Speciation and phylogeography: Coalescent-based models applied to the Cape plant genus *Pauridia*

Graham Rowe

Submitted in partial fulfilment of the requirements of the degree Master of Science (Systematics and Biodiversity Science) in the Department of Botany, University of Cape Town.

September 2005

Supervisor: Associate Professor Terry A.J. Hedderson

Acknowledgements

I would like to thank the David and Elaine Potter Charitable Trust for generous financial support throughout the course of my degree and for providing supplementary funding for fieldwork and lab materials. Thanks also to Linda Vranas and Zerina Davids at the Postgraduate Funding Office for their helpfulness.

John Manning originally suggested studying *Pauridia* and I am very grateful for this recommendation, as well as many fascinating conversations. Thanks to Dee Snijman for providing stimulating advice and hard-to-find literature on the Hypoxidaceae. Finding *Pauridia* in the field proved challenging and I am very grateful to (amongst others) the Duckitts, Nick Helme, Gordon Summerfield and the many collectors who lodged their *Pauridia* sheets in the Bolus and Compton herbaria. Dave Gwynne-Evans and Erika Mias gave invaluable practical assistance during fieldwork.

I am very grateful for generous (and accurate) technical guidance from fellow students in the Hedderson lab, especially Tracey Nowell. Students at the Lesley Hill Molecular lab were also very generous with their time. Rene van Zyl of Genecare provided a rapid and accurate DNA sequencing service that made data gathering much easier than it might have been.

Thanks to Gail Reeves for her advice and guidance during the early stages of the project and for the generous provision of lab facilities. I am extremely grateful for Terry Hedderson's enthusiastic, challenging and encouraging supervision throughout the course of this project. It's great to find a teacher that guides you into teaching yourself.

On a personal note I would like to thank my father Neville, mother Ceri and sister Nicola for their continuing love, support and care (not to mention the occasional hot meal!). Thanks also to Maurice, Ant, Lauren and Ryan for good times outside of work hours. Finally, to my love Erika: 'You good thing!'.

Abstract

The Cape Floristic Region (CFR) is an exceptional centre of plant diversity. This diversity is largely concentrated in a profusion of geographically restricted endemic species within a few diversified lineages. The CFR is a rich and dynamic setting for the molecular study of plant speciation, but until recently studies have focused on factors influencing diversification of whole lineages, rather than on the details of the speciation process within species and between sister species pairs. A reconstruction of speciation in the local endemic genus *Pauridia* (consisting of the sister species *P. minuta* and *P. longituba*) based on phylogeographical analysis of chloroplast sequence data is presented. Nested clade analysis and diversity statistics are used to describe patterns of gene flow and diversity within the species. Although genetic data offers great potential for reconstructing past demographic processes, it is necessary to fully consider the stochasticity of the evolutionary process in order to avoid over-interpretation of the data. The 'isolation with migration' model of speciation is therefore applied to the data using coalescent-based tools in order to estimate and obtain confidence intervals on the demographic parameters of gene flow, population size, time of divergence and population size change. These analyses suggest that selection has played a role in driving divergence, because estimates of gene flow and divergence time show that neutral drift could not have resulted in lineage divergence. Multiple lines of evidence also suggest that the more geographically widespread of the species, *P. minuta* was founded by a small population and grew exponentially in population size following speciation. These inferences are placed in the context of the exceptional diversification of the Cape flora and directions for future investigation are suggested.

Table of contents

1. Introduction

1.1 The Cape Floristic Region: Dynamics of diversity in space and time

Situated at the southern tip of Africa, the Cape Floristic Region (CFR) is a climatic and edaphic island that has long been recognized as a highly distinctive phytogeographical unit (Goldblatt 1978). The climate of the region is typically Mediterranean, defined by hot, dry summers and cool, wet winters. Topographically complex sandstone mountain ranges that run parallel to the coast on north-south and east-west axes support heathy fynbos vegetation on oligotrophic soils. Coastal lowlands with a more nutrient rich, shale-derived substrate support the geophyte-rich renosterveld vegetation (Cowling & Holmes 1992). Wide bands of marine sands as well as scattered limestone and granite outcrops occur along the coast, each with their own distinctive species composition. Nutrient poverty and summer drought result in frequent fires that periodically return rare nutrients to the soil, a process that is especially important in the fynbos (Cowling 1992). The region is thus divided into a complex mosaic of edaphic habitats that are subject to regular disturbance and extreme climatic seasonality.

Although similar to other Mediterranean regions in terms of edaphic complexity and gross climate, the CFR is unusually species rich (Linder 2003). The flora consists of approximately 9000 species (68% of which are endemic) in an area of 90 000 km^2 (Goldblatt & Manning 2000). At the local scale however, patches of less than 1 ha are moderately rich, in line with other Mediterranean regions and less than half as rich as tropical rainforest sites (Cowling *et al.* 1996). Local diversity in the CFR is therefore not of outstanding interest.

Differentiation diversity refers to changes in species composition along habitat and geographical gradients. This component of diversity is exceptionally high in the CFR, especially in the west (Cowling & Lombard 2002). Almost complete species turnover between patches along geographical and ecological gradients results in regional species diversity that is twice as high in the CFR as compared to other Mediterranean regions (Cowling *et al.* 1996).

Cowling et al. (1996) dissected the relative roles of soil fertility, topographical and climatic heterogeneity, fire frequency and climatic stability in species richness patterns between Mediterranean regions. They concluded that extreme fire regime (linked to summer drought severity and nutrient poverty) and mild Quartenary climate best explain patterns of diversity. Regional species diversity is also very high in the southwestern Australian Mediterranean climate zone, which shares a mild southern hemisphere climatic history and frequent fire regime with the CFR. Topographical and climatic heterogeneity is as high in other, less diverse regions such as California and the Mediterranean basin and is not strongly correlated with regional species richness.

Cowling and Lombard (2002) have extended this approach to patterns of diversity within the CFR. It has long been recognised that the western region is more speciose and supports more local endemics than the eastern region. On a yearly cycle rainfall is both more seasonal and more predictable in the western region, and moving east this seasonality becomes less pronounced as the frequency of unreliable rain bearing summer storms increases. Research into long-term climatic patterns in the CFR (Barrable *et al.* 2002) suggests that the eastern region cyclically shifted to a more xeric climate during glacial periods as summer rainfall was interrupted. The western region was not affected by summer rain reduction and may even have been more mesic.

Although eastern and western regions exhibit similar patterns of environmental (ie. climatic and topographical) heterogeneity Cowling and Lombard (2002) note that regional diversity is twice as high in the west. This pattern is consistent across montane and lowland habitats. Climatic stability is therefore invoked as the key explanatory variable. Climatic stability on short (seasonal) and long (glacial cycle) time scales therefore emerges as an important correlate of high regional diversity both within the CFR and between Mediterranean regions. I will explore why this might be the case in section three.

1.2 Insights from phylogenetics

Plant taxonomists in the CFR have long been intrigued by the concentration of species diversity in only a few lineages. Just 13 of 988 genera occurring in the CFR contribute 25% of species diversity and 12 of 173 families include 64% of species taxa (Goldblatt & Manning 2000). Linder (2003) has called these diversified lineages "Cape clades". Cape clades have long been recognized as bearing the signature of adaptive radiations, but only recently have phylogenetic analyses begun to reveal the details of the CFR's deep history (Linder & Hardy 2004).

Like other Mediterranean climate regions the CFR was home to a tropical flora during the earl and mid Miocene, and heathy elements were confined to nutrient poor montane habitats (Linder 2003). As global climatic patterns shifted with the growth of Antarctic ice sheets and the establishment of a circum-Antarctic cold-water current, Mediterranean climates began to develop. At the Cape the upwelling of cold bottom water off the west coast resulted in a decrease in summer rainfall and increasing xerification, eventually resulting in highly seasonal rainfall (Linder & Hardy 2004).

Phylogentic analysis of several Cape clades has revealed that a series of adaptive radiations correspond to the opening of the winter-rainfall mega-niche (eg. Klak *et al.* 2004). Phylogenetic dating of these groups suggests that waves of diversifications in diverse clades began at different time points between 42 and 7 million years ago (mya), implying a gradual displacement of the tropical flora as xerification and seasonality intensified (Linder 2003). Within individual clades more recent radiations appear to have occurred in the western summer-dry region. A phylogeny of the genus *Pelargonium*, for example, shows that the most recent radiation occurred along the west coast and is associated with seasonal drought adaptation (Bakker et al. 2004).

The most extreme example of this phenomenon has occurred in the Suculent Karoo, an arid region north of the western CFR that has been subject to similar climatic trends. This region is dominated by the family Aizoaceae which is composed largely of a highly diversified clade of over 1500 species, the core Ruschioideae. A recent phylogenetic study shows that this lineage has radiated spectacularly since the onset of winter-rainfall (3-8 mya) while sister clades failed to radiate (Klak *et al.* 2004)). This burst of diversification is associated with key innovations in leaf shape and cell type that reduce water loss.

Phylogenetic studies have proved extremely valuable in corroborating limited but suggestive evidence from a scarce fossil record and ancient climatic reconstruction (Richardson 2004). Although they improve our understanding of the broad picture of diversification, phylogenetic studies tell us little about the details of the speciation process itself.

1.3 Thinking about speciation in the CFR

Relying on their knowledge of the geographical distribution of morphologically based species taxonomies, early CFR botanists supported a primarily geographic and drift-driven concept of speciation (Goldblatt 1978). This was in line with the dominant ideas of Mayr concerning bottlenecks and geographic isolation and the allopatric incompatibility theory of Dobzhansky (Turelli *et al.* 2001). Under this model species are maintained by internal gene flow and diverge when geographically isolated. Lack of gene flow during isolation leads to the random development of reproductive incompatibility that maintains species if they shift back into sympatry. This process can be speeded by the isolation of small, peripheral populations that undergo loss of genetic diversity. In this view this study of geographic barriers and vicariance is central to understanding speciation.

The work of Linder (1985) signalled the move towards an ecological perspective in CFR thinking on speciation. This was largely motivated by the growing realisation that gene flow within plant species is generally low and by the need for a less stochastic explanation for very rapid speciation in the CFR. The typical CFR species, split into multiple isolated populations with no obvious long-distance dispersal mechanisms, appears to represent an extreme case of low gene flow (Cowling & Pressey 2001). Since gene flow is too restricted to play a significant role, speciation occurs as populations are subjected to divergent selection pressures. Geography is only significant insofar as it corresponds to edaphic and climatic variations in space. These ideas stress the importance of species response to ecological pressures in explaining rapid diversification.

A recent summary of ecological thinking on speciation in the CFR is presented by Cowling and Pressey (2001). They present a 'micro-geographic' model of speciation. In this model a typical CFR species is composed of multiple geographically isolated populations weakly connected by gene flow and subject to periodic disturbance. Disturbance occurs both through periodic fires on the scale of decades and glacial-cycle driven range shifts over thousands of years. Patterns of range shift and population extinction potentially can exacerbate low gene flow by increasing inter-population distance. Disturbance also forces populations through periodic bottlenecks that allow for the rapid fixation of genetic differences.

Given the environmental complexity of the CFR a peripheral isolated population that survives will often be subject to a different set of ecological pressures to its parent. These might manifest as pollination, edaphic or climatic differences for instance (Linder 2003). The isolated population will tend to adapt to this habitat, randomly accumulating reproductive incompatibility differences during this process. This model accounts for the fact that much species diversity is concentrated in range-restricted habitat specialist species within lineages that possess the key adaptations necessary to compete in the CFR.

The role of bottlenecks and disturbance in generating diversity is by no means clear and recent thinking indicates that gene flow at neutral loci and geographical factors may not be as significant as previously thought to species maintenance or splitting (Rieseberg *et al*. 2003). In contrast, the concept of selection-driven or ecological speciation (McKinnon et al. 2004; Rundle & Nosil, 2005) is gaining growing support. More detailed analyses of species concepts and speciation processes are presented in the following section, and a modification of the CFR-specific model of Cowling and Pressey (2001) is used to generate a set of testable hypotheses for *Pauridia* in section 8.

Testing speciation ideas requires that we gather information on inter- and intra-species demographic history (Hey & Wakeley 1998). This includes data on population size dynamics, the direction and extent of gene flow and the timing of splitting events. This information must be placed in its geographical, ecological and historical context in order to be interpreted in a meaningful way (Hewitt, 2001). Moving towards a set of testable hypotheses requires that we first define exactly what we imagine species and speciation to be (section 4) and then select a coherent model to link data, demographic parameters and evolutionary process (section 5). We can then formulate a strategy for gathering data and statistically testing alternative phylogeographical hypotheses of speciation (section 6).

1.4 Imagining species and speciation

When referring to 'species' I follow the analysis of Hey (2001) and Hey *et al.* (2003). Taxonomic species are categories with associated definitions and therefore exist only in the realm of ideas. They are useful in that they can serve as hypotheses that may lead us to a deeper understanding of the complex history underlying the formation and maintenance of evolutionary lineages, which are real entities that exist in space and time. Although the observation of very clear patterns might cause us to reformulate our hypotheses (the taxonomic species) the central aim of any study of species and speciation should be as complete a description as possible of the underlying evolutionary process. This will necessarily entail boundaries that are indistinct and processes that are very subtle, because real lineages undergo constant evolutionary modification and adaptation. What would the key features of such a description be?

Many students of speciation have focused on the geographical mode of speciation because of its obvious link to the limitation of gene flow (Turelli *et al.* 2001). Speciation processes are therefore classified according to the relative geographical position of the splitting lineages. The reason for this focus is that speciation seems to require that gene flow be limited between lineages in order for differences to evolve independently. Fundamental to this view is the idea that the whole individual and its entire genome are the unit of speciation. If the component parts of the genome are mostly co-adaptive, then any gene flow will tend to homogenize divergent populations and prevent speciation. Reproductive isolation (RI) itself must therefore arise in the absence of gene flow and, logically, the absence of selection for RI. Since the defining feature of speciation is a neutral process, speciation itself must be a neutral process.

Genic species

The above analysis of Mayrian speciation is due to Wu (2001) who presents an alternative view of the speciation process. Recent empirical evidence shows that in many taxa the history of species and speciation is shaped by only a few loci, while most of the genome tracks broad demographic processes. In maize, attempts to track the history of the domestication process have been hampered by the fact that most loci show uninformative polyphyly relative to potential wild ancestors, the teosinte grasses (Wang *et al.* 1999). Given recent divergence and coalescent stochasticity this is to be expected (Funk & Omland 2003). In contrast to neutral expectations, a phylogeny of one particular genomic region (*tb1*) related to the unique morphological structure of domesticated maize resolves domesticated maize as monophyletic. This combined with unexpectedly low diversity within this region led the investigators to conclude that this particular region has been the subject of intense selection for the desirable high-yield morphology of maize. Genomic regions only 2kb away from *tb1* are polyphyletic relative to teosinte species, indicating that the combined forces of selection and recombination have precisely dissociated this particular region from the rest of the genome. 'Speciation genes' have also been discovered in some *Drosophila* species (eg. Greenberg *et al.* 2003) and are similarly isolated from the surrounding genome. More generally, many plant and animal species show strong correlation between a small set of genes and phenotype (Orr, 2001).

This growing empirical evidence leads to a view of speciation that is driven by selection at a relatively small number of loci that contain information about speciation, while the majority of neutral 'marker loci' track broad demographic trends, at least during early speciation (Morjan & Rieseberg 2004). Absolute barriers to gene flow are not required for speciation to proceed. Rather, divergent selection at particular loci maintains species identity in the face of gene flow. As the number of loci linked to divergent traits grows with increasing adaptive divergence, it is expected that RI will evolve to completely isolate two divergent lineages. Far from being seen as neutral, the emerging view of RI is that it is a complex and directed state involving many loci in complex relationships, with variable selection shaping these relationships (Wu, 2001).

Note that genic speciation is not equivalent to the retention of ancestral polymorphism followed by a selective sweep in one population. Although this might leave a similar genetic signature, divergent selection need not be invoked. The selective sweep is not the cause of divergence, but rather a result. Wu (2001) explains that in the case of genic speciation, the speciation loci will have the most ancient divergence, because their origin precedes neutral isolation in the rest of the genome. A selective sweep will follow the isolation of populations, leaving a different signature.

Genic integration

The corollary to the idea of 'genic speciation' is the 'genic view of species integration' (Rieseberg & Burke 2001; Rieseberg *et al.* 2003).Observations of species maintenance in the face of gene flow and strong evidence for ecologically-mediated parapatric speciation in plants have stimulated interest in selection-driven concepts of speciation. Botanists have had little difficulty in imagining a role for selection in speciation (Grant 1981) but in the face of low gene flow the question of species maintenance has been raised. Once species have split they must cohere as units by internal gene flow. It is difficult to see how this could occur given observed measures of gene flow in many natural plant populations (Linder 1985). To resolve this difficulty it is necessary to imagine gene flow at the level of the gene, rather than the whole individual.

Concerns that gene flow is too limited to maintain species are based on the assumption of neutrality. If strong selection acts to promote the spread of certain loci, then this problem falls away. Strong selective advantage for specific loci is obviously the case for the known 'speciation genes' , but there is also growing evidence from QTL studies (Morjan & Rieseberg 2004; Orr 2001) that natural selection acts strongly and directionally on multiple loci in diverse taxa to promote phenotypic change. Even very subtle selection-driven shifts in allele frequencies result in pronounced phenotypic effects in at least some species (Latta 2003). If this is generally true, then phylogeographical patterns of neutral markers will only describe crude gene flow, while the more meaningful patterning of adaptive alleles goes undetected.

The size and variability of the units of selection (ie. non-recombining loci) is obviously highly relevant to genic speciation and maintanence. Although the study of speciation genes in maize (Wang *et al.* 1999) and *Drosophila* (Greenberg et al. 2003) have shown that selection can act specifically on small sections of the genome, the extent of the protected area is dependent on the surrounding rate of recombination. Work in sunflowers has shown that recombination modifiers (specifically chromosomal rearrangements) can protect larger sections of the genome (Rieseberg 2001). Selective reproductive isolation then becomes mosaic within the genome, with certain co-adaptive regions potentially becoming protected from recombination.

Application

A genic view of the maintenance and divergence of species is a necessary condition for ecological (Schluter 2001) models of speciation. The active role of natural selection in both spreading and limiting the movement of particular loci in divergent adaptive environments gives rise to the phenotypic divergence we observe in nature. Neutral loci will initially not reflect these patterns, but as they become linked to selected loci by recombination modifiers and eventually general reproductive isolation they will coalesce within species boundaries. Ideally a study of speciation would include both directly selected loci relevant to the pressures driving adaptation to divergent environments and neutral loci linked, by varying degrees, to 'speciation genes'. When this information is placed in a phylogeographical context and interpreted in terms of model-based coalescent analysis, it should be possible to observe mosaic patterns of gene flow, isolation and demographic dynamics according to selective pressures acting variably across the genome.

The ideal analysis described above is only possible in the very few taxa for which genes involved in speciation are known, and multiple neutral markers are available. In the vast majority of taxa it is only immediately possible to use the uniparentally inherited organellar genomes, the chloroplast and mitochondria. These are generally assumed to be single non-recombining neutral loci (but see Ballard & Whitlock 2004). The usual approach is therefore to use the single neutral marker available and treat taxonomic species boundaries as initial hypotheses. The case for species status is strengthened if identity is maintained in sympatry. Neutral (phylogeographic) and selected (directly observable) patterns can therefore be compared and contrasted. Examples of this type of study and the associated tools will be presented in section 7, but first I discuss the theoretical basis for linking DNA polymorphism data, demographic parameters and models of speciation.

1.5 Testing speciation hypotheses using DNA data: coalescent models and phylogeography

The coalescent

In order to test alternative historical hypotheses using extant molecular data we need a model that utilises all the historical information in the data but at the same time accounts for the stochastic variance in the underlying evolutionary process. Because evolution is an extremely complex process and molecular data is information rich it is very likely that patterns will emerge from any given analyses. Failure to assess the significance of these patterns in terms of well-defined biological models can lead to pronounced over-interpretation of the data (Rosenberg & Nordborg 2002; Knowles & Maddison 2002; Nichols 2001; Maddison 1997). The stochastic process known as the coalescent is a simple, adaptable and statistically rigorous model for the simulation of genetic polymorphism data according to biological parameters relevant to intra-specific variation.

The coalescent is based on the insight, due to Kingman (1982), that genealogical processes (i.e. patterns of descent) are much easier to model backward in time than forward. Imagine a sample of related, selectively neutral genes picked from a population. As we trace descent forward in time each individual gives rise to some random number of offspring. If it produces zero offspring the lineage terminates, if not it continues and varies in frequency in the population. Every time a copy is made there is some probability that a mutation will occur. The processes of descent and mutation therefore give rise to patterns of variation in related genes. The difficulty with this process is that modelling an entire population is extremely computationally intensive.

Now imagine the same process tracked backward in time. We start with a set of extant individuals. As we trace lineages backward in time each individual picks a parent randomly from the previous generation. Occasionally two individuals will pick the same parent, with a probability that is dependent on, amongst other factors, population size. The joining of two lineages, coalescence, is more likely in a given generation in a small population and less likely in a large population (Nordborg 2001). As this process proceeds backward in time and lineages continue to coalesce we eventually arrive at the most recent common ancestor of our sample (see Figure 1.1). Expressing this model mathematically gives us a simple stochastic algorithm for constructing possible genealogies of extant individuals depending on a biological parameter, population size. The model is computationally tractable because we only model the lineages relevant to our sample.

Because genes are selectively neutral and accumulation of mutations is random in the model, genealogical and mutational processes are decoupled. This means that expected patterns of current polymorphism can be generated by stochastically 'dropping' mutations forward in time onto simulated genealogies, to create a set of polymorphism data and corresponding gene trees (Rosenberg & Nordborg 2002). All parameters in coalescent models are therefore scaled according to μ, a measure of the mutation rate. The beauty of the coalescent model is that it is not limited to an ideal population. Many features of real populations, such as age structure and sex ratio can be accommodated simply by scaling the effective population mutation rate (θ), a function of effective population size. Factors such as population substructure and size change affect the shape of the genealogy, but do so in a way that can be predicted and modelled. The action of selection, however, cannot be accommodated because it results in a coupling of mutation accumulation and descent, violating the assumptions of the model. The coalescent therefore provides a framework for linking extant polymorphism data, a set of possible gene trees and the biological forces that shaped these gene trees.

By simulating a sample of gene trees using identical parameters we can immediately see the danger of directly interpreting a single gene tree. This problem is especially severe when we attempt to model divergence between two populations (see Figure 1.2). Stochastic variation in topology and branch length tends to overwhelm the biological signal. The extant polymorphism data that we collect and its associated 'best' gene tree is just one of many realisations of the stochastic processes of descent and mutation

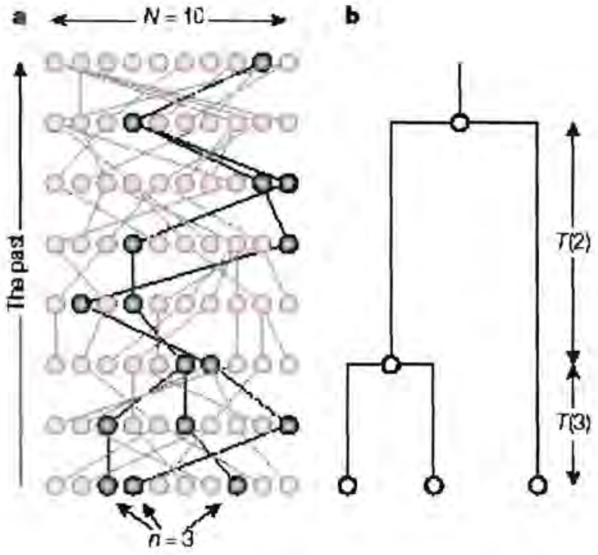


Figure 1.1: Modelling genealogies back in time. The diagram on the left shows the complex ancestor-descendant relationships in a population of 10 individuals. By simply tracking when individuals pick the same parent (i.e. when lineages coalesce), we can derive a very simple description of the history of the sample. This is given by the genealogy on the right. Note that the probability that two lineages coalesce depends on the number of lineages and the number of individuals and is highly stochastic. Randomly inserting mutations onto the genealogy allows us to simulate extant polymorphism data.

This figure is taken directly from Figure 4 of Rosenberg and Nordborg (2002)

(Nordborg 2001; Hey & Machado 2003). How is it possible then to infer evolutionary processes from polymorphism data?

Collecting more individuals for a given gene is unlikely to be of much use. This is because new individuals are sampled from the same realised genealogy. Genealogies by nature have many shallow branches and few deep branches, so increased sampling will tend to accumulate shallow branches and add little historical information. The decline on investment in new individuals actually declines exponentially (Nordborg 2001). Adding gene regions is much more productive, because we gain independent genealogical realisations of the underlying evolutionary forces.

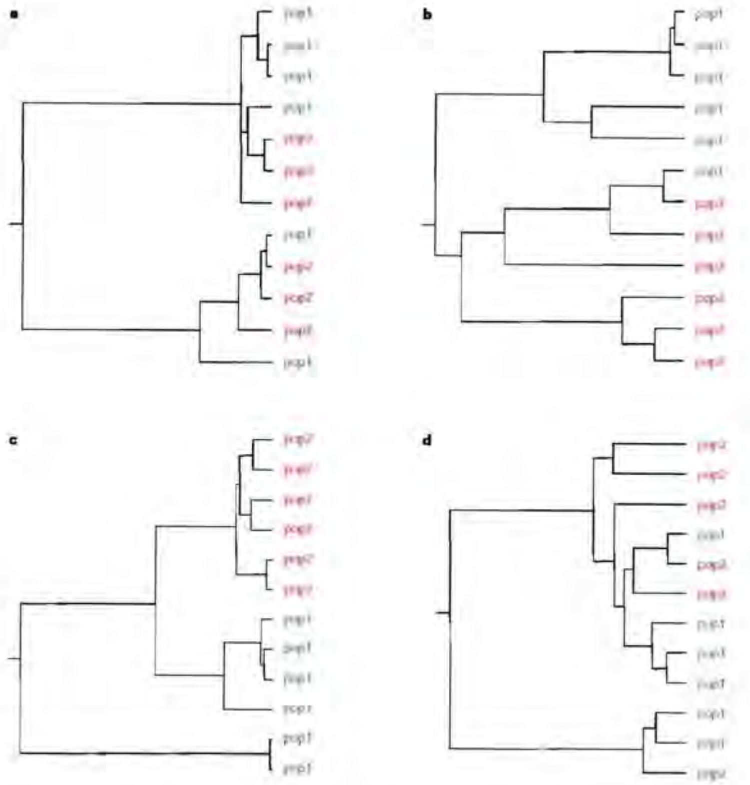


Figure 1.2: Coalescent stochasticity in the genealogies of two diverging populations (pop1 and pop2). All four gene trees are simulated with the same divergence time (4N generations), population size (2N) and migration rate (0.5 migrants per generation in both directions). The very widely varying genealogical patterns do not actually give any directly accessible information about the divergence of the two populations.

This figure is taken directly from Figure 3 of Hey and Machado (2003)

Coalescent based models

Any strategy for inferring biological parameters from polymorphism data must incorporate uncertainty about individual gene trees. Ideally this can be achieved by integrating over all possible gene trees, each weighted by its fit to the data. This is possible if there are very few tips to the tree, but any reasonable number of samples results in an extremely large number of trees, far more than could ever be computed. It is therefore practically impossible to get an exact solution to our problem, but a good approximation is possible. Out of the enormous universe of possible gene trees, many will be of very poor fit to the data and very few will fit well. Because trees are weighted according to their fit, most of the information is contained in the small and tractable set of 'good' trees (Stephens 2001). The problem then becomes how to find and sample a representative set of the best trees.

We can imagine how to deal with this problem if we visualise a tree space. This landscape contains all the possible trees and each tree is surrounded by neighbours that are related to it by a single branch-swapping step. Related high probability trees form 'peaks' in the landscape and areas of low probability correspond to 'valleys'. We need to find and sample the peaks while avoiding the valleys in the landscape. One useful way of exploring the tree space is the Markov Chain Monte Carlo (MCMC) method (Stephens 2001).

Imagine that we randomly pick a point in the tree landscape. The goodness of fit to the data of the randomly chosen tree is measured. A neighbouring tree is then chosen and its goodness of fit measured. Roughly speaking, if this tree is better then the Markov chain (the linear tree-sampling algorithm) will move to the new tree, if not it will stay in place. As this process is repeated the chain migrates through the landscape, hopefully sampling better and better trees as it moves. There are some potential problems with this approach. Firstly, it is difficult to know how long the chain must run before it reaches the best part of the landscape and gathers adequate samples there, this is the problem of convergence. Secondly, if the landscape is complex then the chain might become stuck on a local peak (i.e. a sub-optimal set of trees) and fail to ever cross the valleys separating it from the global peak, the problem of autocorrelation (Hey, 2005b).

The convergence problem can be partially solved by starting the chain at many different starting points and checking if the same end result is reached. Autocorrelation (ie. local peaks) can be reduced by running multiple chains with varying sensitivity to the landscape and measured by tracking the number of independent trees that have been sampled. 'Hot' chains see only the very obvious peaks and valleys and tend to overcome smaller barriers, moving quickly away from local peaks. 'Cold' chains see a detailed surface and explore slowly. By running both types of chains and swapping chain temperatures it is possible to achieve the apparently contradictory goals of a broad search with fine precision. This strategy is known as Metropolis coupling and has proven very useful in MCMC coalescent-based simulation packages (See Won *et al.* 2005 for an extreme example).

In order to infer evolutionary parameters from polymorphism data we require a set of parameters that fit a biological model, a description of how these parameters affect the coalescent and a set of likely genealogies corresponding to our data. A simple example is the fitting of a single parameter, population size. Firstly, the parameter has to be converted to a form that is compatible with the coalescent. It is therefore scaled by the mutation rate, which is outside the model and could have any value. Gene trees are mutated by a branch-swapping algorithm (eg. Beerli & Felsenstein 1999) and sampled by MCMC according to their fit to the data. The value of the parameters is calculated for each sampled tree. The true stationary distribution of each parameter is approximated by its weighted average over the sampled gene trees, with each tree weighted according to its fit to the data. This approach has been extended to incorporate the exponential rate of population growth or decline (Kuhner *et al.* 1998), the migration rate between populations (Beerli & Felsenstein 2001) and the time since divergence of two populations (Nielsen & Wakeley 2001) as well as combinations of these parameters (Hey 2005a). Within a defined model of how these parameters correspond to biological reality we can infer historical processes and associated statistical confidence from extant polymorphism data.

Nested clade phylogeographic analysis

Nested clade phylogeographic analysis (NCPA) is probably the most widely used statistical tool for inferring evolutionary processes from genealogical data (Templeton 2004). NCPA utilises a single reconstructed gene tree (a haplotype network) to group haplotypes into clades according to degree of relatedness and ancestor-descendent relationships. An initial test of the association of haplotypes and geography is conducted with a null model of random geographical distribution of haplotypes (Templeton *et al.* 1995). If there is a significant association between clades and geographical position, then a heuristic tool (the inference key) is used to link these patterns to evolutionary processes of gene flow and range dynamics. The key is based on expectations from coalescent theory, as described by Templeton (1998). NCPA therefore infers evolutionary history from polymorphism data without any prior expectations as to what that history might be.

The most fundamental criticism of NCPA is that it ignores the stochasticity of the coalescent process (Knowles & Maddison 2002). This makes it impossible to determine the potential error associated with a particular inference and the relative strength of various evolutionary dynamics. These complaints aside the randomization strategy has been criticised as being incorrect (Petit & Grivet 2002) because it fails to account for population homogeneity, leading to false inferences of pattern in species with little variation within populations. The only simulation study of NCPA generated disappointing results (Knowles & Maddison 2002). Templeton (2004) has however explained how these simulations violate the assumptions of NCPA. This debacle highlights the importance of correctly interpreting the verbal instructions of the inference key and not extending NCPA to biological situations in which it is inappropriate, such as microvicariance (Templeton et al. 1995).

The broadest test of NCPA showed that is was able to predict expected patterns of range expansion and fragmentation in a wide variety of organisms with a low rate of false positives (Templeton 2004). The validity of this study of course depends on whether expected patterns correspond to true patterns. More detailed coalescent-based simulations that fall within the recommended sampling strategy and biological ambit of NCPA are needed to properly assess its usefulness as a phylogeographic tool.

1.6 Model based hypothesis testing

Background

Phylogeographers have long been aware of the need to incorporate the effects of coalescent stochasticity in their analyses (see Milot *et al.* 2000 and Knowles 2001 for early examples). Until recently, limited and inaccessible theoretical tools have hampered the development of an explicit, model-based approach to analysing data. New theoretical tools and associated software have recently become available (Pearse & Crandall 2004; Hey 2005a). Combined with an increasing understanding of the importance of coalescent stochasticity (Knowles & Maddison 2002; Edwards & Beerli 2000), this has resulted in a recent increase in the number of phylogeographical studies incorporating model-based hypothesis testing under the coalescent.

When estimating biological parameters it is crucial that the underlying biological model be realistic. A long-standing problem in phylogeography, for example, is the difficulty in distinguishing between the role of shared ancestral polymorphism and recent migration in shaping patterns of shared variation between species (Nielsen & Wakeley 2001). It might be possible to obtain a very precise estimate of gene flow if we assume that ancestral variation is irrelevant, but this estimate will obviously be inaccurate if the two populations do share recent common ancestry. This difficulty applies to other combinations of parameters and is not easily solved. Ideally we would create a model including all possible parameters, but this has to be traded-off against computational demands and theoretical limitations.

The basic parameters that are estimated by coalescent packages include effective population size (present and ancestral), migration rate, time of divergence and exponential rate of population growth or decline. All these parameters are scaled according to the mutation rate in line with coalescent theory. The available software packages estimate varying combinations of parameters under different models, but they can essentially be viewed as limiting cases of the 'isolation with migration' (IM) model.

The IM model describes a scenario in which two populations diverge from a common ancestor at a time t ago in the past. The ancestor is of size N_a, and the two derived populations are initially of size sN_a and $(1-s)N_a$, where s is some number between zero and one. These populations then grow or decline exponentially according to an exponential constant g over the course of time t to reach their current sizes, N_1 and N_2 (Hey 2005a). After divergence the populations exchange genes at two potentially unequal rates, m_1 and m_2. More detailed explanation of the model and associated parameters is given in the Methods.

Current state of the field

Coalescent-based models have been applied in a variety of phylogeographic scenarios and I give a representative sample here. This summary of the recent literature is intended to provide an overview of the sort of dynamics and inferences that can be accommodated within a statistical phylogeographic framework (Knowles 2004) and not as an in-depth analysis of the correct approach to application of the techniques. Because of the very recent growth of the field, there is no review linking recent applications of this approach and the interested student is required to perform an intensive search to find appropriate literature.

Fluctuate (Kuhner *et al.* 1998) estimates population size growth rate in a single isolated population. It has been used to infer population growth and associated range expansion from refugia following climatic change in desert insects (Smith & Farrell 2005) and forest amphibians (Carstens *et al.* 2004b) as well as recovery from a bottleneck induced by a volcanic eruption on an island (Moya *et al.* 2004). Ruokonen *et al.* (2005) used Fluctuate to show that population sizes have been constant in size in Pink-footed geese and therefore that their tundra habitat is resilient to climatic cycles. Extension of the growth parameter to dating allowed Provan et al. (2005) and Wares and Cunningham (2001) to date population bottleneck events.

Migrate (Beerli & Felsenstein 1999) and Migrate-n (Beerli & Felsenstein 2001) estimate the population sizes and migration rates of two populations of constant size that diverged very long ago (shared variation due to shared ancestry is not considered). These models have been applied to a wide variety of phylogeographical scenarios. Estimation of migration rates allowed Moya *et al.* (2004) to confirm the isolation of two peripheral populations from each other and their connection to a central population in a fragmented forest habitat. Studies on two species of grouse (Barrowclough *et al.* 2004) showed that low migration rates were consistent with reproductive isolation and resulted in the detection of a third cryptic species. Another bird species, the Red Pochard, was shown to have very low migration between vicariant populations (Gay *et al.* 2004), necessitating separate conservation strategies for the populations. Bernardi (2005) utilised migration patterns as well as growth estimates to establish the relative influence of vicariance and competive exclusion in shaping disjunct Californian surfperch distribution. Work on Little Greenbulls (Smith *et al.* 2004) showed that these tropical forest birds diverged morphologically between habitats despite high levels of gene flow. Pinceel et al. (2005) demonstrated the absence of gene flow between two European slug lineages and Rawson *et al.* (2003) established a link between the dispersal of the coronulid barnacle and its host turtle using Migrate. Zeh *et al.* (2003) used Migrate to infer the directionality of colonisation events and the existence of a cryptic species hybrid zone in a Panamanian pseudoscorpion.

Co-estimation of divergence time and migration rate is necessary to estimate either parameter correctly in recently diverged populations. This probably includes almost all intra-specific cases and most recently diverged species pairs. MDIV (Nielsen & Wakeley 2001) has been used for this purpose by Carstens *et al.* (2004b) to estimate the timing of a climatically induced vicariance event and distinguish it from an alternative explanation of recent migration in a disjunct salamander species. In cactus beetles Smith and Farrell (2005) linked demographic changes with Pleistocene climate change using estimated divergence times of two species. Barrowclough *et al.* (2005) used MDIV to confirm that an unusual haplotype distribution in the blue grouse was a result of incomplete lineage sorting and not long range gene flow.

The program Mesquite (Maddison & Maddison 2005) adopts a different approach to coalescent simulation. The program essentially calculates the probability that a certain pattern of shared variation could be due to drift, given a population branching history and population sizes. If populations are reciprocally monophyletic, then the distribution of times necessary for coalescence can be calculated. If they are non-monophyletic and divergence is known to be ancient, then migration can be inferred. Coalescence rates can also be used to infer population size given a known time to divergence. This approach therefore requires an outside estimate of either effective population size or divergence time as well as some knowledge of the extent of gene flow. Mesquite was used by Russell *et al.* (2005) to show that incomplete lineage sorting has played a central role in shaping patterns of variation in Mexican free-tailed bats. Kotlik *et al.* (2004) were able to show that present diversity in *Barbus* fishes could only be the result of survival in multiple glacial refugium, rather than recolonization from a single source. Using accurate dating of glacial maxima, Carstens *et al.* (2005) were able to distinguish alternative possible refugial hypotheses for Idaho giant salamanders. A study of three bat species (Carstens *et al.* 2004a) showed in conjunction with a phylogenetic approach that inter-island migration must be shaping the pattern of diversity in this group, contrary to expectations from ecological studies. Mesquite is very flexible, but because it does not use the data to distinguish between migration and common ancestry it is sensitive to incorrect estimates of the outside parameters.

The most recently developed and complete package is IM, an extension of MDIV that accommodates, amongst other additions, the analysis of multiple loci and exponential population growth. Except for a few recent papers, published research using this software has been restricted to the Hey laboratory in which it was developed. Because it can utilise information from multiple loci, IM has the potential to reveal the mosaic phylogeographical structure of the genome predicted by genic species concepts (Wu 2001; Rieseberg & Burke 2001). Indeed, studies of gene flow dynamics between *Drosophila pseudoobscura* and *D. persimilis* (Hey & Nielsen 2004) and within the moss *Ceratodon purpureus* (McDaniel & Shaw 2005) show divergent patterns of gene flow across different loci. This is attributed by the authors to the differential effects of selection on various loci. When their histories are equivalent multiple loci should improve estimates of divergence time and migration rates by evening out variable genealogical histories. Won and Hey (2005) used 48 loci to infer these parameters at intra and inter-specific level in the chimpanzee/bonobo clade, detecting an unusual unidirectional pattern of gene flow between chimpanzee subspecies. This may potentially be attributed to a missing population, highlighting the potential weaknesses in a model limited to two populations (Won & Hey 2005).

The rapid diversification of cichlid fishes is one of the classic topics of speciation studies. IM has been applied to both a single HapSTR locus (Hey *et al.* 2004) and 6 HapSTR loci (Won *et al.* 2005) to infer divergence times and population size dynamics in these organisms. HapSTRs are loci that include a short tandem repeat (STR) region and some adjacent sequence (haplotype), thereby incorporating two different time scales of mutation accumulation (Hey *et al.* 2004). The multi-locus study infers extremely recent divergence times (1000 to 17 000 years) between species. Conversion from coalescent time is accomplished by using outgroup rooting to determine the mutation rate. A huge estimate of ancestral population size in one cichlid species in this example may actually indicate that the ancestral population was itself not genetically isolated, again showing an opportunity for possible extension of the model (Won *et al.* 2005). Recently Cassens *et al.* (2005) have used IM to study population size, inter-population divergence and trans-oceanic dispersal in dusky dolphins.

The inclusion of a splitting parameter makes IM the most complex of all the statistical phylogeography packages. Use of this parameter is demonstrated in Hey's (2005a) study of the colonization of the Americas by humans. This is a situation that must have involved tremendous growth in population size, so inclusion of a growth parameter is essential for realistic modelling. A nine locus study of Asian and Amerind-speaking people shows that the New World was found by a small fraction of the Asian population, corresponding to an effective population size of about 80 people. Although accurate estimation of divergence times is difficult with such a complex model, divergence times are consistent with archeological and climatic predictions of a recent colonisation via the Beringian land bridge.

Although there are difficulties relating to the implicit a priori assumptions inherent in the above models (Templeton, 2005) these are being constantly resolved and improved as investigators gain experience in model-based approaches to phylogeography. Simulation tests are being performed to assess the accuracy of coalescent-based software packages under different conditions (Abdo *et al.* 2004). Most of the studies discussed above incorporate traditional phylogenetic, nested clade and population genetic statistical approaches either to generate hypotheses or to compare with model-based inferences. In all cases models are linked to directly to outside ecological and historical data. The rapid development of the field shows great promise for future insight into the mosaic nature of genomes as they respond to multiple ecological, geographical and historical evolutionary forces.

1.7 The genus *Pauridia*: a model system for studying speciation in the Cape Floristic Region

Systematics and distribution

Pauridia Harv. (Hypoxidaceae) is a geophytic genus endemic to the Cape Floristic Region and occurs primarily in lowland, nutrient-rich habitat (renoster shrubland, Goldblatt & Manning 2000). *Pauridia* is therefore a member of an extraordinarily diverse summer-drought adapted flora that incorporates over 1500 species with underground storage organs. Analysis of geophytic flora of the CFR (Proches & Cowling 2004) shows that most of the diversity (80%) is concentrated within the order Asparagales, and it is to this order and the family Hypoxidaceae that *Pauridia* belongs. Also included in this family are the genera *Empodium*, *Rhodohypoxis* and *Hypoxis*, which are not CFR endemic and *Spiloxene* which is largely endemic to the CFR (26 out of 29 species). There is strong morphological evidence that *Paurida* is nested within *Spiloxene* (D. Snijman, pers. comm.). These two genera therefore form a small (28 species) 'Cape clade' (sensu Linder 2003). This is not exceptional in CFR terms (the endemic geophytic genus *Lachenalia* consists of 110 species) but is nevertheless a notable radiation.

Genus level systematics of the Hypoxidaceae is largely based on the morphology of the underground storage organ (the corm) (Thompson 1976). *Pauridia* is indistinguishable from *Spiloxene* on the basis of corm characters and the genera are therefore distinguished by floral morphology (Thompson 1978). In *Spiloxene*, as in the other Hypoxidaceae genera, the flowers have 6 stamens, a 3-lobed stigma and the perianth segments are free. In *Pauridia* there are 3 stamens and a 6-lobed stigma and perianth segments are fused.

The *Pauridia* species are defined on the basis of perianth tube length. *Pauridia minuta* is defined by a perianth tube shorter than the lobes, between 2 and 4mm long. In *Pauridia longituba* the perianth tube is at least twice as long as the lobes and between 8 and 30 mm in length (Thompson 1979). The species are illustrated in Figure 1.3. The geographical range of the genus (Figure 2.1) covers most of the lowland CFR. The species are

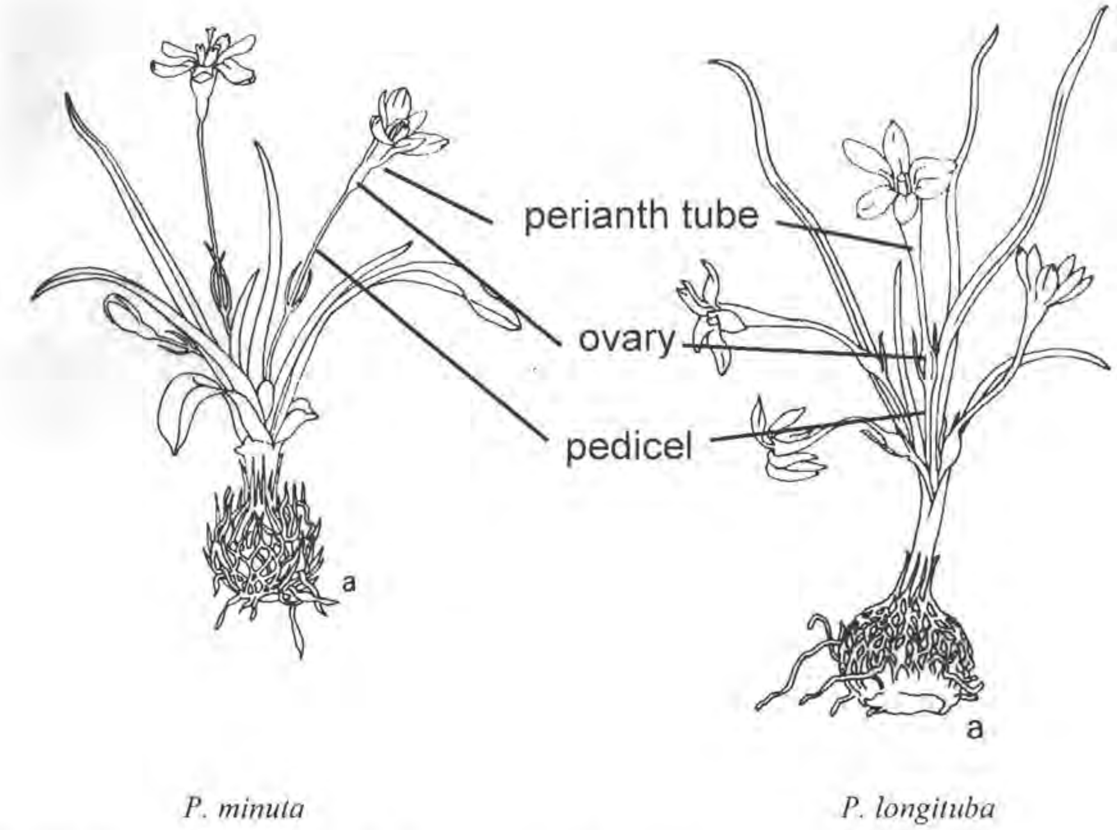


Figure 1.3: Characteristic features of *P. minuta* and *P. longituba*

The perianth tube is much longer in *P. longituba*. The pedicel is therefore reduced and occurs entirely underground, along with the ovary. Seeds are released below the soil surface. *P. minuta* has a long above-ground pedicel that allows seeds to spill onto the surrounding soil surface.

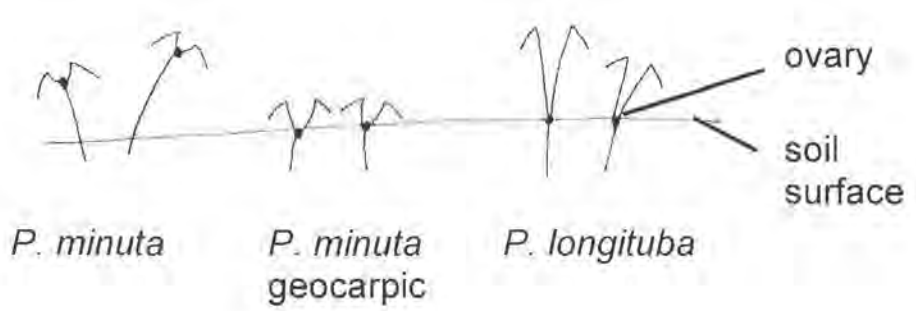


Figure 1.4 Geocarpy in *Pauridia*

Some populations of *P. minuta* at the species boundary have a shortened pedicel that makes them geocarpic. They share this character with *P. longituba*, but not with other *P. minuta* populations.

currently approximately parapatric in distribution and co-occur at two sites at the species boundary.

Ecological features

Pauridia shows strong adaptation to the seasonal Mediterranean climate of the CFR. Plants are dormant during summer-drought and emerge in early autumn in response to the first significant rains. A series of white flowers is produced immediately following leaf emergence and flowering is completed within 4-8 weeks, before most autumn flowering species begin to bloom. *Pauridia* is able to complete flowering before most other autumn-flowering geophytes because it is limited to micro-habitats such as rock crevices, shallow pavements and depressions that accumulate moisture first and can therefore begin growth immediately after rains. The plants are very small (5-50 mm above ground) and are able to reach the flowering stage before other geophytes in the same habitat (pers. obs.).

Population features

A result of adaptation to very specific micro-habitats is that densities of up to several thousand plants per m^2 are often reached, and up to 10 000 individuals per m^2 are achieved in some small rock pockets. This corresponds to a tightly packed mass of bulbs, each approximately 1 cm in diameter, completely dominating a patch. *Pauridia longituba* is confined to the granite outcrops near Vredenburg on the West Coast (see Figure 2.1) a total area of less than 1000 km^2. It is confined to pockets within boulders that have accumulated soil and to boulder margins, both of which become inundated immediately following rainfall and also dry out very rapidly. Extremely high densities result in population sizes of the order of hundreds of thousands of individuals. Granite outcrops (typically 0.1-100 ha in extent, each with 1-100 m^2 of suitable *Pauridia* habitat) are comparatively small, isolated islands in a sea of lowland vegetation.

Paurida minuta has a much broader distribution and occurs over much of the lowland CFR, an area of approximately 40 000 km^2 (Figure 2.1). The species occupies rock crevices on granite and sandstone, shallow rock pavements and depressions in deep clay soils. All of these habitats accumulate moisture very rapidly after rainfall. Populations in rock pockets and crevices are similar in density and extent to *P. longituba* populations, while clay populations are of lower density (approximately 50 individuals/m^2) and larger in extent, typically covering a few 100m^2. *P. minuta* populations are island-like in distribution and surrounded by habitat that accumulates rainwater less rapidly.

Pollination and seed dispersal

The reproductive anatomy of the two species appear to be indistinguishable to pollinators, and they are visited by a variety of small insect pollinators including butterflys, beetles, flies and bees (pers. obs.). Although tube length has been shown to govern pollinator specificity in some species, the characteristic long perianth tube of *P. longituba* is filled with the stigma and so cannot play a role in pollinator selection. The long perianth tube, in combination with a much reduced pedicel, leads to the development of the ovary and release of seeds underground (geocarpy), directly adjacent to the corm. In *P. minuta* the short perianth tube is countered by a long pedicel, resulting in a presentation of the stamens and stigma that is very similar to *P. longituba*. Upon pollination and seed development the pedicel bends to spill the mature seeds a few centimetres away from the mother plant.

Intra-species variation

Both species show great variation in overall size. Individuals vary in size over an order of magnitude both within and between populations. Leaf morphology is highly variable within *P. minuta* (Thompson 1979). Anthers are yellow and erect in *P. longituba* and vary from yellow to white in *P. minuta*. Petal tips can be purple, green or black. None of these features shows any particular geographical or ecological pattern (pers. obs.). Although not noted by Thompson (1978) populations of *P. minuta* at the species boundary have slightly longer perianth tubes and shorter pedicels than other con-specific populations, resulting in a geocarpic (ie. seed-burying) form. This variation is illustrated in Figure 1.4. Geocarpic *P. minuta* populations inhabit the typical *P. longituba* habitat of granite boulder pockets at sites that are shared with or adjacent to *P. longituba* populations (see Figure 2.1 '*P. minuta* geocarpic'). Individuals typically vary from having ovaries completely buried to merely having short pedicels and there is no sharp distinction bewteen geocarpic and non-geocarpic forms. Completely 'normal' *P. minuta* populations also occur on granite outcrops in the same geographical area.

Inter-species variation and interactions

The species are known to co-occur at two populations at the species boundary. At Witteklip Beacon (Population 18, Figure 2.1) individuals are entirely intermingled, occurring within centimetres of each other. At Witteklip Farm (Population 17, Figure 2.1) *P. longituba* is confined to boulder edges, while *P. minuta* occurs exclusively in boulder pockets. Individuals occur within a few centrimetres of one another. In both populations individuals of each species are very clearly distinguishable on the basis of overall size, perianth tube length and perianth tube shape. There are no morphologically intermediate individuals. The two sympatric populations are unusual in that they contain a number of individuals with aberrant floral morphology, including lack of anthers, multiple petal whorls and exceptionally large size. This type of aberrant morphology is known to be a fairly common feature of Hypoxidaceae species (J. Manning pers. comm.), but was only observed in these two populations during field work for this study.

Synthesis

Both *Pauridia* species occur in isolated, island-like seasonally inundated habitats and levels of seed flow are expected to be low (*P. minuta*) to very low (*P. longituba* and *P. minuta* geocarpic), while pollen flow via small insects is not likely to be great between populations. Population sizes of 1E4 - 1E5 individuals are common and overall species population sizes are at the minimum of the order of 1E6 (*P. longituba*) and 1E8 (*P.*

minuta) individuals. Because much of the lowland CFR habitat has very recently been converted to agriculture the present population size is reduced, but this will not be reflected in molecular diversity. There is much intra-specific morphological variation but only the feature of geocarpy near the species boundary in *P. minuta* shows obvious ecological and geographic pattern. There appears to be a continuous intergradation between geocarpic and ‘normal’ forms of *P. minuta* in this area. At the species level, the maintenance of distinct morphological identity despite co-occurrence at two sites is strong initial evidence for some form of reproductive isolation between the taxonomic entities.

1.8 Testable hypotheses in *Pauridia*: Aims and objectives

With the combined tools of coalescent based analyisis of neutral loci and our taxonomic and ecological knowledge of *Pauridia* we can develop a testable set of hypotheses. There are few precedents in using DNA sequence data to test phylogeographical plant speciation hypotheses and almost all of these derive their predictions from interpretations of a single realized genealogy, failing to consider coalescent stochasticity. Phylogeographical work has been done on the role of hybridization in plant speciation (Widmer and Baltisberger 1999; Trewick *et al* 2002; Matos and Schaal 2000), the effect of ecological clines (Holman *et al* 2003; Lumaret *et al* 2002) and the effect of isolation by distance (Cavers *et al* 2003; Coleman and Abbott 2003). Some studies have even inferred multiple processes shaping present species boundaries (Maskas and Cruzan 2000; Dobes *et al* 2003). These efforts all reflect a classical, descriptive approach to phylogeography. Although this may not be a complete list of plant phylogeographical approaches to speciation, further literature was not found after an intensive search. This is in line with a general paucity of plant phylogeographical studies, due to the low variability of the chloroplast genome (Schaal & Olsen 2000; Schaal *et al.* 1998).

Most of the above studies draw their model of plant speciation from Verne Grant's classic hypothesis of 'quantum speciation' (Grant 1981), as does Cowlings 'micro-geographic' CFR speciation model (Cowling & Pressey 2001). Under this model new plant species bud off from the parent species initially as isolated peripheral populations. They are isolated by geographical or ecological specialization leaving the parent species unaffected by the speciation process. The evidence for 'quantum speciation' has recently been reviewed by Gottlieb (2003). Although well supported in the botanical community some aspects of the theory remain untested. It is generally assumed that the derivative species is specialized (probably to an extreme environment) and geographically local, while the progenitor is the more widespread and generalist of the pair. The relative role of selection versus drift is not defined, but this is obviously of interest in the CFR given very rapid speciation and the putative driving selective forces (Cowling & Pressey 2001).

In contrast to the other plant speciation phylogeography studies listed above, the works of Holman *et al* (2003), Lumaret *et al* (2002) and Matos and Schaal (2000) focus on the role of divergent selection and geographical vicariance in affecting both species. This alternative model of vicariant speciation differs from 'quantum speciation' in that derivative species are of roughly equivalent founding sizes and diverge in response to range shifts and ecological clines, rather than adaptation to a particular peripheral site.

A key feature of CFR diversity is the occurrence of very many geographically isolated, specialist species. *Pauridia longituba* is a typical example of a range-restricted edaphic endemic, while *P. minuta* is widespread and a relative habitat generalist. We therefore wish to test the validity of the quantum speciation hypothesis in this candidate species pair. Whether quantum speciation or vicariance is supported by the data we also wish to assess the relative role played by drift and selection in shaping the speciation process.

Describing the speciation process in *Pauridia*

The first stage in testing these hypotheses is to establish that the species are sufficiently recently diverged to provide information on the speciation process. We therefore construct a gene tree in order to establish roughly the degree of divergence (ie. polyphyly or monophyly) and establish if there is any evidence for current reproductive isolation at sympatric populations. We then test for evidence of geographical structuring of variation and attempt to infer rough evolutionary patterns of gene flow and range dynamics within species. Diversity within the species can also be described using various summary statistics that estimate levels of genetic diversity and potentially allow for the detection of change in population size and the effect of selection on our putatively neutral marker.

Under vicariant speciation it is expected that patterns of neutral genetic variation will correspond roughly to the species boundaries, while under quantum speciation the derived species will be one of many geographically isolated fragments of the progenitor species and show little internal structure itself. Expansion into a new niche following speciation may leave the genetic signature of population growth in the derived species.

Hypothesis testing

The broad picture given by descriptive methods will be necessarily flawed because it is based on interpretation of a single realized genealogy. The next stage of analysing the data is therefore to use coalescent-based models to estimate demographic parameter values and compare these estimates to those expected under alternative hypotheses. The 'isolation with migration' model can be used to represent the speciation process. Software implementations of the model and its limiting cases allow us to estimate current and ancestral population size, migration rate and exponential growth rate and determine the error bounds on these estimates.

Under a vicariant speciation scenario it is expected that the founding population sizes of the two species will be approximately equal, but if quantum speciation has occurred the founding population sizes will be very unequal. If the derived species has expanded into a new niche following speciation there should also be signs of exponential population growth.

Migration rate and divergence time bear directly on alternative hypotheses of drift and selection driven divergence. Under a drift-driven model divergence occurs only when gene flow is low and divergence time (relative to population size) is relatively ancient. 'Low' gene flow and 'ancient' divergence can be quantified using expectations from population genetic and coalescent theory, as described in the Methods. It is only necessary that one of these conditions not be satisfied in order for us to favour selection-driven speciation because either recent ancestry or high gene flow will lead to failure to diverge at neutral loci.

Unlike descriptive methods such as NCPA, coalescent-based models are quite likely to lead to an inconclusive outcome. If the bound on estimated migration rate, for instance, spans a range from zero to very high then we obviously cannot use this parameter to distinguish between alternative hypotheses of low and high gene flow. When constructing the tests we therefore establish conservative criteria that will lead us to an inconclusive outcome unless there is an unambiguous signal in the data.

After testing alternative speciation hypotheses I place the coalescent-based description of *Pauridia* in geographical and ecological context and assess the broader implications of this study for the investigation of Cape plant diversity.

The aims of this study are therefore to:

- Collect DNA sequence data for a geographically representative set of individuals of both species.
- Construct a gene tree to assess the approximate level of divergence between the two species.
- Use nested clade phylogeographical analysis to test for the existence of geographical structuring of variation and to infer demographic patterns.
- Assess the level of genetic diversity within the species using diversity statistics.
- Use tests of population expansion to test the data for signals of population growth or stationarity.
- Use coalescent-based analyses to obtain demographic parameter estimates under the various cases of the 'isolation with migration' model.
- Test alternative hypotheses of vicariant/quantum and selection/drift speciation using the obtained parameter estimates.
- Synthesize the available lines of evidence into an overall view of the speciation process in *Pauridia.*
- Interpret the results obtained from these analyses in terms of the ecology and history of the CFR.

2. Materials and methods

2.1 Data collection

Sample collection

Specimens were collected from 23 sites during autumn of 2004 and 2005. The geographical distribution of these sites is shown in Figure 2.1. These sites cover the known distribution range of both species. *Paurida longituba* was collected from 7 sites and *Paurida minuta* from 18 sites. Two sites (#17 and #18) contained populations of both *P. minuta* and *P. longituba*. At each site a number of individual plants (excluding the corm) were placed in dehydrating silica gel for later DNA extraction.

Representative whole plants were collected and corresponding herbarium sheets lodged in the Bolus Herbarium, University of Cape Town (BOL). The number, name, geographical location and herbarium sheet (lodged in BOL) of each collection site are given in Table 2.1. Sites were selected by surveying the available taxonomic literature (Thompson, 1979), searching the collections of the Bolus and Compton herbariums and consulting with individuals with field knowledge of the region. Further sites were identified by searching potentially suitable habitat encountered during field-work.

DNA extraction

For each individual sample 10mg of dried leaf material was removed and combined with 50mg polyvinylpyrrolidone (PVP) in a ceramic pestle. Liquid nitrogen was added directly to the sample and ground until completely evaporated. The sample was then re-suspended in 700µl CTAB/mercaptoethanol extraction buffer at 65°C and transferred to a 1.5ml microfuge tube. Samples were incubated at 65°C for 30 minutes before being mixed with

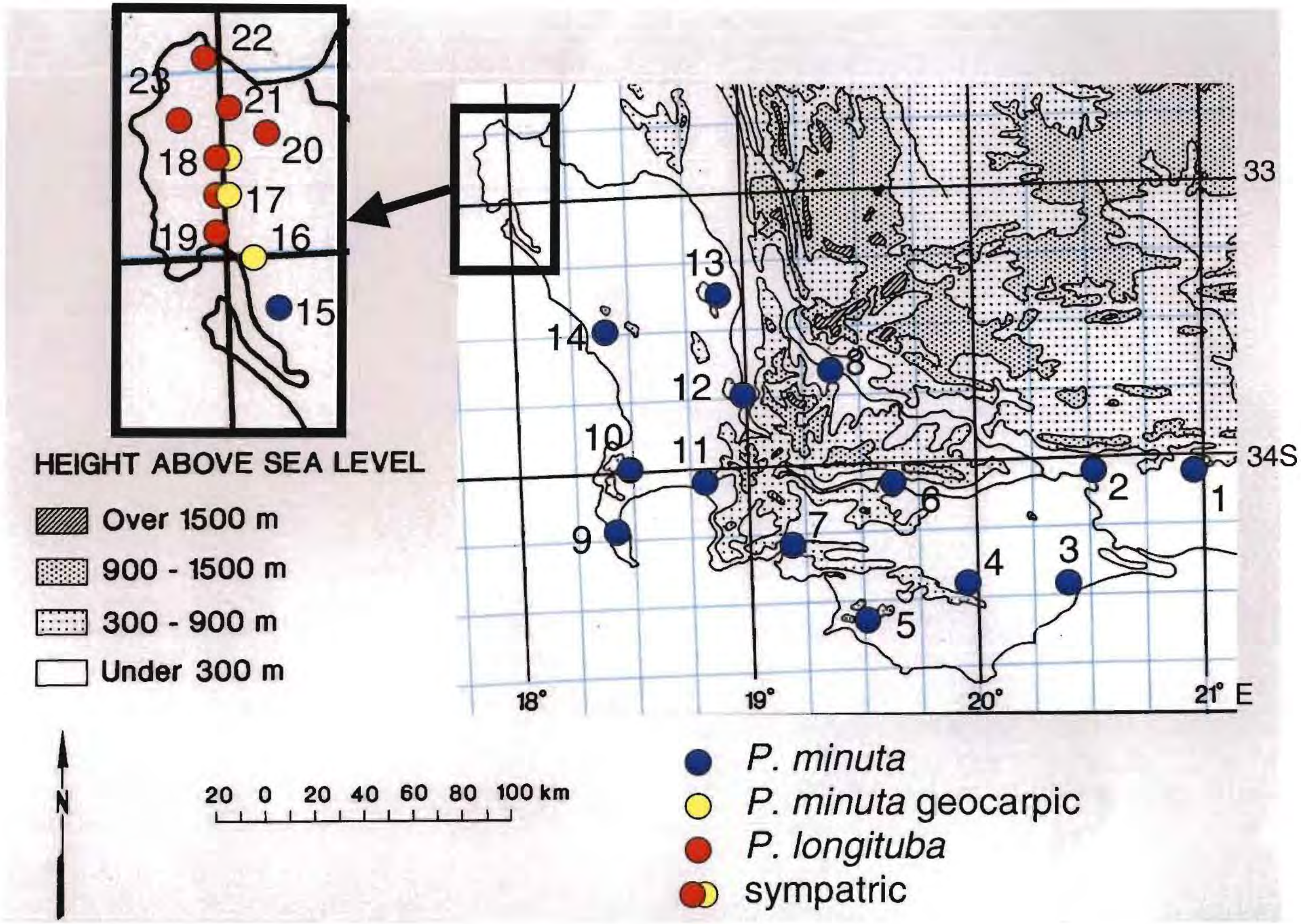


Figure 2.1 Distribution map of *Pauridia* collection sites with population numbers. The location of *P. minuta* (blue) and *P. longituba* (red) sites are marked. Three *P. minuta* sites consisted of individuals with a geocarpic habit (yellow). Two sites consisted of individuals of both species. Site names and geographical co-ordinates are given in Table 2.1.

Number	Name	Co-ordinates	Herbarium sheet
1	Heidelberg	34 02 S 20 56 E	G. Rowe 1
2	Swellendam	34 04 S 20 27 E	G. Rowe 2
3	De Hoop	34 27 S 21 25 E	G. Rowe 3
4	Napier	34 28 S 19 55 E	G. Rowe 4
5	Gaansbaai	34 35 S 19 29 E	G. Rowe 5
6	Greyton	34 05 S 19 32 E	G. Rowe 6
7	Botrivier	34 22 S 19 07 E	G. Rowe 7
8	Worcester	33 37 S 19 22 E	G. Rowe 8
9	Simonstown	34 11 S 18 23 E	G. Rowe 9
10	Rondebosch	33 57 S 18 29 E	G. Rowe 10
11	Somerset West	32 50 S 18 03 E	G. Rowe 11
12	Paarl	33 45 S 18 58 E	G. Rowe 12
13	Allesverloren	33 24 S 18 51 E	G. Rowe 13
14	Rondeberg	33 25 S 18 16 E	G. Rowe 14
15	Langebaan	33 06 S 18 03 E	G. Rowe 15
16	Bonteheuwel	33 03 S 18 03 E	G. Rowe 16
17*	Witteklip Farm	32 56 S 17 58 E	G. Rowe 17
18*	Witteklip Beacon	32 55 S 17 58 E	G. Rowe 18
19	Saldanha	32 59 S 17 56 E	G. Rowe 19
20	Kleinberg	32 53 S 18 07 E	G. Rowe 20
21	Patrysberg	32 50 S 18 03 E	G. Rowe 21
22	Saint Helena	32 46 S 18 02 E	G. Rowe 22
23	Paternoster	32 48 S 17 55 E	G. Rowe 23

Table 2.1: Numbers, names and localities of *Pauridia* collection sites. Herbarium sheets are lodged in the Bolus Herbarium (BOL) at the University of Cape Town. Populations marked '*' contained individuals of both species.

600μl of chloroform:isoamyl alcohol (24:1, v/v) by inversion for 5 minutes. Samples were centrifuged at 12 000 rpm for 5 minutes and the supernatant transferred to a clean 1.5ml microfuge tube. An equal volume of isopropanol at -10°C was added and mixed briefly by inversion. Sample DNA was stored overnight at -10°C to precipitate. Precipitated samples were centrifuged at 12 000 rpm for 5 minutes to recover DNA. After discarding most of the solvent, tubes were carefully drained on tissue paper. The DNA pellet was washed by centrifugation at 12 000 rpm for 3 minutes with 250μl 75% ethanol. Ethanol was discarded and the tubes were dried for one hour in a silica-gel dessicating chamber. After complete drying the DNA was re-suspended in 50μl nanopure water. DNA stocks were stored at -20°C.

Amplification and sequencing

Two chloroplast regions, psbA-trnH, a highly plastic intergenic spacer (Sang *et al.* 1997) and rps16, an intron of a ribosomal protein encoding gene (Oxelman *et al.* 1997) were amplified by PCR. Amplification was carried out in 30μl volumes. The reaction mixture consisted of 0.15 units BIOTAQ DNA polymerase (Bioline), 1XNH4 buffer, 5mM MgCl2, 0.1mM of each dNTP, 0.3μM of each primer and 3μl of unquantified DNA at an empirically determined concentration. PCR cycling was performed on a GeneAmp PCR System 2700 (Applied Biosystems). Samples were denatured at 97°C for 2 minutes followed by 30 cycles of 1 minute at 94°C, 1 minute at 52°C and 2 minutes at 72°C. Samples underwent a final extension step of 7 minutes at 72°C followed by an indefinite hold at 4°C. PCR products were cleaned using the GFX PCR DNA purification kit (Amersham Bioscience) and re-suspended in 30μl nanopure water.

Cycle sequencing was performed on a GeneAmp PCR System 2700 (Applied Biosystems). Reaction volumes of 10μl consisted of 1ul BigDye Terminator v3.1 5X Sequencing Buffer (Applied Biosystems), 2μl BigDye Terminator v3.1 Cycle Sequencing RR-100, 0.14μM primer, 1 to 6.8ul of DNA template and the balance of nanopure water. Products were run on an ABI PRISM 3100 Genetic Analyzer in both the 3' and 5' directions.

Sequence processing

Raw trace data was edited using SeqMan (Lasergene System Software, DNAStar). Sequences were aligned in MegAlign (Lasergene System Software, DNAStar). Because the chloroplast genome is non-recombining psbA-trnH and rps16 sequences were concatenated for subsequent analyses. Potentially homoplasious polynucleotide repeat regions were excluded from the analysis and several large indels were recoded as single nucleotide gaps. Unique haplotypes were identified using TCS v1.13 (Clement *et al.* 2000).

2.2 Descriptive data analysis

Nested clade phylogeographic analysis (NCPA)

NCPA was used to assess the significance of the correlation of haplotypes and geography. A haplotype network was estimated using TCS v1.13 (Clement *et al.* 2000). Haplotypes within the network were grouped into clades according to the procedure of Templeton *et al.* (1987, 1992) and symmetrically stranded clades were resolved using the rules given in Templeton *et al.* (1993).

Unweighted geographical distances were calculated from latitude and longitude and GEODIS v2.0 (Posada et al., 2000) was used to calculate the average distance of individuals from the geographical center of their clade (D_c), the average distance to the center of the clade at the higher nesting level (D_n) and the difference between ancestor (interior) and descendant (tip) clades for D_c and D_n. GEODIS also employs a Monte Carlo procedure to identify clades with geographical distribution significantly different from a random distribution. The inference key (Templeton, 2004) was applied to nested clades with significant associations of haplotype and geography.

Diversity statistics

We calculated the number of haplotypes, nucleotide diversity (π) and the number of polymorphic sites (s) as approximate indicators of the extent of genetic diversity sampled in each species. Mismatch analysis was performed to provide an indication of potential population growth. This test compares observed frequencies of pairwise differences to those expected under demographic models of population growth and stationarity. Alternative hypotheses of stationarity and exponential change were tested, using 1000 coalescent simulations to assess the significance of the raggedness (rg) statistic (Harpending et al. 1993).

Comparison of tests of neutrality have been used to distinguish between population growth and the effect of selection in phylogeographical studies by Russel et al. (2005) and Hoffman and Blouin (2004). Following their approach Fu and Li's (1993) F^*, D^* and Fu's (1997) *Fs* were calculated and their significance assessed with 1000 coalescent simulations. If F^* and D^* are significant but *Fs* is not then selection is indicated. If the opposite is true then population expansion is indicated. We also used the R2 test of Ramos-Onsins and Rozas (2002) to test for the signal of population growth. Analyses were performed with DNASP v4.10 (Rozas et al., 2003).

2.3 Coalescent-based data analysis

Models and parameters

We attempted to reconstruct the history of the speciation process in *P. minuta* and *P. longituba* using a variety of models. All six models are limiting cases of the 'isolation with migration' model. In this model an ancestral population founds two separate populations which then potentially exchange genes and grow or contract. Simplified limiting cases were investigated because there is no way of easily choosing the best model, and we had no prior evidence as to which model would be the most biologically relevant.

In each case we arbitrarily defined *P. longituba* as population 1 and *P. minuta* as population 2. The models are defined and described in Figure 2.2, except for Model A. Model A consists of a single population of current size θ. The population size at time t in the past is given by $\theta_t = \theta\ e^{-gt}$. Positive values of g therefore indicate exponential growth going forward in time and negative values indicate exponential decline. The coalescent parameters of the various models are defined in Table 2.2. The coalescent parameters estimated by the algorithms can be used to establish biological parameters if the mutation rate (μ) is known from an outside source. The biological interpretation of these parameters is given in Table 2.3. If the mutation rate is not accurately known, then it is still possible to derive some biologically meaningful results by obtaining ratios for which μ cancels out, as shown in Table 2.4. The software packages that implement the models and a summary of the parameters that they incorporate are listed in Table 2.5. The notation and terminology used here follows Hey (2005a).

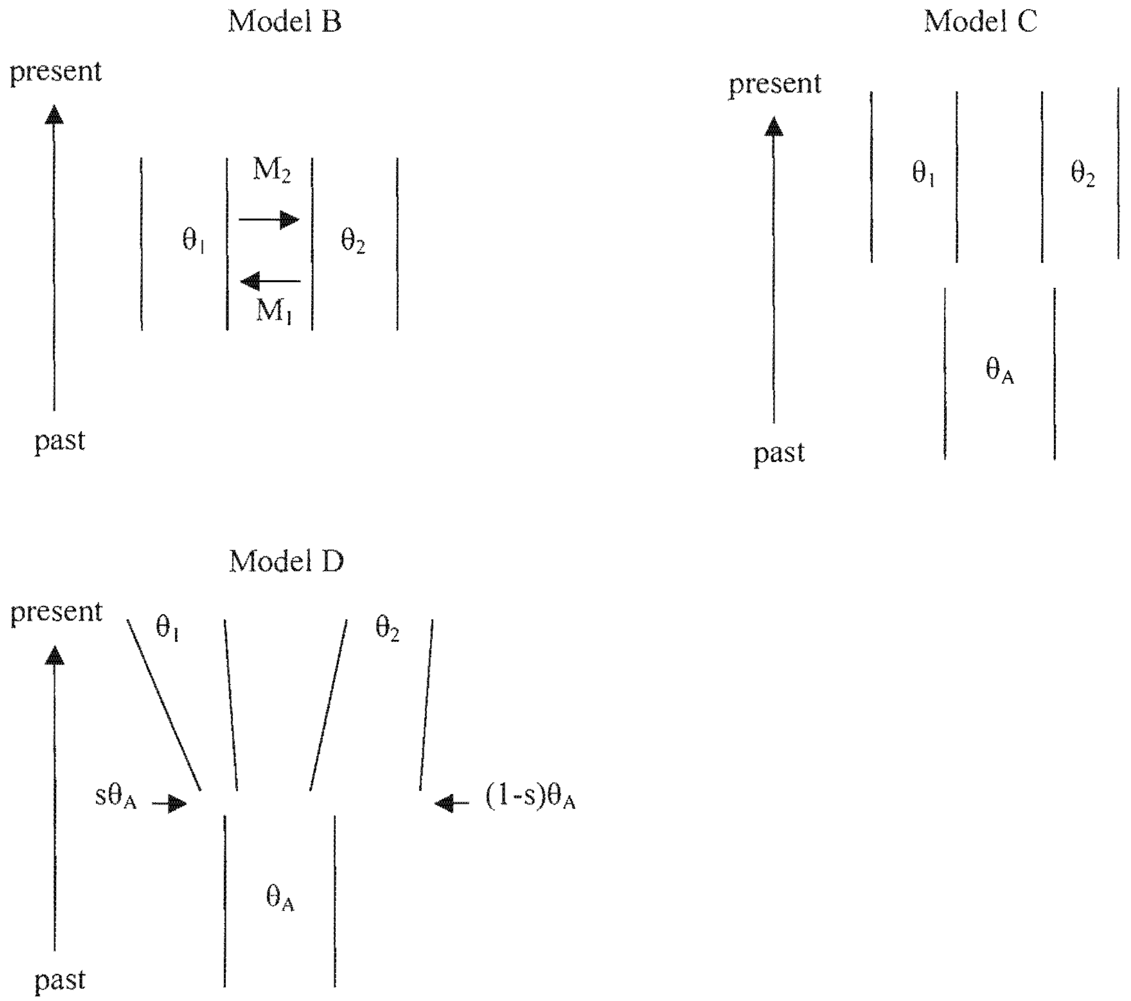


Figure 2.2 (continued overleaf): The 'isolation with migration' model and limiting cases. Coalescent parameters θ: population size, M: migration rate, s: splitting parameter.

Model B:

Both populations are of constant size θ_1 & θ_2. Shared ancestry is assumed to be irrelevant. Migrants move from population 1 into population 2 at a rate M_2 and in the opposite direction at rate M_1.

Model C:

Population size is constant and the populations split from a common ancestor of size θ_A at a time t generations in the past. There is no migration between the populations.

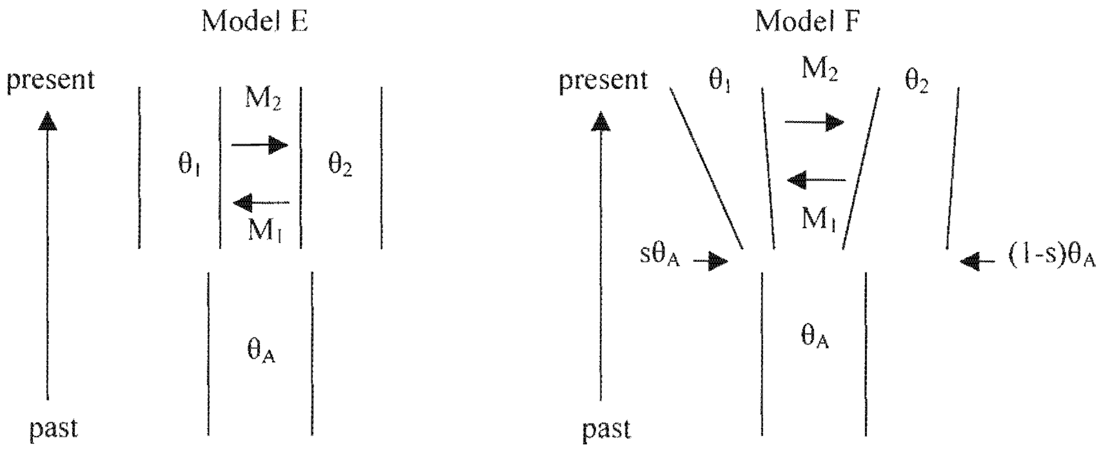


Figure 2.2 (continued): The 'isolation with migration model' and limiting cases.
Coalescent parameters θ: population size, M: migration rate, s: splitting parameter

Model D:

Similar to Model C except that the descendant populations are founded by some proportion of the ancestral population. Population 1 is founded by $s\theta_A$ and population 2 by $(1\text{-}s)\theta_A$, where s is any number between 0 and 1. The total size of the two founding populations is therefore equal to the size of the ancestral population. After divergence each population is free to grow or shrink exponentially over time t to reach θ_1 & θ_2, the current population sizes.

Model E:

Similar to Model C, except that migration is allowed between the populations.

Model F:

The full model. This model includes divergence from a common ancestor, growth and migration.

Parameter	Description	Scaling factor
θ_1	Population mutation rate for population 1	$\theta_1 = 4N_1 \mu$
θ_2	Population mutation rate for population 2	$\theta_2 = 4N_2 \mu$
θ_A	Population mutation rate for the ancestral population	$\theta_A = 4N_A \mu$
M_1	Migration rate per mutation from population 2 into population 1	$M_1 = m_1/ \mu$
M_2	Migration rate per mutation from population 1 into population 2	$M_2 = m_2/ \mu$
T	Number of mutations per gene since divergence	$T = t \mu$
s	Fraction of the ancestral population that founded population 1	
1-s	Fraction of the ancestral population that founded population 2	
g	The exponential growth rate per mutation	
μ	The mutation rate per gene per generation	

Table 2.2: Definitions of the coalescent parameters estimated under the models. Some coalescent parameters are related to a biological parameter (described in Table 2.3) by a scaling factor.

Parameter	Description
N_1	Effective population size of population 1 (p1)
N_2	Effective population size of population 2 (p2)
N_A	Effective population size of the ancestral population
m_1	Probability of migration from p2 to p1 per gene copy per generation
m_2	Probability of migration from p1 to p2 per gene copy per generation
t	Number of generations since divergence

Table 2.3: Definitions of the biological parameters related to the coalescent parameters.

Parameter	Description	Formula
N_1/N_2	The ratio of effective population sizes of population1/population2	θ_1/θ_2
N_1/N_{1A}	The ratio of current population size to ancestral population size (pop1)	$\theta_1/s\theta_A$
N_2/N_{2A}	The ratio of current population size to ancestral population size (pop1)	$\theta_2/(1\text{-}s)\theta_A$
$2N_x m_x$	The effective number of gene migrants into population x per generation	$(\theta_x M_x)/2$
t/N_x	The number of generations per effective population size since divergence	$4T/\theta_x$

Table 2.4: Biologically relevant parameters can be estimated independently of the mutation rate (μ), and some of these are shown here. The relevant formula incorporating the coalescent parameters is given in the column 'Formula'.

Model	migration	growth	ancestry	Parameters	Software
A	-	x	-	4	F
B	x	-	-	4	M, IM
C	-	-	x	4	IM
D	-	x	x	5	IM
E	x	-	x	6	IM
F	x	x	x	7	IM

Table 2.5: Parameters estimated (x) and excluded (-) under the models. The software packages used are: F, Fluctuate; M, Migrate; IM.

Hypothesis testing

As stated in the introduction, we quantify expectations of gene flow, divergence time, and ancestral size ratio under the alternative hypotheses.

In attempting to distinguish between vicariant and quantum speciation we have three criteria. Firstly if the speciation process was vicariant in nature, then the ratio of one ancestral population to the other will not be very small. We define 'very small' as less than 1%. We therefore favour quantum speciation if there is strong evidence that the splitting parameter (s) is less than 0.01 or greater than 0.99. Secondly, NCPA may provide supporting evidence. If geographical structuring of haplotypes is much better defined in one species, it is more likely that it is the progenitor. Thirdly, under a situation where the derived species expands into a new niche we might see strong exponential growth. We would not expect to see this under vicariant speciation. The ratios of current to ancestral population size are $\theta_1 / s\theta_A$ and $\theta_2 / (1\text{-}s)\theta_A$ for species 1 and 2 respectively under the IM model. Fluctuate gives a direct estimate of the growth parameter.

Selection driven speciation may result in a variety of patterns of neutral diversity, depending on how recently speciation occurred. Drift driven speciation should results in reciprocal monophyly of many neutral loci when speciation is complete. Drift to monophyly requires both low gene flow between incipient species and sufficient time for coalescence within one lineage. The effective number of gene migrants into population x per generation is given by $2\ N_x\ m_x$. Accoarding to Hey and Nielsen (2004) (following the work of Sewall Wright) if this value is greater than 1, then lineages will not diverge at neurtral loci. Since $(\theta_x\ M_x)/2 = 2N_x\ m_x$ we define high gene flow as $(\theta_x\ M_x)/2 > 1$. The required time for coalescence within lineages is a function of population size and number of generations. The number of generations per effective population size since divergence is $t\ /\ N_x$ which equals $4T/\ \theta_x$. According to coalescent simulations about $2\ N_e$ generations is normally required for divergence at neutral loci, but may occur at $1\ N_e$ for a limited number of loci (Rosenberg 2003). We therefore treat $4T/\ \theta_x < 1$ as recent divergence.

If there is not unambiguous evidence for either high gene flow or recent divergence, then it is impossible to distinguish between drift driven speciation and selection-driven speciation that occurred long ago. Alternatively, strong evidence for either high gene flow or recent divergence implies that selection must have acted to drive morphological and ecological divergence of the sister species.

Model implementation

FLUCTUATE: Model A

Runs in fluctuate were initiated on Watterson's estimate of theta and population size was allowed to change freely. We conducted ten independent runs of 10 short chains (2000 steps) and 5 long chains (200000 steps) with a sampling increment of 20 steps. The maximum likelihood estimates and their standard deviations were averaged over the ten independent runs.

MIGRATE: Model B

Initial experimentation with MIGRATE showed that multiple heated chains were necessary in order to explore the likelihood space properly. There was also a significant problem of 'attraction to zero', where either one of the migration parameters was estimated to be zero in any given run. This problem was solved by averaging over multiple runs using the 'replicates' option. Final runs consisted of 10 replicates, each of 10 short chains (100 000 steps) and 3 long chains (2 000 000 steps) with 8 Metropolis coupled chains. Sampling was performed every 20 steps. Heating settings were 1; 2; 3; 4; 6; 9; 12; 16.

IM: Models B-F

IM adopts a Bayesian approach to parameter estimation, whereby a pre-defined bound on parameter estimates (the prior distribution) influences the estimate of the value of the parameters (the posterior distribution). If the entire posterior distribution falls within the prior distribution then the smoothed peak of the posterior distribution is equivalent to the maximum likelihood estimate of the parameter and it is possible to establish confidence intervals on the estimated parameter values. In this case the posterior distribution is only a function of the data and not the prior distribution. If the posterior distribution does not fall entirely within the prior distribution (ie. is non-zero at the edges) then the prior estimates are influencing the outcome of the analysis. Because we wish to use only the molecular data to estimate parameter values we should ideally set priors wide enough that they do not influence the posterior distribution (Won & Hey 2005).

With this dataset it was not always possible to set priors wide enough because some parameters had very long tails of low probability. In these cases multiple runs were performed with increasing upper bounds to ensure that prior settings were not altering the peak estimates. Aside from checking that prior settings do not influence the estimates we also need to ensure that the Markov chain has adequately explored the parameter space. The use of multiple Metropolis coupled chains was useful in ensuring convergence with this dataset. We used ten Metropolis coupled chains for all runs, with heating settings either determined empirically or by using the adaptive heating function.

As recommended by Hey and Nielsen (2004) we monitored ESS (estimated sample size) values during runs and only accepted runs in which the lowest ESS value was greater than 100 and autocorrelation values were lower than 0.03 after at least 3 million steps. ESS is a measure of the independence of samples taken over a run for a parameter and ESS values of less than 50 indicate inadequate sampling of the search space. Because parameter estimates are not independent of one another it is the lowest ESS value that is critical in assessing the validity of a run. If autocorrelation is not monitored chains can become 'stuck' at a point in the genealogy-parameter space and fail to sample properly even if the chain is run for a very long time. In this case the parameter estimates do not actually reflect the true stationary distribution. In addition to monitoring autocorrelation statistics, we only accepted results that were consistent across a minimum of three independent runs.

Input datasets and run parameters used in the above analyses are included in the supplementary data supplied on CD. The infinite-sites model was used for all IM analyses.

3. Results

3.1 Data collection

Sampling and sequence curation

Two chloroplast regions (rps16 and psbA-trnH) of total length 1188bp were successfully sequenced in 125 individuals from 23 populations. Of 23 populations, only 5 had more than 1 haplotype. The number of individuals sequenced per population and the haplotype frequencies are given in Table 3.1 (*P. minuta*) and Table 3.2 (*P. longituba*). Sequences will be deposited in Genbank.

3.2 Descriptive data analysis

Haplotype network

The estimated haplotype network for the *Pauridia* haplotypes is shown in Figure 3.1. The species share one haplotype (I). This haplotype occurs in many populations in both species and shows the highest connectivity in the network. Wide geographical spread and high connectivity are both expected in the ancestral haplotype (Templeton 1998), so we assigned haplotype I ancestral status for subsequent NCPA analyses.

Population #	Number of individuals	Haplotype frequency	Random sample
1	5	5P	P
2	5	5Q	Q
3*	4	3P, 1R	P
4	3	3P	P
5	5	5P	P
6	3	3P	P
7	4	4O	O
8*	5	4K, 1L	K
9	5	5H	H
10	5	5J	J
11	4	4J	J
12	4	4I	I
13	5	5I	I
14*	10	5J, 4G, 1I	G
15	9	9C	C
16	6	6B	B
17*	4	3M 1N	M
18	5	5M	M
Total	92 individuals	14 haplotypes	11 haplotypes

Table 3.1: *Pauridia minuta* sequencing results. For each population the number of individuals and the haplotype frequency is given. See section 3.4 'metapopulation structure' for an explanation of the column 'Random sample'. Only populations marked with a * contain more than one haplotype.

Population #	Number of individuals	Haplotype frequency	Random sample
17	5	5I	I
18	10	10I	I
19	2	2I	I
20	5	5F	F
21	5	5I	I
22*	3	2E 1D	E
23	3	3A	A
Total	33 individuals	5 haplotypes	4 haplotypes

Table 3.2: *Pauridia longituba* sequencing results. For each population the number of individuals and the haplotype frequency is given. See section 3.4 'metapopulation structure' for an explanation of the column 'Random sample'. Only populations marked with a * contain more than one haplotype.

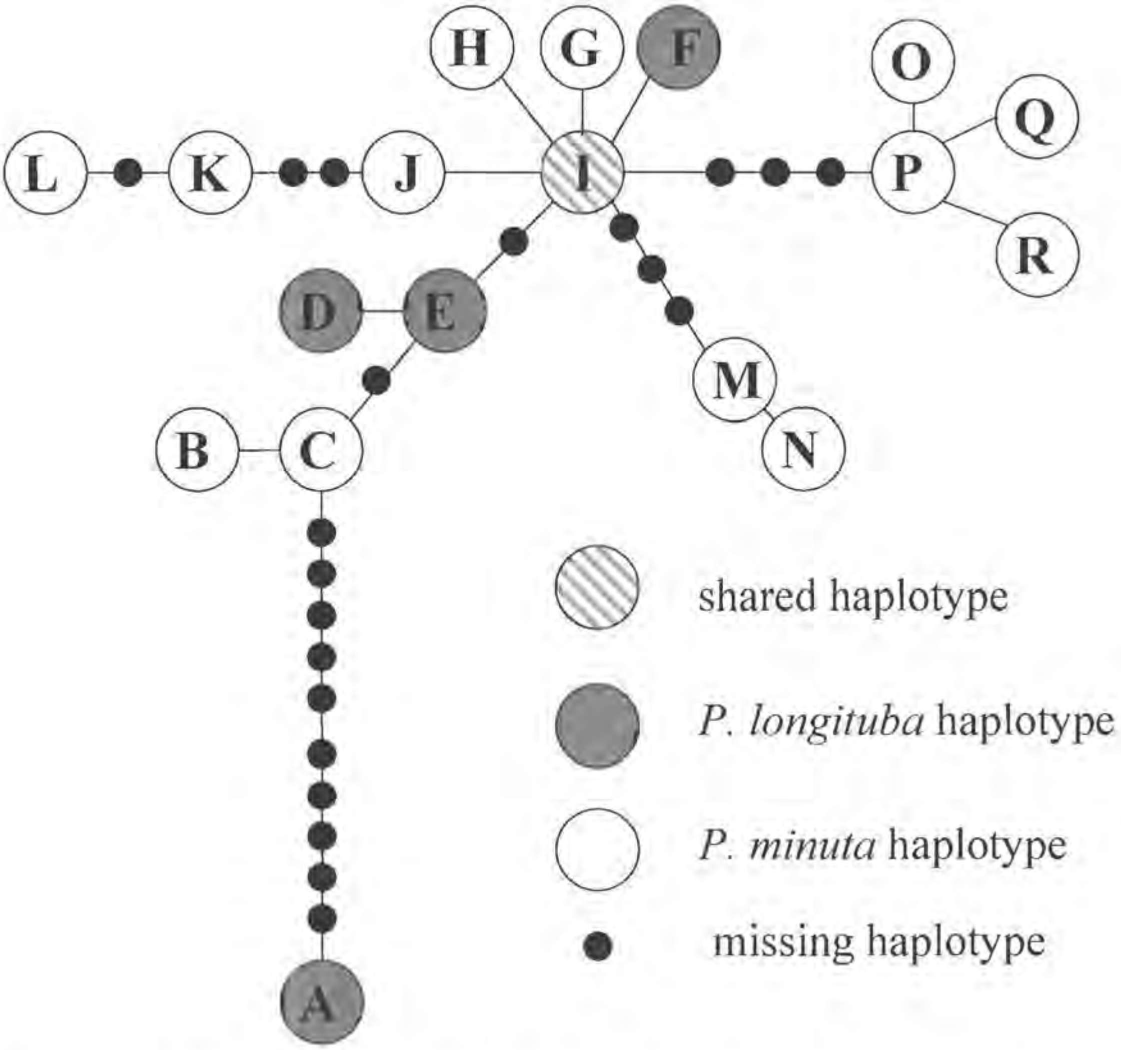


Figure 3.1: Pauridia haplotype network. Nodes on the network are haplotypes. These either belong to one species, or both, or were not found in the sample (missing) and are inferred. Lines connecting haplotypes indicate a single mutational event on the 1188bp sequence.

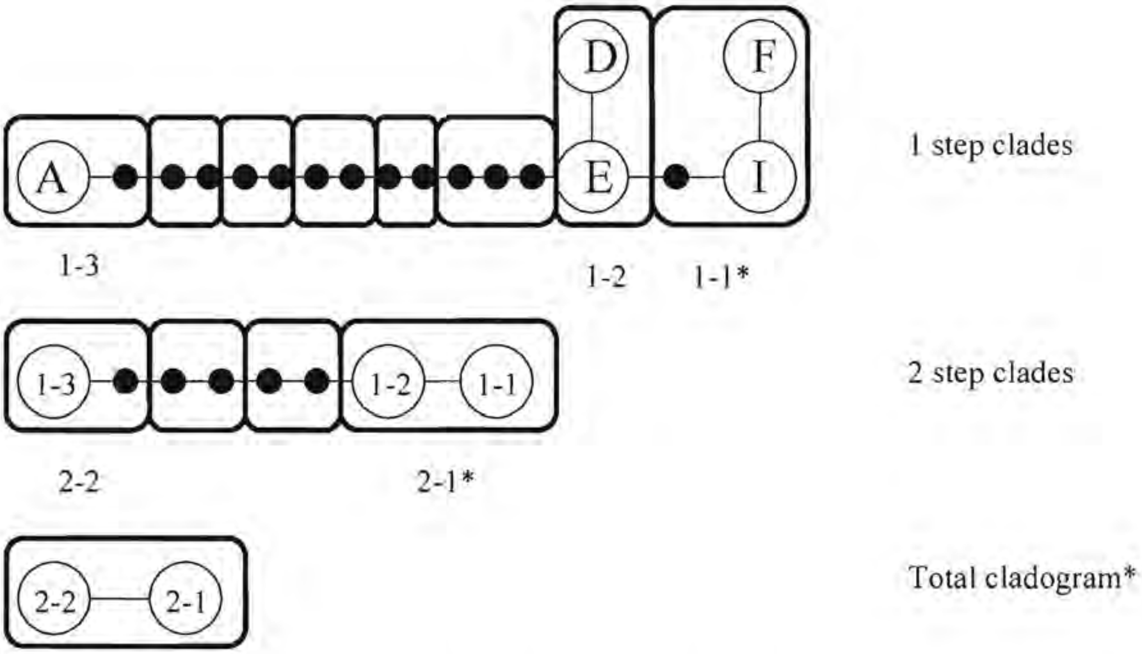


Figure 3.2.1: *Pauridia longituba* haplotype network and nesting clades. Clades in which a significant association of haplotypes and geography was detected are labelled with a *. Geographical distribution of clades and inferred allopatric breaks is given in Figure 3.2.2

Nested clade phylogeographic analysis

In the *P. longituba, P. minuta* and the total *Pauridia* analyses we were able to reject a random association of haplotypes with geography in multiple clades. The nesting structure for *P. longituba* is shown in Figure 3.2.1. In *P. longituba* geographical structure was inferred for clades 1-1, 2-1 and the total cladogram. In all cases the inference was one of allopatric fragmentation. In the context of *P. longituba* distribution this corresponds to multiple fragmentation events between granite outcrops (see Figure 3.2.2).

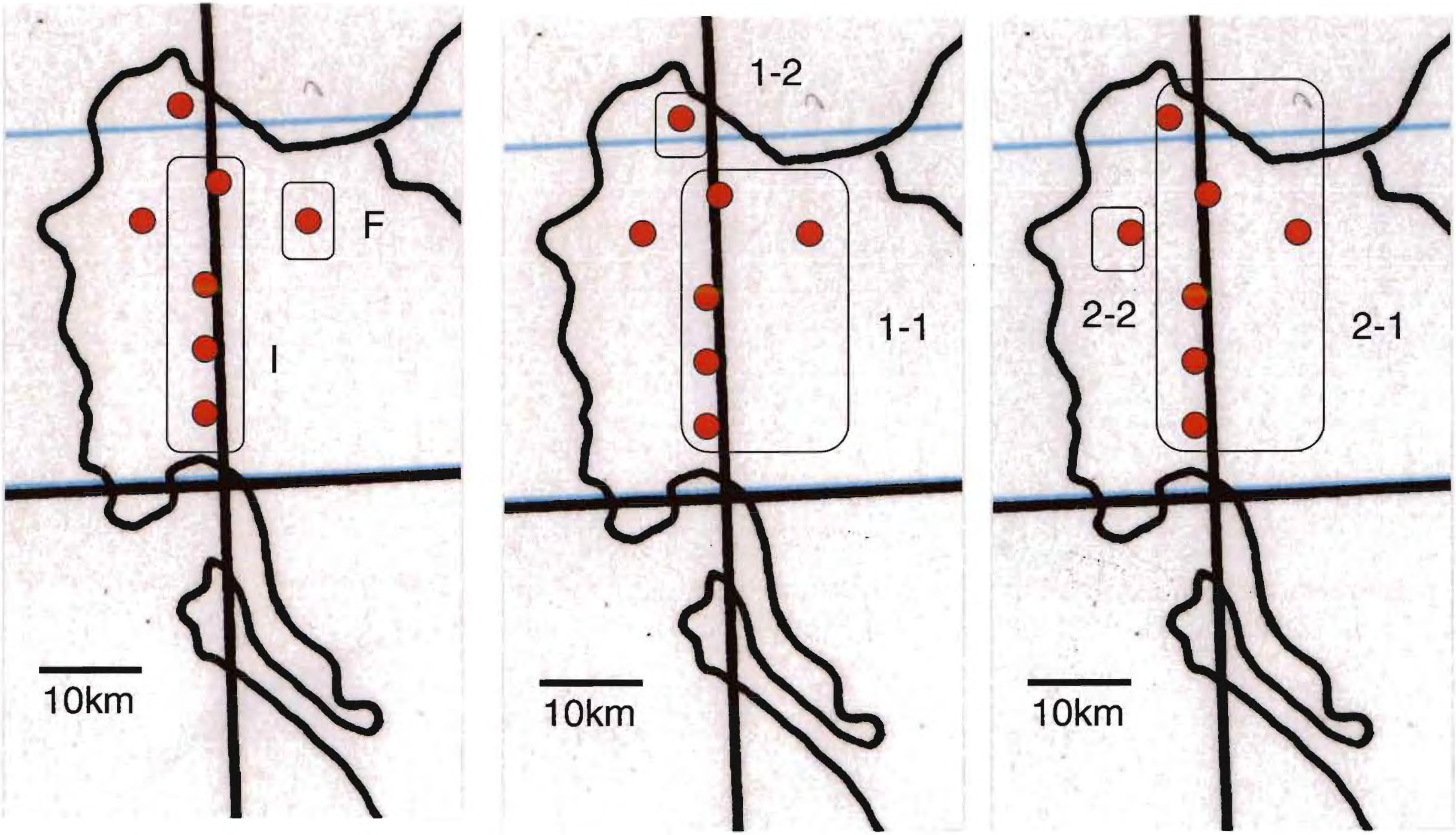


Figure 3.2.2: The association of clades with geography in *P. longituba*. Allopatric breaks are inferred between the clades I/F, 1-1/1-2 and 2-1/2-2. This corresponds to multiple allopatric breaks between isolated granite outcrops. Sampling sites are shown as red dots and all haplotypes within a clade are found exclusively within the populations denoted by the rectangular boxes.

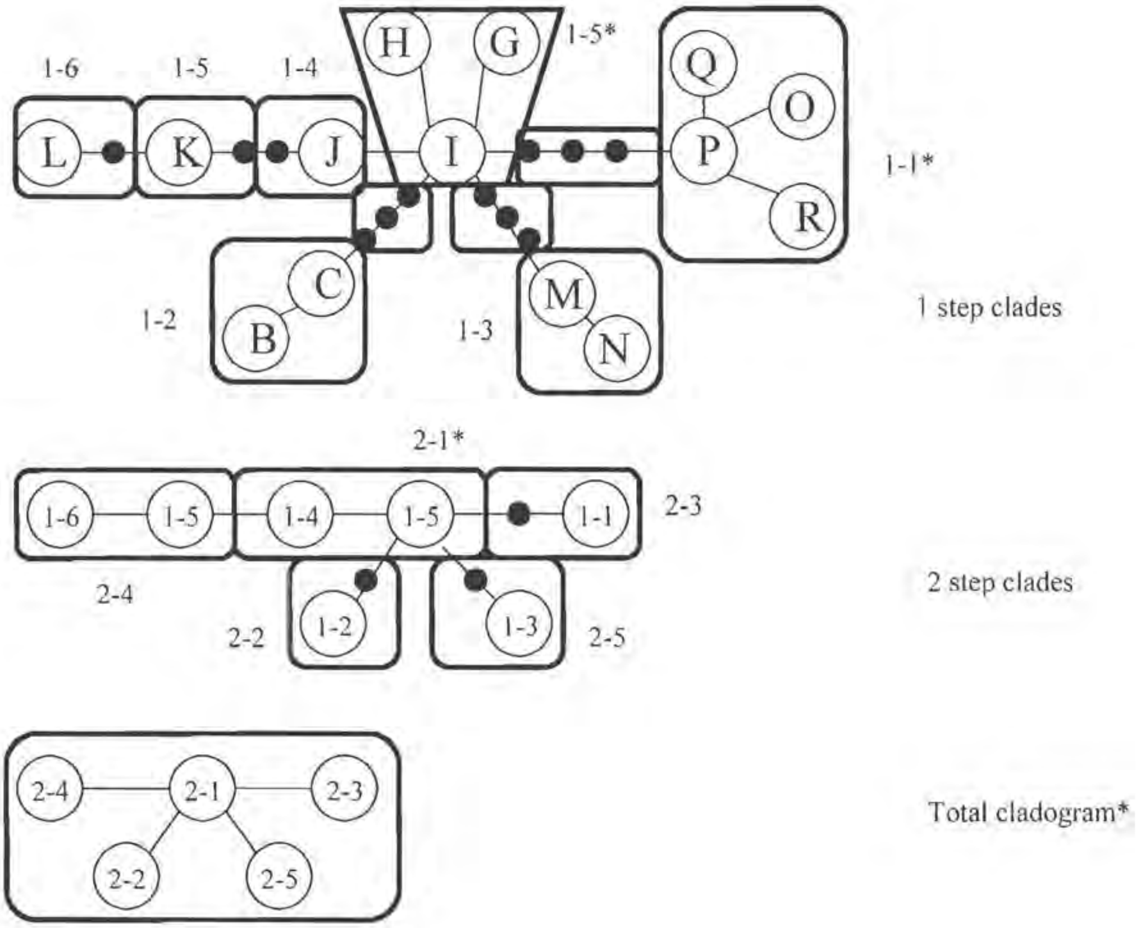


Figure 3.3.1: *Pauridia minuta* haplotype network and nesting clades. Clades in which a significant association of haplotypes and geography was detected are labelled with a *. Geographical distribution of 2-step clades and inferred allopatric breaks is given in Figure 3.3.2.

In *P. minuta* geographical structure was inferred in clades 1-1, 1-5, 2-1 and the total cladogram. The nesting structure is shown in Figure 3.3.1. In 1-5 and 2-1 contiguous range expansion was inferred and in clade 1-1 an inference of restricted gene flow was made. In the total cladogram, allopatric fragmentation was detected between the five 2-step clades (see Figure 3.3.2). Clade 2-4 consists of only a single population and sampling is too sparse to comment on this clade. Clades 2-1 and 2-3 are found in extensive lowland areas separated by a mountain range. Clades 2-2 and 2-5 are found on separate granite outcrop complexes respectively near to and sympatric with *P. longituba*.

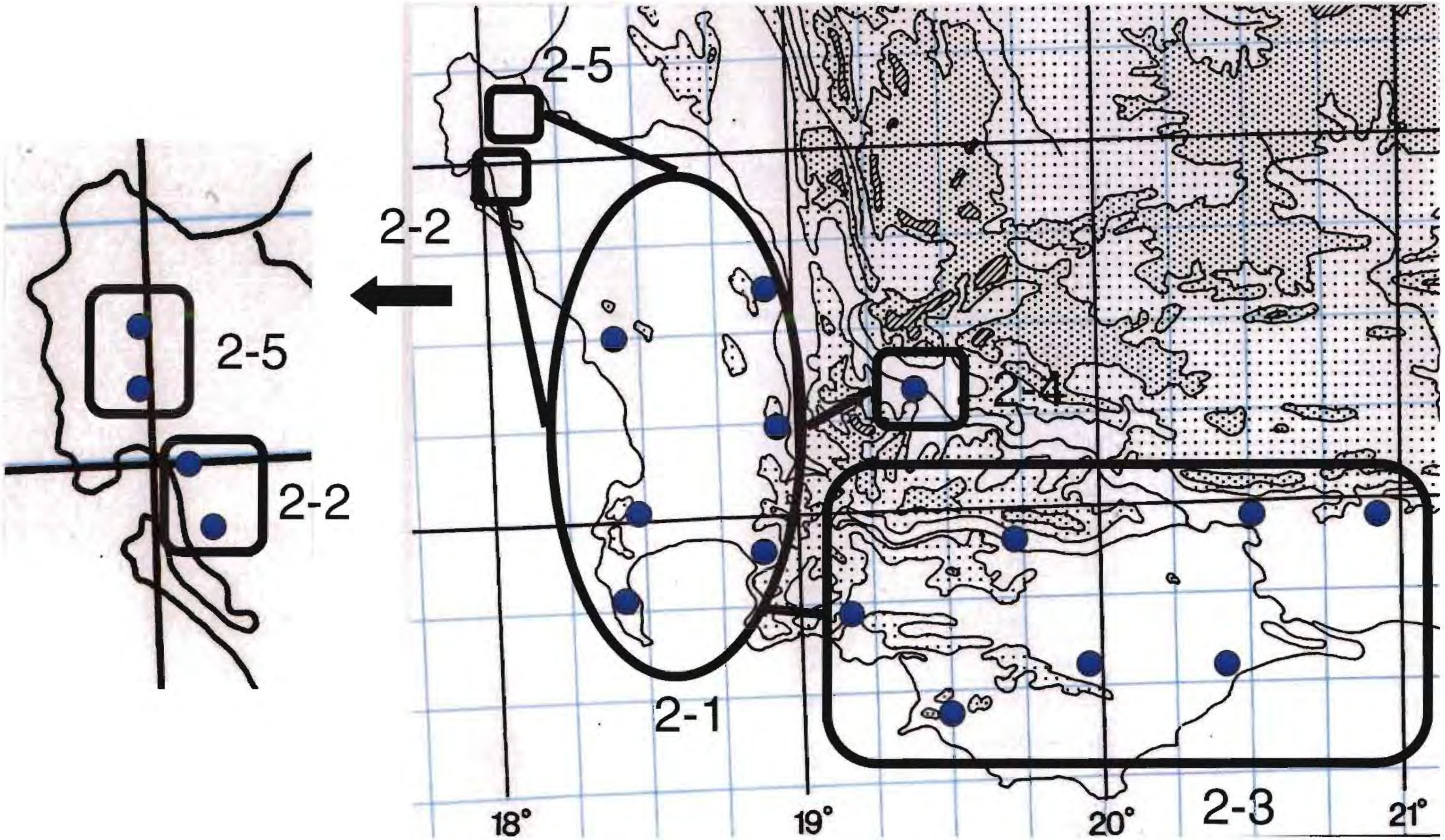


Figure 3.3.2: The association of 2-step clades with geography in *P. minuta*. Allopatric breaks are inferred between the 5 clades. Range expansion is inferred in clade 2-1 and restricted gene flow in clade 1-1. Clade 1-1 is geographically equivalent to clade 2-3. Sampling sites are shown as blue dots and all haplotypes within a clade are found exclusively within the populations denoted by the boxes.

The total *Pauridia* analysis revealed geographical structure in 6 clades (nesting structure not shown). In clade 1-6 no conclusive outcome was possible. Clades 2-3, 3-1 and 3-2 showed evidence of range expansion. In clade 1-15 restricted gene flow by distance was inferred and in clade 1-8 it was not possible to distinguish between range expansion and restricted gene flow. No conclusive outcome is reached for the total cladogram. These results are not shown graphically as they do not add to the picture established analyzing each species separately.

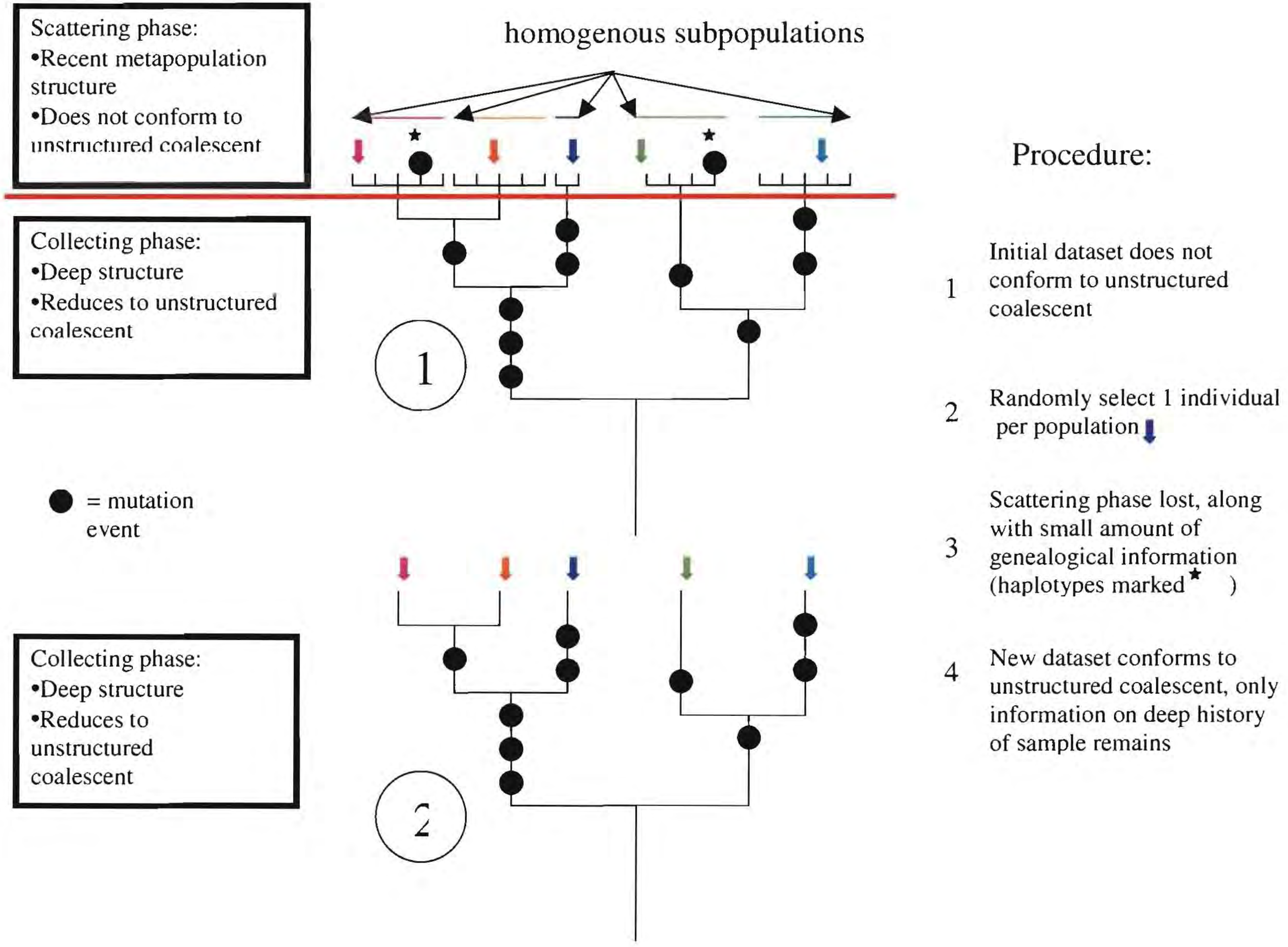
Although it does not give any additional information on spatial structuring of diversity relative to the species analysis, the *Pauridia* analysis does show isolation between the two species at the sympatric sites. All intra-species populations share haplotypes from a single two-step clade. In contrast the sympatric populations of *P. minuta* and *P. longituba* are distinct from one another at the two-step clade level, and haplotypes are separated by at least four mutational steps. There is therefore unambiguous isolation of the two species at the sites of contact, even though they do share the ancestral haplotype (I) at physically separate sites.

The NCPA of each species gives clear and geographically sensible signals of restricted gene flow and fragmentation. It is clear that the distribution of genetic diversity is highly structured in both species. This is especially true between *Pauridia* populations on the granite outcrops within and adjacent to the *P. longituba* range. The sympatric populations are genetically isolated from one another and their haplotypes are distinct at a deep nesting level. However, because there are no confidence estimates associated with specific NCPA inferences it is uncertain whether the particular inferences of range expansion and allopatric fragmentation are significant.

Sampling strategy

Coalescent-based models assume that within each subpopulation sampling is performed at random from a freely mixing population. Our NCPA results show that there is strong geographical substructure in both species. In fact, most populations (18 out of 23) contain only a single haplotype, whereas inter-population diversity is reasonably high. Aside from implying strongly restricted gene flow, this may also indicate metapopulation processes such as population extinction and recolonization (Whitlock 2003). In either case, within population sampling captures information that is not modeled when making the necessary simplifying assumptions that *P. minuta* and *P. longituba* are single, unstructured populations. Wakeley (2004) refers to these recent dynamics as the 'scattering phase'. During this phase only lineages within a subpopulation (ie. a *Pauridia* colony) can coalesce. Rare events such as occasional migration eventually bring these isolated lineages together (going backward in time). Although rare, these events are rapid on the coalescent time scale, so the average structure of the metapopulation is equivalent to the unstructured coalescent. This long-term history of metapopulations is the 'collecting phase'. When attempting to model speciation we wish to model the 'collecting phase' rather than recent metapopulation processes.

We can imagine the effect that biased sampling would have on our results by visualising a gene tree (see Figure 3.4). The data set that includes within-population sampling has many identical haplotypes within populations. This corresponds to many short or zero-length terminal branches on simulated genealogies. Because the models we use do not account for the true source of these short branches (strong metapopulation structure) they must be incorporated with the available parameters. The most obvious result should be that estimates of growth rate and population size will be much lower, in order to accommodate the many very recent coalescent events. It is not immediately clear what the effects on migration rate and divergence time will be, but these are intimately related to estimates of population size and will therefore be affected.



Scattering phase:
- Recent metapopulation structure
- Does not conform to unstructured coalescent

homogenous subpopulations

Procedure:

Collecting phase:
- Deep structure
- Reduces to unstructured coalescent

1 Initial dataset does not conform to unstructured coalescent

2 Randomly select 1 individual per population ⬇

● = mutation event

3 Scattering phase lost, along with small amount of genealogical information (haplotypes marked ★)

4 New dataset conforms to unstructured coalescent, only information on deep history of sample remains

Collecting phase:
- Deep structure
- Reduces to unstructured coalescent

Figure 3.4: Metapopulation structure and sampling strategy. See text for detailed explanation.

These problems can be partly solved by altering our sampling strategy. By simply sampling one individual from each population we lose the 'false' scattering phase information. There are still short terminal branches in the genealogy, but these reflect the actual history of the sample and not recent processes. The downside of this approach is that we have reduced sampling. In this case, we randomly pick one haplotype from each population and move from an overall sample of 125 individuals to 25 individuals (the reduced data set, see Table 3.1 and 3.2: random sample). This reduces our chance of sampling the *Pauridia* MRCA from 0.984 to 0.917 (the probability of sampling the MRCA = (n-1)/(n+1) (Rosenberg & Nordborg 2002)). There should be little overall reduction in haplotype diversity, nucleotide diversity and the number of polymorphic sites, because haplotypes within a population are closely related in this particular data set. The loss of power is therefore not as significant as phylogeographical intuition might suggest, and given the potentially confounding effects of metapopulation dynamics this seems justified. We therefore lose some precision, offset by the loss of a large confounding signal in the data.

Diversity statistics

Mismatch distribution tests and the R2 test for population size change did not give statistically significant results and therefore did not allow us to confirm or refute population size change (results not shown). Diversity statistics (Table 3.3) show that levels of diversity and polymorphism are approximately equal in the two species, although fewer individuals and haplotypes were sampled in *P. longituba*. This result is independent of which sampling strategy is used. Results from test of neutrality and expansion were not significant, except for the D^* test in the full P. longituba data set ($p<0.02$). This statistic measures the proportion of singletons relative to nucleotide differences in order to detect selection. This is most likely a result of almost complete haplotype homogeneity within populations (low singleton occurence) with large haplotype distance between populations (many nucleotide differences). This signal does not appear at all in the reduced data set, indicating that it is probably an artifact of sampling strategy and geographical structuring of variation.

dataset	*n*	*prob.*	*#haps*	*s*	*π*	*D*	*D**	*F**	*Fs*
P. minuta	92	0.98	14	22	0.004± 0.0002	-0.190	-0.129	-0.181	1.06
P. minuta<	18	0.89	11	21	0.004± 0.0005	-0.978	-0.523	-0.758	-2.37
P. longituba	33	0.94	5	17	0.003± 0.0007	-0.302	**1.61!**	1.15	3.29
P. longituba<	7	0.75	4	17	0.004± 0.002	-1.36	-1.37	-1.51	1.83

Table 3.3: Diversity statistics by dataset. The reduced (<) and complete datasets for each species, the probability of capturing the deepest coalescent event (*prob.*), the number of haplotypes (*#haps*), the number of polymorphic sites (*s*), the nucleotide diversity (π) ± one standard deviation, Tajima's *D*, Fu& Li's D^* and F^*, and Fu's *Fs* are shown. The statistically significant vaule of D^* is marked (!, p<0.02).

The number of polymorphic sites (*s*), haplotype diversity and nucleotide diversity (π) do not differ much between the complete and reduced data sets, confirming that most of the genealogical information in the dataset is retained with the reduced sampling strategy.

2.3 Coalescent-based data analysis

In order to explore the potential metapopulation bias in the complete sample both data sets were analysed for models A-D. Results from these analyses confirmed that there was downward bias in population size and migration estimates for the full data set (results not shown), and models E and F were therefore only applied to the reduced data set. Migrate and IM returned similar estimates for M and θ under model B for the complete data set (results not shown). On the reduced data set Migrate occasionally estimated extremely large values for θ_1, the size parameter for *P .longituba*. The results of the Migrate analysis are therefore not shown and the results for model B reflect only the IM analysis.

Fluctuate and Migrate return maximum likelihood estimates and associated confidence intervals, while IM generates posterior distributions. The smoothed peak of the curve and 90% highest posterior density (HPD) give the most likely estimate of the parameter and associated confidence interval respectively (Hey & Nielsen 2004). In cases where the curve was not smooth or was not adequately contained within the prior bounds an exact estimate of the 90% HPD was not possible. Parameter estimates are shown in Tables 3.4 and 3.5. Graphs of the posterior distributions are shown in Figure 3.5(a-h).

Parameter	θ_1	θ_2	θ_A	M_1	M_2	t
Model A:						
MLE	4.64	12.29				
lower	3.32	8.46				
upper	5.96	16.12				
Model B						
MLE	15.3	9.35		1.785	0.025	
lower	5.9?	1.85		0.485?	0.005	
upper	*	36.65		8.815?	7.235	
Model C						
MLE	30.95	26.25	17.15			0.877
lower	15.65?	11.05	6.75			0.222
upper	*	71.55	50.45			1.748
Model D						
MLE	38.41	42.78	23.42			0.885
lower	13.42?	15.64	9.47			0.185
upper	*	134.26	52.77			1.625
Model E						
MLE	17.17	10.31	22.80	1.75	0.01	0.1215
lower	5.7?	1.11	4.48	0.01?	0.01?	0.0165?
upper	*	101.20	102.94	17.45?	17.57?	*
Model F						
MLE	*	7.19	16.55	1.87	0.095	1.475
lower	*	1.41	0.10?	0.055?	0.005?	0.175?
upper	*	92.00	132.91?	8.10?	9.99?	9.985?

Table 3.4 Results of coalescent-based analyses: all parameters excluding growth.

The parameters and their relationship to biological measures are described in Tables 2.2-2.4. The models are as described in Figure 2.1. For each parameter per model the maximum likelihood estimate (MLE) (corresponding to the smoothed peak of the posterior distribution) and the upper and lower bound of the 90% highest posterior density (HPD) are given. MLE estimates marked '*' could not be determined, and bound estimates are marked with a '?' in cases for which the 90% HPD was not well defined.

Parameter	θ_c	θ_f	θ_c / θ_f
Model A			
P. longituba	4.64	4.20	1.10
P. minuta	12.29	2.52	4.88
Model D	s = 0.758		
P. longituba	17.17	17.75	0.97
P. minuta	10.31	5.67	1.82
Model F	s = 0.997		
P. longituba	?	16.55	?
P. minuta	7.19	0.050	143.8

Table 3.5 Results of coalescent-based analyses: growth parameters

IM peak estimates for s in Model D and F are shown. These are used to calculate the estimates of the current (θ_c) and founding (θ_f) population mutation parameter. This leads to the ratio of the current effective population size to founding effective population size (θ_c / θ_f) is given.

Model A estimates: These estimates are based on the MLE estimates of g given in the text. For a rough estimate of past population size we assume a founding time of T=2 and apply the formula $\theta_t = \theta\ e^{-gt}$.

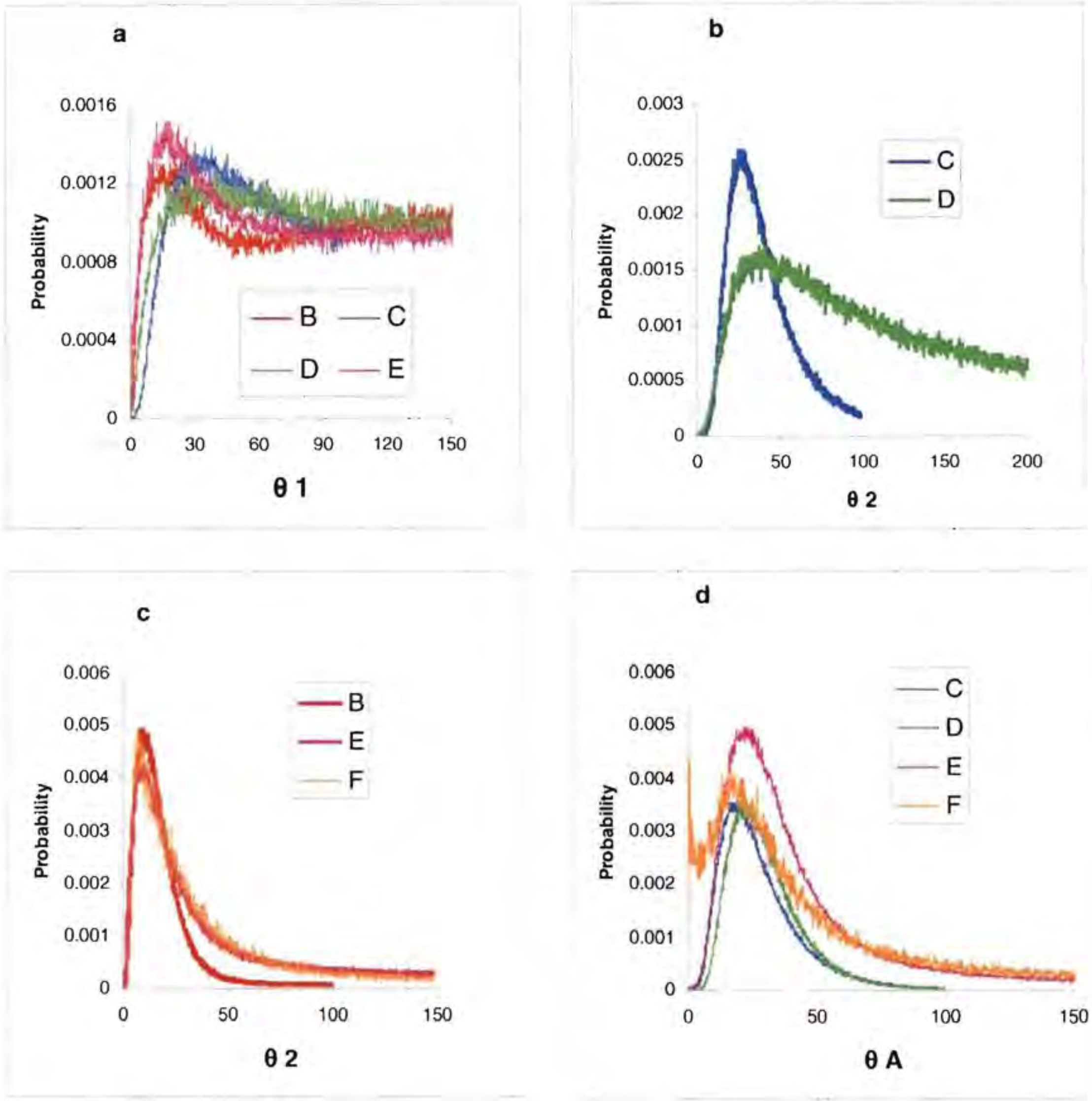


Figure 3.5: Posterior distributions of parameters generated by IM

3.5a - θ_1: Population mutation rate of *P. longituba*

3.5b&c - θ_2: Population mutation rate of *P. minuta*

3.5d – θ_A: Population mutation rate of the ancestral population

Curves are drawn for each model under which the relevant parameter was estimated.

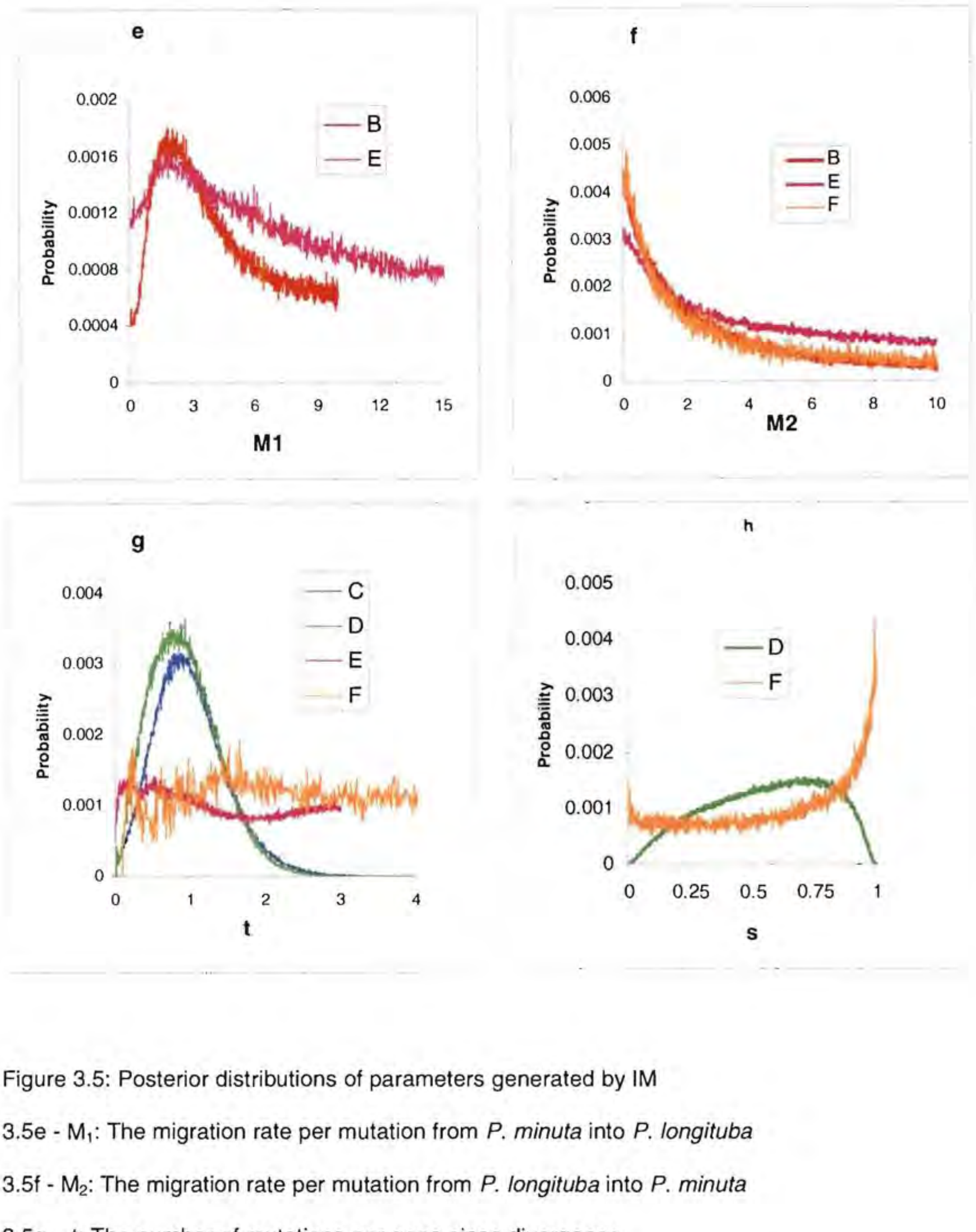


Figure 3.5: Posterior distributions of parameters generated by IM

3.5e - M_1: The migration rate per mutation from *P. minuta* into *P. longituba*

3.5f - M_2: The migration rate per mutation from *P. longituba* into *P. minuta*

3.5g – t: The number of mutations per gene since divergence

3.5h – s: The fraction of the ancestral population that founded *P. longituba*. The remaining fraction founded *P. minuta*

Curves are drawn for each model under which the relevant parameter was estimated.

Hypothesis testing

M_2, the migration rate from *P. longituba* into *P. minuta* was adjacent to the origin in all our analyses, with a peak estimate of not greater than 0.1 under any of the models. The peak estimates of θ_2 are of the order of 10. There are wide bounds on M_2 and we therefore cannot use this parameter to distinguish between alternative speciation hypotheses.

Migration in the opposite direction, M_1, is estimated to be higher. Under model B, the lower bounds on M_1 and θ_1 are 0.485 and 5.9. This gives a very conservative lower bound on the number of gene migrants into *P. longituba* $[(\theta_x M_x)/2]$ of about 1.5. There is therefore convincing evidence for high gene flow from *P. minuta* into *P. longituba* under model B. The lower bound on M_1 is not well defined for models E and F.

Estimates of divergence time were problematic. Only in the case of models C and D was T bounded by the prior. In this case we can calculate the maximum number of generations (in units of effective population size) since divergence. We use the upper bound on T and the lower bounds on θ_1 and θ_2 to establish a conservative upper bound on this parameter with the formula $4T/\theta_x = t / N_x$ (see Table 2.4). In the case of model C for *P. longituba* the maximum number of generations is 0.45 Ne and for *P. minuta* 0.63 Ne. For model D the maximum number of generations is 0.48 Ne for *P. longituba* and for *P. minuta* 0.42. These upper bounds are much lower than the number of generations required for neutral loci to reach reciprocal monophyly.

Models E and F provide approximately equal peak estimates to those generated by the other models (see Figure 3.5). The bounds on the parameters are not well defined and we cannot apply the tests described above to these models. We therefore proceed to test hypotheses relating to the founding sizes of the two species.

Posterior distributions of s under models D and F are shown in Figure 3.5h. In both cases the 90%HPD covers almost the entire prior range. The MLE for model D is 0.758, implying that the founding population of *P. minuta* was smaller than that of *P. longituba*. Under this model the probability of very unequal founding sizes is low (ie. the probability of s near to the boundary is near zero). Model F gives a very different picture. There is a very sharp peak near s = 1 and the smoothed peak estimate is 0.997. This implies that the founding population of *P. minuta* was very small and that the founding population of *P. longituba* was equal in size to the ancestral population. If this estimate is reliable it provides strong evidence for quantum splitting, but see the comments below. Models A, D and F all estimate growth in population size in *P. minuta* and population stability in *P. longituba* as shown in Table 3.5. In order to estimate ancient population size using Model A we guess a rough divergence time of T=2, near to the upper bound of T estimates (Figure 3.5g). These estimates provide weak evidence that *P. minuta* population has expanded and that *P. longituba* has been constant in size, results that are consistent with quantum budding of *P. minuta* and expansion into a new niche.

The simple models A, B, C and D provide useful and well-bounded estimates of the parameter values. Models incorporating both ancestral divergence and migration were more problematic. A visual inspection of Figure 3.5 shows that the curves for models E anf F were flatter and less informative than the other models. This is to be expected, given the limited data set. Model E gave good estimates for some parameters, but the parameters critical to hypothesis testing (T and M) were not well bounded. In the case of model F ESS values were too low (<35) for us to have much confidence in the posterior distributions. The results are included here because the curve for s shows a very strong peak, and the same result was obtained in 4 independent short runs, none of which was longer than 1.5 million updates.

4. Discussion

4.1 Methodological issues

NCPA

The confirmation of species status is the most unambiguous and useful of the haplotype network outcomes. Although the species share a haplotype globally, they are distinct at the sympatric sites. At these sites haplotypes are closely related within species, while these haplotype sets are deeply separated from one another. More broadly, there is strong evidence for restricted gene flow throughout the range of the genus. Because this is a single locus study, the inference is actually of restricted seed flow and it is possible that pollen is more easily dispersed. The inference of multiple allopatric breaks between populations on adjacent granite outcrops is striking. Because NCPA depends on branch lengths within a particular gene tree, this inference may be the result of more mutations accumulating on these branches by chance, and no assessment of this possibility is made by the method. Both species at these sites are geocarpic (ie. seeds are released underground) and it is possible that the inferred allopatric breaks simply represent extremely limited seed flow between outcrops over a long period of time, while pollen moves freely by small insect dispersal. Data from nuclear loci is needed to resolve this issue.

In contrast to the granite habitat, the lowland *P. minuta* habitat is inferred to be the site of range expansion, with fewer inferred allopatric breaks. The vast lowland habitat is split into only three subunits and the boundaries correspond exactly to inaccesible montane regions. Again, it is possible that these boundaries simply reflect barriers to seed flow, while pollen transmits other loci across these physical barriers. Existing montane barriers are a function of sea level. At periods when sea level was lower there would have been a coastal lowland corridor joining the two regions. This has occurred within the last 25 000 years (Barrable *et al.* 2002). It is therefore possible that the dispersal barrier could be incidental and that the fragmentation pattern is refugial in nature and associated with variable climatic regimes across an east-west cline (Cowling & Lombard 2002). During glacials the eastern CFR becomes more xeric, while the west becomes more mesic, potentially forcing eastern *P. minuta* to retreat to separate refugia even when a lowland corridor links the two regions. Alternatively, the fragmentation event could have occurred subsequent to the most recent rise in sea level.

A naïve overlay of the haplotype network on geography shows that the majority of diversity is concentrated at the granite outcrops and that lowland habitats are dominated by a few widespread, closely related haplotypes. Both this reading of the data and the NCPA are open to the criticism that the apparent signal may simply be an artefact of coalescent stochasticity. Because the complexity of genealogical processes easily exceeds our intuition (Hey & Machado 2003), we cannot exclude the possibility that the concentration of diversity around the *P. longituba* range and homogeneity on lowlands is a plausible random distribution for a locus with limited gene flow and high homogeneity within populations.

Diversity statistics

The diversity statistics that we applied to our data were not particularly informative. Tests of selection and population size dynamics produced no significant results. The fact that diversity is equally high in both species, even though *P. longituba* occupies such a small geographical range is suggestive, but open to the same coalescent stochasticity criticism as NCPA. The most informative outcome of these analyses was that by multiple measures, levels of diversity within species are essentially unchanged by a drastic reduction in sampling. As explained, this is most likely the result of metapopulation processes that tend to homogenize within-population diversity. The prediction that a reduced data set should retain most of the genealogical information was confirmed by the outcome of coalescent-based analyses.

Basic coalescent models: Fluctuate and Migrate

Migrate failed to return consistent results on the reduced dataset, a consequence of a flat likelihood surface for the limited *P. longituba* sample. Results with the larger data set were highly congruent with IM results on the same model. Although this is encouraging from a methodological perspective, these results are probably misleading. This is because the full data set is biased by the effect of metapopulation structure, a trend which is quite clear from the Fluctuate analysis. Estimates of g are significantly lower for the full data set for both species. If the bias were not present we would expect that the outcome of analyses on a subset of the data would be a similar maximum likelihood value with a wider confidence interval.

Isolation with Migration

When attempting to distinguish between alternative hypotheses using a Bayesian approach it is critical to consider the effect of prior choice on the outcome of analyses. Priors should be uninformative if we intend to distinguish between alternative hypotheses using only the molecular data. IM uses flat priors and the range on each prior should ideally be set so that the entire posterior distribution falls within the range. With some parameters and especially when analysing data with little information (ie. single locus data sets) this is not possible because posterior distributions have very long, low tails. It is therefore necessary to choose an upper bound for some parameters. In line with theoretical expectations (Hey 2005b) experience with this particular data set showed that the choice of upper bound on one parameter can have a large influence on all the other parameters. We therefore empirically determined upper bounds for problematic parameters such that bounds were large enough to not bias the final results, especially the estimate of the peak value.

We used ESS (estimated sample size) measures, autocorrelation values and multiple independent runs to monitor convergence and found that running multiple Metropolis-coupled chains was very helpful in increasing chain mixing. This is unusual for a single-locus dataset (Hey 2005b) and it is not clear why this is the case. Whatever the reason, final runs of less than 2 million updates took many days when implementing the most complex model. Longer runs of a few weeks would have been ideal had more computing power and time been available.

Estimates of θ were most successful for θ_A, as could be expected when the divergent populations are very closely related. In most cases the 90%HPD for θ_A and θ_2 could be determined. As in Migrate, there appeared to be too little information to estimate θ_1 and it generally showed a shallow peak followed by an upward trending tail. Setting the prior to include only the peak was found to bias the other parameter estimates, so we generally included a very wide prior on this parameter of approximately 20 times the estimated peak value. As shown in the results, estimates of θ_A and θ_2 were consistent across most models. Overall the estimates support approximately equal effective population sizes in the species, with a slightly larger ancestral population. This is consistent with results from NCPA and the diversity statistics if we assume that retained haplotype diversity is roughly a function of effective population size. Fluctuate also delivers roughly equal estimates of θ for both species. The important distinction is that the coalescent models indicate that we should have little faith in the θ_1 estimate and be cautious when using this parameter to test our hypotheses.

Migration estimates were quite broad and had long upper tails. In all models migration from *P. longituba* into *P. minuta* is estimated as close to zero. Migration in the opposite direction is most likely strong.. We found consistent evidence that migration into *P. longituba* has very likely been greater than 1 migrant per generation and therefore sufficient to prevent divergence at neutral loci. Estimates from model B are emphatic. Models E and F give similar peak estimates, but the curves are shallower and not well bounded.

Divergence time estimates are generally the most problematic of all IM parameters (McDaniel & Shaw 2005; Hey 2005a) and this study was no exception. The peak of the posterior distribution varied over a range from approximately 0.2 to 2 across different models and showed a rising tail for most models. The exception are models C and D, which exclude migration. Under these models t is well bounded allowing us to show that under models C and D divergence was too recent to allow for divergence at neutral loci

Analysis of the data with the full IM model with splitting provides the most unambiguous of all the parameter estimations. The value of s shows a single strong peak near s=1. Biologically this implies that *P. minuta* was founded by a very small subset of the ancestral population. The present size of *P. minuta* is estimated at roughly the ancestral population size. This implies that *P. minuta* has undergone exponential growth since splitting. This is consistent with the Fluctuate analysis. Although bounds on parameters under Model F are not well defined the key features of the other models, unidirectional migration and very recent splitting time relative to θ appear to follow the same trend. As mentioned, ESS values were low and it will be necessary to conduct much longer runs with different search strategies to confirm the s estimate. It may be possible that there is simply too little data to estimate so many parameters. For the short runs that were performed ESS values appeared to rise steadily and very slowly after 1 million steps, implying that long runs might eventually sample the search space sufficiently. On our system the required run time for a single Model F run of 20 million updates with 10 Metropolis coupled chains would be about six weeks.

4.2 Hypothesis testing

We attempted to distinguish between vicariant and quantum modes of species splitting and to distinguish between selection- and drift-driven speciation. There are several lines of ecidence that quantum speciation has occurred in *Pauridia*. Firstly, if the results of the model F IM analysis are correct then there is a high probability that the founding population of *P. minuta* was very small relative to the founding population of *P. longituba*. Secondly, Fluctuate predicts that *P. minuta* has grown exponentially, implying that the *P. minuta* population was much smaller in the past. Descriptive phylogeography, although not as powerful as the other methods, provides supporting evidence. High diversity and structuring of variation is detected in the very small geographical range of *P. longituba*, implying stability over a long period of time. *P. minuta* shows evidence of recent range expansion with lower structuring and diversity over much of its lowland range. The populations of *P. minuta* that do show fine geographical structuring are near or within the current *P. longituba* range.

This evidence allows us to reject an hypothesis of quantum speciation in which *P. longituba* is the derived species. All the evidence points towards the opposite case, in which *P. minuta* was founded by a subpopulation of *P. longituba*. Rejecting a vicariant scenario would firstly require that we have confidence that the inferred growth in *P. minuta* is both biologically meaningful (ie. not an artefact of the model assumptions) and is related to its founding and expansion into a new niche. The fact that IM supports the growth estimate enhances confidence that the Fluctuate output is not an artefact of excluding ancestral variation and migration from the model. If exponential growth is real it may point to some process other than species founding, such as expansion out of a refugium. A refugial model would require that all extant *P. minuta* variation is derived from a single, recent refugium. This seems biologically unreasonable given its broad range and the climatic and topographical complexity of the CFR. Our data also suggest that this is unlikely, because *P. minuta* shows a high degree of geographical structuring of haplotypes that would be atypical of expansion from a recent refugium

Rejecting vicariant speciation requires that we have confidence in the IM estimate of the splitting parameter. Hey (2005a) has stated that estimates of all parameters are difficult with few loci and that estimation of s is especially difficult when migration is high. If the data is ambiguous we should therefore see no clear peak. The contrary is true, implying that there is a signal of uneven splitting in the data. This signal occurred over all independent runs of model F, although low ESS values (<35) for the parameter T were cause for concern. Although each individual line of evidence is not very strong, all the analysis are consistent with one interpretation. On the basis of limited single locus data there appears to be a signal of quantum speciation, with *P. minuta* derived from *P. longituba*.

Our second hypothesis relates to the relative role of drift and selection in shaping the divergence of the two species. As shown in the results, under models B and D migration from *P. minuta* into *P. longituba* is sufficiently high to homogenize neutral loci. Under model C almost no loci would have coalesced since splitting. Whether due to recent ancestry or migration, there is strong evidence that neutral loci could not have diverged in the two species.

In the model that includes the splitting parameter the initial effective size of *P. minuta* is predicted to have been very small. It is not possible to reject the hypothesis that some loci coalesced by neutral drift within *P. minuta* while the population size was small and this allowed for reproductive isolation to develop. Whatever the initial cause migration appears strongly unidirectional, implying a barrier to gene flow into *P. minuta* but not into *P. longituba*. In either case, selection has acted to maintain certain loci in *P. longituba* in the face of ongoing gene flow. We therefore have too little information to discern whether or not selection was involved in the initiation of speciation, but it seems very likely that it has been involved in the maintenance of species identity in at least *P. longituba*.

The models therefore provide strong evidence that selection has been involved at some stage of speciation, maintaining loci that define the phenotypic divergence of the two species. The individual pieces of evidence for quantum speciation are somewhat weaker but multiple lines of evidence support a model in which *P. minuta* recently budded off from *P. longituba* and expanded its range.

4.3 Implications

This work has obvious implications for phylogenetic studies of the Cape flora. CFR phylogenetic studies have often failed to resolve species relationships into bifurcating trees within clades (see Linder 2003 and refs. therein). If speciation within *Pauridia* is typical of CFR elements, then this polyphyly reflects the real processes of ongoing gene flow and shared ancestry in closely related species rather than lack of data. This implies that coalescent approaches may be more useful than phylogenetics in understanding CFR diversification even at the genus level within clades that have diversified rapidly or in which reproductive isolation is incomplete. This also implies that intra-specific studies will benefit from the inclusion of sister taxa because this makes it possible to separate the effects of internal geographical structure and introgression of lineages from other species.

On a methodological level, adjusting the sampling strategy usually required for phylogeography (ie. multiple individuals per population) appears to have been very helpful in recovering the underlying history. Sampling of only one individual per population allowed for quite detailed reconstruction of evolutionary history, given that only a single locus was used. Nested clade analysis and diversity statistics were useful heuristic tools, but did not add much to the overall analysis of the data. As suggested by Templeton (2005), NCPA will probably continue to be useful as an indicator of which alternative hypotheses should be tested in cases where there is no prior expectation of a particular pattern. It should be acknowledged that coalescent models that analyse specific evolutionary hypotheses involve a number of simplifying assumptions and should be applied cautiously, as demonstrated by the metapopulation structure sampling bias discovered in this study.

We found multiple lines of evidence for a counter-intuitive case of quantum speciation, in which the derived species is much more widely dispersed than its progenitor. Although not in line with most studies of this mode of speciation (Gottlieb 2003) this seems to be a necessary mode if quantum speciation actually contributes to overall diversification of plant lineages. If all derived lineages remained restricted and specialised then novel progenitor species would not appear. Although this is an obvious implication of the theory, it is surprising that the signal of speciation is still retrievable at the time when the derived species (*P. minuta*) already shows some evidence of beginning to fragment. This is generally encouraging for the phylogeographical study of plant speciation, because the normally confounding effects of shared ancestry and ongoing migration become useful in a coalescent context. Coalescent analysis also does not require as detailed a data set as methods that directly interpret a gene tree. Lack of polymorphism has generally retarded plant phylogeography, but coalescent-based analyses are more dependant on independent genealogies than the details of particular realised genealogies.

In this study we have avoided attempts to infer the geographical mode of speciation, largely because potential range shifts make inferences on ancestral distribution very difficult (Losos & Glor 2003). If *P. longituba* is the ancestral species, then there is a very defined range in which the peripheral isolated population could have been founded and there is a case for parapatric speciation. This chance of long distance dispersal to a granite outcrop separate from the isolated Vredenburg granite complex is very remote, given the data on restricted gene flow in *P. longituba*. When discussing parapatric speciation Coyne & Orr (2004: 124) state: "If an edaphic endemic species of plant – one restricted to a small patch of aberrant habitat – has its sister species in an adjacent area, one can conclude that speciation was parapatric." This description appears to describe *Pauridia* quite well and the *P. minuta* – *P. longituba* split may represent an example of parapatric speciation. This geographical classification of the speciation process is problematic because parapatry in a very limited dispersal organism such as *P. longituba* may be a very different evolutionary phenomenon to parapatry in, for instance, a migratory bird.

4.4 *Pauridia* in the CFR

Without an accurate measure of the mutation rate it is not possible to establish real values corresponding to parameter estimates. We can however establish rough estimates of population size and divergence time using the known range of mutation rates. In the plant chloroplast noncoding regions mutate at a rate of between 0.02 and 0.11 % / million years (Zhang & Hewitt 2003). Peak estimates of T (about 1) and θ_A (about 20) were quite well bounded in models C and D and similar estimates were obtained under models B, E and F. This translates to a range of 760 000 to 4.2 million years for divergence time and 3.8 to 21 million for the effective ancestral population size. This divergence time falls roughly within estimated rates of lineage diversification in the CFR (Linder 2003). Taking the bounds on θ_A (5-100) with the upper and lower mutation rates effective ancestral population size could be anywhere between 1 million and 100 million. These estimates are approximately equal to the *P. longituba* census size, but much smaller than the estimated *P. minuta* census size (see section 1.7). This provides further support for the hypothesis that the ancestor of the two species was *P. longituba*. The current estimate of θ for *P. minuta* could be low because the mutation process has not yet had time to reflect the demographic change in the species.

It is not possible to say exactly what caused the initial isolation of the founding population but we do know that the defining features of *P. minuta* are its longer pedicel, shorter perianth tube and increased seed dispersal. *P. minuta* is in fact the only species in the Hypoxidaceae without a narrow perianth tube, providing further evidence that it is the derived species (D. Snijman, pers. comm.). *P. longituba* is highly adapted to a seasonally moist habitat, but has very restricted seed dispersal. It is therefore possible to imagine a situation in which new seasonally moist habitat opened up and selection acted for increased dispersal of a subtype of *P. longituba* into this open niche. At the same time selection acted to maintain the identity of *P. longituba* in response to increased seed dispersal from the new species back into the parent range, implying that geocarpy is somehow advantageous for *P. longituba*. It is interesting that current *P. minuta* populations in the same area tend towards geocarpy. The difference in seed dispersal capacity may also explain why gene flow (at least seed flow) of *P. longituba* into *P. minuta* is low.

Given the relatively broad array of habitats inhabited by *P. minuta* it seems that only a broad climatic transition would be sufficient to create a suitable niche. The very approximate estimate of divergence time given above (0.8 – 4.2 my) corresponds to the development of a full Mediterranean climate in the CFR and Pleistocene glacial cycling (Linder 2003). Either one of these factors could have opened up new seasonally moist habitat for *P. minuta* and fluctuations in climate may have played a role in isolating peripheral populations of the ancestral species.

This rough estimate of divergence time is very relevant to the NCPA inferred fragmentation of eastern and western populations of *P. minuta*. If the species has existed for much of the Pleistocene, then there would have been many long periods when there was a wide (>100km) lowland corridor linking the currently isolated lowland regions (Barrable *et al.* 2002). If the NCPA result is accurate, then this implies distinct refugia in eastern and western regions during climatic cycles or that the two populations are cryptic species. This deserves further study, including detailed morphological analysis.

If it is true that *P. minuta* arose in the fashion described above then another question, beyond the scope of this study, is 'Why is *P. longituba* not extinct, given its extreme specialisation and range restriction?'. This question could be equally asked of the hundreds of micro-endemic species in the CFR and the answer may be linked to the unusual climatic stability of especially the western CFR (Cowling & Lombard 2002)

4.5 Summary

This study is one of the first attempts to study plant phylogeography in the CFR and the first in this flora to include more than one species and to apply coalescent based likelihood techniques. The very recent development of the theoretical tools to deal with reconstructing quantum speciation in plants makes this possibly this first study to apply IM to this particular question. A counter-intuitive suggestion of 'reverse' quantum speciation has been obtained with a variety of modelling techniques. Given the limited sampling undertaken and the fact that only a single locus was used, future studies with better sampling strategies and more loci should gain even greater insight into CFR evolutionary processes. There are many hundreds of fascinating species pairs in the flora and potential examples of adaptation to pollinators, fire, topography, edaphic diversity and novel dispersal strategies provide a rich substrate for future studies of speciation in this flora.

There is obviously much room for improvement of this study. Immediately, it should be possible to achieve narrower confidence intervals and check for convergence by running more thorough analyses. This is especially required for the full IM model with the splitting parameter (model F). In the longer term, given that interesting results has been obtained it might be more productive to expand the number of loci in this study group rather than gather more chloroplast data for other species pairs. This may reveal the mosaic nature of adaptation predicted by genic theories of species and suggested by the role for selection in shaping *Pauridia*. Gathering nuclear data from plants is difficult (Zhang & Hewitt 2003). This study included a failed attempt to use ribosomal loci, but multiple paralogous copies appeared to be present within individuals and further analysis was not possible. A variety of new methods are now appearing for gathering nuclear DNA data (Zhang & Hewitt 2003) and perseverance on this front will be necessary if real progress is to be made.

It would also be interesting to examine ploidy levels within the genus, as hybridisation is a significant process in plant speciation (Rieseberg 2001). Cross-pollination and breeding studies may also be useful, but these will likely be confounded by the long generation time of at least three years. Given the evidence for gene flow from *P. minuta* into *P. longituba* it would be interesting to determine if hybrids are infertile or whether ecological factors prevent their survival at the sympatric sites. More detailed morphological studies are needed for *Paurida* and for the Cape Hypoxidaceae more broadly. Field observation of morphological variation within many species suggests that there may be much taxonomically informative variation still to be described (D. Snijman pers. comm.).

Aside from increased data quality there is much room for the development of novel theoretical and simulation tools. IM models that incorporate more than three populations should greatly enhance the power of estimates in very dynamic systems, where the ancestral population itself is subdivided (see Won & Hey 2005). Also required are models that incorporate and estimate metapopulation structure. Current methods push the boundaries of computational power, so model complexity will probably follow the increase in available processing speed.

Although inferring fine evolutionary processes is a difficult science it is also very exciting. This study has confirmed the usefulness of coalescent-based tools in studying CFR speciation and will hopefully form part of an increasing trend towards empirical reconstruction of the fine processes that have generated the diversity of the magnificent Cape flora.

References

Abdo Z, Crandall KA, Joyce P (2004) Evaluating the performance of likelihood methods for detecting population structure and migration. *Molecular Ecology*, **13**, 837-851.

Bakker FT, Culham A, Marais E, Gibby M (2004) Nested radiations in Cape Pelargonium. In: Plant species-level systmeatics: new perspectives in pattern and process (ed. Bakker FT *et al.*) Konigstein, Koeltz.

Ballard JWO, Whitlock MC (2004) The incomplete natural history of mitochondria. *Molecular Ecology*, **13**, 729-744.

Barrable A, Meadows ME, Hewitson BC (2002) Environmental reconstruction and climate modeling of the Late Quarternary in the winter rainfall region of the Western Cape, South Africa. *South African Journal of Science*, **98**, 611-616.

Barrowclough GF, Groth JG, Mertz LA, Gutierrez RJ (2004) Phylogeographic structure, gene flow and species status in blue grouse (*Dendragapus obscurus*). *Molecular Ecology*, **13**, 1911-1922.

Barrowclough GF, Groth JG, Mertz LA, Gutierrez (2005) Genetic structure, introgression, and a narrow hybrid zone between northern and California spotted owls (*Strix occidentalis*). *Molecular Ecology*, **14**, 1109-1120.

Beerli P, Felsenstein J (1999) Maximum likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, **152**, 763-773.

Beerli P, Felsenstein J (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences USA*, **98**, 4563-4568.

Bernardi G (2005) Phylogeography and demography of sympatric sister surfperch species, *Embiotica jacksoni* and *E. lateralis* along the California coast: historical versus ecological factors. *Evolution*, **59**, 386-394.

Carstens BC, Sullivan J, Davalos LM, Larsen PA, Pedersen SC (2004a) Exploring population genetic structure in three species of Lesser Antillean bats. *Molecular Ecology*, **13**, 2557-2566.

Carstens BC, Stevenson AL, Degenhardt JD, Sullivan J (2004b) Testing nested phylogenetic and phylogeographic hypotheses in the *Plethodon vandykei* species group. *Systematic Biology*, **53**, 781-792.

Carstens BC, Degenhardt JD, Stevenson AL, Sullivan J (2005) Accounting for coalescent stochasticity in testing phylogeographical hypotheses: modeling Pleistocene population structure in the Idaho giant salamander *Dicamptodon aterrimus*. *Molecular Ecology*, **14**, 255-265.

Cassens I, van Waerebeek K, Best PB, Tzika A, van Helden AL, Crespos EA, Milinkovitch MC (2005) Evidence for male dispersal along the coats but no migration in pelagic waters in dusky dolphins (*Lagenorhynchus obscurus*). *Molecular Ecology*, **14**, 107-121.

Cavers S, Navarro C, Lowe AJ (2003) Chloroplast DNA phylogeography reveals colonization history of a Neotropical tree, *Cedrela odorata* L., in Mesoamerica. *Molecular Ecology*, **12**, 1451-1460.

Coleman M, Abbot RJ (2003) Possible causes of morphological variation in an endemic Moroccan groundsel (*Senecio lecanthemifolius* var. *casablancae*): evidence from chloroplast DNA and random amplified polymorphic DNA markers. *Molecular Ecology*, **12**, 423-434.

Cowling RM (1992) The ecology of fynbos: Nutrients, fire and diversity. Oxford University Press, Cape Town.

Cowling RM, Holmes PM (1992) Flora and vegetation. In: The ecology of fynbos: Nutrients, fire and diversity (ed. Cowling RM), pp 23-61. Oxford University Press, Cape Town.

Cowling RM, Rundel PW, Lamont BB, Arroyo MK, Arianoutsou M (1996) Plant diversity in mediterranean-climate regions. *Trends in Ecology and Evolution*, **11**, 362-366.

Cowling RM, Pressey RL (2001) Rapid plant diversification: planning for an evolutionary future. *Proceedings of the National Academy of Sciences USA*, **98**, 5452-5457.

Cowling RM, Lombard AT (2002) Heterogeneity, speciation/extinction history and climate: explaining regional plant diversity patterns in the Cape Floristic Region. *Biodiversity Research*, **8**, 163-179.

Coyne JA, Orr HA (2004) Speciation. Sinauer, Sunderland.

Dobes CH, Mitchell-Olds T, Koch MA (2004) Extensive chloroplast haplotype variation indicates Pleistocene hybridization and radiation of North American *Arabis drummondii*, *A. x divaricarpa* and *A. holboellii* (Brassicaceae). *Molecular Ecology*, **13**, 349-370.

Edwards SV, Beerli P (2000) Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution*, **54**, 1839-1854.

Excoffier L (2004) Special issue: Analytical methods in phylogeography and genetic structure. *Molecular Ecology*, 13, 727.

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

Fu Y, Li WH (1997) Statistical tests of neutrality of mutations. *Genetics*, **133**, 693-709.

Funk DJ, Omland KE (2003) Species level paraphyly and polyphyly: Frequency, causes, and consequences, with insights from animal mitochondrial DNA. Annual Review of Ecology, Evolution and Systematics ??, 397-423.

Gay L, Du Rau PD, Mondain-Monval JY, Crochet PA (2004) Phylogeography of a game species: the red-crested pochard (*Netta rufina*) and consequences for its management. *Molecular Ecology*, **13**, 1035-1045.

Goldblatt P (1978) An analysis of the flora of Southern Africa: its characteristics, relationships and origins. *Annals of the Missouri Botanical Gardens*, **65**, 369-436.

Goldblatt P, Manning J (2000) Cape plants. A conspectus of the Cape flora of South Africa. National Botanical Institute, Pretoria.

Gottlieb LD (2003) Rethinking classic examples of recent speciation in plants. *New Phytologist*, **161**, 71-82.

Grant V (1981) Plant speciation, 2nd edn. Columbia University Press, New York.

Greenberg AJ, Moran JR, Coyne JA, Wu CI (2003) Ecological adaptation during incipient speciation revelaed by precise gene replacement. *Science*, **302**, 1754-1757.

Harpending HC, Sherry ST, Rogers AR, Stoneking M (1993) The genetic structure of human populations. *Current Anthropology*, **34**, 483-496.

Hewitt GM (2001) Speciation, hybrid zones and phylogeography – or seeing genes in space and time. *Molecular Ecology*, **10**, 537-549.

Hey J, Wakeley J (1998) Testing speciation models with DNA sequence data. In: Molecular approaches to ecology and evolution (eds. De Salle R & Schierwater B). pp. 157-176. Birkhauser, Berlin.

Hey J (2001) The mind of the species problem. *Trends in Ecology and Evolution*, **16**, 326-329.

Hey J, Waples RS, Arnold ML. Butlin RK, Harrison RG (2003) Understanding and confronting species uncertainty in biology and conservation. *Trends in Ecology and Evolution*, **18**, 597-603.

Hey J, Machado CA (2003) The study of structured populations – new hope for a difficult and divided science. *Nature Reviews Genetics*, **4**, 535-543.

Hey J, Won Y-J, Sivasundar A, Nielsen R, Markert JA (2004) Using nuclear haplotypes with microsatellites to study gene flow between recently separated Cichlid species. *Molecular Ecology*, **13**, 909-919.

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.

Hey J (2005a) On the number of New World founders: A population genetic portrait of the peopling of the Americas. *PloS Biology*, **3**, e193.

Hey J (2005b) IM documentation. http://lifesci.rutgers.edu/~heylab/ProgramsandData/Programs/IM/IMdocumentation_Sept_5_05.pdf

Hoffman EA, Blouin MS (2004) Evolutionary history of the northern leopard frog: reconstruction of phylogeny, phylogeography, and historical changes in population demography from mitochondrial DNA. *Evolution*, **58**, 145-159.

Holman JE, Hughes JM, Fensham RJ (2003) A morphological cline in *Eucaplyptus*: a genetic perspective. *Molecular Ecology*, **12**, 3013-3025.

Kingman JFC (1982) On the genealogy of large populations. *Journal of Applied Probability*, **19A**, 27-43.

Klak C, Reeves G, Hedderson T (2004) Unmatched tempo of evolution in Southern African semi-desert ice plants. Nature, 427, 63-65.

Knowles LL (2001) Did the Pleistocene glaciations promote divergence? Tests of explicit refugial models in montane grasshoppers. *Molecular Ecology*, **10**, 691-701.

Knowles LL, Maddison WP (2002) Statistical phylogeography. *Molecular Ecology*, **11**, 2623-2635.

Knowles LL (2004) The burgeoning field of statistical phylogeography. *Journal of Evolutionary Biology*, **17**, 1-10.

Kotlik P, Bogutskaya NG, Ekmekci FG (2004) Circum Black Sea phylogeography of *Barbus* freshwater fishes: divergence in the Pontic glacial refugium. *Molecular Ecology*, **13**, 89-95.

Kuhner MK, Yamato J, Felsenstein J (1998) Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, **149**, 429-434.

Latta RG (2003) Gene flow, adaptive population divergence and comparative population structure across loci. *New Phytologist*, **161**, 51-58.

Linder HP (1985) Gene flow, speciation, and species diversity patterns in a species-rich area: the Cape flora. In: Species and speciation (ed. Vrba ES), pp. 53-57. Transvaal Museum Monograph No. 4, Traansvaal Museum, Pretoria.

Linder HP (2003) The radiation of the Cape flora, southern Africa. *Biological Reviews*, **78**, 597-638.

Linder HP, Hardy CR (2004) Evolution of the species-rich Cape flora. *Philosophical Transactions of the Royal Society of London, Series B*, **359**, 1623-1632.

Losos JB, Glor RE (2003) Phylogenetic comparative methods and the geography of speciation. *Trends in Ecology and Evolution*, **18**, 220-227.

Lumaret R, Mir C, Michaud H, Raynal V (2002) Phylogeographical variation of chloroplast DNA in holm oak (*Quercus ilex* L.). *Molecular Ecology*, **11**, 2327-2336.

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

Maddison WP, Maddison DR (2005) Mesquite: a modular system for evolutionary analysis. Version 1.06 http://mesquiteproject.org

Maskas SD, Cruzan MB (2000) Patterns of intraspecific diversification in the *Piriqueta caroliniana* complex in southeastern North America and the Bahamas. *Evolution*, **54**, 815-827.

Matos JA, Schaal BA (2000) Chloroplast evolution in the *Pinus montezumae* complex: a coalescent approach to hybridization. *Evolution*, **54**, 1218-1233.

McDaniel SF, Shaw AJ (2005) Selective sweeps and intercontinental migration in the cosmopolitan moss *Ceratodon purpureus* (Hedw.) Brid. *Molecular Ecology*, **14**, 1121-1132.

McKinnon JS, Mori S, Blackman BK, David L, Kingsley DM, Jamieson L, Chou J, Schluter D (2004) Evidence for ecology's role in speciation. *Nature*, **429**, 294-298.

Meadows ME, Baxter AJ (1999) Late Querternary paleoenvironemnts of the southwestern Cape, South Africa: a regional synthesis. *Quarternary International*, **57/58**. 193-206.

Milot E, Gibbs HL, Hobson KA (2000) Phylogeography and genetic structure of northern populations of the yellow warbler (*Dendroica petechia*). *Molecular Ecology*, **9**, 667-681.

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

Moya O, Contrera-Dias HG, Oromi P, Juan C (2004) Genetic structure, phylogoegraphy and demography of two ground-beetle species endemic to the Tenerife laurel forest (Canary Islands). *Molecular Ecology*, **13**, 3153-3167.

Nichols R (2001) Gene trees and species trees are not the same. *Trends in Ecology and Evolution*, **16**, 358-364.

Nielsen R, Wakeley J (2001) Distinguishing migration form isolation: a Markov Chain Monte Carlo approach. *Genetics*, **158**, 885-896.

Nordborg M (2001) Coalescent theory. In: Handbook of Statistical Genetics (ed. Balding JD et al.) John Wiley & Sons, London.

Orr HA (2001) The genetics of species differences. *Trends in Ecology and Evolution*, **16**, 343-350.

Oxelman B, Liden M, Berglund D (1997) Chloroplast rps16 intron phylogeny of the tribe *Sileneae* (*Carophyllaceae*). *Plant Systematics and Evolution*, **206**, 393-410.

Pearse DE, Crandall KA (2004) Beyond Fst: Analysis of population genetic data for conservation. *Conservation Genetics*, **5**, 585-602.

Petit RJ, Grivet D (2002) Optimal randomization strategies when testing the existence of a phylogeographic structure. *Genetics*, **161**, 469-471.

Pinceel J, Jordaens K, Pfenninger M, Backeljau T (2005) Rangewide phylogeography of a terrestrial slug in Europe: evidence for Alpine refugia and rapid colonization after the Pleistocene glaciations. *Molecular Ecology*, **14**, 1133-1150.

Proches S, Cowling RM (2004) Cape geophytes: putting the pieces together. In: Proceedings 10th MEDECOS Conference (ed. Arianoutsou & Papanastatis)

Provan J, Wattier RA, Maggs CA (2005) Phylogeographic analysis of the red seaweed Palmaria palmate reveals a Pleistocene marine glacial refugium in the English Channel. *Molecular Ecology*, **14**, 793-803.

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

Rawson PD, MacNamee R, Frick MG, Williams KL (2003) Phylogeography of the cornulid barnacle, *Chelonibia testudinaria*, from the loggerhead sea turtles, *Caretta caretta*. *Molecular Ecology*, **12**, 2697-2706.

Richardson JE, Weitz FM, Fay MF, Cronk QCB, Linder HP, Reeves G, Chase MW (2001) Rapid and recent origin of species richness in the Cape flora of South Africa. Natue, 412, 181-183.

Rieseberg LH (2001) Chromosomal rearrangements and speciation. *Trends in Ecology and Evolution*, **16**, 351-358.

Rieseberg LH, Burke JM (2001) A genic view of species integration. *Journal of Evolutionary Biology*, **14**, 883-886.

Rieseberg LH, Church SA, Morjan CL (2003) Integration of populations and differentiation of species. *New Phytologist*, **161**, 59-69.

Rosenberg NA, Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, **3**, 380-390.

Rosenberg NA (2003) The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. *Evolution*, **57**, 1465-1477.

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DNASP, Dna polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19**, 2496-2497.

Rundle HD, Nosil P (2005) Ecological speciation. *Ecology Letters*, **8**, 336-352.

Ruokonen M, Aarvak T, Madsen J (2005) Colonization history of the high-arctic pink-footed goose *Anser brachyrhynchus*. *Molecular Ecology*, **14,** 171-178.

Russel AL, Medellin RA, McCracken GF (2005) Genetic variation and migration in the Mexican free-tailed bat (*Tadarida brasiliensis mexicana*). *Molecular Ecology*, **14**, 2207-2222.

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

Schaal BA, Hayworth DA, Olsen KM, Rauscher JT, Smith WA (1998) Phylogeographic studies in plants: problems and prospects. *Molecular Ecology*, **7**, 465-474.

Schaal BA, Olsen KM (2000) Gene genealogies and population variation in plants. *Proceedings of the National Academy of Sciences of the USA*, **97**, 7024-7029.

Schluter D (2001) Ecology and the origin of species. *Trends in Ecology and Evolution*, **16**, 372-380.

Slatkin M, Maddison WP (1989) A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics*, **123**, 603-613.

Smith CI, Farrell BD (2005) Range expansion in the flightless longhorn cactus beetles, *Moeilema gigas* and *Moneilema armatum*, in response to Pleistocene climate changes. *Molecular Ecology*, **14**, 1025-1044.

Smith TB, Calsbeek R, Wayne RK, Holder KH, Pires D, Bardeleben C (2005) Testing alternative mechanisms of evolutionary divergence in an African rain forest passerine bird. *Journal of Evolutionary Biology*, **18**, 257-268.

Stephens M (2001) Inference under the coalescent. In: Handbook of Statistical Genetics (ed. Balding DJ *et al.*) John Wiley & Sons, London.

Templeton AR, Routman E, Phillips CA (1995) Seperating population structure from population history: A cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. *Genetics*, **140**, 767-782.

Templeton AR (1998) Nested clade analysis of phylogeographic data: testing hypotheses about gene flow and population history. *Molecular Ecology*, **7**, 381-397.

Templeton AR (2001) Using phylogeographic analyses of gene trees to test species status and processes. *Molecular Ecology*, **10**, 779-791.

Templeton AR (2004) Statistical phylogeography: methods of evaluating and minimzing enference errors. *Molecular Ecology*, **13**, 789-809.

Thompson MF (1976) Studies in the Hypoxidaceae. I. Vegetative morphology and anatomy. *Bothalia*, **12**, 111-117.

Thompson MF (1978) Studies in the Hypoxidaceae. II. Floral morphology and anatomy. *Bothalia*, **12**, 429-435.

Thompson MF (1979) Studies in the Hypoxidaceae. III. The genus *Pauridia*. *Bothalia*, **12**, 621-625.

Trewick SA, Morgan-Richards M, Russell SJ, Henderson S, Rumsey FJ, Pinter I, Barrett JA, Gibby M, Vogel JC (2002) Ploypoidy, phylogeography and Pleistocene refugia of the rockfern *Asplenium ceterach*: evidence from chloroplast DNA. *Molecular Ecology*, **11**, 2003-2012.

Turelli M, Barton NH, Coyne JA (2001) Theory and speciation. *Trends in Ecology and Evolution*, **16**, 330-343.

Wakeley J (2004) Metapopulation models for historical inference. *Molecular Ecology*, **13**, 865-875.

Wang RL, Stec A, Hey J, Luken L, Doebley J (1999) The limits of selection during maize domestcation. *Nature*, **398**, 236-239.

Wares JP, Cunningham CW (2001) Phylogeography and historical ecology of the North Atlantic intertidal. *Evolution*, **55**, 2455-2469.

Whitlock MC (2003) Dispersal and the genetic properties of metapopulations. In *Dispersal* eds. Clobert J et al., Oxford Univerity Press, UK.

Widmer A, Baltisberger M (1999) Molecular evidence for allopolyploid speciation and a single origin of the narrow endemic *Draba ladina* (Brassicaceae). *American Journal of Botany*, **86**, 1282-1289.

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

Won Y-J, Sivasundar A, Wang Y, Hey J (2005) On the origin of Lake Malawi cichlid species: A population genetic analysis of divergence. *Proceedings of the National Academy of Sciences USA*, **102**, 6581-6586.

Won Y-J, Hey J (2005) Divergence population genetics of chimpanzees. *Molecular Biology and Evolution*, **22**, 297-307.

Zeh JA, Zeh DW, Bonilla MM (2003) Phylogeography of the harlequin bettle-riding pseudoscorpion and the rise of the Isthmus of Panama. *Molecular Ecology*, **12**, 2759-2769.

Zhang D, Hewitt GM (2003) Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. *Molecular Ecology*, **12**, 563-584.