	.	G	.	A	.	.
	1	.	.	.	G	.	.	.	.	C	.	T	.	.	.	.	.	.	.	.	.	.	C	.	.	A	-	.	.	.	.	.	.	.	C	.	.	C	.	.	G	.	A	.	.
	1	.	.	.	C	.	.	.	.	C	.	C	.	.	.	.	.	.	.	.	.	.	T	.	.	G	-	.	.	.	.	.	.	.	C	.	.	T	.	.	A	.	C	.	.
	1	.	.	.	G	.	.	.	.	T	.	T	.	.	.	.	.	.	.	.	.	.	C	.	.	A	-	.	.	.	.	.	.	.	C	.	.	T	.	.	G	.	A	.	.
	1	.	.	.	C	.	.	.	.	C	.	C	.	.	.	.	.	.	.	.	.	.	T	.	.	G	-	.	.	.	.	.	.	.	C	.	.	T	.	.	A	.	C	.	.
AJP0816		C	T	C	G	G	G	X	C	C	C	Y	T	A	C	C	A	G	C	C	G	C	T	C	G	R	-	A	T	T	C	C	C	C	Y	G	G	T	G	C	G	A	M	S	G
	2	.	.	.	.	.	.	G	.	.	.	T	.	.	.	.	.	.	.	.	.	.	T	.	.	G	-	.	.	.	.	.	.	.	T	.	G	.	.	.	.	.	C	C	.
	2	.	.	.	.	.	.	G	.	.	.	T	.	.	.	.	.	.	.	.	.	.	C	.	.	A	-	.	.	.	.	.	.	.	C	.	G	.	.	.	.	.	A	C	.
	1	.	.	.	.	.	.	-	.	.	.	C	.	.	.	.	.	.	.	.	.	.	T	.	.	G	-	.	.	.	.	.	.	.	T	.	G	.	.	.	.	.	C	C	.
	1	.	.	.	.	.	.	G	.	.	.	T	.	.	.	.	.	.	.	.	.	.	T	.	.	G	-	.	.	.	.	.	.	.	C	.	G	.	.	.	.	.	A	C	.
	1	.	.	.	.	.	.	G	.	.	.	T	.	.	.	.	.	.	.	.	.	.	T	.	.	G	-	.	.	.	.	.	.	.	C	.	G	.	.	.	.	.	C	C	.
	1	.	.	.	.	.	.	-	.	.	.	C	.	.	.	.	.	.	.	.	.	.	T	.	.	G	-	.	.	.	.	.	.	.	C	.	A	.	.	.	.	.	C	C	.
AJP0873		C	T	C	G	G	G	X	C	C	C	Y	T	A	C	C	A	G	C	C	G	C	T	C	G	G	-	A	T	T	C	C	Y	C	C	G	G	T	G	C	G	R	C	S	G
	7	.	.	.	.	.	.	-	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	A	.	C	.
	1	.	.	.	.	.	.	-	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	A	.	G	.
	1	.	.	.	.	.	.	-	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	A	.	G	.
AJP0969		C	T	C	G	G	G	X	C	C	C	Y	T	A	Y	C	A	G	C	C	G	C	T	C	G	G	-	A	T	T	C	C	C	S	C	X	G	T	S	C	G	A	C	C	G
	3	.	.	.	.	.	.	G	.	.	C	T	.	.	C	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	G	.	.	G	.	.	.	.	.	.
	2	.	.	.	.	.	.	G	.	.	C	C	.	.	T	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	G	.	.	G	.	.	.	.	.	.
	2	.	.	.	.	.	.	G	.	.	T	C	.	.	C	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	G	.	.	G	.	.	.	.	.	.
	1	.	.	.	.	.	.	G	.	.	C	C	.	.	C	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	-	.	.	C	.	.	.	.	.	.
	1	.	.	.	.	.	.	-	.	.	C	C	.	.	C	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	G	.	.	G	.	.	.	.	.	.
AJP0980		C	T	C	G	G	G	G	C	C	Y	Y	T	A	C	C	R	G	C	C	G	C	T	C	G	G	-	A	T	T	C	C	C	C	C	X	G	T	G	C	G	A	C	C	G
	5	.	.	.	.	.	.	.	.	.	C	T	.	.	.	.	A	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.
	2	.	.	.	.	.	.	.	.	.	T	C	.	.	.	.	G	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.
	1	.	.	.	.	.	.	.	.	.	C	C	.	.	.	.	G	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	.

Table A.5. Character states of ITS clones at variable sites in clone sequences from *Schotia afra* and *Schotia latifolia*. Direct sequences could not be obtained from these samples due to an excess of insertion/deletion events in ITS variants. Clones are shown for each sample with identical clone variants bundled together and the number of clones per sequence is indicated at the beginning of each line. The range of bases found in the remaining 81 successfully direct-sequenced samples are shown at the bottom of the table.

		Nucleotide position																																							
		ITS-1																							5.8S				ITS-2													
		17	23	31	36	41	51	52	63	66	67	73	89	90	95	96	101	102	114	192	211	213	218	222	259	267	294	360	430	458	492	504	508	544	550	556	570	577	609	615	623
Consensus		C	-	T	C	C	G	G	G	C	T	C	C	C	C	A	C	G	C	C	C	C	C	C	G	C	C	G	T	C	G	C	G	-	G	-	C	0	C	0	C
AJP0507	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	T	.	A	.	.	G	.	.	.	.
S. afra	1	.	.	.	.	.	.	.	.	T	C	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	1	.	.	.	.	.	.	.	.	T	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	1[a]	.
	1	.	.	.	.	.	.	.	.	T	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	1[a]	.
AJP0600	6	.	.	.	A	.	A	.	.	.	.	.	.	.	.	G	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S. afra	1	.	.	.	A	.	A	.	.	.	.	.	.	.	.	G	.	.	-	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.
	1	.	.	.	A	.	A	.	.	.	.	.	.	.	.	G	.	.	-	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.
	1	.	.	.	A	.	A	.	.	.	.	.	.	.	T	G	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
AJP0899	4	T	C	.	.	T	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	T
S. afra	2	.	.	C	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.
	1	.	.	.	.	.	.	.	.	T	C	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.
AJP0928	2	T	.	.	.	T	.	A	.	.	.	.	T	.	.	.	A	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S. afra	1	T	.	.	.	T	.	A	.	.	.	.	T	.	.	.	A	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.
	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	G	.	T	.	.
	1	.	.	.	.	T	.	.	.	T	C	.	.	.	.	.	-	.	.	.	.	.	.	A	.	.	T	.	.	.	.	.	.	.	T	.	T	.	.	.	.
	1	.	.	.	.	T	.	.	A	T	C	.	.	.	.	.	-	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.
	1	.	.	.	.	T	.	.	A	T	C	.	.	.	.	.	-	.	.	.	.	.	.	A	T	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.
AJP0930	4	.	.	C	.	.	.	.	.	.	.	.	.	T	.	C	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	T	.	.
S. afra	1	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.
	1	.	.	.	.	T	.	.	A	T	C	.	.	.	.	.	-	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	1	.	.	.	C	T	.	.	A	T	C	.	.	.	.	.	-	.	.	.	.	.	.	A	.	T	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.
	1	.	.	.	.	T	.	.	A	T	C	.	.	.	.	.	-	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.
	1	.	.	.	A	.	.	.	.	T	C	.	.	.	.	.	.	C	.	.	T	.	G	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	T	.	.
NB1973	3	.	.	.	.	.	.	.	.	T	C	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S. latifolia	2	T	C	.	.	T	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	T
	2	.	.	.	.	.	.	.	.	T	C	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	T
Base variability		C	0	T	C	T	G	G	G	C	T	C	C	C	C	A	C	G	C	C	C	C	C	G	G	C	C	G	T	C	G	C	G	0	G	0	C	0	C	0	C
in direct		T	X	C	A	C	R	A	R	T	C	Y	T	Y	Y	G	A	A	T	T	Y	T	A	A	K	Y	Y	K	Y	M	A	Y	R	C	K		T		Y	1[a]	T
sequences		Y		Y	M	Y		R		Y	Y		Y			C	G	C	Y	Y		Y	G	R							R		S	X			Y			X	Y
																R	M	R				K	T																		
																M		S				S	Y																		
																S							M																		

All clone sequences are compared to the reference consensus sequence. Number ‘0/1’ in the sequences indicate the absence/presence of length polymorphisms whereby the superscripts identify corresponding character states. A dash (-) denotes a single-nucleotide indel.
a, CAAA

Figure A.15. Phylogeny reconstructions of ITS *Pappea capensis* sequences. Methods used for reconstructions are Neighbour Joining (NJ), Maximum Parsimony (MP) and Maximum Likelihood (ML) with intra-individual site polymorphisms treated as informative characters (see Chapter 3). Dotted branches are reduced by a factor of 5. Samples AJP0006 and AJP0007 were considered to be anomalies as they had exceptionally high number of mutations and were removed from subsequent network and sPCA analyses. Clusters identified in the NeighbourNet network (Figure 5.6 Pg. 163) are shown.

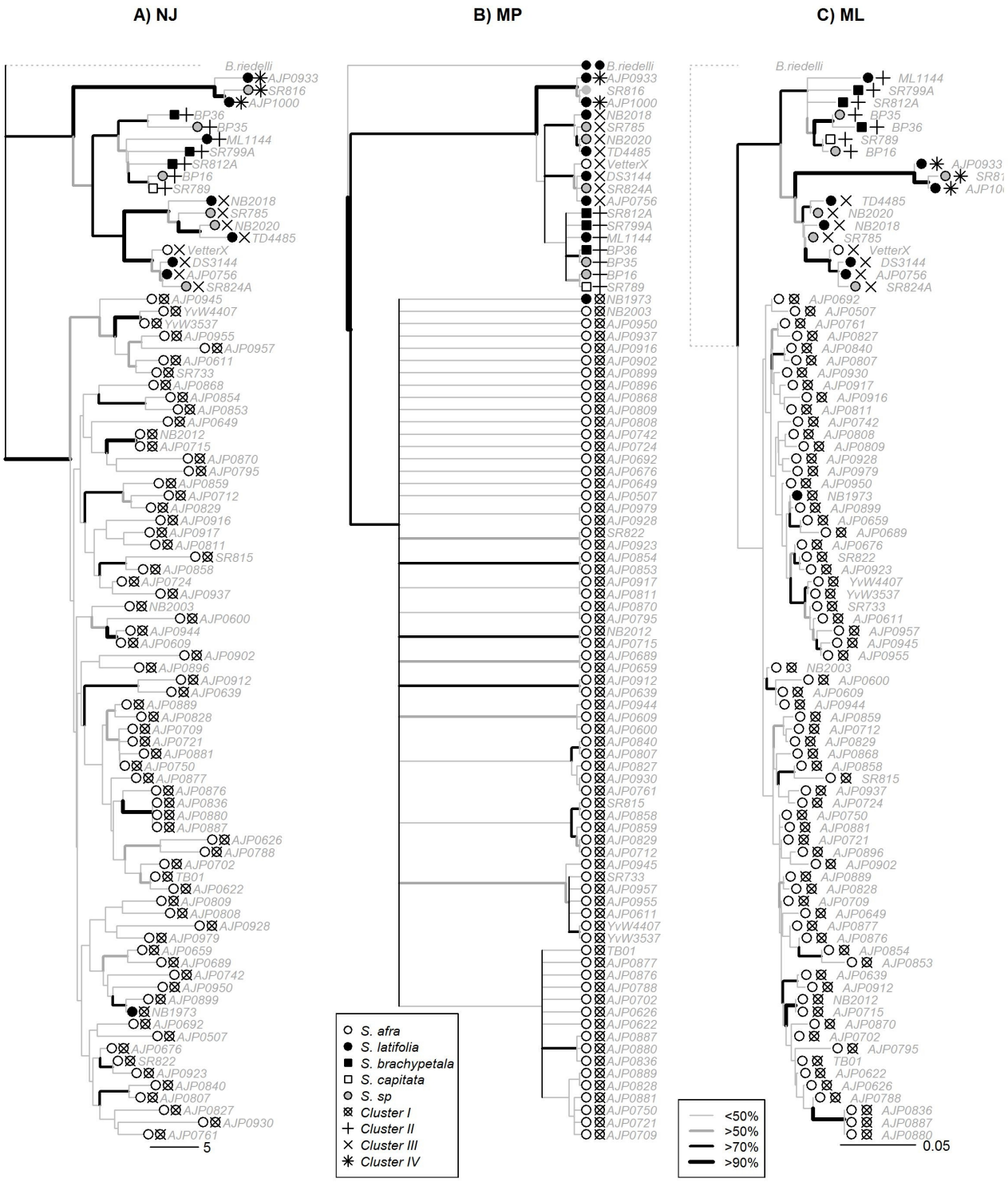

Figure A.16. Phylogeny reconstructions of ITS *Schotia* sequences. Methods used for reconstructions are Neighbour Joining (NJ), Maximum Parsimony (MP) and Maximum Likelihood (ML) with intra-individual site polymorphisms treated as informative characters (see Chapter 3). Dotted branches are reduced by a factor of 5. Clusters identified in the NeighbourNet network (Figure 5.8, Pg. 165) are shown.

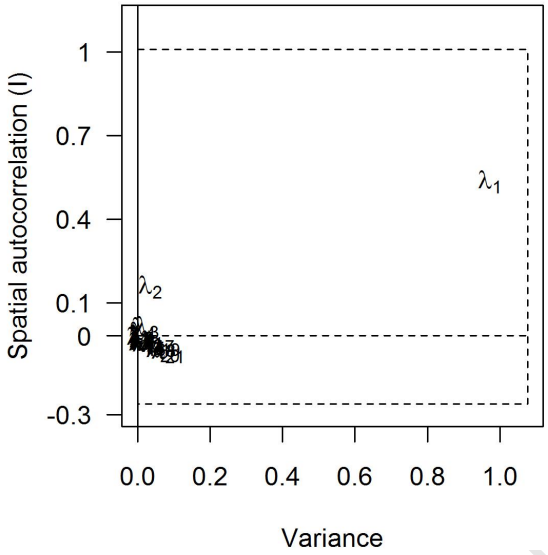

Figure A.17. Spatial and variance components of the Eigenvalues of the sPCA analysis of ITS sequences of *Pappea capensis* samples from the Albany Subtropical Thicket.

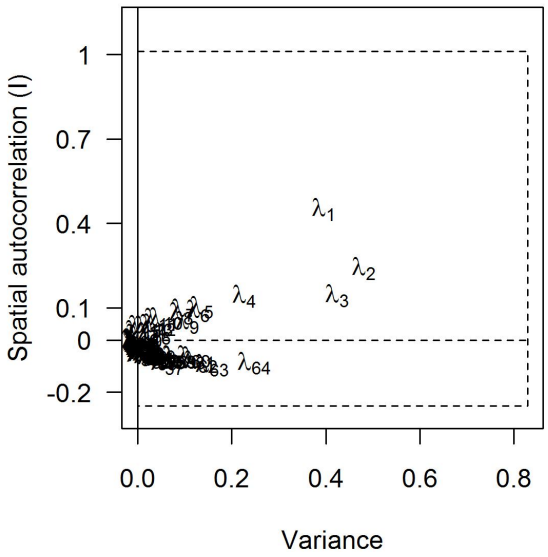

Figure A.18. Spatial and variance components of the Eigenvalues of the sPCA analysis of ITS sequences of *Schotia afra* samples from the Albany Subtropical Thicket.