Evolution of associations between *Cymothoe* butterflies and their *Rinorea* host plants in tropical Africa

Robin van Velzen

Thesis committee

Promotor
Prof. Dr M.S.M. Sosef
Professor of Biosystematics
Wageningen University

Co-promotor
Dr F.T. Bakker
Assistant professor, Biosystematics Group
Wageningen University

Other members
Prof. Dr M. Dicke, Wageningen University
Prof. Dr S.B.J. Menken, University of Amsterdam
Prof. Dr B.J. Zwaan, Wageningen University
Dr N. Wahlberg, University of Turku, Finland

This research was conducted under the auspices of the Graduate School of Production Ecology & Resource Conservation

Evolution of associations between *Cymothoe* butterflies and their *Rinorea* host plants in tropical Africa

Robin van Velzen

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 11 December 2013
at 11 a.m. in the Aula.

Robin van Velzen
Evolution of associations between *Cymothoe* butterflies and their *Rinorea* host plants in tropical Africa
248 pages

PhD thesis, Wageningen University, Wageningen, NL (2013)
With references, with summaries in Dutch and English

ISBN 978-94-6173-778-6

To Marleen

for her love and support

To Jean-Louis Amiet

for laying the foundation for my research on Cymothoe *and* Rinorea

Contents

1 General Introduction **11**

1.1 Evolution of insect–host plant associations . . . 12
1.2 *Cymothoe* . . . 14
1.3 *Rinorea* . . . 15
1.4 *Cymothoe*-*Rinorea* host plant associations . . . 16
1.5 Methodologies & Analytics used in this thesis . . . 17
1.5.1 Field work . . . 17
1.5.2 Species identification and DNA barcoding . . . 18
1.5.3 Phylogenetics and divergence time estimation . . . 19
1.5.4 Statistics . . . 20
1.5.5 Phytochemistry . . . 22
1.6 Research Questions . . . 24
1.7 Outline of this thesis . . . 25

2 Systematics of *Cymothoe* butterflies **27**

2.1 Introduction . . . 29
2.1.1 Theoretical framework . . . 31
2.2 Materials and methods . . . 33
2.2.1 Taxon sampling . . . 33
2.2.2 Molecular methods . . . 33
2.2.3 An integrative taxonomic decision pipeline . . . 34
2.3 Results . . . 38
2.3.1 Sequencing and quality control . . . 38
2.3.2 Confirmation of current classification . . . 40
2.3.3 Delimitation of candidate species . . . 45
2.3.4 Corroboration of candidate species and revision of taxonomic classification . . . 48
2.3.5 Candidates requiring more data . . . 51

2.3.6 Rejected candidates . . . 58
2.3.7 Additional taxonomic findings . . . 63
2.3.8 Identification of immature specimens . . . 66
2.4 Discussion . . . 67
2.4.1 Integrative taxonomic decision pipeline . . . 67
2.4.2 Niger Delta endemism . . . 68
2.4.3 Technical issues . . . 68
2.4.4 Incomplete lineage sorting and introgression . . . 69
2.4.5 Future directions . . . 71
2.5 Conclusions . . . 72

3 DNA barcoding of recently diverged species 75
3.1 Background . . . 77
3.2 Materials and methods . . . 82
3.2.1 Data simulation . . . 82
3.2.2 Empirical data sets . . . 83
3.2.3 Species 'barcode gap' and monophyly . . . 84
3.2.4 Method performance . . . 84
3.2.5 Statistical tests . . . 86
3.3 Results . . . 86
3.3.1 Data simulation . . . 86
3.3.2 Species 'barcode gap' and monophyly . . . 86
3.3.3 Method performance . . . 87
3.3.4 Empirical data sets . . . 92
3.4 Discussion . . . 93
3.4.1 Method performance . . . 93
3.5 Conclusions . . . 97

4 Species diversification in *Cymothoe* tropical forest butterflies 99
4.1 Background . . . 101
4.2 Materials and methods . . . 103
4.2.1 Taxon sampling . . . 103
4.2.2 Molecular methods . . . 103
4.2.3 Phylogenetic inference . . . 108
4.2.4 Timing of divergences . . . 109
4.2.5 Diversification analyses . . . 110

4.3 Results 113
4.3.1 Phylogenetic inference 113
4.3.2 Timing of divergences 116
4.3.3 Diversification analyses 116
4.4 Discussion 119
4.4.1 Diversification and host association 120
4.4.2 Diversification and climate change 122
4.4.3 Speciation versus extinction 123
4.4.4 Diversification rate versus phylogenetic resolution 123
4.5 Conclusions 124

5 Phylogenetics and historical biogeography of *Rinorea* 125
5.1 Background 127
5.2 Materials & methods 129
5.2.1 Phylogenetic analysis 138
5.2.2 Divergence time estimation 139
5.3 Results 141
5.3.1 Phylogenetic results 142
5.3.2 Divergence timing results 142
5.4 Discussion 143
5.4.1 Trans-Atlantic dispersal 149
5.4.2 Dispersal to Asia 151
5.4.3 Multiple independent colonizations of Madagascar 151
5.5 Conclusions 152

6 Asynchronous divergence of *Cymothoe* forest butterflies and their *Rinorea* host plants in tropical Africa 155
6.1 Background 157
6.1.1 Host shifts 158
6.1.2 Synchronicity 159
6.1.3 Species as fundamental units 160
6.1.4 *Cymothoe* butterflies and their *Rinorea* host plants 160
6.1.5 Objectives 162
6.2 Materials and methods 162
6.2.1 Sampling 162
6.2.2 Phylogeny estimation and divergence timing 164

6.2.3 Cophylogeny estimation 167
6.3 Results 172
6.3.1 Host plant observations 172
6.3.2 Phylogeny estimation and distribution of associations 173
6.3.3 Comparative dating 176
6.3.4 Comparison of pairwise phylogenetic distances 176
6.3.5 Cophylogenetic analyses 177
6.4 Discussion 178
6.4.1 An evolutionary scenario for *Cymothoe* host plant associations 181
6.5 Conclusion 184

7 General discussion **185**
7.1 Summarizing results 187
7.2 A chronology of reconstructed historical events in *Cymothoe* and *Rinorea* 190
7.3 Unruliness of rapid radiations 192
7.4 The end of cospeciation theory? 194
7.5 Chemical ecology 196
7.6 *Wolbachia* and insect speciation 198
7.7 Herbivorous insects and food security 198
7.8 A plea for integration of taxonomic data 200

References **203**

Summary **233**

Samenvatting **237**

Acknowledgements **241**

Publications **244**

Curriculum Vitae **245**

Education statement **246**

Chapter 1

General Introduction

1.1 Evolution of insect–host plant associations

Insects are by far the most diverse group of multicellular organisms on earth (Ødegaard 2000, Price 2002). Most insect species are herbivores and plants constitute the vast majority of terrestrial biomass. Understanding the evolution of interactions between herbivorous insects and their host plants is therefore crucial to comprehending global patterns in terrestrial biodiversity (Mitter, Farrell *et al.* 1988, Farrell and Mitter 1998, Price 2002, Lewinsohn and Roslin 2008, Futuyma and Agrawal 2009, Novotny, Miller *et al.* 2010).

One important observation is that herbivorous insects are generally highly host specific (Futuyma and Moreno 1988). They may restrict feeding to particular clades, species or plant tissues (Weiblen, Webb *et al.* 2006). This is probably due to the diversity of mechanisms with which plants defend themselves against herbivore attacks (Coley and Barone 1996). In the context of this wide variety of chemical defences, specialization is one way to overcome at least some of these barriers. Phylogenetically, one therefore might expect generalist → specialist transitions to be more common and specialist clades to be more species rich compared with their non-herbivore sister groups. This has indeed be found in various insect clades (Kelley and Farrell 1998, Janz, Nylin *et al.* 2006). In addition, host plant-specific toxins may be sequestered for the defence of the insect, reducing mortality from natural enemies (Nishida 2002).

Most phylogenetic studies show that related insect species tend to feed on related plant species (Mitter, Farrell *et al.* 1991, Janz and Nylin 1998, Ronquist and Liljeblad 2001, Braby and Trueman 2006, Lopez-Vaamonde, Wikström *et al.* 2006, Wilson, Forister *et al.* 2012). This pattern can be explained by two alternative processes. First, in accordance with Fahrenholz' rule that parasites follow the speciation events of their hosts (Fahrenholz 1913), if the association between two species is very close, they may speciate in parallel. Such concomitant occurrence of speciation in hosts and their parasites is referred to as 'cospeciation' (Page 2003), the associations between pocket gophers and their chewing lice being a well-known example (Hafner and Nadler 1988). Highly specialized insect herbivores can be regarded as parasites and may therefore also speciate in parallel with their host plants (Farrell and Mitter 1990, Farrell and Mitter 1998). The assumption is that when host plant populations become isolated, so will the associated specialist insect herbivores (Janz 2011). A cospeciation process therefore predicts that the divergence times of the insects and associated herbivores are synchronous. I note that this process is sometimes regarded as a special case of 'coevolution'. However, although this term has been used to label a wide variety of different processes (Janzen 1980), in a narrow sense it usually assumes reciprocal selection pressures and resulting micro-evolutionary changes in the two actors (here: insects and plants) (Smith, Godsoe *et al.* 2008, de Vienne, Refrégier *et al.* 2013). As this cannot always be clearly demonstrated, I therefore prefer the term cospeciation, which is more specific

and indicates concomicant speciation events of the actors without assuming any reciprocal selection. Evidence for insect host plant cospeciation has been found in for example *Phyllobrotica* leaf beetles feeding on Lamiales (Farrell and Mitter 1990), *Tetraopes* longhorn beetles feeding on *Asclepias* (Farrell and Mitter 1998) and *Blepharida* beetles feeding on Burseraceae (Becerra 2003).

Alternatively, insect - host plant associations may be shaped by host shifts, either by substituting one host species for another or by incorporating additional host species in their diet. In general, host shifts can be expected to occur most often between related plant species because they are likely to share similar secondary phytochemistry (Ehrlich and Raven 1964, Menken 1996, Becerra 1997, Agosta 2006, Agrawal 2007). Such phylogenetic constraint in host shifts can be expected to generate a pattern of related insects feeding on related plants. Host shifts assume divergence times of insect species to postdate those of their associated host plants. Evidence for predominance of host shifts over cospeciation has been found for example in leaf-mining sawflies (Leppänen, Altenhofer *et al.* 2012), flies (Brändle, Knoll *et al.* 2005), and moths (Lopez-Vaamonde, Wikström *et al.* 2006).

Whether common or rare, it is unclear whether host shifts are an important driver for diversification, and if so, how. At higher taxonomic levels, a shift to a plant lineage distant from the ancestral host(s) could promote species diversification in herbivorous insects because it allows entering a new ‘empty’ niche (Simpson 1953, Mayr 1963, Ehrlich and Raven 1964, Braby and Trueman 2006, Janz, Nylin *et al.* 2006, Wheat, Vogel *et al.* 2007, Winkler, Mitter *et al.* 2009). Indeed, correlations between increased rates of species diversification and host plant shifts have been found in various butterfly clades (Weingartner, Wahlberg *et al.* 2006, Fordyce 2010).

At the insect species level, after a host shift has occurred, divergent selection on herbivores could ultimately lead to their reproductive isolation and, hence, ecological speciation (Emelianov, Dres *et al.* 2001, Berlocher and Feder 2002, Drès and Mallet 2002, Nosil, Crespi *et al.* 2002, Stireman, Nason *et al.* 2005, Singer and McBride 2010). When insect populations are sympatric, divergent selection can only occur when reproduction occurs preferentially between mates that use the same host use. The basis for such assortative mating can be positional (when mating occurs on the host), time-dependent (when mating times diverge between individuals on different hosts), or chemical (when host-specific phytochemicals are sequestered by the herbivore and used as mate recognition signals) (Drès and Mallet 2002). However, for most insect herbivores mating is probably not tightly coupled with host plant use. In addition, some clades of herbivorous insects appear to have diversified without any evidence for any ecological divergence (e.g. Imada, Kawakita *et al.* 2011). Therefore, it remains controversial whether such processes are common enough to provide an explanation for the high levels of insect diversity (Nyman, Vikberg *et al.* 2010).

Alternatively, speciation rates can be influenced by environmental factors such as geological events (Hall 2005, Mallarino, Bermingham *et al.* 2005, Wahlberg and Freitas 2007, Casner and Pyrcz 2010) and changing climate (Peña and Wahlberg 2008,

Aduse-Poku, Vingerhoedt *et al.* 2009, Müller and Beheregaray 2010, Condamine, Sperling *et al.* 2012). When such factors generate persistent barriers to gene flow (e.g. through formation of mountain ranges or climate-driven habitat fragmentation) populations may genetically diverge, due to either adaptive divergence or genetic drift and, ultimately, become reproductively isolated (allopatric speciation). Such processes are assumed to be common and allopatric speciation is consequently often considered the dominant mode of diversification for sexual organisms (Mayr 1963, Fitzpatrick, Fordyce *et al.* 2009, Butlin, Debelle *et al.* 2012). Allopatric speciation is therefore a logical null hypothesis when examining mechanisms of speciation.

Much of the Earth's biodiversity is concentrated in tropical forests (McKenna and Farrell 2006), and the largest clades are tropical specialist insect herbivores (Duffey and Stout 1996). Consequently, answers to questions about processes shaping global species diversity should preferably be sought in tropical systems of specialist insect herbivores and their host plants. The subject of this thesis is the evolution of trophic interactions between *Cymothoe* butterflies (Nymphalidae, Limenitidinae) and their *Rinorea* host plants (Violaceae) in tropical Africa. More specifically, it addresses the phylogenetic patterns of insect - host plant associations and whether host shifts may have driven herbivore species diversification.

1.2 *Cymothoe*

The genus *Cymothoe* comprises roughly 78 butterfly species that are confined to the forested regions of tropical Africa and Madagascar (Ackery, Smith *et al.* 1995, Larsen 2005, Williams 2012). The species are visually attractive and show a high degree of sexual dimorphism, with males exhibiting sometimes spectacular coloration. Thus, *Cymothoe* is regularly considered a 'flagship' for African butterflies, a status attested by its prominence on the cover of textbooks on the subject (d'Abrera 2004, Larsen 2005, Vande weghe 2010).

Nevertheless, despite over a hundred year of taxonomic endeavors, many species groups and sections within *Cymothoe* remain taxonomically difficult. This is in part due to the fact that either males or females of different species can be morphologically highly similar. For example, males in section *Sangaris* are morphologically indistinguishable while females are highly variable (Berger 1981, Larsen 2005), and, while males of *C. ogova* and *C. harmilla* can be easily distinguished from each other, their females have been confused for a long time (Gompert, Forister *et al.* 2008). The last taxonomic monograph on *Cymothoe* included only 39 of the 78 currently recognized species and was written by Overlaet (1952) who, shortly before his death in 1956, commented to Lucien Berger that at least for some sections "C'est tout encore à refaire" [it all needs to be redone] (Berger 1981). Since then, nobody dared to take up this task. In summary, it is clear that traditional lines of evidence such as morphology and biogeography alone have failed to provide decisive evidence for a reliable and stable classification of *Cymothoe*.

Harma constitutes the monotypic sister genus to *Cymothoe*. With respect to most ecological and morphological traits, species of *Cymothoe* and *Harma* are highly similar: they are forest butterflies, frugivorous and sexually dimorphic. This suggests a high degree of niche conservatism and indeed multiple species can usually be found living together at the same locality (Amiet and Achoundong 1996, Larsen 2005). Nevertheless, *Harma* and *Cymothoe* differ markedly in their species diversity (a single *Harma theobene* versus 78 *Cymothoe* species), suggesting a markedly differential species diversification between these sister lineages.

At the start of the present project, there was only one molecular phylogenetic analysis including *Harma* and *Cymothoe*. In their paper describing the phylogenetic relationships of 400 genera of Nymphalidae, Wahlberg *et al.* (2009) included *Harma theobene* and *Cymothoe caenis*, which appeared as sisters on an early diverging branch within Limenitidinae. However, with only a single *Cymothoe*, their sampling does not allow making inferences about the evolution within the *Harma* - *Cymothoe* clade.

1.3 *Rinorea*

Rinorea is a pantropical genus of shrubs and small trees within the family Violaceae, order of Malpighiales. Africa accommodates the largest number of species (110 to 150 spp.; Achoundong 2000); Cameroon and Gabon being particularly species-rich (55 and 49 species, respectively; Achoundong 1996, Bakker, van Gemerden *et al.* 2006, Sosef, Wieringa *et al.* 2006). African *Rinorea* are often abundant or even dominant in the understory of humid or semi-deciduous forests (Achoundong 1996, Kenfack, Thomas *et al.* 2007, Chuyong, Kenfack *et al.* 2011), possibly constituting a reliable resource for herbivorous insects. In addition, as *Rinorea* species are usually restricted to specific environments (Achoundong 1996, Achoundong 2000, Adomou, Sinsin *et al.* 2006, Mwavu and Witkowski 2009, Tchouto, de Wilde *et al.* 2009, Djuikouo, Doucet *et al.* 2010), African *Rinorea* are considered useful bioindicators for forest typification (Achoundong 1996, Achoundong 2000, Tchouto, de Wilde *et al.* 2009). Despite their importance, as is the case for *Cymothoe*, species identification of African *Rinorea* is difficult and specimens are regularly unidentified in ecological (e.g. Kenfack, Thomas *et al.* 2007, Tchouto, de Wilde *et al.* 2009) and taxonomic studies (e.g. Robson 1960, Hawthorne and Jongkind 2006, Sosef, Wieringa *et al.* 2006). The most recent overview of all African *Rinorea* species was published in 1914 (Brandt), and confident assignment of species to the recognized infrageneric groups is often difficult (Bos 1989, Wahlert and Ballard 2012). Although Dowsett-Lemaire & White (1990) stated that "*Rinorea* badly needs a critical pan-African revision" this has not yet been achieved today, however, rendering *Rinorea* an urgent case taxonomically.

Phylogenetic analyses have shown that *Rinorea* is a relatively early diverging lineage within Violaceae (Tokuoka 2008, Wurdack and Davis 2009). Within *Rinorea*,

studies based on plastid data have shown that there is an early split between a Neotropical and a Palaeotropical clade. Within the Palaeotropical clade, African *Rinorea* (with the inclusion of some closely related Malagasy taxa) are monophyletic (Wahlert and Ballard 2012). At least 4 Malagasy taxa appear within separate clades from mainland Africa, suggesting independent dispersals from mainland Africa to Madagascar (Bakker, van Gemerden *et al.* 2006, Wahlert and Ballard 2012). In spite of the fact that these studies have collectively progressed our understanding of African *Rinorea* systematics, several issues remain: i) the lack of a nuclear DNA based phylogenetic perspective allowing reconstruction of actual clades instead of plastid haplotypes, and ii) some idea of the absolute time frame in which diversification in *Rinorea* has taken place.

1.4 *Cymothoe-Rinorea* host plant associations

Besides their visual attractiveness, *Cymothoe* are of special interest due to their highly specialized host plant associations. Based on a decade of field observations of *Cymothoe* oviposition behavior in Cameroon, Amiet & Achoundong (1996) found that within *Cymothoe* 27 species (out of 44 species in Cameroon) are associated with species of *Rinorea*. These associations showed a high degree of trophic specialization: 18 species appeared monophagous (i.e. feeding on a single species), the other 9 stenophagous (feeding on 2–6 species of *Rinorea*) (Fontaine 1982, Amiet and Achoundong 1996, McBride, van Velzen *et al.* 2009). This species-level specificity was confirmed by larval choice assays showing that, as soon as they hatch, *Cymothoe* larvae are able to recognise their *Rinorea* host plant (Amiet and Achoundong 1996). Another 13 species of *Cymothoe* feed exclusively on species of Achariaceae (also Malpighiales), which are also host to *Harma* (van Son 1979, Kielland 1990, Larsen 1991, Pringle, Henning *et al.* 1994, Amiet and Achoundong 1996), suggesting that Achariaceae are the ancestral host plant group.

The above renders the *Cymothoe* host plant system highly suitable to investigate processes shaping evolutionary patterns of host plant associations and species diversity, more specifically because of the following reasons. First, the large number of related species involved (34 herbivores and 33 hosts) allows reconstructing the evolution of a high number of associations at the species-level (compare for instance with the *Pieris-Brassica* model system), as well as to quantify host range in terms of associated species instead of arbitrary higher taxa. This enables a higher level of precision for assessment of specialization (Kaartinen, Stone *et al.* 2010), phylogenetic constraints in the evolution of host plant associations, and diversification (Funk, Filchak *et al.* 2002, Rabosky, Slater *et al.* 2012).

If we would assume that *Cymothoe* have remained associated with the same *Rinorea* species over macroevolutionary time-scales, it seems plausible that they would diverge in concert, as observed in the high level of specificity typical for

specialized ectoparasites that usually show a high degree of cospeciation (e.g. Hafner and Nadler 1988, Hughes, Kennedy *et al.* 2007). On the other hand, current monophagy obviously does not rule out a more dynamic history of associations.

Secondly, the observed dichotomy in host plant use – either Achariaceae or *Rinorea* (Violaceae) – allows the reconstruction of host shifts to distantly related plants and, hence, addressing the question whether such shifts promoted species diversification. It is not known if *Rinorea* was colonized once or multiple times independently. In any case, when assuming that shifts to distantly related hosts signify entrance into new niches, it is possible that the shift(s) from Achariaceae to *Rinorea* explain the elevated diversification rate of *Cymothoe* compared with that of *Harma*. Alternatively, diversification in *Cymothoe* may be simply due to allopatric speciation. Over geological time scales, Africa has experienced large fluctuations in climatic conditions (Coetzee 1993, Jacobs 2004, Segalen, Lee-Thorp *et al.* 2007), resulting in cycles of fragmentation and expansion of the areas occupied by rain forest (Dupont, Jahns *et al.* 2000, Cohen, Stone *et al.* 2007, Dupont 2011). Given the apparent niche conservatism of *Cymothoe* with respect to forest habitats (Larsen 2005), and given the approximate age of Limenitidinae at around 57 My (Wahlberg, Leneveu *et al.* 2009), climate-driven habitat fragmentation may therefore had a major influence on *Cymothoe* diversification (see Chapter 3). At what time scale this would have happened, i.e. whether at a Miocene rather than Pleistocene, should be discernable based on the amount of DNA sequence divergence observed among *Cymothoe* species.

1.5 Methodologies & Analytics used in this thesis

1.5.1 Field work

Various studies reported differences between the potential or even 'ideal' host plant range under laboratory conditions and the realised range in the field (e.g. Janz 2005, Forister, Nice et al. 2009). This tells us that, although allowing better control and replication, laboratory experiments may be a poor proxy for species interactions in ecosystems. Consequently, field observations are imperative for any evolutionary study on insect - host plant interactions. In addition, an insect's repertoire may include additional host plants that have not been recorded simply because of insufficient field work (Jermy and Szentesi 2003). Because nearly all current data on *Cymothoe* - *Rinorea* associations were recorded in Cameroon (Amiet and Achoundong, 1996), it is not clear whether they would have covered possible geographic variability in host use among *Cymothoe* species. Therefore, during this project, additional host plant associations were recorded in the field in Nigeria, Ghana, Gabon and Kenia; representing most of the geographical range of *Cymothoe*. Sometimes this concerned an observation of oviposition by females, but more often

finding eggs or caterpillars on a host plant. Such immature stages can be recognised as *Cymothoe*, but morphological species identification is impossible for eggs and difficult for most caterpillars.

1.5.2 Species identification and DNA barcoding

Obviously, when plant and insect species identifications are incorrect, conclusions that are drawn based on them will be also incorrect. As was outlined above, both *Cymothoe* and *Rinorea* are taxonomically difficult groups, complicating the task of generating reliable identifications. In addition, identification of immature stages is problematic and up to now generally requires rearing eggs or caterpillars to adulthood (Amiet and Achoundong 1996). Rearing is a time-consuming and laborious process, further complicated by a high incidence of parasitism in *Cymothoe* material. Likewise, an individual host plant may lack flowers or fruits that usually serve as species-diagnostic characters. Indeed, *Cymothoe* often select sterile shoots of young treelets that aretherefore nearly impossible to ID based on morphology. DNA sequence-based identifications provide a good solution for these problems, as they allow comparison with those from identified adult or flowering specimens, providing accurate identification of immature stages such as eggs or caterpillars and sterile plant specimens. Obviously, identification also requires an accurate taxonomic classification.

In the last decade, the use of standardized loci for species identification (Hebert, Cywinska *et al.* 2003) has proven instrumental in species identification, especially in insects and vertebrates (Hebert, Stoeckle et al. 2004, Hajibabaei, Janzen et al. 2006, Ward, Hanner et al. 2009). Recently, DNA barcodes have proven to be effective in discovering cryptic species as well as to resolve taxonomic issues within selected *Cymothoe* lineages (van Velzen, Bakker *et al.* 2007, McBride, van Velzen *et al.* 2009, van Velzen, Larsen *et al.* 2009). Here, I extend my scope to genus-wide sampling with the aim to improve species delimitation and assist identification of immature stages in *Cymothoe*.Identification based on DNA barcodes has been reported to be problematic in some cases where species are assumed to have diverged only recently (Wallman and Donnellan 2001, Meyer and Paulay 2005, Kaila and Stahls 2006, Dexter, Pennington *et al.* 2010, Lou and Golding 2010, Yassin, Markow *et al.* 2010). Problems are mostly due to incomplete lineage sorting and probably related to large effective population size and/or low mutation rate (Meyer and Paulay 2005, Elias, Hill *et al.* 2007, Wiemers and Fiedler 2007, Ross, Murugan *et al.* 2008, Austerlitz, David *et al.* 2009, Hollingsworth, Graham et al. 2011, McFadden, Benayahu *et al.* 2011. On the other (analytical) hand, alternative methods for matching DNA barcodes to reference libraries have been published (e.g. DasGupta, Konwar *et al.* 2005, Meier, Shiyang *et al.* 2006, Munch, Boomsma *et al.* 2008, Bertolazzi, Felici *et al.* 2009, Little 2011), but systematic comparisons of the relative performance of these methods are few.

1.5.3 Phylogenetics and divergence time estimation

Answering questions about the evolution of insect - host plant associations requires detailed phylogenetic information about both the insects and their hosts. For example, to address whether current host associations are shaped through processes that are phylogenetically constrained, we need to know which species are phylogenetically related. Likewise, assessing whether parasites have cospeciated with their host, or rather whether they have shifted betweenthem, requires careful reconciliation of host and parasite phylogenetic trees. In this case, phylogenetic congruence is taken as evidence supporting a cospeciation hypothesis, whereas incongruence would favor host shifts (de Vienne, Refrégier *et al.* 2013). For *Cymothoe*, no phylogenetic hypothesis based on DNA sequence data is available so far. For *Rinorea*, previous phylogenetic studies relied on plastid data only and did not include most of the species that act as host for *Cymothoe* (Bakker & al. 2006). We therefore aimed at generating molecular phylogenetic trees for *Cymothoe*, as well as for (African) *Rinorea*, based on organellar and nuclear DNA sequences.

Divergence time estimates play an important role in this thesis, because they enable testing explicit hypotheses. First, an estimate of time is required to quantify species diversification rates. Assuming a constant rate, the number of species will accumulate exponentially over time. Because measures of diversification are usually based on extant species only, diversification estimates are the result of the rate of speciation minus the rate of extinction. Without fossil data (which are unavailable for *Cymothoe* or *Rinorea*) separate estimation of speciation and extinction is problematic (Rabosky 2010). Therefore, we report net diversification rates throughout this thesis, noting that (i) it is uncertain whether a current rate is primarily shaped by speciation or by extinction (Rabosky 2010), and (ii) lineages that are currently species poor may have been highly diverse in the past (Crisp and Cook 2005). Secondly, distinguishing among various historical scenarios for insect - host plant evolution also requires time estimates (Percy, Page *et al.* 2004, Sorenson, Balakrishnan *et al.* 2004, De Vienne, Giraud *et al.* 2007). For example, to determine whether insects and their hosts diverged in synchrony or sequentially, i.e. insects colonized already existing plant clades, or to correlate different patterns from e.g. the fossil record and paleoclimatic reconstructions. Studies of insect - host plant evolution based on accurately dated phylogenies are few, however, and relative timing of divergence in insects and their associated hosts is controversial (Wheat, Vogel *et al.* 2007, Nyman 2010, Kergoat, Le Ru *et al.* 2011). Finally, divergence time estimations are also important for testing historical biogeographic hypotheses. For instance, the disjunction between Neo- and Palaeotropical *Rinorea* may either be explained by a Gondwanan vicariance or by a more recent long-distance dispersal event (Queiroz, 2005). Likewise, it remains unknown if the putative independent dispersals from Africa to Madagascar were synchronous. Within *Cymothoe*, species divergences may be correlated with climatic events such as global cooling or with mountain uplift.

For a long time, divergence time estimations have been based on the assumption that the rate of evolutionary change is approximately constant over time and over different lineages (molecular clock; Zuckerkandl and Pauling 1962). In reality, however, many data sets show considerable departures from clocklike evolution (Britten 1986) and rate variation among lineages can seriously mislead divergence date estimation (Yoder and Yang 2000). In the past decade, increasingly sophisticated methods have been developed to relax the molecular clock assumption and allow rates to vary over lineages. For example, by estimating different rates for specific lineages (relaxed local clocks; Federal Ministry of Environment, United Nations Development Programme *et al.* 2011), or by modelling rates among lineages as varying in an autocorrelated manner within a Bayesian statistical framework (Bayesian relaxed-clock; Aris-Brosou and Yang 2002). More recently, methodological innovations implemented in the software package BEAST (Drummond and Rambaut 2007) have enabled simultaneous estimation of phylogeny and lineage-specific rates using the Bayesian relaxed-clock (Phil-Eze and Okoro 2009).

In principle, divergence dating analyses estimate relative rather than absolute ages. In order to estimate absolute divergences, calibrations are of crucial importance. Ideally, calibrations are based on the fossil record or dated biogeographic events for which there is a priori knowledge about how these data correspond with lineages in the tree (Funk and Omland 2003). Fossils of *Cymothoe* or *Rinorea* are unknown, however, and generally one should in my opinion aim to reconstruct (rather than assume) biogeographical events. I therefore use known fossil-based divergence estimates of related lineages (Wahlberg, Leneveu *et al.* 2009, Bell, Soltis *et al.* 2010) to (secondarily) calibrate phylogenetic trees using BEAST. To avoid overestimation of precision I carefully transfer uncertainty in the original estimates to the secondary calibrations.

1.5.4 Statistics

The study of macroevolution is inherently difficult, mainly because experimental approaches to validate macro-evolutionary hypotheses (both pattern and process) are logically impossible; all we are left with is corroborating competing hypotheses using as much as evidence as possible (Popper 1963). We therefore aim to ensure that our conclusions are firmly supported by empirical evidence, in two principal ways. First, at various levels we accommodate uncertainty in our analyses as much as possible, in order to avoid over-interpretation. For example, our Bayesian estimations include posterior probabilities of clades as well as confidence intervals for divergence times and rates of evolution. We also developed a novel method to account for unsampled species in our diversification analyses (see Chapter 3). Secondly, we apply statistical tests to answer our research questions, wherever possible. Ideally, statistical tests should be tailored to the particular question at hand. Researchers

are usually dependent on the availability of methods and software implementations, however. Because different software programs may be (and often are!) incompatible with each other, and may be unable to handle the available data formats, statistical testing is often problematic. Fortunately, in the last years, the open source R software environment for statistical computing has become a prominent tool. The R environment provides a coherent system, rather than a collection of very specific and inflexible tools, as is frequently the case in data analysis software. It can easily be extended via packages, which are available for a plethora of statistical tests. By consequence, these tests are compatible and can be applied to the same data structure (i.e. within R). Because of the open source nature of R and its packages, users can even tailor tests to their specific needs. In addition, because it is a computer language, R facilitates handling of large data sets. For example, in order to compare relative performance of different DNA barcode matching methods, we develop an R script that aggregates the output from different software programs, assesses relevant statistics, and formally tests systematic differences in matching performance (see Chapter 2). For assessing whether related *Cymothoe* feed on related *Rinorea* hosts, we modify the R script implementing a permutation test described by Hommola *et al.* (2009) so that it checks the concistency of the input data and returns correlation coefficients (see Chapter 5).

Simulations also play an important part in this thesis. Simulation is a way to model random events, such that simulated outcomes closely resemble those in the real world, generally allowing better understanding of system dynamics. One advantage of modelling is that the user knows and controls the ‘truth’. For example, in order to assess whether different methods accurately match DNA barcodes from recently diverged species to those in reference libraries, we apply these methods to simulated DNA barcode data. Such simulations are based on a population genetic model that allows for incomplete lineage sorting, applied to species with known divergence times. By consequence, for each DNA barcode we know what species it belongs to, and whether that species is recently diverged or not, thereby permitting unambiguous assessment of matching accuracy of the different methods (chapter 2). Simulations are also useful in generating null distributions for significance testing. For instance, an observed correlation between insect and host phylogenetic trees can be due to their topology rather than to the actual host plant associations (e.g. when the trees are smaller, their topologies have a higher chance of being congruent). Therefore, in order to assess whether the observed correlation coefficient is significant, it is compared with a null distribution of coefficients based on the same data but with randomized associations. The observed coefficient is then considered significant only if it falls outside that null distribution (Legendre, Desdevises *et al.* 2002, Hommola, Smith *et al.* 2009; see Chapter 5). A similar approach is taken in statistical tests for shifts in diversification rates and for conflict between gene partitions (Nichols 2001, Ross, Murugan *et al.* 2008; see Chapter 3).

1.5.5 Phytochemistry

As outlined before, insect - host plant associations are most likely mediated through plant phytochemistry, as females usually locate suitable oviposition sites based on specific host plant chemical cues (Renwick and Chew 1994, Honda 1995, del Campo, Miles *et al.* 2001, Nishida 2005, Schoonhoven, van Loon *et al.* 2005). Likewise, insect larvae must be physiologically adapted to the particular phytochemicals present in their diet (Berenbaum and Zangerl 1998). Host shifts therefore require adaptations to different chemicals at various levels, and it is expected that host phytochemistry is an important factor in the evolution of host plant associations. In addition, the notion of chemical signalling has become paramount in current evolutionary ecological research (Vet and Dicke 1992, Dicke 2000, Dicke and van Loon 2000, Allmann, Spathe *et al.* 2013). Ideally, we would therefore incorporate phytochemical data in this thesis.

The ancestral *Cymothoe* host plant family Achariaceae is known to be cyanogenic due to the presence of cyclopentanoid glycosides that act as precursor for cyanide production. These glycosides have been found in all biochemically examined Achariaceae genera reported as *Cymothoe* host plants: *Caloncoba* (Cramer, Rehfeldt *et al.* 1980), *Kiggelaria* (Raubenheimer and Elsworth 1988), *Lindackeria* (Jaroszewski, Ekpe *et al.* 2004), and *Rawsonia* (Andersen, Clausen *et al.* 2001). Outside Achariaceae, cyclopentanoid glycosides were originally only known to occur in the closely related families Passifloraceae and Turneraceae (Spencer and Seigler 1985, Clausen, Frydenvang *et al.* 2002). Achariaceae, Passifloraceae and Turneraceae are early diverging lineages in the Parietal clade within the Malpighiales, together with Violaceae (Xi, Ruhfel *et al.* 2012), suggesting that Violaceae, although non-cyanogenic, could be phytochemically similar. In addition, *Acraea* butterflies (Nymphalidae, Heliconiinae) are known to feed on *Rinorea* as well as on Achariaceae and Passifloraceae (van Someren 1974, Ackery 1988). It may therefore be expected that the host plant associations of *Cymothoe* and *Acraea* that both feed on *Rinorea* as well as Achariaceae have a common chemical basis. Following this hypothesis, Clausen *et al.* (2002) discovered a novel cyclopentanoid glycoside in the non cyanogenic *Rinorea ilicifolia*. This led them to hypothesize that cyclopentanoid glycosides are the common chemical basis for the host plant associations of *Cymothoe* (and *Acraea*) butterflies (Clausen, Frydenvang *et al.* 2002).

Experimental confirmation of the role of cyclopentanoid glycosides is a logical next step. We have attempted to perform choice assays to assess whether cyclopentanoid glycosides applied to artificial substrates act as an oviposition stimulant in female *Cymothoe egesta* (a species that oviposits on *Rinorea ilicifolia* in the wild). This appeared challenging mainly because females would not accept artificial oviposition substrates. We also compared cyclopentanoid glycoside levels in *Rinorea ilicifolia* leaves that were accepted and rejected by ovipositing females and found no significant differences. To assess whether a different metabolite caused the

differential oviposition we performed an untargeted comparison of metabolites in *Rinorea ilicifolia* leaves that were accepted and rejectedby *Cymothoe egesta* females in the field. This comparison highlighted many differences, but these are most likely due to the fact that ovipositing females generally accept young leaves and reject old ones. Given these practicalities and the time constraint of a PhD project, this thesis does not further pursue unraveling of the phytochemical basis for oviposition in *Cymothoe*.

1.6 Research Questions

Based on what is outlined above, I identify nine specific research questions that will be addressed in this thesis. They are divided into three groups based on whether they mainly apply to *Cymothoe*, *Rinorea*, or specifically focus on the associations between the two:

Cymothoe

Rq1. Do DNA barcodes facilitate species delimitation and identification in *Cymothoe*?

Rq2. What are the best matching-methods when using DNA barcodes from recently diverged species?

Rq3. Does the variation in net diversification rate between *Cymothoe* and *Harma* correlate with host shift(s) from Achariaceae to *Rinorea* or rather with (presumed) climate-driven habitat fragmentation?

Rinorea

Rq4. Does a nuclear genomic perspective confirm previous classifications for *Rinorea* based on plastid DNA?

Rq5. Does the split between Neo- and Palaeotropical *Rinorea* result from Gondwanan vicariance or from transatlantic dispersal?

Rq6. Were independent dispersals of *Rinorea* from Africa to Madagascar synchronous?

Associations

Rq7. Did *Cymothoe* lineages diverge in synchrony with their hosts, or rather colonize already-existing clades within African *Rinorea*?

Rq8. Do related *Cymothoe* feed on related *Rinorea*?

Rq9. What are the levels of congruence (i.e. supporting cospeciation) and incongruence (i.e. supporting host switches) between *Cymothoe* and *Rinorea* phylogenetic trees?

1.7 Outline of this thesis

In Chapter two I present an extensive dataset of 1204 DNA *Cymothoe* barcode sequences and assess whether they facilitate species delimitation and identification in *Cymothoe* (Rq1) within an integrative taxonomic approach. Although such an approach is advocated by many, it is often unclear what the approach entails, exactly. We therefore designed a practical decision pipeline for integrating DNA barcodes, morphology and biogeography within a taxonomic framework.

In Chapter three I compare the relative performance of six matching methods in their ability to correctly identify recently diverged species with DNA barcodes (Rq2). We analyze simulated data assuming different effective population sizes as well as selected published empirical data sets. In general, data simulations allow for replication and, hence, statistical testing of method performance.

Chapter four deals with one of the central questions in this thesis, namely species diversification. In this Chapter the first species-level molecular phylogenetic tree of *Cymothoe* and *Harma* is presented and discussed. To test whether rates of net species diversification are best explained by shifts to novel host plants or by palaeoclimatic factors (Rq3) we calibrate the tree within an absolute time-frame, identify significant shifts in species diversification rates, and assess correlations of estimated diversification with reconstructed host plant associations and with trends in global temperature variation at geological timescales.

In Chapter five, focus shifts from *Cymothoe* to *Rinorea*. It describes the updated phylogeny of *Rinorea* with increased taxonomic sampling, using plastid as well as nuclear DNA sequences (Rq4). To answer historical biogeographic questions about possible Gondwanan origin (Rq5) and synchronicity of independent dispersals from Africa to Madagascar (Rq6), we estimate lineage divergence within an absolute time-frame.

Chapter six aims to distinguish between alternative scenarios for the evolution of insect-host plant associations. It presents an integration of the time-calibrated phylogenetic evidence from *Cymothoe* and *Rinorea* presented in Chapters four and five with updated host association records from the field. To gain insight into the degree of (a)synchronicity in butterfly and host-plant diversification (Rq7), we compare divergence time estimates for associated clades. To see if closely related herbivores use closely related hosts (Rq8) we compare pairwise phylogenetic distances among *Cymothoe* with those among *Rinorea* species. In order to assess the relative contribution of cospeciation versus host shifts to butterfly speciation (Rq9), we reconstruct event-based historical scenarios of associations using phylogenetic reconciliation methods.

Finally, in Chapter seven, I address answers to research questions posed in this thesis and discuss their implications to other scientific fields such as biological regulation and agriculture. Challenges and possibilities for future research are identified and the conclusions of the current research highlighted.

Chapter 2

Systematics of *Cymothoe* butterflies (Nymphalidae, Limenitidinae)

Integrating DNA barcodes in a formal taxonomic decision pipeline

Robin van Velzen[1,2], Philippe Oremans[3], Thierry Bouyer[4], Dieuwko Knoop[5], Gael Vande Weghe[6], Steve Collins[7], Marc SM Sosef[1,2] & Freek T Bakker[1]

[1] Biosystematics Group, Department of Plant Sciences, Wageningen University and Research Centre, Wageningen, The Netherlands
[2] Naturalis Biodiversity Centre (Section NHN), Wageningen University, Generaal Foulkesweg 37, 6703 BL Wageningen, The Netherlands
[3] 35 rue des Jacinthes, 6110 Montigny-le-Tilleul, Belgium
[4] 57 rue Genot, B-4032 Chênée, Belgium
[5] Rijsberkamperweg 7, 8392 TP Boijl,The Netherlands
[6] 88 Louis Mettewie, 1080 Brussels, Belgium
[7] African Butterfly Research Institute, P.O. box 14308, 0800 Nairobi, Kenya

Abstract

A reliable taxonomic classification of species is of the essence for any biological study. Some species groups are notoriously problematic, however. The Afrotropical butterfly genus *Cymothoe* is such a group, as species are highly sexually dimorphic and associating males with females can be problematic because either males or females can be morphologically similar between species. Another issue is that, while species of *Cymothoe* are of special interest due to their highly specialized associations with *Rinorea* host plants, morphology-based identification of immature stages found on host plants in the field is problematic.

We therefore generated an extensive dataset of 1204 DNA barcode sequences and assessed whether it facilitates species delimitation and identification in *Cymothoe*. We applied a novel practical decision pipeline for integrating DNA barcodes, morphology and biogeography within a taxonomic framework, which proved instrumental for solving taxonomic problems in *Cymothoe* as five taxa within *Cymothoe* could be confidently raised to species level. In addition, our DNA barcode data set allowed for the identification of 42 immature specimens from 6 different countries, significantly increasing the data on *Cymothoe* host plant associations. Nevertheless, our results demonstrate that such an integrative approach cannot diagnose all species of *Cymothoe*, probably because of incomplete lineage sorting between recently diverged species as well as introgression (the latter possibly mediated through *Wolbachia* endosymbionts).

Nuclear DNA markers may provide a solution to some of the problems, but given the recent divergence of many *Cymothoe* species and the larger effective size of nuclear markers compared with mitochondrial ones, success is not guaranteed. Morphology of immature stages could provide important auxiliary evidence to help solve taxonomic issues where adult morphology and DNA sequences are inconclusive.

2.1 Introduction

A reliable classification and taxonomy of species is critical for any biological study. However, some species groups are notoriously problematic because traditional (morphological) characters used for taxonomic classification are inconclusive, lack resolution or disagree (Dayrat 2005, Pillon, Fay *et al.* 2006, Devictor, Mouillot *et al.* 2010). At the same time, such species groups usually represent cases of recent or insipient speciation, introgression, etc., rendering them especially interesting and suitable for evolutionary ecological studies (Knowles and Carstens 2007, Gavrilets and Losos 2009).

The use of standardized loci for species identification (DNA barcoding; Hebert, Cywinska *et al.* 2003) and delimitation has become established as an important taxonomic tool. For animals and many other eukaryotes the standard DNA barcode locus is the mitochondrial Cytochrome *c* oxidase subunit 1 (COI) gene (Hebert, Cywinska *et al.* 2003); for land plants, it is the concatenation of the plastid *rbcL* and *matK* genes (Hollingsworth, Forrest *et al.* 2009); for fungi it is the nuclear ribosomal internal transcribed spacer (ITS) region (Schoch, Seifert *et al.* 2012). Advantages of using organellar loci are ease of amplification from a wide variety of taxa and, because they are haploid, sequences can be obtained without cloning. In addition, because organelle sequences generally exhibit an effective population size of approximately one-quarter of that of nuclear markers, it allows reconstruction of more recent speciation events (Hurst and Jiggins 2005). Because DNA can be extracted from most biological tissues, DNA barcoding facilitates accurate identification of e.g. natural products in commerce, vegetative plant materials, or immature stages such as insect eggs of larvae (e.g. Stur and Ekrem 2011). Although not methodologically new (COI, *rbcL*, *MatK* and ITS have been standard loci for many years), DNA barcoding has now embraced a high-throughput ‘big data’ approach which, combined with substantial funding, allowed the generation of unprecedented high volumes of genetic sequence data at a reduced cost (Hajibabaei, DeWaard *et al.* 2005, Ivanova, DeWaard *et al.* 2006). As a result, population-level genetic sampling has become feasible for most taxonomic projects.

Here we combine DNA barcodes with morphological and biogeographical data with the aim to improve species delimitation in the Afrotropical butterfly genus *Cymothoe* Hübner, 1819 (gliders). *Cymothoe* constitutes a clade of 78 currently recognized species confined to all types of forested regions in tropical Africa and Madagascar (Ackery, Smith *et al.* 1995, Larsen 2005, van Velzen, Wahlberg *et al.* 2013). Phylogenetic estimations indicated that *Cymothoe* is a young clade (late Miocene) exhibiting an elevated rate of species diversification, compared with its sister genus *Harma* Doubleday, 1848, that was found to correlate with climatological oscillations and global cooling (van Velzen, Wahlberg *et al.* 2013). Within *Cymothoe*, most species were inferred to have diverged in the Pliocene and Pleistocene (van Velzen, Wahlberg *et al.* 2013). Delimitation and diagnosis of recently-diverged species

can be particularly difficult, because characters are not yet fixed and haplotypes may not be sorted according to lineages (Knowles and Carstens 2007, Shaffer and Thomson 2007, Hollingsworth, Forrest *et al.* 2009). Indeed, despite over a hundred year of taxonomic endeavors based on morphology and biogeography, many *Cymothoe* species groups and sections remain 'difficult' in the following way:

Species of *Cymothoe* are highly sexually dimorphic and associating males with females can be problematic because either males or females can be morphologically similar between species. For example, males in section *Sangaris* are morphologically indistinguishable while females are highly variable (Berger 1981, Larsen 2005) and, while males of *C. ogova* Plötz, 1880 and *C. harmilla* Hewitson, 1874 can be easily distinguished from each other, their females have been confused for a long time (Bouyer and Joly 1995). Indeed, some species later turned out to represent the associated sex of a previously described species (e.g. *C. adelina* Hewitson, 1869 is the female of *C. caenis* Drury 1773), and some allotypes were later found to be specifically distinct from the holotype. For example, the female allotype of *C. herminia gongoa* Fox, 1965 is now classified as *Cymothoe weymeri mulatta* Belcastro, 1990 (Larsen 2005).

To make things worse, the pages on *Cymothoe* in the textbook "Butterflies of the Afrotropical Region" by d'Abrera (2004) contains an excess of identification errors. The last monograph on *Cymothoe* included only 39 of 78 currently recognized species and was written by Overlaet (1952) who, shortly before his death in 1956, commented to Lucien Berger that, at least for some sections "C'est tout encore à refaire" [everything needs to be redone] (Berger 1981). In summary, it is clear that traditional lines of evidence such as morphology and biogeography alone have failed to provide decisive evidence for a reliable taxonomy of *Cymothoe*.

At the same time, *Cymothoe* are of special interest due to their highly specialized host plant associations. Based on a decade of field observations of *Cymothoe* oviposition behavior in Cameroon, Amiet & Achoundong (1996) found that roughly half the species are highly specialized, utilizing particular species of *Rinorea* Aubl. (Violaceae, Malpighiales) as host. Most are even monophagous (Fontaine 1982, Amiet and Achoundong 1996, McBride, van Velzen *et al.* 2009). Other species of *Cymothoe* feed exclusively on Achariaceae (also Malpighiales), which are also host to *Harma* (van Son 1979, Kielland 1990, Larsen 1991, Pringle, Henning *et al.* 1994, Amiet and Achoundong 1996). Because *Harma* is sister to *Cymothoe*, these patterns would suggest that Achariaceae are the ancestral host plant group(van Velzen, Wahlberg *et al.* 2013). Species-level specificity was confirmed by larval choice assays showing that, as soon as they hatch, *Cymothoe* larvae are able to recognize their *Rinorea* host plant (Amiet and Achoundong 1996). This high level of host plant specificity enables investigation of processes driving patterns of host plant associations and species diversity. However, such studies are hindered by the fact that morphological identification of immature *Cymothoe* specimens is problematic, and generally require rearing eggs or caterpillars to adulthood (Amiet and Achoundong 1996). This is

a time-consuming and laborious process, further complicated by a high incidence of parasitism. Therefore, a DNA sequence-based identification system would allow for a more efficient way to associate immature *Cymothoe* specimens (and their host plants) to species.

We use DNA barcodes to help (at least partially) resolve these issues. Our choice for DNA barcodes is not because we feel it provides characters superior to e.g. morphology but because it provides a relatively large number of characters not often considered before in *Cymothoe* taxonomy. Previously, we demonstrated DNA barcodes to be effective in discovering cryptic species as well as resolve taxonomic issues within some selected *Cymothoe* lineages (van Velzen, Bakker *et al.* 2007, McBride, van Velzen *et al.* 2009, van Velzen, Larsen *et al.* 2009). The present study extends the scope to a genus-wide sampling and, though not meant as an exhaustive revision, brings together new data in order to resolve taxonomic problems in *Cymothoe*.

2.1.1 Theoretical framework

Obviously, any taxonomic project should clearly state what species concept is assumed. We consider species to be separately evolving metapopulation lineages (sensu Wiley 1978) and adhere to a unified species concept that regards secondary species criteria (such as reproductive isolation, monophyly, diagnosability) as independent lines of evidence that may support lineage separation (De Queiroz 2007). As such, a species name constitutes a hypothesis about lineage separation, and current taxonomic classifications may be regarded as a collection of congruent species hypotheses. We call these hypotheses 'taxonomic species'. When delineation is difficult, some data may support alternative, competing species hypotheses. We call these hypotheses 'candidate species' (i.e. a set of organisms identified as putative new species; Padial, Miralles *et al.* 2010).

In accordance with the framework proposed by DeSalle *et al.* (1995) candidate species based on one line of evidence must be corroborated by other lines of evidence before they can be used for species delineation. Within this framework, named the 'Unholy Trinity', taxonomy therefore involves hypothesis testing, corroboration, reciprocal illumination and revision (DeSalle, Egan *et al.* 2005). Including DNA sequence data for corroboration of taxonomic hypotheses, usually entails integrating methods from different biological fields such as phylogenetics, comparative morphology, population genetics, and phylogeography; an approach that has become known under the name 'integrative taxonomy' (Dayrat 2005). In general, this usually means that independent data are separately applied to the same problem and when congruence is assessed, the hypothesis is accepted ('integration by congruence' sensu Padial, Miralles *et al.* 2010, 'iterative taxonomy' sensu Yeates, Seago *et al.* 2011). The rationale behind this approach is that concordant patterns of divergence among independent lines of evidence would indicate full lineage separation. Taxonomists

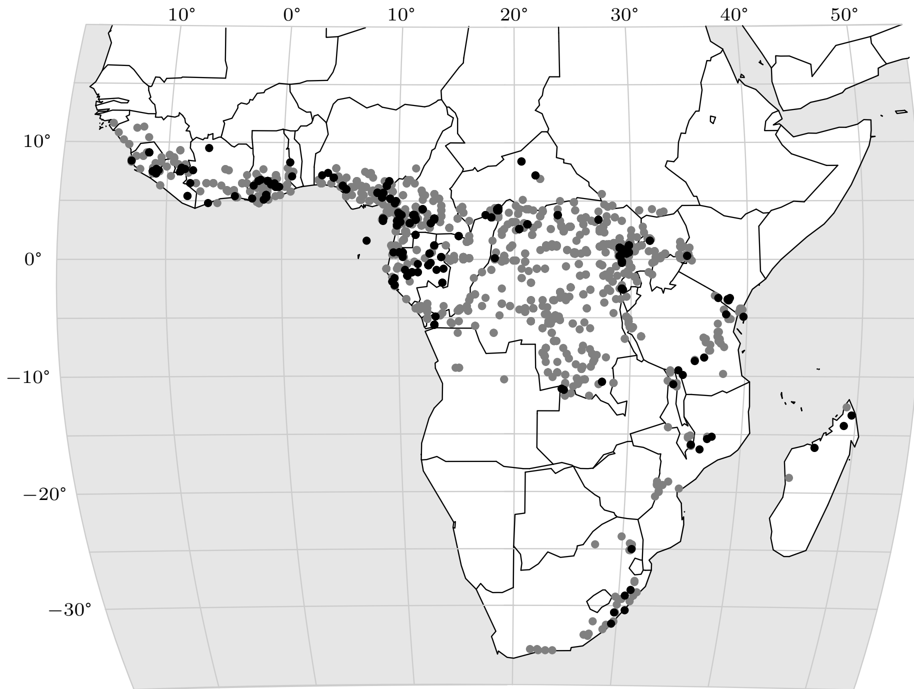


Figure 2.1: Geographical distribution of *Cymothoe*. Localities of DNA barcoded *Cymothoe* specimens (black dots) and general distribution of *Cymothoe* (grey dots). Distribution is based on specimens at the Royal Museum for Central Africa (Tervuren, Belgium), Natural History Museum (London, UK), African Butterfly Research Institute (Nairobi, Kenya) and the research collection of RvV.

also expect that species so discovered will more often correspond with distinct evolutionary units because it is improbable that a coherent pattern of divergence emerges by chance (Padial, Miralles *et al.* 2010).

In conclusion, we endorse that taxonomy should be pluralistic and multidisciplinary in order to ensure the best and maximally corroborated species delimitations (Dayrat 2005, DeSalle, Egan *et al.* 2005, Padial, Miralles *et al.* 2010). The most obvious way to achieve such pluralism is through collaboration between experts from different fields (Fisher and Smith 2008). Our fruitful collaboration between experts on butterfly taxonomy and morphology (DK, TB, GvdW, PhO, SC), biogeography (RvV, DK, MS), molecular evolution (FTB, RvV) and the ecology and immature stages of *Cymothoe* (SC, RvV) nicely reflects this notion.

2.2 Materials and methods

2.2.1 Taxon sampling

It is not clear how many individuals per species are sufficient in a DNA barcode reference library (Morando, Avila *et al.* 2003, Zhang, He *et al.* 2010) though this will obviously vary between species with different ecology and/or biogeographic distributions. We therefore sampled as many individuals as possible, especially from geographically widespread species, trying to cover their wide geographic distribution (see Figure 2.1). This resulted in a sampling comprising 1521 specimens from 64 taxonomic species. Specimens were newly collected, either by RvV or kindly donated by a network of collectors (see acknowledgements), or accessed at the African Butterfly Research Institute (ABRI; Nairobi), the Royal Museum for Central Africa (RMCA; Tervuren), the Zoological Museum of the Jagiellonian University (MZUJ; Cracow) and the Natural History Museum (NHM; London). Initial identifications were based on Larsen (1991, 2005), Vande weghe (2010), and Berger (1981). We sampled single legs for DNA sequencing. Most specimens are deposited at ABRI, others are part of the various authors' research collections and are available upon request.

2.2.2 Molecular methods

The vast majority of DNA barcode sequences were generated at the Canadian Center for DNA Barcoding (Guelph, Canada), where total genomic DNA was extracted using the NucleoSpin 96 Tissue kit (Macherey-Nagel Duren, Germany), following manufacturer's protocols, or silica-based Pall plates. A 658-bp region near the 5' terminus of the CO1 gene (DNA barcode fragment) was amplified using published primers LepF1 and LepR1 (Hebert, Penton *et al.* 2004). In cases where a 658-bp product was not successfully generated, internal primer pairs (LepF1-mLepR2 or LepF1-LepR1; Hajibabaei, Janzen *et al.* 2006) were used to generate shorter overlapping sequences that allowed the creation of a 658-bp contig. PCR was performed in volumes of 12.5 µl with a thermocycling profile of 1 cycle of 1 min at 94°C, 5 cycles of 40 s at 94°C, 40 s at 45°C, and 1 min at 72°C, followed by 35 cycles of 40 s at 94°C, 40 sec at 51°C, and 1 min at 72°C, with a final step of 5 min at 72°C. Samples showing clean single bands on agarose gel were bidirectionally sequenced using BigDye version 3.1 on an ABI 3730 DNA Analyzer (Applied Biosystems).

An additional 83 sequences were generated at Naturalis Biodiversity Center (Leiden, The Netherlands). There, total genomic DNA was extracted using an automated magnetic bead method on a KingFisher flex 96 (Thermo Scientific) with a NucleoSpin 96 Tissue kit (Macherey-Nagel Duren, Germany), following

manufacturers protocols. The DNA barcode fragment was amplified using published primers LepF1 and LepR1 ligated to M13 tails for sequencing. In cases where a product was not successfully generated, internal primer pairs LepF1 and mLepR1 and mLepF1-LepR1) were used to generate shorter overlapping sequences that allowed the creation of a contig. PCR was performed in volumes of 25 µl with a thermocycling profile of 1 cycle of 3 min at 94°C, 40 cycles of 15 s at 94°C, 30 s at 50°C, and 40 s at 57°C, with a final step of 5 min at 57°C. Samples showing clean single bands on agarose gel were bidirectionally sequenced on an ABI 3730 XL DNA Analyzer (Applied Biosystems) by Macrogen Europe.

All sequences were aligned using MAFFT (Katoh and Toh 2008), version 6.864b using E-INS-i strategy and gap opening costs of 1.0 (mafft.bat --op 1 --ep 0 --maxiterate 1000 -- 1 --genafpair --reorder input > output).

2.2.3 An integrative taxonomic decision pipeline

Although an integrative taxonomic approach is advocated by many, it is often unclear what the approach entails, exactly (Padial, Miralles *et al.* 2010). We therefore designed a practical decision pipeline for integrating DNA barcodes, morphology and biogeography within a taxonomic framework. Our pipeline is comparable to the protocol for integrative taxonomy by Schlick-Steiner *et al.* (Schlick-Steiner, Steiner *et al.* 2010) but differs in that, where their general pipeline uses different lines of evidence equally, we take current taxonomic opinion as our starting point. This choice is motivated by the notion that current classification is usually already based on an evaluation of multiple lines of evidence (i.e. morphology, biogeography) by taxonomic experts. Consequently, if the current taxonomy is confirmed by DNA barcodes, it is not necessary to re-evaluate all data. Instead, taxonomists can focus on the problematic cases where classification and DNA barcodes disagree, aiming to find a biological interpretation reconciling all available data. When taxonomic species and DNA barcodes are incongruent, we delineate candidate species based on DNA barcode patterns and try to find independent evidence supporting their specific status. Our pipeline can be applied to any clade for which an a priori taxonomy is available and comprises four distinct steps (see Figure 2.2):

Confirmation of current classification

In the first step DNA barcode data are used to confirm current taxonomic species hypotheses. Our confirmation criteria are 1) the existence of a 'DNA barcode gap' and 2) reciprocal species monophyly in a DNA barcode haplotype tree. To assess the existence of a barcode gap in our data, we extracted within- and between-species K2P (Kimura 1980) distances for all species. We then scored a species as having a 'barcode gap' when the minimum between-species sequence distance exceeded the maximum within-species distance (Meyer and Paulay 2005, Meier,

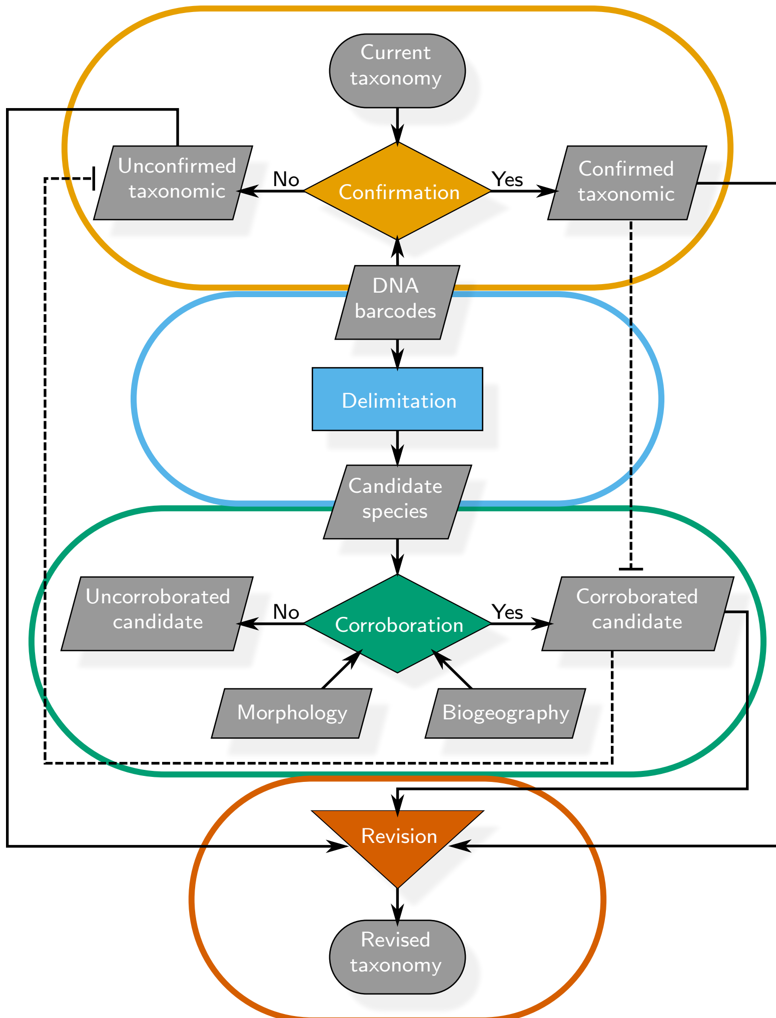


Figure 2.2: Integrative taxonomic decision pipeline. Figure illustrates the four steps, from top to bottom: (i) confirming taxonomic species with DNA barcodes; (ii) delimiting DNA barcode-based candidate species; (iii) corroborating candidate species with independent lines of evidence; and (iv) revision of taxonomy when current taxonomic opinions are not confirmed and alternative candidate species delineated based on DNA barcodes are corroborated by other disciplines.

Zhang *et al.* 2008). To assess reciprocal species monophyly we built a DNA barcode haplotype tree based on pairwise K2P distances between all specimens using the Neighbor Joining clustering algorithm (NJ) (Saitou and Nei 1987). NJ is the most widely used method for representing DNA barcodes in literature, and implemented in e.g. the Barcode Of Life Database (Ratnasingham and Hebert 2007). In addition, NJ was found to outperform maximum parsinomy in accurately matching DNA barcode queries to an existing database of classified sequences (van Velzen, Weitschek *et al.* 2012). For NJ tree building, the input order of sequences was randomly shuffled before clustering; pairwise distances between short sequences with an overlap <350 bp were considered unreliable and substituted with estimates based on the ultrametric procedure (Makarenkov and Lapointe 2004). We assessed whether species corresponded with monophyletic clusters in the NJ tree using the 'is.monophyletic' function implemented in the R. package APE 2.5–3 (Paradis, Claude *et al.* 2004). We note that clusters in a NJ tree based on a single marker represent haplotype similarities and not necessarily species ancestry (clades). We therefore use the terms 'monophyly', 'paraphyly' and 'polyphyly' in an operational sense, referring to the NJ tree shape, without making statements about shared ancestry.

There are multiple scenarios in which taxonomic species are rejected based on DNA barcodes: 1) Multiple taxonomic species have identical barcodes, leading to absence of a DNA barcoding gap and species monophyly; 2) A taxonomic species can comprise deep conspecific lineages that are more divergent to each other than to another species (i.e. no barcode gap); 3) A taxonomic species can appear polyphyletic, suggesting separately evolving lineages (i.e. splitting) with regards other such lineages; 4) Multiple taxonomic species can appear mixed within a single monophyletic cluster suggesting a single species (i.e. merging). Obviously, taxonomic species can be monophyletic while at the same being nested within another non-monophyletic species (i.e. rendering the second paraphyletic) suggesting that it may represent a population rather than a species. In this case, monophyly is not reciprocal and both taxonomic species are considered unconfirmed. We note that paraphyletic patterns can be expected in recently diverged species (Funk and Omland 2003; this chapter). That is why, in our pipeline, taxonomic species are not rejected in such cases, but remain unconfirmed.

Delimitation of candidate species

For those taxonomic species that are confirmed by the DNA barcode data, the pipeline stops here and their taxonomy is left unaltered. For taxonomic species that are unconfirmed, alternative hypotheses for species delimitation are generated, based on the DNA barcode data. We used the automatic barcode gap discovery (ABGD) approach to assign specimens to candidate species, based on pairwise distance data (Puillandre, Lambert *et al.* 2012). While our confirmation step depended on finding a barcode gap for predefined species, ABGD estimates barcode gaps from the data,

irrespective of any prior taxonomy. Instead of simply classifying DNA barcodes on an arbitrary distance threshold (e.g. Hebert, Stoeckle *et al.* 2004), ABGD infers cluster-specific distance thresholds from the data. First, ABGD uses a range of prior intraspecific divergence thresholds to infer a model-based one-sided confidence limit for intraspecific divergences from the data. Subsequently, it detects a 'global barcode gap' based on the distribution of pairwise distances and uses this gap to partition the data into groups. Inference of the limit and gap detection are then recursively applied to previously obtained groups to get finer partitions that are considered candidate species (Puillandre, Lambert *et al.* 2012). The advantage of the ABGD over other species-delineation methods such as the General Mixed Yule Coalescent (Pons, Barraclough *et al.* 2006, Monaghan, Wild *et al.* 2009) is that it does not rely on phylogenetic inference and diversification models, that arguably require more information than a single DNA barcode locus for reliable estimation. In addition, it is much faster to compute, while producing qualitatively similar results (Puillandre, Modica *et al.* 2012). We applied ABGD to pairwise K2P distances between sequences >350bp, set a range of 100 prior intraspecific divergence thresholds between 0.1 and 0.001 and required any barcode gap to be 1.5 times the maximum within-species distance ($P_{min} = 0.001$, $P_{max} = 0.100$, Steps = 100, X = 1.5). Setting different values for the arbitrarily chosen default value X produced qualitatively similar results (data not shown), suggesting that results are robust. In order to avoid sampling bias, the ABGD analysis was based on all data, including those from previously confirmed taxonomic species.

Corroboration of candidate species

DNA barcode-based candidate species are then corroborated with independent lines of evidence, in our case morphology and biogeography. Morphological evidence came from previous taxonomic work, photographs of all sampled specimens, data on larval morphology based on Amiet (2000) as well as new observations, and male genital dissections of selected specimens. Basically, when any character could be inferred that would corroborate candidate species, we 'took it'.

Biogeographical evidence came from geo-references of all sampled specimens and those associated with collection data from NHM, RMCA, ABRI and personal research collections of RvV, PhO and ThB. These were amalgamated in order to represent a more complete record of species distributions. We compared geographic distance with genetic variation based on species distribution maps that were created using the 'Aitoff equal area' projection. In panmictic species (i.e. with random mating), there is no geographic structure in the genetic diversity. Most *Cymothoe* species are not expected to be panmictic, however, because their forest habitat is fragmented, causing reduced gene-flow between populations. Consequently, a pattern of isolation by distance is expected, where populations that are geographically closer have the tendency to be more similar than populations that are further apart (Wright 1943).

Conversely, when geographically close populations are genetically more different than those further apart, this could be the result of long-term reproductive isolation and, hence, suggest distinct species, although there are alternative explanations (e.g. occasional dispersal over longer distances).

Candidate species merging multiple taxonomic species suggest that the latter may be oversplit. When the taxonomic species concerned are sympatric, this may be due to over-interpretation of morphological characters (e.g. seasonal or sexual dimorphism, ecological variability). When allopatric, this may be due to confounding morphologically divergent populations (i.e. subspecies) with species. Such mergers are corroborated when biogeographical patterns correspond to panmixis or isolation by distance, suggesting conspecifity.

Conversely, candidate species corresponding to a subset of a single taxonomic species suggest that the latter should be split into multiple species. These candidate species may be morphologically divergent (and possibly correspond with currently recognized subspecies), or highly similar (i.e. cryptic species). Either way, such splits are corroborated when biogeographical patterns are inconsistent with panmixis or with isolation by distance, suggesting reproductive isolation.

Revision of taxonomic classification

By integrating results from all three steps outlined above, taxonomic decisions can be made. Taxonomic species that were confirmed in step 1 are left unaltered. For those that were rejected, alternative, candidate species were delimited based on DNA barcodes and corroborated with independent evidence in step 3. Corroborated candidate species likely represent independently evolving lineages (i.e. species) and their taxonomic classification is changed accordingly.

2.3 Results

2.3.1 Sequencing and quality control

Out of the 1521 specimens, 1204 samples could be successfully sequenced for COI (80%). From the latter, 383 required the use of additional internal primers to sequence the 658 bp DNA barcode marker, sometimes resulting in partial sequences, see Figure 2.3. Partial DNA barcodes were not excluded from the analyses as they may allow discrimination of species (Hajibabaei, Smith *et al.* 2006, Meusnier, Singer *et al.* 2008) and did not constitute a large part of our data (average sequence length was 600 bp). As most samples were collected in hot and humid environments we attribute these low success rates to poor specimen preservation after collection and subsequent DNA degradation.

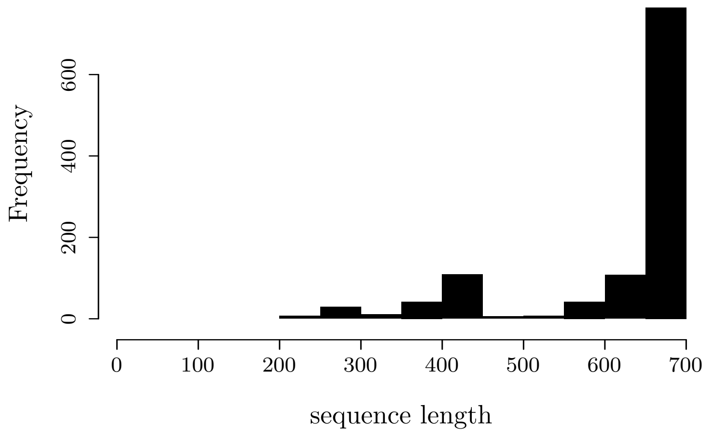


Figure 2.3: DNA barcode sequence length. Histogram of *Cymothoe* and *Harma* DNA barcode sequence lengths in the final data set.

In a total of 155 cases sequences were identical to those of a different species than initially assigned. In some cases this could be attributed to cross-contamination, due to sampling errors or lab contamination of low-quality samples with DNA from neighboring samples in the same plate (14 sequences). Two sequences were identical to that of species in another subfamily (*Acraea jodutta* in Heliconiinae) and two others clearly belonged to *Wolbachia* α–proteobacteria endosymbionts probably belonging to supergroup A (results not shown). Most non-contamination cases could be attributed to identification problems in taxonomically complex groups (76 sequences). In some, such errors could be ruled out, however, because the donor and recipient species were not sampled nor processed together. In addition, such anomalous results were sometimes replicated in different independently processed sampling sets, suggesting that the identical sequences are reliable (65 sequences).

After exclusion of these erroneous sequences, our final data set comprised 1093 DNA barcode sequences; constituting 72% of our original sampling. For four species we obtained a single sequence only: *C. adela* Staudinger, 1890 from the Liberian region, *C. crocea* Schultze, 1917 from Cameroon, *C. magambae* Rydon, 1980 and *C. melanjae* Bethune-Baker, 1926 from mount Mulanje in Malawi, *C. magambae* Rydon, 1980 from Pare and Usambara mountains in northern Tanzania, and *C. meridionalis* Overlaet, 1944 from Cameroon. All results below are based on this final data set.

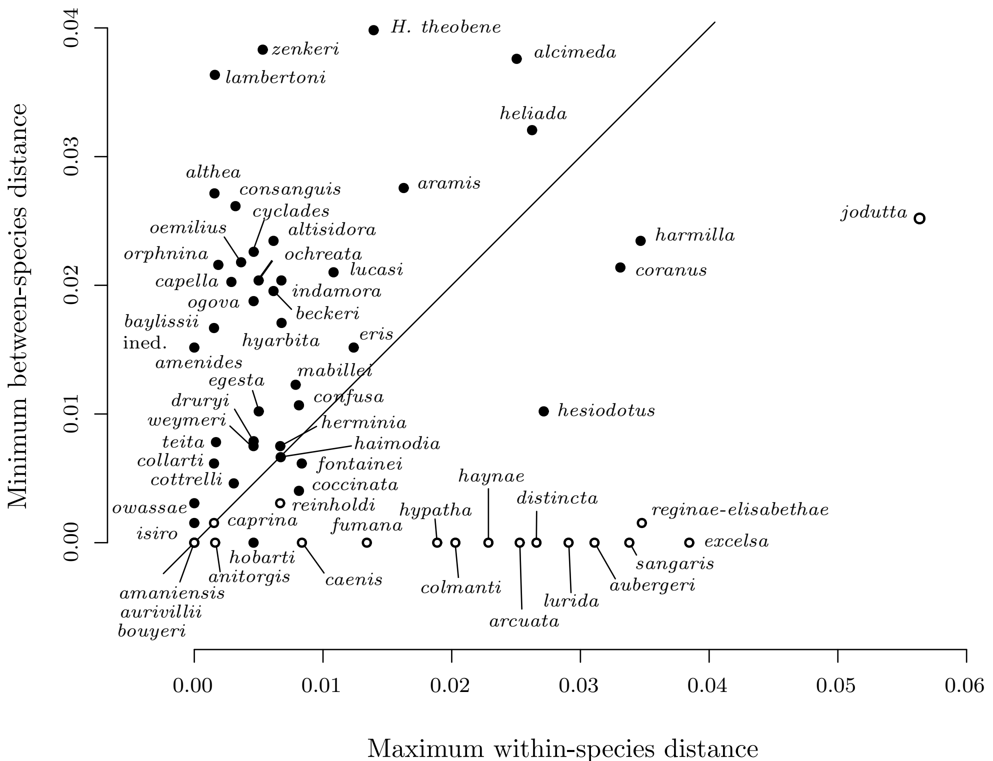


Figure 2.4: DNA barcode gap. Scatterplots of maximum within- and minimum between-species distance (based on the Kimura 2-parameter model) for the 59 taxonomic species sampled more than once. Species plotted above the diagonal line have a barcode gap; closed circles represent species that are monophyletic in the NJ haplotype tree, open circles represent species that are not. Note that 16 species have sequences that are identical to those of another (bottom).

2.3.2 Confirmation of current classification

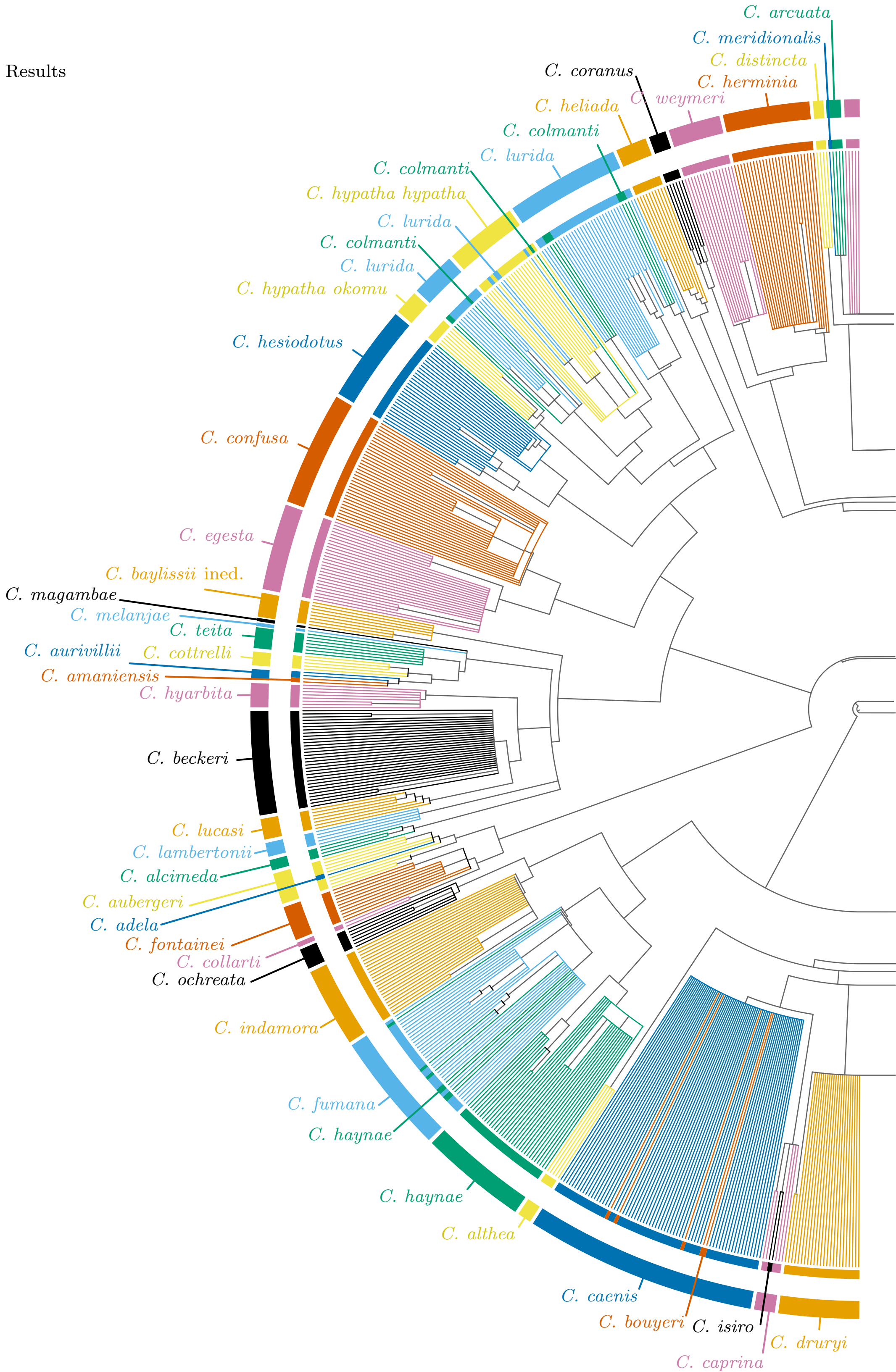
Of the 59 taxonomic species sampled more than once, 33 (56%) showed a barcode gap and 37 (63%) appeared as reciprocally monophyletic clusters in our NJ tree, see Table 2.1 and figures 2.4 & 2.5. As expected, the 33 taxonomic species with a barcode gap also appeared as monophyletic. However, two of these species were nested within others and thus remained unconfirmed; *C. isiro* within *C. caprina* Aurivillius, 1897, and *C. owassae* Schultze, 1916 within *C. reinholdi* Plötz, 1880. Therefore, 31 taxonomic species (53%) were confirmed according to the first step in our pipeline. This low confirmation rate of just over half of all species considered was mainly due to multiple taxonomic species having identical DNA barcodes (16 species). Some cases of shared sequences could be attributed to identification problems in

'difficult' species groups such as *C. anitorgis* Hewitson, 1874, *C. arcuata* Overlaet, 1945, *C. distincta* Overlaet, 1944, *C. excelsa* Neustetter, 1912, and *C. meridionalis* Overlaet, 1944 from section *Aramis* and *C. colmanti* Aurivillius, 1898, and *C. lurida* Butler, 1871 from section *Lurida*. In other cases, however, identical sequences were found in morphologically well-characterized species. For example, sequences of the recently described *C. bouyeri* Vande weghe, 2011 are identical to those of widespread *C. caenis* Drury, 1773, despite clear morphological differences (Robe, Cordeiro *et al.* 2010, Vande weghe 2011) and the same is true for *C. fumana* Westwood, 1850 versus *C. haynae* Dewitz, 1887, and for *C. hypatha* Hewitson, 1866 versus *C. lurida* (not shown).

Eight taxonomic species comprised deep conspecific COI lineages more divergent to each other than to another species. Six of them were monophyletic species, but due to large within-species distances, without a barcode gap: distances within *C. coccinata* Hewitson, 1874 exceed the smallest distance to closely related *C. sangaris* Godart, 1824, distances within *C. fontainei* Overlaet, 1952 exceed those to *C. collarti* Overlaet, 1942, distances within *C. coranus* Grose-Smith, 1889 exceed those to *C. herminia* Grose-Smith, 1887, distances within *C. hesiodotus* Staudinger, 1890 exceed those to *C. hypatha*; distances within *C. haimodia* Grose-Smith, 1887 exceed those to *C. crocea* but this is most likely an artefact of the short sequence length of one of the samples (GW_3527; 317 bp); distances within *C. harmilla* exceed the smallest distance to many other species. Two of them (*C. caprina*, *C. reinholdi*) were paraphyletic due to the nested species mentioned earlier (see Table 2.1).

Another eight taxonomic species appeared polyphyletic; *C. arcuata*, *C. colmanti*, *C. distincta*, *C. excelsa*, *C. hypatha*, *C. jodutta* Westwood, 1850, *C. lurida*, *C. reginae-elisabethae* Holland, 1920, and *C. sangaris*. Seven of them come from taxonomically 'difficult' sections, and also appear in mixed clusters due to shared sequences, suggesting identification problems (see above). *Cymothoe jodutta* therefore is the only species for which polyphyly is certain.

Figure 2.5: Species monophyly (next pages). Radial representation of NJ haplotype tree based on *Cymothoe* and *Harma* DNA barcodes. Branch lengths are transformed based on Grafen (2009) with node heights raised to power 0.001 for better representation of terminal clusters.


C. arcuata
C. meridionalis
C. distincta
C. herminia
C. coranus
C. weymeri
C. heliada
C. colmanti
C. lurida
C. colmanti
C. hypatha hypatha
C. lurida
C. colmanti
C. lurida
C. hypatha okomu
C. hesiodotus
C. confusa
C. egesta
C. baylissii ined.
C. magambae
C. melanjae
C. teita
C. aurivillii
C. cottrelli
C. amaniensis
C. hyarbita
C. beckeri
C. lucasi
C. lambertonii
C. alcimeda
C. aubergeri
C. adela
C. fontainei
C. collarti
C. ochreata
C. indamora
C. fumana
C. haynae
C. haynae
C. althea
C. caenis
C. bouyeri
C. isiro
C. druryi
C. caprina

C. reginae-elisabethae
C. distincta
C. arcuata
C. excelsa
C. distincta
C. coccinata
C. crocea
C. haimodia
C. reginae-elisabethae
C. excelsa
C. anitorgis
C. aramis
C. mabillei
C. harmilla
C. orphnina
C. sangaris
C. hobarti
C. jodutta jodutta
C. ogova
C. altisidora
C. reinholdi
C. owassae
C. oemilius
C. cyclades
H. theobene
C. jodutta ciceronis
C. consanguis
C. capella
C. eris
C. amenides

Table 2.1: Confirmation of taxonomic species.

Taxonomic species	#	MWD	MBD	Barcode gap	Monophyly	Confirmation
C. adela	1	n.a.	0.000	n.a.	n.a.	n.a.
C. alcimeda	4	0.025	0.038	True	True	confirmed
C. althea	5	0.002	0.027	True	True	confirmed
C. altisidora	8	0.006	0.023	True	True	confirmed
C. amaniensis	2	0.000	0.000	False	False	identical seqs
C. amenides	6	0.000	0.015	True	True	confirmed
C. anitorgis	3	0.002	0.000	False	False	identical seqs
C. aramis	19	0.016	0.028	True	True	confirmed
C. arcuata	14	0.025	0.000	False	False	identical seqs
C. aubergeri	9	0.031	0.000	False	False	identical seqs
C. aurivillii	2	0.000	0.000	False	False	identical seqs
C. baylissi	8	0.002	0.017	True	True	confirmed
C. beckeri	31	0.006	0.020	True	True	confirmed
C. bouyeri	5	0.000	0.000	False	False	identical seqs
C. caenis	63	0.008	0.000	False	False	identical seqs
C. capella	9	0.003	0.020	True	True	confirmed
C. caprin.a.	5	0.002	0.002	False	False	Paraphyletic
C. coccinata	44	0.008	0.004	False	True	DCL
C. collarti	2	0.002	0.006	True	True	confirmed
C. colmanti	11	0.020	0.000	False	False	identical seqs
C. confusa	34	0.008	0.011	True	True	confirmed
C. consanguis	3	0.003	0.026	True	True	confirmed
C. coranus	6	0.033	0.021	False	True	DCL
C. cottrelli	5	0.003	0.005	True	True	confirmed
C. crocea	1	n.a.	0.016	n.a.	n.a.	n.a.
C. cyclades	10	0.005	0.023	True	True	confirmed
C. distincta	31	0.027	0.000	False	False	identical seqs
C. druryi	32	0.005	0.008	True	True	confirmed
C. egesta	26	0.005	0.010	True	True	confirmed
C. eris	2	0.012	0.015	True	True	confirmed
C. excelsa	63	0.038	0.000	False	False	identical seqs
C. fontainei	11	0.008	0.006	False	True	DCL
C. fuman.a.	29	0.013	0.000	False	False	identical seqs
C. haimodia	4	0.007	0.007	False	True	DCL
C. harmilla	9	0.035	0.023	False	True	DCL
C. haynae	39	0.023	0.000	False	False	identical seqs
C. heliada	10	0.026	0.032	True	True	confirmed
C. herminia	26	0.007	0.008	True	True	confirmed
C. hesiodotus	29	0.027	0.010	False	True	DCL
C. hobarti	7	0.005	0.000	False	True	identical seqs
C. hyarbita	8	0.007	0.017	True	True	confirmed
C. hypatha	24	0.019	0.000	False	False	identical seqs
C. indamora	24	0.007	0.020	True	True	confirmed
C. isiro	2	0.000	0.002	True	True	nested
C. jodutta	42	0.056	0.025	False	False	polyphyletic
C. lambertoni	5	0.002	0.036	True	True	confirmed

(Continued on next page)

Table 2.1: Confirmation of taxonomic species, continued.

Taxonomic species	#	MWD	MBD	Barcode gap	Monophyly	Confirmation
C. lucasi	7	0.011	0.021	True	True	confirmed
C. lurida	43	0.029	0.000	False	False	identical seqs
C. mabillei	22	0.008	0.012	True	True	confirmed
C. magambae	1	n.a.	0.017	n.a.	n.a.	n.a.
C. melanjae	1	n.a.	0.020	n.a.	n.a.	n.a.
C. meridionalis	1	n.a.	0.000	n.a.	n.a.	n.a.
C. ochreata	7	0.005	0.020	True	True	confirmed
C. oemilius	13	0.004	0.022	True	True	confirmed
C. ogova	11	0.005	0.019	True	True	confirmed
C. orphnina	6	0.002	0.022	True	True	confirmed
C. owassae	2	0.000	0.003	True	True	nested
C. reginaeelisabethae	15	0.035	0.002	False	False	polyphyletic
C. reinholdi	10	0.007	0.003	False	False	Paraphyletic
C. sangaris	159	0.034	0.000	False	False	identical seqs
C. teita	7	0.002	0.008	True	True	confirmed
C. theobene	27	0.014	0.040	True	True	confirmed
C. weymeri	16	0.005	0.008	True	True	confirmed
C. zenkeri	12	0.005	0.038	True	True	confirmed

MWD = maximum within-species distance; MBD = minimum between-species distance; DCL = deep conspecific lineages; # = number of samples; n.a. = not applicable because only a single specimen was sampled.

2.3.3 Delimitation of candidate species

Our ABGD analysis suggested two global barcode gaps, coinciding with roughly 1% and 2.5% K2P distances (see Figure 2.6). After recursion, ABGD found a variable number of candidate species, based on the range of prior intraspecific divergences P. These numbers fell into four distinct classes with roughly 40 (P = 0.0064–0.0051), 70, (P = 0.0049–0.0032) 83 (P = 0.0031–0.0017), and 300 (P = 0.0015–0.0010) candidate species. Within each class, candidate species numbers were relatively constant and robust to differences in P (see Figure 2.7) and the relative gap size X (data not shown), suggesting that the delimitations are based on real structure in the pairwise distance data and not on arbitrary parameter values of the ABGD analysis. Because delimitations with 40 and 70 candidate species lumped some morphologically well-defined species (e.g. *C. caenis* with *C. caprina*; *C. herminia* with *C. weymeri*), and 300 candidate species probably is a gross overestimation of the actual taxonomic diversity, we selected the delimitation with 83 candidate species (P = 0.0031) as the best representation of our data.

As expected, most taxonomic species already confirmed by both barcode gap and haplotype monophyly were also present in the set of 83 candidate species (24 out of 31 confirmed taxonomic species). Exceptions were *C. alcimeda*, *C. cottrelli*,

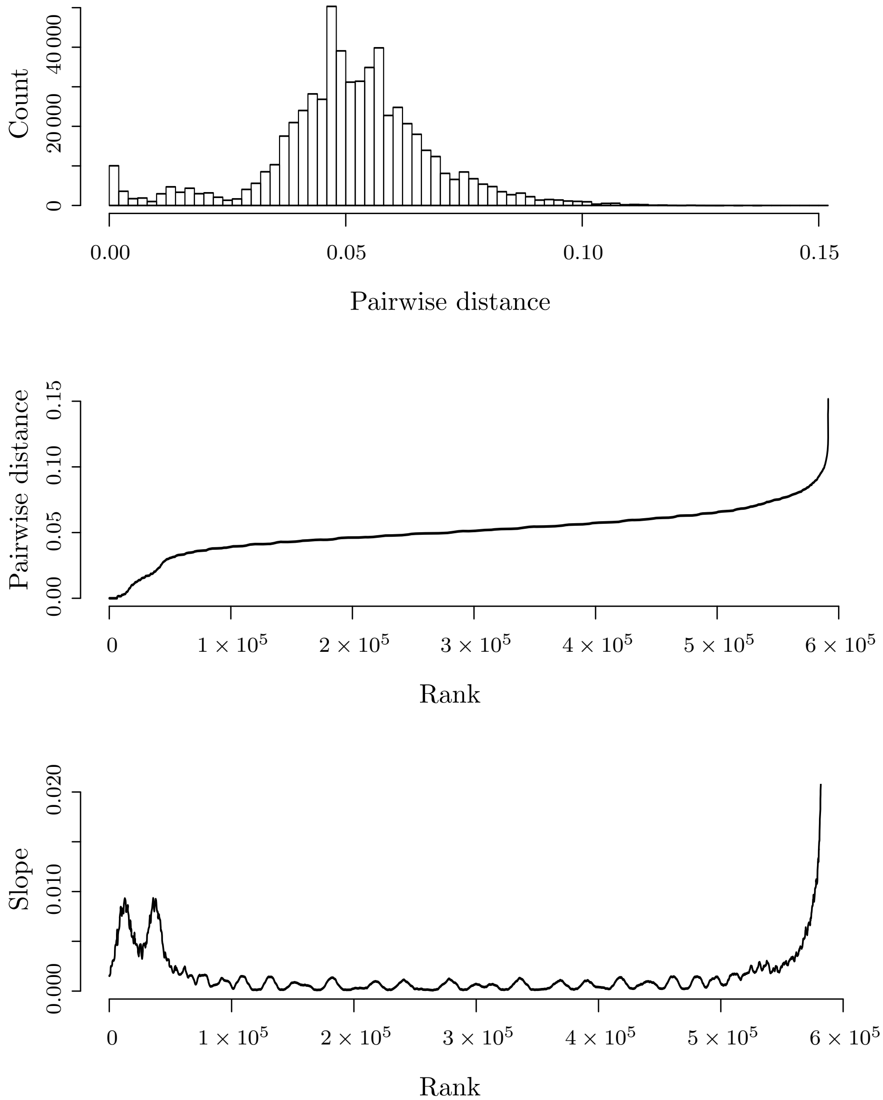


Figure 2.6: Automatic Barcode Gap Discovery. ABGD inference of distance thresholds. Distribution (top), ranked ordered values (middle), and slope of the ranked ordered values (bottom) of pairwise differences between *Cymothoe* and *Harma* DNA barcodes. Two peaks in the left part of the slope correspond with potential DNA barcode gaps in the data.

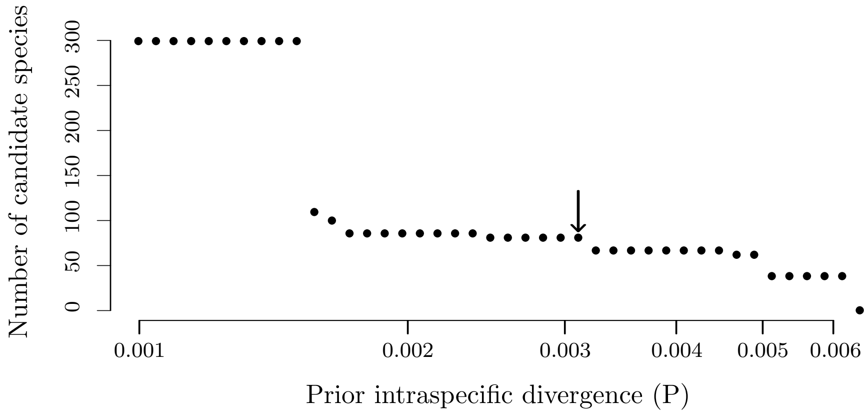


Figure 2.7: Candidate species delimitation. Number of candidate species after ABGD partitioning based on a range of prior intraspecific divergences P (on log scale). Arrow indicates partition selected as best representation of our data (P = 0.0031)

C. eris, *C. heliada, C. lucasi*, *C. mabillei* and *C. oemilius* that were each split into two candidate species. Nevertheless, because they were already confirmed, the latter species are maintained according to our pipeline. In addition, all four taxonomic species that were sampled only once and could therefore not be confirmed by barcode gap or by monophyly (i.e. *C. adela*, *C. crocea*, *C. magambae*, *C. melanjae*) also came up as candidate species in the ABGD analysis, thus confirming their current status. Of those taxonomic species that appeared monophyletic without a barcode gap, only *C. haimodia* corresponded with a candidate species (probably because the sort sequence was excluded before delimitation). Thus, in total, 29 candidates corresponded with a taxonomic species.

Of the remaining 54 candidate species, 4 were mergers of multiple taxonomic species, suggesting that they may be oversplit: *C. fumana* was merged with *C. haynae*, *C. aurivilii* with *C. amaniensis*, *C. caprina* with *C. isiro,* and *C. owassae* with *C. reinholdi.* An additional 7 candidate species were also mergers but of subsets of taxonomic species (i.e. they are mergers as well as splits), requiring a more complex explanation or, possibly, suggesting misidentifications: *C. excelsa* with *C. anitorgis; C. arcuata* with *C. coccinata, C. distincta, C. excelsa,* and *C. reginae-elisabethae; C. arcuata* with *C. meridionalis* and *reginae-elisabethae, C. bouyeri* with *C. caenis; C. colmanti* with *C. lurida*; *C. colmanti* with *C. lurida* and *C. hypatha*; *C. hobarti* with *C. sangaris.*

Fifty candidate species corresponded to a subset of taxonomic species, suggesting that the latter should be split into multiple species: *C. alcimeda* (split into 2 candidates), *C. arcuata* (2), *C. aubergeri* (3), *C. caenis* (2), *C. colmanti* (2), *C. coranus* (4), *C. cottrelli* (2), *C. distincta* (2), *C. eris* (2), *C. excelsa* (3), *C. fontainei*

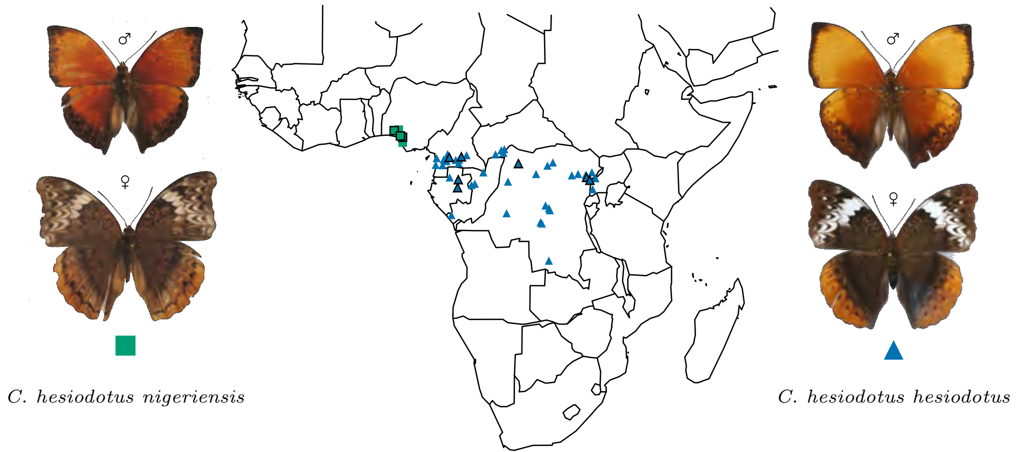


Figure 2.8: ***Cymothoe hesiodotus.*** Specimen photographs and collection localities of *C. hesiodotus hesiodotus* (blue triangles), and *C. hesiodotus nigeriensis* (green squares); localities of sampled specimens have a black outline.

(4), *C. harmilla* (3), *C. heliada* (2), *C. hesiodotus* (2), *C. jodutta* (4), *C. lucasi* (2), *C. lurida* (2), *C. mabillei* (2), *C. oemilius* (2), *C. reginae-elisabethae* and *C. sangaris* (7). Note that members of *C. sangaris*, *C. reginae-elisabethae*, *C. lurida*, *C. excelsa*, *C. arcuata* and *C. colmanti* sometimes also co-occur with other taxonomic species in the same candidate species (i.e. they are mergers as well as splits; see above). In total, 43 candidate species are strict splits of single taxonomic species.

2.3.4 Corroboration of candidate species and revision of taxonomic classification

Of the 59 candidate species not corresponding with taxonomic species, 7 were corroborated by morphology and/or biogeography as independent lines of evidence (see Table 2.2). Six of these candidate species correspond to a subset of a single taxonomic species (i.e. splitting), and 1 comprises a merger.

Cymothoe hesiodotus was split into 2 candidate species, mainly because of its large intraspecific divergence (2.7% K2PD). These candidates were corroborated by clear morphological differences in mainly the females, corresponding with two current subspecies. The nominate subspecies *C. hesiodotus hesiodotus* occurs from central Cameroon to Gabon, the Central African Republic, and parts of DR Congo, and subspecies *C. hesiodotus clarior* Overlaet, 1952 in north-eastern (Ituri) and eastern DR Congo. Subspecies *C. hesiodotus nigeriensis* Overlaet, 1952 is only found on the western side of the Niger Delta, and seems to be absent in eastern Nigeria

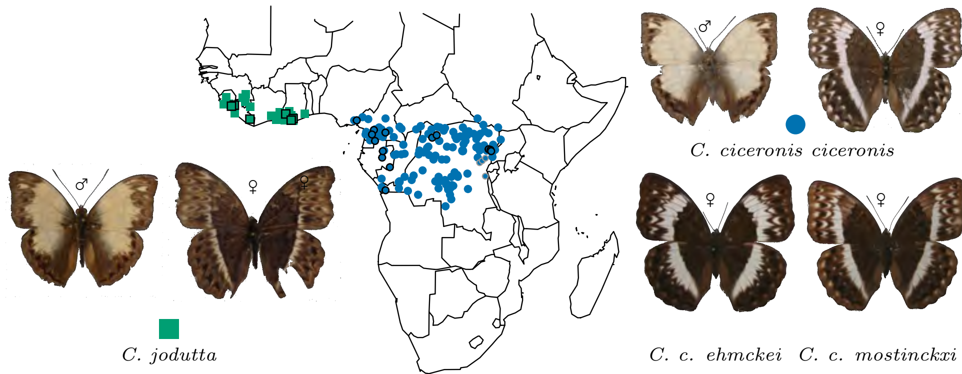


Figure 2.9: *Cymothoe jodutta* and *C. ciceronis*. Specimen photographs and collection localities of *C. jodutta jodutta* (green squares), *C. ciceronis ciceronis*, *C. ciceronis ehmckei* and *C. ciceronis mostinckxi* (blue dots); localities of sampled specimens have a black outline.

and western Cameroon (see Figure 2.8). *C. hesiodotus nigeriensis* was originally described as a variety of *C. lurida* from Warri (Overlaet 1952), but was reclassified by Larsen as a western subspecies of *C. hesiodotus* based on the ochreous-orange color of the males (Larsen 2005) which is characteristic for that species. In our DNA barcode haplotype tree, all specimens of *C. hesiodotus hesiodotus* and *C. hesiodotus clarior* cluster together, even though they are spanning a geographical distance of over 2000 km. Surprisingly, specimens of *C. hesiodotus nigeriensis* cluster together and separately from the other subspecies, even though they are only 600 km away from the nominate, refuting an isolation by distance explanation and suggesting long-term reproductive isolation between the two candidates corresponding with *C. hesiodotus hesiodotus* and *C. hesiodotus nigeriensis*. We therefore raise *C. nigeriensis* Overlaet, 1952 to species level (**stat. rev.**).

Cymothoe jodutta was split into 4 candidate species, mainly because of its extraordinary intraspecific divergence (5.6% K2PD). Independent corroboration was found for two pairs of candidate species that each exhibit morphological and biogeographical uniformity. The first pair corresponds with subspecies *C. jodutta jodutta*, occuring in West Africa from Sierra Leone to Ivory Coast. The second pair corresponds with subspecies *C. jodutta ciceronis* Ward, 1871 ocurring from Cross River in eastern Nigeria through Cameroon to Gabon, the Central African Republic, and DR Congo; subspecies *C. jodutta mostinckxi* Overlaet 1952 in eastern DR Congo and Uganda, and subspecies *C. jodutta ehmckei* Dewitz, 1887 in southern DR Congo (see Figure 2.9). Within each pair, biogeographical patterns suggest panmixis as candidates within each pair contain specimens from the same locality. Additionally, divergence within each pair is much smaller (0.8% and 1.6% K2PD)

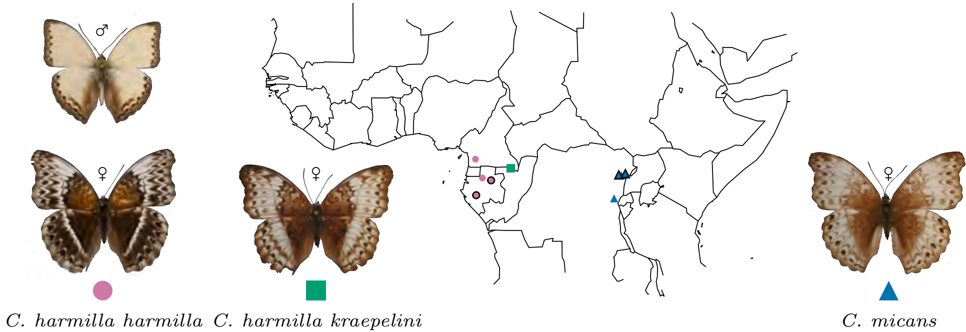


Figure 2.10: *Cymothoe harmilla* and *C. micans*. Specimen photographs and collection localities of *C. harmilla harmilla* (purple circles), *C. harmilla kraepelini* (green square) and *C. micans* (blue triangles); localities of sampled specimens have a black outline.

than between (5.6% K2PD), suggesting that pairs are conspecific. *C. jodutta ciceronis* was originally described as a species, but subsequently usually treated as a subspecies of *C. jodutta* (e.g. Larsen 2005, Vande weghe 2010). While the male does not appear to differ significantly from the nominate subspecies, the female differs by having a broad white discal band and white postdiscal chevrons (see Figure 2.9), present also in *C. jodutta mostinckxi* and *C. jodutta ehmckei*. In our DNA barcode haplotype tree, all specimens of the nominate cluster together. Subspecies *ciceronis* and *mostinckxi* cluster together but are genetically so different from the nominate that they do not even appear close in the haplotype tree and, indeed, correspond with separate candidate species. The genetic, morphological and biogeographical data therefore collectively support long-term reproductive isolation and we revert *C. ciceronis* to species status (**stat. rev.**), with three subspecies: *C. ciceronis ciceronis*, *C. ciceronis mostinckxi*, and *C. ciceronis ehmckei*. Consequently, *C. jodutta* now comprises the West African populations only. We note that Berger (1981) treated ciceronis as a distinct species but Larsen (2005) considered it to be a subspecies of jodutta as did Vande weghe (2010).

Cymothoe harmilla was split into 3 candidate species, mainly because of its extraordinary intraspecific divergence (3.5% K2PD). According to Bouyer & Joly (2008), *C. harmilla* comprises three subspecies: nominate subspecies *C. harmilla harmilla* occurs from south-western Cameroon to Gabon (it was originally described from "Calabar", Nigeria but to our knowledge it has never been caught there again, and we suspect that the type specimen was mislabelled), subspecies *C. harmilla kraepelini* Schultze, 1912 appears to be endemic to south-eastern Cameroon, and subspecies *C. harmilla micans* Bouyer & Joly, 1995 from eastern DR Congo. The males are similar in all three subspecies, but the females differ considerably: The nominate is dark and relatively similar to the female of *C. ogova*. Indeed, females of *C. harmilla harmilla* and *C. ogova* are often confused (Bouyer and Joly 1995) even in the

textbook by d'Abrera (2004). In comparison, females of subsp. *kraepelini* are much whiter. Females of subsp. *micans* are even lighter than those of subsp. *kraepelini* but in addition have a much more angular-shaped forewing apex (see Figure 2.10). One candidate species combines samples of *C. harmilla harmilla* from Gabon and samples of *C. harmilla kraepelini* from Cameroon. Within this candidate, both subspecies appear reciprocally monophyletic but with only 2 samples each we cannot conclude whether this is consistent. The other two candidates correspond with *C. harmilla micans* from eastern DR Congo, and because both constitute samples from the same locality (Mt. Hoyo) and are genetically close (0.6% divergence) we believe they represent a single species. Biogeographical patterns do not exclude isolation by distance, but given the large geographical distance between these subspecies we assume they are de facto reproductively isolated lineages. In conclusion, morphology (e.g. wing shape) as well as biogeography corroborate specific status of *C. micans* (**stat. nov.**).

Specimens identified as *Cymothoe excelsa* were non-monophyletic and appear in three separate candidate species, only one of which could be corroborated by independent lines of evidence, suggesting it warrants specific status. This candidate species merges specimens identified as *C. excelsa* and *C. anitorgis*, from Nigeria, Cameroon, Gabon and DR Congo that exhibit good morphological uniformity in both males and especially females. In addition, the candidate appears to be panmictic based on the DNA barcodes, further corroborating its specific status. *Cymothoe anitorgis* is the oldest name and therefore has priority. However, while the male type specimen of *C. excelsa* (from Cameroon) closely matches two sampled males from Gabon and Kivu, eastern DR Congo, the female type specimen of *C. anitorgis* (from Gabon) does not match any of the sampled females morphologically (see Figure 2.11). In addition, all Cameroonian specimens identified as *C. excelsa* occur in this candidate species, suggesting that the type may indeed be included. However, pending more information about the affinities of *C. anitorgis* we cannot conclude whether the name *C. excelsa* is the correct species name, or that it should be a synonym of *C. anitorgis* (Vande weghe 2010). In any case, because we associate the name *C. excelsa* with this candidate species only, it is clear that other candidates including this name require another. But given the general taxonomic disarray of many species in section *Aramis* we cannot even speculate about their delineation and which names would be appropriate.

2.3.5 Candidates requiring more data

In addition to the confirmed candidate species described above, we have found 18 candidates that could represent good species but for which final confirmation of their status must await more material.

Cymothoe fontainei was split into 4 different candidate species, mainly due to its level of intraspecific COI sequence divergence (0.8%) which exceeds the distance

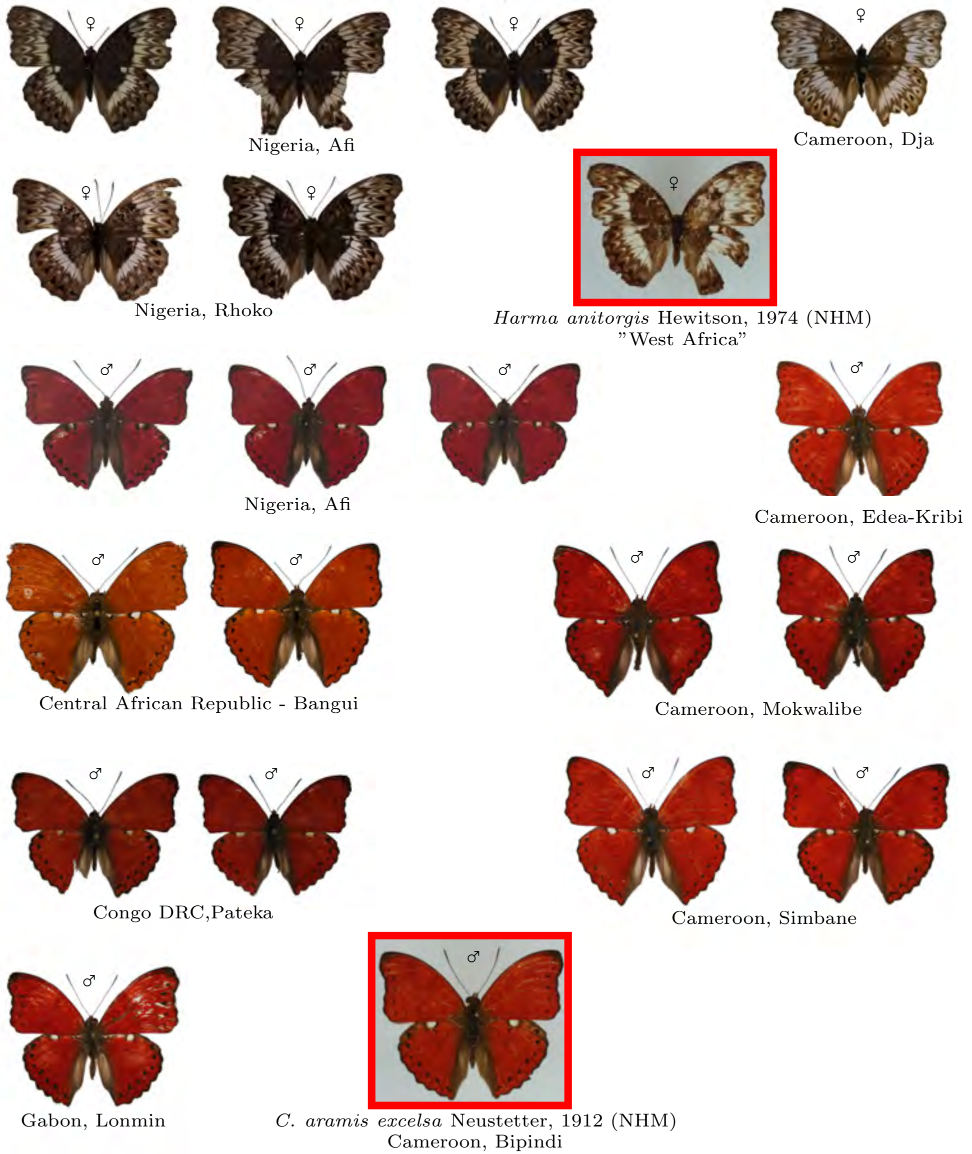


Figure 2.11: Candidate species 04. Morphology of specimens identified as *C. excelsa* and *C. anitorgis* that are merged in a single candidate species. Red outlined images show holotypes(not sampled).

to closely related *C. collarti* (0.6%; i.e. there is no barcode gap). We note, however, that our *C. collarti* samples are from subspecies *werneri* only. Therefore, the COI distance to the nominate remains unknown. One candidate species corresponds with *C. fontainei debauchei* Overlaet, 1952 from Cameroon and Gabon suggesting that there is morphological and biogeographical corroboration. The other 3 candidate species correspond with the nominate *C. fontainei fontainei* from DR Congo, however, and all samples come from the same locality (Bobeta). These candidates are therefore not corroborated. Because the *C. fontainei debauchei* candidate is nested within the genetic variation of the nominate we cannot conclude whether it warrants specific status or not.

There were two candidate species merging specimens identified as *C. lurida* and *C. colmanti* One of these merges all specimens of West African endemic *C. lurida lurida* with some specimens identified as *C. lurida* or *C. colmanti* from Cameroon and the Central African Republic. Indeed, females of *C. lurida lurida* and *C. colmanti* are very similar; the males, however, are quite different. Biogeography suggests panmixis between the West African and Cameroonian populations and possibly isolation by distance of the population in the Central African Republic. Given the genetic as well as morphological divergence between the two candidate species it seems clear that they are distinct. However, whether *C. colmanti* is conspecific with *C. lurida lurida* within this candidate remains unclear.

East African endemic *C. coranus* was split into 4 candidate species, mainly because of its extraordinary intraspecific COI sequence divergence (3.3% K2PD). Three of the candidate species split the nominate *C. coranus coranus* occurring along the coast, suggesting that they could reveal cryptic species. The fourth candidate represents a single specimen of *C. coranus murphyi* Beaurain, 1988 from eastern Malawi (Beaurain 1988), providing morphological and biogeographical corroboration. On the other hand, as the species is widespread in East Africa, significant population-level structuring can be expected and divergences could be conspecific. Moreover, because most candidate species comprise single samples it is not clear whether these patterns are consistent. Additional samples are needed from throughout its distribution to be able to test these competing hypotheses.

Tanzanian mountain endemics *C. aurivillii* and *C. amaniensis* were merged in a single candidate species and constitute a morphologically homogeneous group. In addition, their morphological differences and available sequences do not refute isolation by distance. However, we sampled only few specimens and some closely related species were not sampled (i.e. *C. vumbui* Bethune-Baker, 1926 and *C. zombana* Bethune-Baker, 1926), suggesting that these patterns may change with increased sampling effort.

C. aubergeri was split into 3 candidate species, mainly because of its extraordinary intraspecific COI sequence divergence (3.1% K2PD). Surprisingly, all three candidates contain specimens from the same locality (e.g. Kakum, Ghana), and are hence sympatric. Given the morphological uniformity of *C. aubergeri* they would

therefore suggest morphologically cryptic sympatric species. Such a hypothesis cannot be confirmed based on the currently available data, however. In addition, the morphologically similar *C. adela* (endemic to Guinea, Sierra Leone, Liberia and western Ivory Coast) is nested within *C. aubergeri*'s intraspecific variation. *C. adela* was sampled only once, however, and we therefore cannot confirm its monophyly nor intraspecific divergence. Pending more data (in terms of samples as well as detailed morphological and ecological information), our hypothesis is that *C. adela* and *C. aubergeri* are conspecific. The fact that a female *C. aubergeri* was collected in Guinea confirms that the two taxa may represent intraspecific variability. Still, the high intraspecific divergence is difficult to reconcile with the fact that *C. aubergeri* and *C. adela* are not common and widespread but endemic to specific regions within West Africa. Possibly, this reflects 'local' increased COI substitution rates caused by particular demographic changes in these species.

Cymothoe sangaris was split into 7 candidate species. The first candidate merges some specimens of *C. sangaris* with *C. hobarti*. However, the candidate does not have a barcode gap and we could find no morphological characters supporting it. Instead, we find that within this candidate, the *C. hobarti* and *C. sangaris* specimens are reciprocally monophyletic (except for some short sequences which are probably clustering poorly). In addition, while females of *C. hobarti* are white with brown, those within the *C. sangaris* cluster are either red or orange. We therefore reject this merger. The second candidate comprises specimens from West Africa (Guinea, Sierra Leone and Ghana); females being white with brown and black. The third candidate comprises specimens from Cameroon, Gabon, DR Congo and Zambia with highly variable females. Dark-brown females from north-eastern DR Congo (Ituri) matching subspecies *C. sangaris mwami-kazi* Overlaet, 1952, dark-red females from Gabon and DR Congo matching *C. sangaris rubrior* Overlaet, 1945, and white with brown females from Gabon and Cameroon matching *C. sangaris euthalioides* Kirby, 1889 have identical DNA barcode sequences suggesting that they are panmictic. Orange with brown females from Zambia matching subspecies *C. sangaris luluana* Overlaet, 1945 are also included. The fourth candidate comprises specimens from Ghana, Nigeria and Cameroon. Females are white with black and brown and often have an orange tone on the proximal half of all wings. The three remaining candidate species comprise single specimens from Gabon that are genetically similar to each other and to the *C. hobarti* / *C. sangaris* candidate described earlier. As they are all males (which are morphologically indistinguishable) we could not assess morphological differences, but we suspect that they may constitute a single reproductively isolated species.

Cymothoe reginae-elisabethae was split into 2 separate candidate species. The first is a merger of all specimens from DR Congo with 5 females identified as *C. arcuata* and *C. meridionalis* that are morphologically clearly different from *C. reginae-elisabethae*, rejecting its specific status. The second is a merger of specimens with identifications matching 5 different taxonomic species from section *Aramis* (*C.*

arcuata, *C. coccinata*, *C. distincta*, *C. excelsa*, *C. reginae-elisabethae*), despite clear morphological differences in males and especially females, as well as biogeographical structure rejecting isolation by distance. Moreover, the candidate is not monophyletic, nor has a barcode gap. Closer examination of this candidate revealed that it comprises three monophyletic clusters: (i) a cluster corresponding with the species *C. coccinata*; (ii) a cluster containing all specimens that were identified as *C. excelsa* but do not belong to the *C. excelsa* / *C. anitorgis* candidate described above, together with some male specimens identified as *C. distincta* and *C. arcuata*; and (iii) a cluster with male and female specimens identified as *C. distincta* (see below). In addition this candidate species comprises a single male *C. reginae-elisabethae* specimen from Cameroon that is genetically so much diverged from all other members of section *Aramis* in our database, that it must be reproductively isolated. More importantly, this specimen is genetically clearly differentiated from all other *C. reginae-elisabethae* specimens from DR Congo (see previous candidate). Morphologically, it differs from the *C. reginae-elisabethae* type (from Medje, DR Congo) by having white costal spots on the hindwing undersides. Instead, it matches a specimen in NHM labeled by Overlaet as holotype for *C. reginae-elisabethae lomiensis* in 1953, but to our knowledge this name is unpublished and hence unavailable. Formal description of this species requires more samples, including females.

Specimens identified as *C. distincta* occured in 3 different clusters, distributed over 2 candidate species, of which the second is the merger with other species within section *Aramis* (*C. coccinata*, *C. aruata*, *C. excelsa*, *C. reginae-elisabethae*) described above. As both candidate species are genetically well separated (2.7% K2PD), these results strongly suggest that *C. distincta* comprises at least two reproductively isolated lineages. Females of *C. distincta* are tricolorous and similar to those of the confirmed taxonomic species *C. aramis*. The color of the distal part of the forewing is usually thought to be orange in *C. aramis* and red in *C. distincta*, but Amiet (1997) convincingly showed that at least in Cameroon both species are polymorphic (i.e. the female forewing can be either orange or red) and indistinguishable, which is confirmed by our DNA barcode data. Because *C. distincta* males are difficult to distinguish from other species in section *Aramis* (Larsen 2005, Vande weghe 2010) we cannot exclude that the males in the third cluster (combining males identified as *C. excelsa*, *C. distincta* and *C. arcuata*) are misidentified. Therefore, *C. distincta* comprises only two reliable (i.e. containing female specimens) clusters. Surprisingly, these two clusters contain females from the same two localities (Camp Nouna in Gabon and Kwokoro in DR Congo) suggesting that they are sympatric. Corroboration of these clusters requires more material, however.

2

Table 2.2: Corroboration of candidate species.

Candidate	#	MWD	MBD	Barcode gap	Mono-phyly	Taxonomic species	Type	Corroboration
Cand.01	65	0.024	0.027	True	True	*C. fumana* *C. haynae*	merge	reject
Cand.02	5	0.002	0.027	True	True	*C. althea*	confirmed	
Cand.03	25	0.005	0.013	True	True	*C. egesta*	confirmed	
Cand.04	31	0.008	0.007	False	True	*C. anitorgis* *C. excelsa*	merge/split	Yes
Cand.05	26	0.007	0.008	True	True	*C. herminia*	confirmed	
Cand.06	8	0.006	0.031	True	True	*C. altisidora*	confirmed	
Cand.07	31	0.005	0.013	True	True	*C. druryi*	confirmed	
Cand.08	19	0.006	0.008	True	True	*C. arcuata* *C. meridionalis* *C. reginae-elisabethae*	merge/split	reject
Cand.09	64	0.007	0.010	True	True	*C. sangaris*	split	more data
Cand.10	27	0.014	0.065	True	True	*C. theobene*	confirmed	
Cand.11	62	0.005	0.005	True	False	*C. bouyeri* *C. caenis*	merge/split	reject
Cand.12	18	0.003	0.008	True	True	*C. sangaris*	split	more data
Cand.13	23	0.005	0.037	True	True	*C. indamora*	confirmed	
Cand.14	11	0.007	0.022	True	True	*C. owassae* *C. reinholdi*	merge	reject
Cand.15	70	0.028	0.010	False	True	*C. colmanti* *C. hypatha* *C. lurida*	merge/split	reject
Cand.16	2	0.000	0.014	True	True	*C. heliada*	split	
Cand.17	113	0.027	0.011	False	False	*C. arcuata* *C. coccinata* *C. distincta* *C. excelsa* *C. reginae-elisabethae*	merge/split	reject
Cand.18	4	0.000	0.005	True	True	*C. amaniensis* *C. aurivillii*	merge	more data
Cand.19	20	0.005	0.005	False	True	*C. mabillei*	split	more data
Cand.20	31	0.006	0.033	True	True	*C. beckeri*	confirmed	
Cand.21	32	0.008	0.013	True	True	*C. confusa*	confirmed	
Cand.22	14	0.016	0.011	False	True	*C. hesiodotus*	split	Yes
Cand.23	11	0.005	0.038	True	True	*C. zenkeri*	confirmed	
Cand.24	2	0.000	0.008	True	True	*C. jodutta*	split	reject
Cand.25	6	0.002	0.017	True	True	*C. teita*	confirmed	
Cand.26	9	0.003	0.020	True	True	*C. capella*	confirmed	
Cand.27	4	0.008	0.020	True	True	*C. harmilla*	split	Yes
Cand.28	10	0.005	0.023	True	True	*C. ogova*	confirmed	
Cand.29	7	0.008	0.014	True	True	*C. heliada*	split	
Cand.30	1	0.000	0.012	n.a.	n.a.	*C. eris*	split	
Cand.31	6	0.000	0.003	True	True	*C. fontainei*	split	more data

(Continued on next page)

Table 2.2: Corroboration of candidate species, continued.

Candidate	#	MWD	MBD	Barcode gap	Mono-phyly	Taxonomic species	Type	Corroboration
Cand.32	7	0.005	0.030	True	True	*C. hyarbita*	confirmed	
Cand.33	11	0.000	0.004	True	True	*C. oemilius*	split	
Cand.34	18	0.016	0.028	True	True	*C. aramis*	confirmed	
Cand.35	14	0.011	0.016	True	True	*C. jodutta*	split	reject
Cand.36	8	0.008	0.010	True	True	*C. colmanti* *C. lurida*	merge/split	more data
Cand.37	6	0.000	0.017	True	True	*C. amenides*	confirmed	
Cand.38	11	0.002	0.011	True	True	*C. hesiodotus*	split	Yes
Cand.39	5	0.002	0.044	True	True	*C. lambertoni*	confirmed	
Cand.40	3	0.007	0.011	True	True	*C. haimodia*	confirmed	
Cand.41	8	0.002	0.027	True	True	*C. baylissii* ined.	confirmed	
Cand.42	1	0.000	0.017	n.a.	n.a.	*C. magambae*	confirmed	
Cand.43	10	0.005	0.027	True	True	*C. cyclades*	confirmed	
Cand.44	4	0.008	0.009	True	False	*C. lucasi*	split	
Cand.45	6	0.002	0.023	True	True	*C. orphnina*	confirmed	
Cand.46	1	0.000	0.005	n.a.	n.a.	*C. adela*	confirmed	more data
Cand.47	3	0.000	0.019	True	True	*C. coranus*	split	more data
Cand.48	3	0.011	0.009	False	True	*C. aubergeri*	split	more data
Cand.49	7	0.005	0.025	True	True	*C. ochreata*	confirmed	
Cand.50	1	0.000	0.023	n.a.	n.a.	*C. crocea*	confirmed	
Cand.51	1	0.000	0.012	n.a.	n.a.	*C. eris*	split	
Cand.52	4	0.000	0.006	True	True	*C. harmilla*	split	Yes
Cand.53	1	0.000	0.013	n.a.	n.a.	*C. coranus*	split	more data
Cand.54	1	0.000	0.009	n.a.	n.a.	*C. lucasi*	split	
Cand.55	3	0.003	0.026	True	True	*C. consanguis*	confirmed	
Cand.56	3	0.002	0.023	True	True	*C. alcimeda*	split	
Cand.57	1	0.000	0.013	n.a.	n.a.	*C. coranus*	split	more data
Cand.58	1	0.000	0.019	n.a.	n.a.	*C. coranus*	split	more data
Cand.59	1	0.000	0.023	n.a.	n.a.	*C. alcimeda*	split	
Cand.60	2	0.000	0.016	True	True	*C. jodutta*	split	Yes
Cand.61	1	0.000	0.020	n.a.	n.a.	*C. melanjae*	confirmed	
Cand.62	1	0.000	0.007	n.a.	n.a.	*C. excelsa*	split	reject
Cand.63	16	0.005	0.008	True	True	*C. weymeri*	confirmed	
Cand.64	4	0.005	0.008	True	True	*C. distincta*	split	more data
Cand.65	1	0.000	0.005	n.a.	n.a.	*C. caenis*	split	reject
Cand.66	7	0.003	0.006	True	True	*C. caprina* *C. isiro*	merge	reject
Cand.67	53	0.017	0.008	False	True	*C. sangaris*	split	more data
Cand.68	1	0.000	0.005	n.a.	n.a.	*C. sangaris*	split	more data
Cand.69	3	0.000	0.003	True	True	*C. cottrelli*	split	
Cand.70	2	0.000	0.003	True	False	*C. cottrelli*	split	
Cand.71	2	0.000	0.005	True	True	*C. mabillei*	split	more data
Cand.72	22	0.003	0.008	True	True	*C. jodutta*	split	Yes
Cand.73	1	0.000	0.003	n.a.	n.a.	*C. fontainei*	split	reject
Cand.74	2	0.000	0.003	True	True	*C. fontainei*	split	reject
Cand.75	2	0.000	0.003	True	True	*C. fontainei*	split	reject

(Continued on next page)

Table 2.2: Corroboration of candidate species, continued.

Candidate	#	MWD	MBD	Barcode gap	Mono-phyly	Taxonomic species	Type	Corroboration
Cand.76	2	0.002	0.006	True	True	*C. collarti*	confirmed	
Cand.77	1	0.000	0.004	n.a.	n.a.	*C. oemilius*	split	
Cand.78	4	0.000	0.005	True	True	*C. aubergeri*	split	more data
Cand.79	1	0.000	0.005	n.a.	n.a.	*C. aubergeri*	split	more data
Cand.80	1	0.000	0.006	n.a.	n.a.	*C. harmilla*	split	reject
Cand.81	25	0.011	0.005	False	True	*C. hobarti* *C. sangaris*	merge/split	reject
Cand.82	1	0.000	0.005	n.a.	n.a.	*C. sangaris*	split	more data
Cand.83	1	0.000	0.005	n.a.	n.a.	*C. sangaris*	split	more data

MWD = maximum within-species distance; MBD = minimum between-species distance; n.a. = not applicable because only a single specimen was sampled.

2.3.6 Rejected candidates

Sixteen candidate species could be rejected (see Table 2.2):

Cymothoe heliada was split into 2 candidate species, mainly because of its large intraspecific COI sequence divergence (2.6% K2PD). Surprisingly, the candidates contain specimens from the same locality (e.g. Bobeta, DR Congo), and hence are sympatric. Given their morphological uniformity the data would therefore suggest morphologically cryptic sympatric species, but without any corroboration from independent lines of evidence such a hypothesis seems improbable. We think that deep conspecific divergence is a more likely explanation.

Cymothoe caprina was merged with *C. isiro*, because the latter is 'genetically nested' within the first. However, both species are morphologically as well as biogeographically clearly distinct (see Figure 2.12). *Cymothoe caprina claireae* occurs from Cameroon and Gabon to the Central African Republic *C. caprina caprina* in western DR Congo; males are white-creamy and females are red (similar to a miniature version of some *C. sangaris* females). *Cymothoe isiro* occurs in the northeastern part of DR Congo; males are more yellow and have more extensive black markings; females are brown with a narrow transversal band (similar to females of *C. althea*). We therefore conclude that these are probably recently diverged but reproductively isolated species, rather than a single species.

Cymothoe caenis and *C. bouyeri* were merged because they have identical sequences. *Cymothoe bouyeri* was recently described based on differences in wing patterning as well as genital structure (Vande weghe 2011, Bouyer and Ducarme 2013) rejecting a hypothesis of conspecificity. We therefore favor current taxonomy and attribute their merge to lack of divergence leading to identical sequences in the case of *C. caenis* and *C. bouyeri*. Indeed, sequences of the localized *C. bouyeri* are identical to only a subset of the haplotypes within the widespread *C. caenis*,

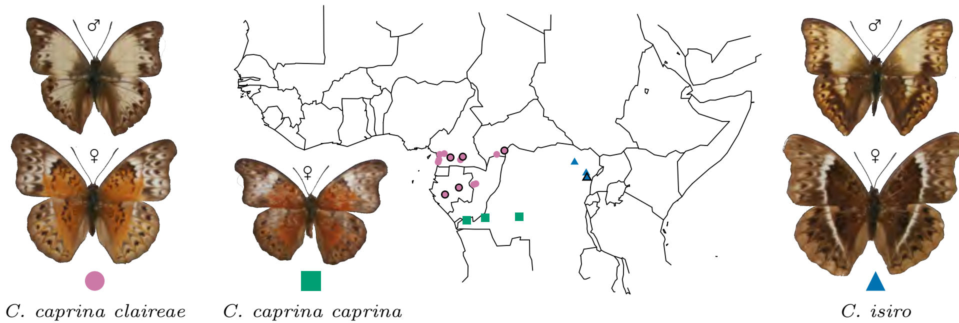


Figure 2.12: *Cymothoe caprina* and *C. isiro*. Specimen photographs and collection localities of *C. caprina claireae* (purple circles), *C. caprina caprina* (green squares) and *C. isiro* (blue triangles); localities of sampled specimens have a black outline.

supporting such a hypothesis. Likewise, the candidate comprising a single *C. caenis* from Nigeria is clearly nested within genetic variation between all other samples.

Cymothoe fumana and *C. haynae* were merged into a single candidate species, despite clear morphological differences in males and especially females, as well as a biogeographical pattern rejecting isolation by distance. Closer examination of this candidate revealed that it comprises two reciprocally monophyletic clusters (see Figure 2.13). Within the first, *C. haynae superba* Aurivillius, 1898 from Cameroon and Gabon appears nested within an apparently panmictic cluster of *C. haynae diphyia* Karsch, 1894 from DR Congo and the Central African Republic, suggesting that they are reproductively isolated. Within the *C. fumana* cluster, two reciprocally monophyletic haplotypes coincide with *C. fumana fumana* from West Africa and *C. fumana balluca* Fox & Howarth, 1968 from Nigeria, Cameroon and Gabon. This overall pattern is largely congruent with current taxonomy except that, surprisingly, some *C. haynae diphyia* have a barcode identical to the *C. fumana balluca* haplotype. A possible explanation for this pattern is that the haplotypes may be ancestral polymorphisms that have been retained within *C. haynae diphyia*. However, this fails to explain why we consistently recovered the *C. fumana balluca* haplotype and not the *C. fumana fumana* haplotype. In addition, it begs the question why the same ancestral polymorphism was not retained within *C. fumana*, a species with an arguably wider distribution and hence perhaps increased population structuring (e.g. across the Dahomey Gap) than *C. haynae*. An alternative, better explanation is introgression of the mitochondrial haplotype from *C. fumana balluca* into *C. haynae diphyia*. As the two taxa have a reasonable biogeographic zone of overlap in Cameroon, Gabon and possibly in the Republic of the Congo, occasional hybridization is not unlikely. And given the apparent panmixis within *C. haynae diphyia* it is not surprising that the introgressed haplotype currently occurs throughout DR Congo, including populations outside the zone of overlap. *C. haynae superba* appears to be

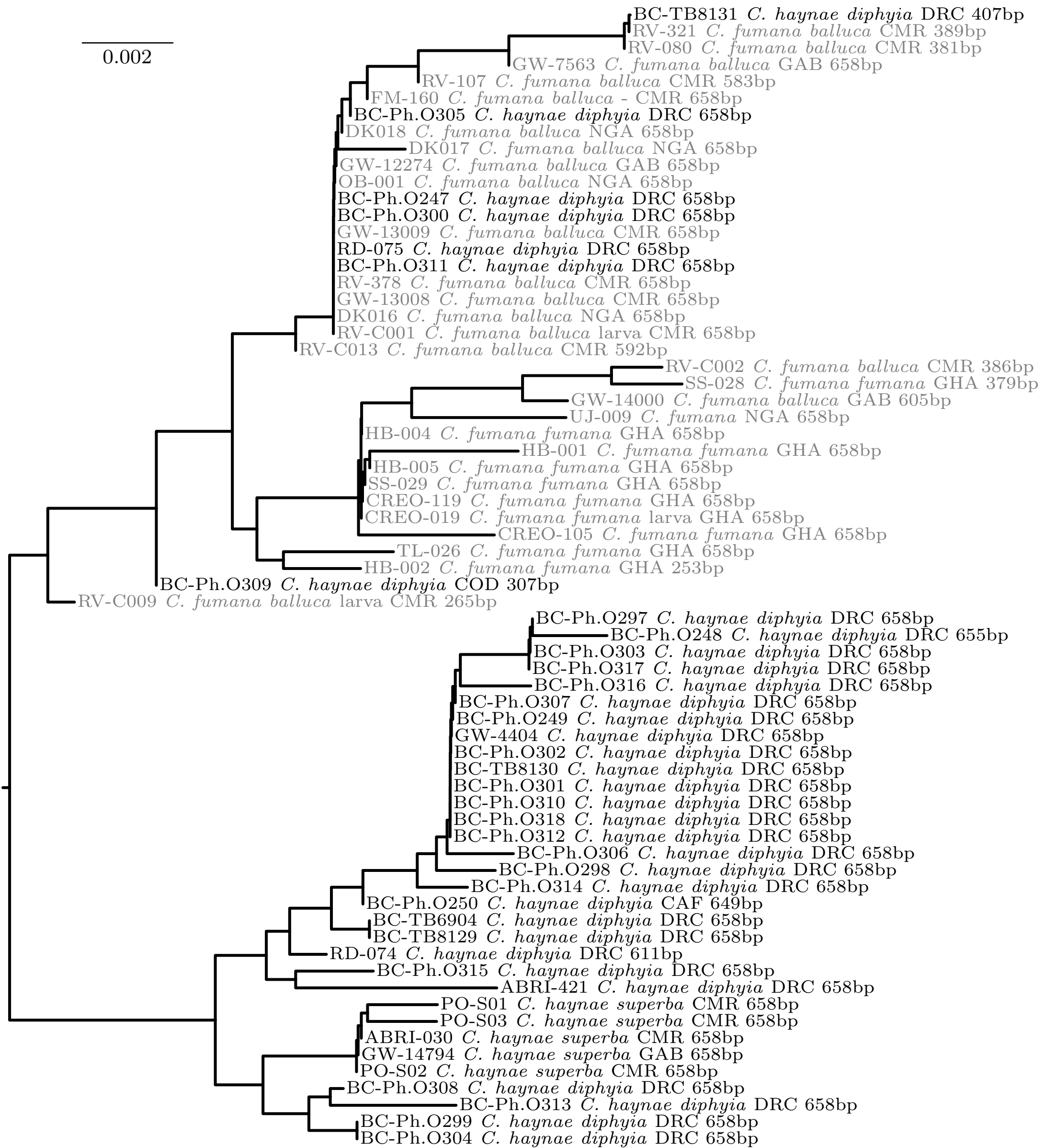


Figure 2.13: Patterns of mitochondrial introgression from *C. fumana* to *C. haynae*. DNA barcode NJ clustering of *C. fumana* (grey) and *C. haynae* (black) samples. Scalebar indicates genetic distance. CMR = Cameroon; DRC = Democratic Republic of the Congo; GAB = Gabon; GHA = Ghana; CAR = Central African Republic; NGA = Nigeria. Note that some *C. haynae* samples contain *C. fumana* haplotypes indicating possible introgression.

reproductively isolated from *C. haynae diphyia*, which is hardly suprising given the striking morphological differences in especially the females. However, as *C. haynae superba* is nested within *C. haynae diphyia* and we do not have samples of the latter subspecies from Cameroon, we cannot rule out an isolation-by-distance pattern.

Cymothoe reinholdi and *C. owassae* were merged into a single candidate species. They are morphologically similar but nevertheless distinct enough to warrant their specific separation. The candidate species shows clear clustering based on DNA barcodes but does not conform to panmixis nor isolation by distance, suggesting that it does not correspond with a single separately evolving lineage (i.e. species). Non-monophyly and absence of a barcode gap appears to be due to large intraspecific COI sequence diversity within *C. reinholdi* (0.7% K2PD) compared with *C. owassae*, suggesting that its coalescent predates the time when *C. owassae* became reproductively isolated (i.e. incomplete lineage sorting). Such explanation seems biogeographically plausible because, while *C. reinholdi* is widespread, occurring from Nigeria to eastern DR Congo, *C. owassae* is strictly endemic to the island of Bioko. Newly collected material of *C. reinholdi* from a mountain in West Cameroon is now available at ABRI and may shed more light on the COI divergence patterns between *C. reinholdi* and *C. owassae*.

There were two candidate species each merging specimens identified as *C. lurida* and *C. colmanti*. The first merges West African *C. lurida lurida* together with *C. colmanti* from Cameroon and the Central African Republic (see previous section). The second merges various morphologically well-characterized subspecies of *C. lurida*, with *C. colmanti* as well as with morphologically distinct *C. hypatha* Hewitson, 1966. It encompasses large COI sequence variation (2.8%) and does not have a barcode gap (divergence from the first candidate is 1%), suggesting that it does not represent a single species. Indeed, we found correlations between morphological characters and clusters in our haplotype tree, suggesting at least four species: First, all specimens identified as *C. lurida hesione* Weymer, 1907 and *C. lurida tristis* Overlaet, 1952 cluster together and appear panmictic. Males of both subspecies have a rather orangey yellow ground colour and a black marginal band that widens dramatically in the hind wings. Females are brown with a single white subapical band on the forewings that is broad in the typical *C. lurida hesione* and much narrower in typical *C. lurida tristis*, but all intermediate phenotypes can be found even in single series of specimens, suggesting that they are a single taxon. Being the oldest name, *C. lurida hesione* is the appropriate name for this taxon. Second, all specimens identified as *C. lurida butleri* Grünberg, 1908 from Kenya, Tanzania and Uganda cluster together. They are genetically more similar to *C. hypatha* than to *C. lurida hesione*, suggesting that they represent a separate taxon. Indeed, *C. lurida butleri* is morphologically well defined and we found that its caterpillars have a characteristic bright blue dorsolateral band (see Figure 2.14). Specimens identified as *C. hypatha* are divided over two separate clusters, with all *C. hypatha hypatha* from eastern Nigeria, Cameroon, Gabon and DR Congo clustering together and closer to *C. lurida*

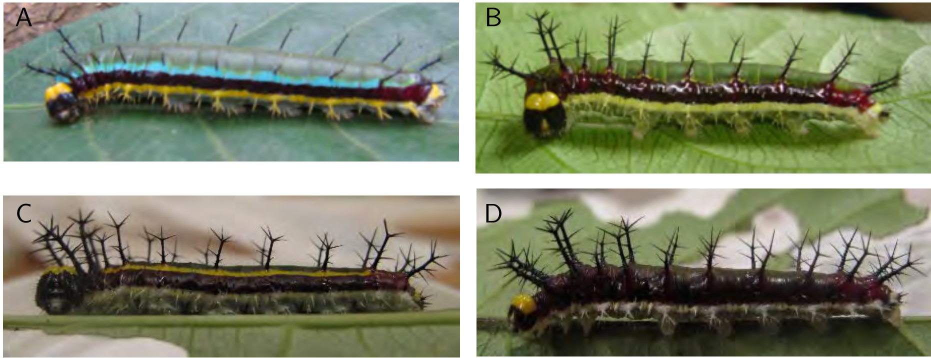


Figure 2.14: Larval morphology of *Cymothoe* in section Lurida. (A) *Cymothoe butleri* found on *Rinorea brachypetala*in Kenya; (B) *C.* cf. *colmanti* found on *R. batesii* in Cameroon; (C) *C. nigeriensis* found on *R. rubrotincta* in Nigeria; (D) *C. okomu* found on *R. welwitschii* in Nigeria.

butleri than to *C. hypatha okomu* Hecq & Larsen, 1997 from the Niger Delta in Nigeria. Despite these general patterns, however, numerous specimens identified as *C. lurida* or *C. colmanti* appear to cluster randomly with most of these clusters. Even worse, DNA barcodes of some of these specimens are identical to those of *C. hypatha hypatha*, despite clear morphological differences. Obviously, this complicates taxonomic interpretation. We are nevertheless confident in the specific status of *C. hypatha okomu*. While the male does not appear to differ significantly from the nominate subspecies, the female differs from the nominate in many respects, but mainly in its much lighter colour and broader wing margins that are almost white (Hecq and Larsen 1997). The nominate subspecies *hypatha* occurs from the Cross River in eastern Nigeria through Cameroon to Gabon, the Central African Republic, and most of northern DR Congo (Larsen 2005, Figure 2.15). Two of the nominate specimens from Cross River in eastern Nigeria (DK_015 and RW_050) are only 300 km away from the Niger Delta (see Figure 2.15). Their clear genetic separation from *C. hypatha okomu* therefore refutes an isolation-by-distance explanation and suggests long-term reproductive isolation. This is further corroborated by the clear morphological differentiation between the females. In addition, the larvae of *C. hypatha okomu* lack the yellow dorsolateral line and the yellow coloration of the 13th tergite present in the nominate (Amiet 2000; see Figure 2.15). We therefore raise *C. okomu* to species level (**stat. nov.**).

Candidate species merging (i) C. sangaris with C. hobarti; (ii) C. arcuata, with C. coccinata, C. distincta, C. excelsa, and C. reginae-elisabethae (from Cameroon); and (iii) C. arcuata, with C. meridionalis, and C. reginae-elisabethae (from DR Congo), as well as candidates splitting some specimens of C. ciceronis, C. excelsa, C. fontainei, C. jodutta, and C. harmilla could also be rejected (see previous sections).

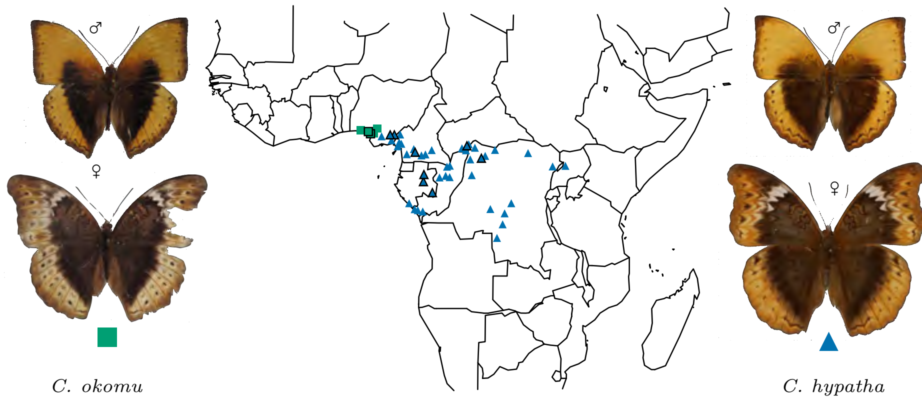


Figure 2.15: *Cymothoe hypatha* and *C. okomu*. Specimen photographs and collection localities of *C. hypatha* (blue triangles), and *C. okomu* (green squares); localities of sampled specimens have a black outline.

2.3.7 Additional taxonomic findings

In addition to corroborating candidate species, our systematic evaluation of type specimens and morphological characters revealed additional taxonomic implication, noted below.

First, we note that *C. serpentina* Kirby, 1889 was synomized with *C. harmilla* in Ackery *et al.* (1995), but that it clearly matches *C. ogova* and should be reclassified as synonym of *C. ogova* (**syn. nov.**). *C. congoensis* Suffert, 1904 was incorrectly synonymized with *C. lurida hesione* Weymer 1907 but closely matches *C. hesiodotus* Staudinger, 1890 with which we deem it synonymous (**syn. nov.**). Similarly, the female aberration *C. similis* Neustetter 1912 was synonymized with *C. reginae-elisabethae belgarum* but is clearly synonymous to *C. excelsa excelsa* / *C. anitorgis* (corroborated by a det slip by Overlaet from 1953 on the type specimen in the Natural History Museum, London).

In addition, we found the type specimen of the currently recognized species *C. rebeli* Neustetter, 1912 in the Natural History Museum, London, to closely match *C. altisidora,* with which we deem it synonymous (**syn. nov.**).

Three confirmed taxonomic species were split into multiple candidate species. According to our taxonomic pipeline, these taxonomic species are retained, but the candidate species may nevertheless warrant further examination:

Cymothoe oemilius was split into 2 candidate species, with all specimens from the mainland (Nigeria, Cameroon, Gabon, DR Congo) having identical sequences and that from Bioko (*C. oemilius fernandina* Hall, 1929) differing by 2 basepairs (bp).

Obviously, such biogeographical pattern refutes an isolation by distance scenario, suggesting reproductive isolation. However, as we have sampled a single *C. oemilius fernandina* specimen only, we cannot assess whether these genetic differences are fixed. Confirmation of the status of *C. oemilius frederica* Distant, 1880, a morphologically divergent subspecies from eastern Cameroon, also requires additional material.

C. mabillei was split into 2 candidate species, separating specimens from Ghana, Sierra Leone and Nigeria from those from Guinea. Because the Guinean specimens are genetically different from those from nearby Sierra Leone (by 2bp) and all other specimens differ by up to a single bp, biogeographical patterns reject an isolation-by-distance hypothesis, suggesting reproductive isolation. However, having only two partial sequences from Guinea we could not assess whether this pattern is consistent, and correlates with morphological characters. Because the type locality of *C. mabillei* is in Ivory Coast, if these candidate species are confirmed, the Guinean taxon would require a new name. In any case, our results confirm that *C. mabillei*, previously considered to occur from Sierra Leone to Ghana and not crossing the Volta River (Larsen 2005) is also present in western Nigeria. A specimen at MZUJ suggests that it is also present in the Central African Republic but we could not amplify DNA barcodes of this specimen for confirmation (it could also be mislabelled, however).

The South African battling glider *Cymothoe alcimeda* was split into two candidate species, corresponding with subspecies *C. alcimeda trimeni* Aurivillius, 1912 from the Eastern Cape and Kwazulu Natal and *C. alcimeda transvaalica* Rydon, 1994 from Transvaal suggesting morphological and biogeographical corroboration. However, as *C. alcimeda transvaalica* was sampled only once and additional subspecies were not sampled at all (i.e. subsp. *alcimeda* from the Western Cape, subsp. *clarki* Stevenson, 1934 from the Eastern Cape and Kwazulu-Natal, subsp. *marieps* Rydon, 1994 from Transvaal, and subsp. *rhodesiae* Stevenson, 1934 from Zimbabwe), we are not certain whether these patterns are consistent with an isolation by distance scenario and which candidate corresponds with the nominate.

Finally, we note that *C. egesta* requires taxonomic reevaluation. After long being considered conspecific, taxonomic species *C. egesta* and *C. confusa* were finally separated based on data from DNA barcodes and morphology of immature stages (Amiet 1997, McBride, van Velzen *et al.* 2009). Caterpillars of *C. confusa* in Cameroon and Tanzania have a yellow dorsolateral stripe, an extensive black mask and feed on *Rinorea ilicifolia* (Welw. ex Oliv.) Kuntze and allied species. The pupa has a blue or purple dorsal keel. Contrastingly, caterpillars of *C. egesta* in Cameroon feed on the closely related host plants *R. lepidobotrys* Mildbr.and *R. breviracemosa* Chipp and have a white dorsolateral stripe and a less extensive black mask. The pupa has a dorsal keel that is red on the thorax. Based on these observations, McBride *et al.* (2009) predicted that *C. breviracemosa* and *R. lepidobotrys* are host plants of *C. egesta* also in West Africa. Surprisingly, however, we found *C. egesta* caterpillars to feed on *R. ilicifolia* in Ghana (see Chapter 6). Moreover, their larval morphology closely matched that of *C. confusa* (see figure 2.16).

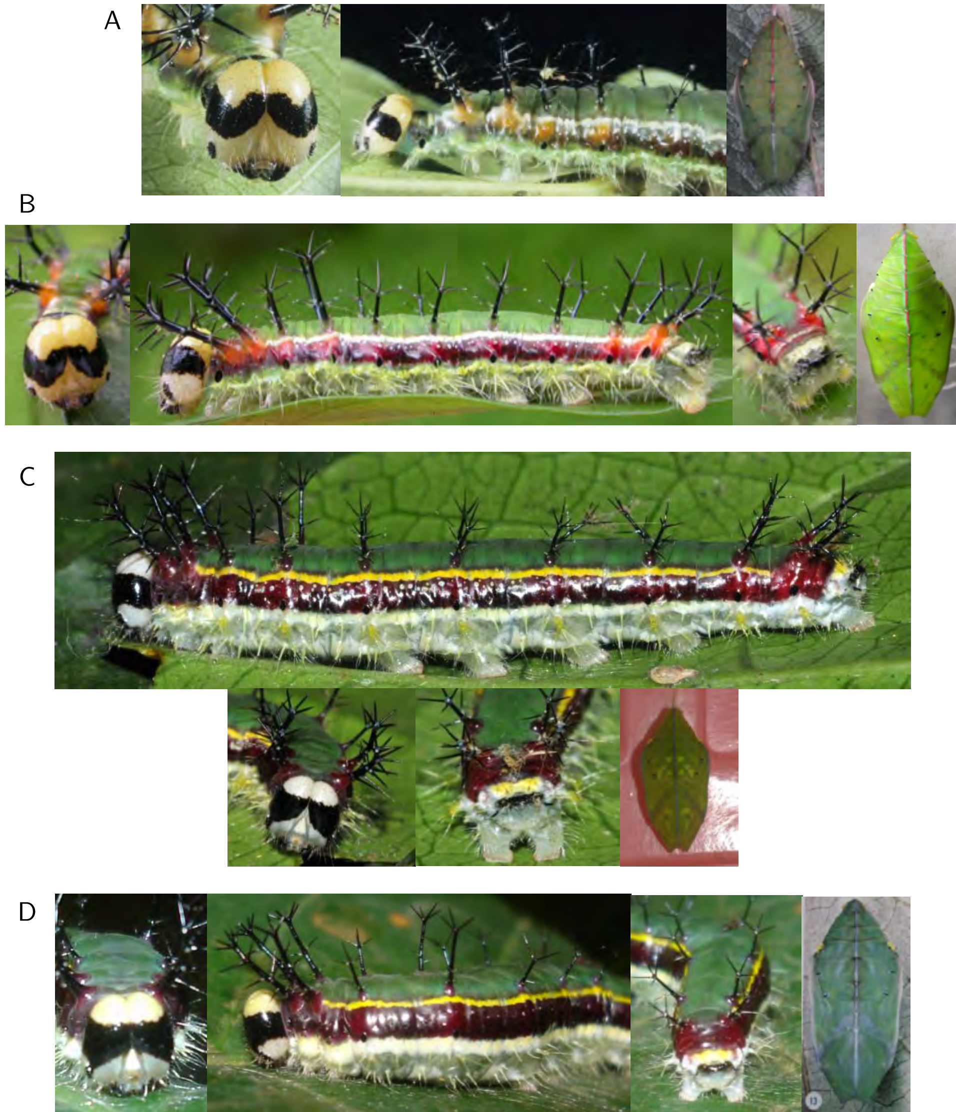


Figure 2.16: Larval morphology in *Cymothoe egesta* and *C. confusa*. (A) *Cymothoe egesta* "orange" found on *Rinorea lepidobotrys* in Cameroon; (B) *C. egesta* "orange" found on *R. breviracemosa* in Cameroon; (C) *C. egesta egesta* found on *R. ilicifolia* in Ghana); (D) *C. confusa* found on *R. ilicifolia* in Cameroon.

Table 2.3: List of revised taxa.

Original taxonomic name	Previous classification	Revised classification
Harma ciceronis Ward, 1871	*C. jodutta ciceronis*	*C. ciceronis*
C. ehmckei Dewitz, 1887	*C. jodutta ehmckei*	*C. ciceronis ehmckei*
C. jodutta mostinckxi Overlaet, 1952	*C. jodutta mostinckxi*	*C. ciceronis mostinckxi*
C. harmilla micans Bouyer & Joly, 1995	*C. harmilla micans*	*C. micans*
C. lurida var. nigeriensis Overlaet, 1952	*C. hesiodotus nigeriensis*	*C. nigeriensis*
C. hypatha okomu Hecq & Larsen, 1997	*C. hypatha okomu*	*C. okomu*
C. serpentina Kirby, 1889	syn. *C. harmilla*	syn. *C. ogova*
C congoensis Suffert, 1904	syn. *C. lurida hesione*	syn. *C. hesiodotus*
C. rebeli Neustetter, 1912	*C. rebeli*	syn. *C. altisidora*

Our DNA barcodes confirm that these specimens are indeed *C. egesta*, refuting the hypotheses of McBride *et al.* (2009). Apparently, *C. egesta* and *C. confusa* immature stages and host plant associations are largely homogeneous throughout tropical Africa, the *C. egesta* populations feeding on *R. lepidobotrys* and *R. breviracemosa* in Cameroon being the only exception. Given the clear morphological differences in immature stages, we therefore reinstate the original hypothesis by Amiet (1997) that these populations (named *C. egesta* "orange" due to the subtly more orange tone of the males) represent a distinct species, and that *Cymothoe megaesta* Staudinger, 1890 (holotype from the Barombi station in Cameroon) is a likely candidate.

2.3.8 Identification of immature specimens

Based on DNA barcodes we could identify 15 eggs and 27 larvae. Immature specimens from Ghana were identified as *C. aubergeri* [found on *Rinorea angustifolia* (Thouars) Baill. subsp. *engleriana* De Wild. & T.Durand], *C. egesta* [*R. ilicifolia* (Welw. ex Oliv.) Kuntze], *C. fumana* [*R. oblongifolia* (C.H.Wright) Marquand ex Chipp], *C. mabillei* [*R. oblongifolia*], *C. sangaris* [*R. longicuspis* Engl.] and *Harma theobene* [*Caloncoba* Gilg]; those from Kenya were identified as *C. butleri* and *C. hobarti* [both *R. brachypetala* (Turcz.) Kuntze]; those from Cameroon were identified as *C. colmanti* [*R. batesii* Chipp], *C. coccinata* [*R. dentata* (P.Beauv.) Kuntze, *R. yaundensis* Engl. and *R. zenkeri* Engl.], *C. confusa* [*R. ilicifolia* and *R. dewitii* Achound.], *C. egesta* [*R. breviracemosa* and *R. lepidobotrys*], *C. excelsa* [*R. oblongifolia*], *C. fumana* [*R. oblongifolia*], *C. sangaris* [*R. batesii* and *R. preussii* Engl.] and *Harma theobene* [*Lindackeria schweinfurthii* Gilg]; those from Nigeria were identified as *C. nigeriensis* [*R. rubrotincta* Chipp], *C. okomu* [*R. welwitschii* (Oliv.) Kuntze], *C. sangaris* [*R. rubrotincta*, *R. welwitschii*]; one from Gabon was identified as *C. lucasi* [*R. gabunensis* Engl.]; those from Guinea were identified as *C. mabillei* [*R. microdon* M.Brandt].

2.4 Discussion

2.4.1 Integrative taxonomic decision pipeline

Our integrative taxonomic decision pipeline integrating DNA barcode sequence data with morphology and biogeography has proven highly suitable for solving taxonomic problems in *Cymothoe*. Using our sequence of 'confirmation' (with existing taxonomy), 'delimitation' (using ABGD), and 'corroboration' (using biogeographical and morphological evidence) steps resulted in solving several taxonomic issues and to an increase of 5 species to a total of 83 species in our updated taxonomy. Four synonyms were reclassified in the process, see Table 2.3. In addition, our expanded DNA barcode data set allowed for the identification of 42 immature specimens from six different countries, significantly increasing the data on *Cymothoe* host plant associations. Our results confirm that *Cymothoe* in general, and sections *Aramis*, *Lurida*, and *Sangaris* in particular, are difficult to interpret taxonomically even with the help of DNA barcodes. We therefore expect that our integrative taxonomic pipeline will perform even better when applied to groups of older species.

Previously published pipelines for integrative taxonomy focus mainly on the theoretical considerations and possible alternative routes for data integration (Padial, Miralles *et al.* 2010, Schlick-Steiner, Steiner *et al.* 2010, Puillandre, Modica *et al.* 2012). In comparison, our pipeline aims at being a practical implementation that can be readily applied to other taxa. Importantly, we integrate current taxonomic classification in our pipeline as an a priori species hypothesis, which has three major implications. First, this means that our pipeline is conservative because taxonomic species that can be confirmed based on independent lines of evidence will be retained even when a particular delimitation method suggests otherwise. Within taxonomy some conservatism is requested, because taxonomic names are the main vehicle for transmitting information about biological entities. Consequently, it must be clear what the names mean and changing this meaning based on only weak evidence is undesirable because it impairs effective transmission of information (McNeill 2000). Second, it means that our pipeline allows researchers to quickly identify and focus on the problematic cases, without revisiting all species in a clade. In our case, for example, 29 candidate species based on DNA barcodes corresponded with taxonomic species and therefore did not require any additional corroboration. As corroboration from independent lines of evidence can require data that may be difficult to get as well as careful interpretation of possible conflicting signals, this saves valuable time and effort. Third, using taxonomic species as a priori hypotheses facilitates assessment of the relative accuracy of the delimitation method chosen. Precisely because a large number of candidate species delimited by our ABGD analyses corresponded with taxonomic species, we have more confidence that these delimitations are generally in line with biological reality.

Our pipeline is flexible and allows implementation of various methodologies in each individual step. For example, in the confirmation step, we applied reciprocal monophyly and the existence of a barcode gap as criteria. Other criteria such as a simple distance threshold (Hebert, Stoeckle *et al.* 2004) or Statistical Parsimony (Templeton, Crandall *et al.* 1992) could also be used. In the delimitation step, we applied the recently developed ABGD method to delimit candidate species. Other methods based on population genetic models (Pons, Barraclough *et al.* 2006, Monaghan, Wild *et al.* 2009, Yang and Rannala 2010, Zhang, Zhang *et al.* 2011) are computationally much more expensive and were found to produce results qualitatively similar to those based on ABGD,(Jorger, Norenburg *et al.* 2012, Puillandre, Modica *et al.* 2012) but they could replace or supplement the ABGD analysis in our pipeline. In any case, the fact that *Cymothoe* candidate species based on ABGD largely correspond with a priori species hypotheses (themselves highly corroborated) suggests that it is an accurate method. In the corroboration step we used biogeography and morphology to corroborate candidate species, but our pipeline can equally include data from e.g. biochemistry, ecology, behavior, etc., when available.

2.4.2 Niger Delta endemism

Of the five taxa we raised to species level, two are endemic to the Niger Delta in Nigeria (*C. nigeriensis* and *C. okomu*). A number of other butterfly taxa are also restricted to this area, namely *Acraea actinotina* Lathy, 1903, *Euptera nigeriensis* Chovet, 1998, and possibly *E. knoopii* Libert & Chovet, 1998. Our finding thus confirms the status of the Niger Delta as an important region of endemism, which is also home to all of Nigeria's endemic or near-endemic mammal species (Powell 1997). However, its biodiversity is under threat due to pollution by the oil and gas industry as well as to increasing population pressure causing habitat degradation and unsustainable harvest of biological resources (Oates John, Bergl Richard *et al.* 2004, Phil-Eze and Okoro 2009, Federal Ministry of Environment, United Nations Development Programme *et al.* 2011).

2.4.3 Technical issues

Our DNA barcoding efforts generally faced low amplification and sequencing success, which we attribute to a combination of DNA degradation and primer-template mismatches. Most of our sampling comes from archival specimens in collections. These specimens were often originally collected in remote field locations under warm and humid conditions and, when finally dried were rehydrated before mounting. These conditions clearly favor growth of fungal and microbial organisms and degradation of DNA. Nevertheless, some samples that were recently collected under conditions suitable for DNA preservation also failed to amplify. Primer mismatches resulting in low annealing temperatures offer an additional explanation for our low sequencing

success in these cases. Indeed, we found multiple mismatches in nearly all internal primers used, indicating parts of the DNA barcode locus that are considered conserved in Lepidoptera are variable in *Cymothoe* (not shown). Such mismatches may have contributed to the low sequencing success rates and high incidence of contaminated sequence. Primers designed specifically for *Cymothoe* may improve success rates but are in principle not part of the highly standardized DNA barcoding approach using "universal" primers (Hebert, Cywinska *et al.* 2003), although primer pairs optimized for several main clades are indeed generally used for routine DNA barcoding (e.g. Hebert, Penton *et al.* 2004).

2.4.4 Incomplete lineage sorting and introgression

In general, DNA barcodes (as well as most other DNA markers) are expected to work well for species that have been reproductively isolated for a long time. The genus *Cymothoe*, however, is an example of a recent and rapidly diversified genus and most species are relatively young (van Velzen, Wahlberg *et al.* 2013). In such cases, the coalescent of haplotypes may be much older than the species containing them, leading to patterns of incomplete lineage sorting (Nichols 2001, Funk and Omland 2003, Ross, Murugan *et al.* 2008, van Velzen, Weitschek *et al.* 2012). This is clearly reflected in our results, where low rates of taxonomic species confirmation are mainly due to lack of interspecific divergence. Indeed, our data suggest that, for at least 6 species, DNA barcodes are not yet sorted. Given the ecological uniformity within *Cymothoe*, widespread species potentially have larger effective population sizes, and hence need more time to get sorted. Indeed, barcodes from *Cymothoe* species with a narrow distribution are usually nested within those with a wide distribution (e.g. *C. owassae* within *C. reinholdi*), confirming this hypothesis. Possibly, character-based DNA barcode matching methods could reliably diagnose such non-monophyletic species (DeSalle, Egan *et al.* 2005, Rach, DeSalle *et al.* 2008, van Velzen, Weitschek *et al.* 2012). Absence of DNA barcode monophyly has been reported for butterflies (Elias, Hill *et al.* 2007, Wiemers and Fiedler 2007) as well as other insects (Meier, Shiyang *et al.* 2006, Boyer, Baker *et al.* 2007, Whitworth, Dawson *et al.* 2007, Trewick 2008).

In some cases, different morphologically and taxonomically well-defined *Cymothoe* species turned out to have identical DNA barcodes, indicating that DNA barcoding resolution limits are reached (Trewick 2008). There are two possible explanations for identical barcodes in different species. First, species may have diverged only recently, so that they have not experienced enough time in reproductive isolation to acquire differences in the COI DNA barcode locus (Shaffer and Thomson 2007, van Velzen, Weitschek *et al.* 2012). Given their morphological similarities, this seems a likely explanation for sharing of identical sequences between *C. caenis* and *C. bouyeri*, and between *C. aurivillii* and *C. amaniensis*.

A second explanation for shared haplotypes is introgression. When species occasionally hybridize, repeated backcrossing of an interspecific hybrid with one of its parent species can lead to the movement of a gene (gene flow) from one species into the gene pool of another (e.g. Schmidt and Sperling 2008). This seems a likely explanation for the identical barcodes shared between morphologically and genetically divergent species such as *C. lurida* and *C. hypatha* as well as *C. fumana* and *C. haynae*. In both cases, the gene flow appears to be asymmetric with backcrossing into only one of the two species involved (i.e. into *C. lurida* and *C. haynae*), further supporting a hypothesis of introgression rather than incomplete lineage sorting. Because the DNA barcode locus is maternally inherited, hybridization seems to be predominant between females of *C. hypatha* and males of *C. lurida* as well as between females of *C. fumana* and males of *C. haynae*. Assuming random sampling, introgressed DNA barcodes have a high incidence in both recipient species (21% in *C. haynae*, 18% in *C. lurida*), which is surprising given their consistent morphological differentiation from the donor species. The most likely explanation for this high incidence of introgressed mitotypes is that rare hybridization events, although producing very little gene flow of nuclear genes, has led to the transfer of parasitic *Wolbachia* α-proteobacteria (Raychoudhury, Baldo *et al.* 2009). *Wolbachia* are maternally transmitted and maximize their spread of infection by inhibiting development of male offspring (Stouthamer, Breeuwer *et al.* 1999). Despite initially being in poorly adapted hybrid individuals, *Wolbachia* can consequently quickly spread through the recipient species. And because *Wolbachia* and mitochondria are co-inherited, the associated mitochondrial haplotypes spread with it (Jiggins 2003, Hurst and Jiggins 2005). Indeed, in both recipient *Cymothoe* species we found multiple instances of the same DNA barcode haplotype suggesting that such spread may occur. On the other hand, in both cases the mitotype has not replaced all the original mitotypes in the recipient species suggesting that there still is selection against the mitotype or the infection. However, a *Wolbachia*-mediated genetic sweep could explain the identical barcodes in *C. arcuata*, *C. excelsa*, and *C. distincta* mentioned above.

Wolbachia infect a high proportion of insects and *Wolbachia*-mediated introgression of mitotypes has been documented for butterflies (Hurst and Jiggins 2005, Narita, Nomura *et al.* 2006) as well as other insects (Whitworth, Dawson *et al.* 2007, Raychoudhury, Baldo *et al.* 2009, Kvie, Hogner *et al.* 2013) and it is suspected in the butterfly genera *Lycaedes* (Gompert, Forister *et al.* 2008), *Hypolimnas* (Charlat, Duplouy *et al.* 2009) and *Erynnis* (Zakharov, Lobo *et al.* 2009). Although we did not confirm presence of *Wolbachia* in the *Cymothoe* species for which introgression is suspected, *Wolbachia* DNA barcodes were amplified from some of our *C. mabillei* and *C. hartigi* samples (see Results), confirming that, in general, *Wolbachia can indeed infect Cymothoe species*. This suggests that *Wolbachia*-mediated introgression is a likely explanation for the shared haplotypes in *Cymothoe*.

A recent study assessing incidence of *Wolbachia* infection and associated confounding patterns in DNA barcode data sets, suggested that *Wolbachia*-mediated introgression is uncommon (Smith, Bertrand *et al.* 2012). Such introgression may be difficult to detect, however, because identical haplotypes in samples of different species are commonly regarded as contaminations and discarded from the databases. Indeed, we found multiple cases of contaminations in our *Cymothoe* data set and it was only because of the multiple independent incidences of haplotype sharing that we could reject contamination as an explanation. Distinguishing contamination from evidence therefore depended on our extensive and long-term sampling effort, which contrasts with most other DNA barcoding studies where in general only few individuals are sampled per species.

2.4.5 Future directions

Some taxonomic issues could not be resolved based on currently available data and require additional, more intensive sampling. For example, additional samples are required to elucidate species boundaries within unconfirmed taxonomic species *C. coranus*, *C. distincta*, *C. haynae*, *C. mabillei*, *C. oemilius*, *C. reginae-elisabethae*, and within the group comprising East African mountain endemics *C. amaniensis*, *C. aurivillii*, *C. collinsi*, *C. cottrelli*, *C. magambae*, *C. melanjae*, and *C. zombana*. With the current advances in high-throughput sequencing, DNA samples will become an ever more important source of biological information. Therefore, in all cases, sampling conditions for DNA preservation should be improved, preferably by direct sampling of DNA samples in alcohol in the field. In addition, sequencing DNA from types could potentially resolve long-standing nomenclatural problems. Types are often poorly or erroneously labeled and in case of cryptic species it is often impossible to know which genotype to associate with a name. For example, the type specimen of *C. sangaris* is labeled “Guinea”, which can mean anything from Guinée to western DR Congo. In case of putative cryptic sympatric species such as *C. egesta* “orange” and *C. confusa* in Cameroon (see Results) genotyping type material may be the only solution to associating the right names. *Cymothoe* are large butterflies and removal of small tissue parts (e.g. a single leg) may be possible without negatively affecting future morphological or molecular characterization.

Sections *Aramis*, *Lurida* and *Sangaris* remain poorly understood even with our increased DNA barcode sampling. Resolution of these cases should therefore be sought in additional lines of evidence.

To resolve the issues involved with incomplete lineage sorting and haplotype sharing, sequencing additional mitochondrial and fast-evolving nuclear loci is necessary. The operational advantage of using nuclear over mitochondrial sequences is that nuclear sequences provide multiple independently transmitted and recombining markers. However, higher effective population size for nuclear markers means higher levels of incomplete lineage sorting effects. In addition, because nuclear DNA exists

in far less copies in a cell than mitochondrial DNA, amplifying nuclear markers requires isolation of larger amounts of DNA. However in today's next-generation, single cell genomics world, (Shapiro, Biezuner *et al.* 2013) high-throughput sequencing is rapidly becoming more feasible for *Cymothoe*. Whole-genome shotgun sequencing of few representative species could allow selection of suitable (i.e. single-copy) nuclear genes. Assessing patterns based on nuclear loci can also allow differentiating between different scenarios of introgression (Whitworth, Dawson *et al.* 2007). If *Wolbachia* mediated such introgressions, the pattern is expected to be restricted to mitochondrial haplotypes. In addition, the donor populations are expected to be infected. Also, because they are not present in all specimens, the introgressed genes apparently have not (yet) become fixed in the recipient *Cymothoe* populations. It can therefore be expected that the recipient populations are either uninfected or host a different *Wolbachia* strain. This could be tested by screening *Cymothoe* specimens for *Wolbachia*-specific markers such as the *WSP* gene encoding the *Wolbachia* surface protein (Narita, Nomura *et al.* 2006).

2

Morphology of immature stages (in particular, of last instar larvae) has proven a highly valuable character to distinguish between sections (Amiet 2000) as well as between closely related species (Amiet 1997; this chapter). Possibly, immature morphology can help solve issues where adult morphology and DNA barcodes are inconclusive. In addition, increased attention for immature stages would generate additional data on the highly specialized associations between *Cymothoe* and their plant hosts (Amiet 1997), the evolution of which is the subject of ongoing studies.(van Velzen, Wahlberg *et al.* 2013). Unfortunately, immature stages of *Cymothoe* species (and Nymphaliid species in general) are usually poorly represented in natural history collections, confounding global delimitation of relevant characters.

Nevertheless, ultimately, joint efforts should be directed toward a taxonomic revision and authoritative classification of the genus, taking into account evidence from different levels of biological organization, and possibly representing a first fine example of a 'Taxonomics' approach.

2.5 Conclusions

In conclusion, our standardized pipeline provides a practical and useful way to integrate DNA barcode data with multiple lines of evidence. Five taxa within *Cymothoe* could be confidently raised to species level, and our expanded DNA barcode data set facilitated identification of 42 immature specimens. Nevertheless, our results demonstrate that such an integrative approach cannot diagnose all species of *Cymothoe*, probably because of incomplete lineage sorting between recently diverged species as well as introgression. Consequently, our findings underline the notion that in practice some species are inherently very difficult to delineate (Dayrat 2005, Pillon, Fay *et al.* 2006). We would like to echo Schlick-Steiner *et al.* (Schlick-

Steiner, Steiner *et al.* 2010) that such cases, where the nature of the evolving groups is poorly understood, may demarcate the limits of present integrative taxonomy. Nuclear DNA markers may provide a solution to some of the problems, but given the recent divergence of many *Cymothoe* species and the larger effective size of nuclear markers compared with mitochondrial ones, success is not guaranteed. Morphology of immature stages could provide important auxiliary evidence to help solve issues where adult morphology and DNA sequences are inconclusive.

Acknowledgements

We are indebted to Paul Hebert and CCDB staff for granting access to their sequencing facilities. Camiel Doorenweerd for generating sequences at Naturalis Biodiversity Center. Guillome Achaz and Sophie Brouillet (UPMC) for help with running ABGD. Renske Onstein collected adult and immature specimens in Ghana. Gaston Achoundong & Olivier Sene Belinga (Herbier National du Cameroun), Oskar Brattström (Cambridge University), Sáfián Szabolcz (University of West Hungary) kindly facilitated fieldwork. Jean-Louis Amiet, Torben Larsen (NHM), Hein Boersma, Dries Bonte (Ghent University), Frans Desmet (RMCA), Carolyn McBride (UCDavis), Erik Koenen, Jan Wieringa (Naturalis & Wageningen University), Robert Ducarme, Tomasz Pyrcz (MZUJ), Julian Bayliss, Oskar Brattstrom & David Lees (Cambridge University), Freerk Molleman (University of Tartu), Dino Martins, Sáfián Szabolcz (University of West Hungary), Eric Vingerhoedt, Haydon Warren-Gash, Robert Warren and Steve Woodhall donated specimens or samples. We thank Blanca Huartas (NHM), Ugo Dall'Asta (RMCA) for facilitating access to collections. Wouter Dekonick, Stefan Kerkhof (KBIN), Teresa di Micco de Santo (ABRI) and Suzanne Rab Green (AMNH) kindly provided specimen photographs.

Chapter 3

DNA barcoding of recently diverged species

Relative performance of matching methods

Robin van Velzen[1,2], Emanuel Weitschek[3,4], Giovanni Felici[3] & Freek T Bakker[1]

PLoS ONE, 2012 7(1), e30490

[1] Biosystematics Group, Department of Plant Sciences, Wageningen University and Research Centre, Wageningen, The Netherlands

[2] Naturalis Biodiversity Centre (Section NHN), Wageningen University, Generaal Foulkesweg 37, 6703 BL Wageningen, The Netherlands

[3] Istituto di Analisi dei Sistemi e Informatica Antonio Ruberti, Consiglio Nazionale delle Ricerche, Rome, Italy

[4] Department of Informatics and Automation, University Roma Tre, Rome, Italy

Abstract

Recently diverged species are challenging for identification, yet they are frequently of special interest scientifically as well as from a regulatory perspective. DNA barcoding has proven instrumental in species identification, especially in insects and vertebrates, but for the identification of recently diverged species it has been reported to be problematic in some cases. Problems are mostly due to incomplete lineage sorting or simply lack of a 'barcode gap' and probably related to large effective population size and/or low mutation rate. Our objective was to compare six methods in their ability to correctly identify recently diverged species with DNA barcodes: neighbor joining and parsimony (both tree-based), nearest neighbor and BLAST (similarity-based), and the diagnostic methods DNA-BAR, and BLOG.

We analyzed simulated data assuming three different effective population sizes as well as three selected empirical data sets from published studies. Results show, as expected, that success rates are significantly lower for recently diverged species (∼75%) than for older species (∼97%) ($p<0.000\,01$). Similarity-based and diagnostic methods significantly outperform tree-based methods, when applied to simulated DNA barcode data ($p<0.000\,01$). The diagnostic method BLOG had highest correct query identification rate based on simulated (86.2%) as well as empirical data (93.1%), indicating that it is a consistently better method overall. Another advantage of BLOG is that it offers species-level information that can be used outside the realm of DNA barcoding, for instance in species description or molecular detection assays. Even though we can confirm that identification success based on DNA barcoding is generally high in our data, recently diverged species remain difficult to identify. Nevertheless, our results contribute to improved solutions for their accurate identification.

3.1 Background

Recently diverged species are frequently of special interest, for example in ecology, regulation or forensics (Armstrong, Cameron *et al.* 1997, Wallman and Donnellan 2001, Dexter, Pennington *et al.* 2010), and hence their accurate identification is warranted. DNA barcoding (Hebert, Cywinska *et al.* 2003, Hebert and Gregory 2005, Schindel and Miller 2005) has proven instrumental in identifying recently diverged species (e.g. species complexes or cryptic species) that are of importance to conservation biology (Bickford, Lohman *et al.* 2007, Neigel, Domingo *et al.* 2007, Lahaye, Van der Bank *et al.* 2008, McBride, van Velzen *et al.* 2009), pest management (Boykin, Shatters *et al.* 2006, Aveskamp, Woudenberg *et al.* 2009, Skoracka and Dabert 2010), fishery (Bucciarelli, Golani *et al.* 2002, Ward, Zemlak *et al.* 2005, Hubert, Hanner *et al.* 2008, Ward, Costa *et al.* 2008, Zemlak, Ward *et al.* 2009, Griffiths, Sims *et al.* 2010), invasive biology (Armstrong and Ball 2005, May, Gelembiuk *et al.* 2006, Hsieh, Wang *et al.* 2007, Yassin, Capy *et al.* 2008, Newmaster and Ragupathy 2009, Bastos, Nair *et al.* 2011) and disease control (Nolan, Carpenter *et al.* 2007, Paredes-Esquivel, Donnelly *et al.* 2009, Azpurua, De la Cruz *et al.* 2010, McKeon, Lehr *et al.* 2010). In some cases, however, identification of recently diverged species using DNA barcodes has been reported to be problematic (Wallman and Donnellan 2001, Meyer and Paulay 2005, Kaila and Stahls 2006, Dexter, Pennington *et al.* 2010, Lou and Golding 2010, Yassin, Markow *et al.* 2010) due to ambiguous barcode matches or the absence of barcode clusters in DNA barcode trees.

Failure of DNA barcodes to properly resolve recently-diverged species can be attributed to population genetic factors of the species involved (Meyer and Paulay 2005, Elias, Hill *et al.* 2007, Wiemers and Fiedler 2007, Ross, Murugan *et al.* 2008, Austerlitz, David *et al.* 2009, McFadden, Benayahu *et al.* 2011). Coalescent theory (Kingman 1982) predicts that the chance that gene sequences sampled from a species are monophyletic is dependent on the age of that species (measured in number of generations since speciation) and reversely dependent on its effective population size (N_e) (Kingman 1982, Hudson, Futuyama *et al.* 1990). This is because species with large N_e are predicted to have larger within-species genetic variation (Kingman 1982, Hudson, Futuyama *et al.* 1990, Nichols 2001). When such species have diverged only recently their gene sequences are likely to have a most recent common ancestor predating the speciation event (incomplete lineage sorting) (Nichols 2001). This results in overlapping within- and between-species genetic distances (lack of a 'barcode gap') and paraphyly or even polyphyly of conspecific samples in gene trees (Nichols 2001, Funk and Omland 2003, Nielsen and Matz 2006, Elias, Hill *et al.* 2007). For example, in Lycaenidae (Blue butterflies) Wiemers and Fiedler (Wiemers and Fiedler 2007) found a general lack of 'barcode gaps' and paraphyly or polyphyly of conspecific DNA sequences, probably caused by incomplete lineage sorting (Wiemers and Fiedler 2007), as did McFadden *et al.* in Octocorals (McFadden, Benayahu *et al.*

2011). Meyer and Paulay (Meyer and Paulay 2005), in their DNA barcode study of marine gastropods, explained non-monophyly of some species by incomplete lineage sorting effects. Elias *et al.* (Elias, Hill *et al.* 2007) reported limited performance of DNA barcoding in two butterfly communities in Ecuador, which they attributed in part to large N_e and associated long coalescent times (Elias, Hill *et al.* 2007). Based on simulated DNA barcode data sets Ross *et al.* (2008) and Austerlitz *et al.* (2009) found that species monophyly and identification success generally decreased with increasing coalescent depth.

Regardless of N_e, recently diverged species have acquired only few genetic differences meaning that there are few characters to discriminate them. The rate at which two sister species genetically diversify is dependent on their effective mutation rate (μ). If μ is sufficiently low, even reciprocally monophyletic species will share identical haplotypes. Indeed, some morphologically well-differentiated species may share identical DNA barcode sequences, preventing accurate identification using DNA barcodes (Wiemers and Fiedler 2007, Lou and Golding 2010, McFadden, Benayahu *et al.* 2011). If μ is higher, identification success depends on the extent of lineage sorting: on the one hand, a single fixed mutation can be enough for successful identification (DeSalle, Egan *et al.* 2005, Rach, DeSalle *et al.* 2008, McFadden, Benayahu *et al.* 2011); on the other hand, non-monophyletic (i.e. incompletely-sorted) species will have overlapping genetic variation even when μ is high. Therefore, we consider the factors governing lineage sorting: time (measured in generations), and N_e, to be the most important factors contributing to DNA barcode identification problems with recently diverged species. Obviously, when given enough time any N_e or μ will ultimately result in high levels of between-species divergence. We therefore emphasize time here and focus on 'recent' versus 'old' species.

Various methods have been proposed to match DNA barcodes to a reference library for identification, amongst which we recognize the following:

Tree-based methods assign unidentified (query) barcodes to species based on their membership of clusters (or clades) in a DNA barcode tree. This approach is usually based on neighbor joining (Saitou and Nei 1987, Munch, Boomsma *et al.* 2008), parsimony (Edwards and Cavalli-Sforza 1963) or Bayesian inference (Huelsenbeck and Ronquist 2001). Tree-based methods assume that samples of distinct species form discrete clusters in a DNA barcode tree (Hebert, Cywinska *et al.* 2003, Munch, Boomsma *et al.* 2008). It is generally acknowledged, however, that gene trees (i.e. DNA barcode trees) do not necessarily reflect organismal history (Nichols 2001), and that the incomplete lineage sorting effects outlined above may lead to incorrect identifications based on such trees (Meyer and Paulay 2005, Elias, Hill *et al.* 2007, Wiemers and Fiedler 2007, Ross, Murugan *et al.* 2008, Austerlitz, David *et al.* 2009).

Similarity-based methods assign query barcodes to species based on how much DNA barcode characters they have in common. Similarity can be calculated directly from nucleotide sites (e.g. using MOTU (Floyd, Abebe *et al.* 2002), nearest neighbor

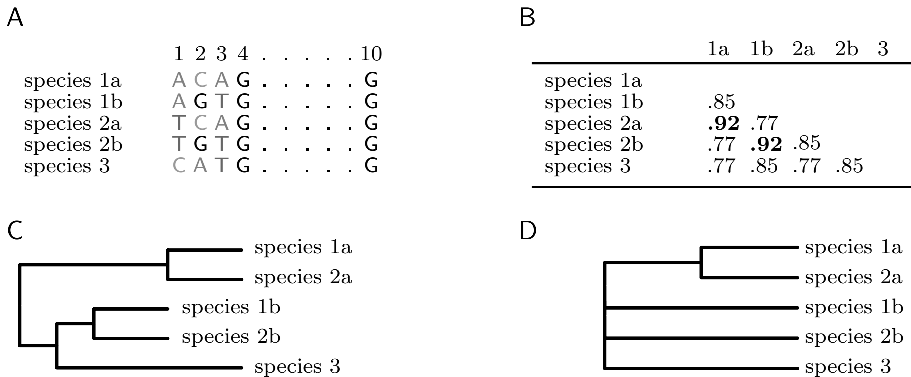


Figure 3.1: Hypothetical DNA barcode sequences where tree-based and similarity methods produce incorrect identifications. (A) Alignment where two recently diverged sister species (species 1 and species 2) have only one diagnostic nucleotide differentiating them from each other (position 1) and at the same time share two polymorphisms (positions 2 and 3). Species 3 is included as outgroup; (B) Pairwise uncorrected similarities based on the alignment with highest pairwise similarities in boldface; (C) Neighbor joining tree; (D) Strict consensus of all maximum parsimony trees.

(Meier, Shiyang *et al.* 2006, Austerlitz, David *et al.* 2009), or BLAST (Altschul, Madden *et al.* 1997)) or from a projection of nucleotides (e.g. Kernel methods (Austerlitz, David *et al.* 2009, Kuksa and Pavlovic 2009, Seo 2010), ATIM (Little, Stevenson *et al.* 2008), BRONX (Little 2011)). Similarity-based methods assume that conspecific samples will be more similar to each other than to samples of any other species. However, this need not be true in all cases. For instance, if we consider two hypothetical sister species that share two polymorphisms and have only one nucleotide differentiating them from each other, tree- and similarity-based methods will fail to correctly identify (some of the) haplotypes in these species, see Figure 3.1.

Statistical methods estimate confidence measures on DNA barcode matches for species identification. These methods typically employ Bayesian estimation based on explicit population genetic or phylogenetic models (Matz and Nielsen 2005, Nielsen and Matz 2006, Abdo and Golding 2007) . Obviously, confidence measures are of great importance when dealing with regulated species, forensics or disease vectors (Nielsen and Matz 2006). However, because statistical methods for species identification are computationally intensive and the appropriate model parameters are not known for the majority of species we will not treat them further.

Diagnostic methods (sometimes included in 'character-based' methods (DeSalle, Egan *et al.* 2005)) rely on the presence/absence of particular characters in DNA barcode sequences for identification, instead of using them all. Diagnostics can be either "simple" when based on a single unique character or "compound" when

based on a unique combination of characters (Sarkar, Planet *et al.* 2008). Some methods use nucleotide data and require a multiple sequence alignment (e.g. CAOS (Sarkar, Planet *et al.* 2002, Sarkar, Thornton *et al.* 2002, Sarkar, Planet *et al.* 2008), BLOG (Bertolazzi, Felici *et al.* 2009). Others use diagnostic nucleotide strings as diagnostics and are therefore alignment-free (e.g. DNA-BAR (DasGupta, Konwar et al.)). Diagnostic methods are analogous to classical taxonomic practices that rely on morphological diagnostic characters (DeSalle, Egan *et al.* 2005, Goldstein and DeSalle 2011). As opposed to other methods, diagnostic methods have the potential to select the differentiating nucleotide only and ignore any within-species variation obscuring that signal (DeSalle, Egan *et al.* 2005, Reid, Le *et al.* 2011, Zou, Li *et al.* 2011). For example, a diagnostic method could correctly identify the two hypothetical species in Figure 3.1 based on the diagnostic nucleotide at position 1.

Our objective was to compare relative performance of six DNA barcode matching methods in correctly identifying barcodes of recently diverged species. Below we provide some motivations for choosing each of these six methods:

1. Tree-based neighbor joining (NJ) (Saitou and Nei 1987) because it is the most widely used method for classifying DNA barcodes in the literature, and implemented in, for instance, the Barcode Of Life Database (Ratnasingham and Hebert 2007). Speed being its main advantage, NJ is a bottom-up clustering algorithm that calculates a single tree from a distance matrix. Results can be dependent on the ordering of the matrix, however, making results sometimes less reproducible. The underlying assumption in NJ barcode matching is that barcode sequences of distinct species form discrete clusters in a NJ tree (Hebert, Cywinska *et al.* 2003). For identification, query sequences are included in the NJ tree to see in which cluster they appear.

2. Tree-based parsimony (PAR) (Edwards and Cavalli-Sforza 1963) as it outperformed other tree-based methods (such as the Statistical Assignment Package SAP (Munch, Boomsma *et al.* 2008)), in a published comparative study (Little 2011). PAR adopts the optimality criterion under which the preferred tree is the tree that requires the least evolutionary change to explain the data. Assessing all possible trees for more than 20 sequences is computationally impossible and therefore PAR methods employ heuristics to find the preferred tree(s).

3. Similarity-based nearest neighbor (NN) because it gave high correct identification rates in previous studies (Meier, Shiyang *et al.* 2006, Austerlitz, David *et al.* 2009). Based on a distance matrix, NN simply assigns a query sequence to the same species membership as its closest sequence in the reference data base. It is equivalent to the 'Best Match' method by Meier *et al.* (Meier, Shiyang *et al.* 2006) and the '1-NN' method used by Austerlitz *et al.* (Austerlitz, David *et al.* 2009).

4. Similarity-based BLAST (Altschul, Madden *et al.* 1997) as it is probably the most commonly used method for classifying DNA sequences in practice. It is an algorithm for comparing query sequences with an unaligned reference data base

calculating pairwise alignments in the process. It is faster than NN, but can give incorrect matches in some cases, especially with incomplete reference data bases (Koski and Golding 2001).

5. The diagnostic method DNA-BAR (DasGupta, Konwar et al.) because it showed higher levels of accurate species identification in previous studies (Little, Stevenson *et al.* 2008, Little 2011) compared to the other diagnostic method CAOS (Sarkar, Planet *et al.* 2008). DNA-BAR first selects sequence substrings (distinguishers) differentiating the sequences in the reference data set, and then records presence/absence of these distinguishers. An advantage of using substrings is that the method does not require an alignment.

6. The recently developed diagnostic logic mining method BLOG (Bertolazzi, Felici *et al.* 2009) because it has not been used in any comparative test before (except (Weitschek, van Velzen *et al.* 2011)). BLOG first selects a number of characters ('features') from the reference data set that optimize discrimination of a particular species, based on an integer programming feature selection method. It then uses the selected features to search for the simplest logic formula that discriminates that species from all others using a learning method based on decomposition techniques (Bertolazzi, Felici *et al.* 2009, Weitschek, van Velzen *et al.* 2011). This process is reiterated for every species in the reference data set. Subsequently, query sequences are screened for their recognition by the formulas for identification. The reader may refer to (Bertolazzi, Felici *et al.* 2008, Bertolazzi, Felici *et al.* 2009, Bertolazzi, Felici *et al.* 2010) and (Felici and Truemper 2002) for a complete description of the mathematical models that constitute the main characteristics of BLOG.

We use simulated and empirical DNA barcode datasets, the latter from published studies. In general, data simulations allow for replication and, hence, statistical testing of method performance. For instance, Austerlitz *et al.* (Austerlitz, David *et al.* 2009) assessed relative performance of NJ, NN, classification and regression trees, random forest, and kernel methods in correctly assigning query barcodes to predefined species. They concluded that, although NN was the most reliable method overall, none was found to be best under all circumstances. However, the authors simulated datasets with only 2–5 species and assumed simultaneous divergence of all species which seems biologically unrealistic (Austerlitz, David *et al.* 2009). Here, we simulated more realistic DNA barcode datasets comprising 50 species along a phylogenetic tree, thus producing more typical levels of sequence divergence. In this regard our approach is similar to that of Ross *et al.* (2008) who tested similarity and tree-based methods of species identification using 'realistic' simulated datasets. They concluded that tree-based methods returned ambiguous identifications. However, they did not take species divergence times explicitly into account, nor did they include diagnostic methods, which we do here.

Our results show that, even though recently diverged species pose a significant problem for effective DNA barcoding, sensitive similarity-based and diagnostic methods can significantly improve identification performance compared with the commonly used tree-based methods such as NJ.

3.2 Materials and methods

Our analytical pipeline started with generating simulated DNA barcode data sets and selection of published empirical data sets. Subsequently, we assessed both 'barcode gap' and monophyly of species and performed matching analyses with tree-based (NJ,PAR), similarity-based (NN, BLAST) and diagnostic (DNA-BAR, BLOG) methods on both types of data. The pipeline concluded with a comparative evaluation of methods used in terms of accuracy of species identification.

3.2.1 Data simulation

DNA barcode datasets were simulated using the Coalescent package in Mesquite version 2.73 build 544 (Maddison and Maddison 2010, Maddison and Maddison 2010). We simulated along two axes: time of species divergence and effective population size (N_e). We started by simulating a random ultrametric species tree for 50 species using the Yule model (Steel and McKenzie 2001), with a total tree depth of 1 million generations. Species were divided into two equally-sized groups (N=25) based on their rank in divergence times: one with 'recently diverged' species and another with 'old' species. Ultrametric gene trees were simulated on the ultrametric species tree according to the coalescence model, generating 20 individuals per species. Gene trees were simulated using $N_e = 1000$, 10 000 and 50 000 with each simulation replicated 100-fold, resulting in 300 gene trees in total. Additive gene trees were then obtained by adding noise to the branch lengths of gene trees in order to ensure more realistic (i.e. non-ultrametric) data structure. Thereby we effectively mimicked heterogeneity of the effective mutation rate (μ) over branches of the gene trees. Noise was normally distributed, with a variance σ of 0.7 times the original branch length.

DNA barcode sequences were then simulated on the additive gene trees according to a HKY substitution model (Hasegawa, Kishino *et al.* 1985), the choice of which was based on the best-fitting model for a representative empirical dataset of 527 Nymphalidae DNA barcodes as selected using JModelTest 0.1.1 (Posada 2008) applying the AIC criterion. Model parameters encompassed a transition/transversion ratio κ of 8.3, nucleotide frequencies of 0.30 (A), 0.15 (C), 0.10 (G), 0.45 (T), and gamma-distributed rate variation over sites with 4 rate categories and a shape parameter α of 0.2. Sequence length was 650 base pairs, approximating the length of the standard DNA barcode for animals (Cytochrome *c* oxidase subunit 1; COI). Simulated sequences were divided over reference data sets (16 sequences per species) and query data sets (4 sequences per species). The reference data sets were considered as DNA barcode reference libraries containing sequences with *a priori* assigned species membership. The query data sets were considered to comprise unknown DNA barcodes, although in our case species membership was known because they were simulated together with the reference data set. Consequently, accuracy of their identification could be evaluated *a posteriori*.

Table 3.1: Summary of selected empirical data sets used.

Data set	Reference	Marker(s)	Length	#seqs	#spp.	#spp. ≥5seqs
Drosophila	Lou & Golding (2010)	COI	663	615	19	15
Inga	Dexter *et al.* (2010)	*trnT–D* ITS	1838	913	56	35
Cypraeidae	Meyer & Paulay (2005)	COI	614	2008	211	112

Length = total sequence length, #seqs = number of sequences, #spp. = total number of species in the data set, #spp.≥5seqs = number of species represented by ≥5 sequences.

3.2.2 Empirical data sets

We selected three published empirical DNA barcode data sets based on the following criteria: 1. Data contain species that are problematic to identify using DNA barcodes because of incomplete clustering in barcode trees; 2. Data encompass high phylogenetic diversity, i.e. from different phyla (Plantae, Mollusca and Arthropoda), to ensure the general applicability of our outcomes; 3. Data come from different markers, i.e. from all three genomic compartments. A summary of the selected data can be found in Table 3.1; details are below:

Drosophila. Lou and Golding (2010) used this data set to test the ability of algorithms to assign sequences to species in the absence of a barcode gap. They found that many species are siblings with low between-species distances and some have no 'barcode gap' (Yassin, Capy *et al.* 2008, Lou and Golding 2010). *Drosophila* species are also known to have relatively large N_e's and associated high within-species divergence (Petit and Barbadilla 2009, Castillo, Mell *et al.* 2011). The data set comprised 615 barcodes from 19 species.

Inga (Fabaceae) is a large genus of tropical leguminous trees. Many morphologically distinct *Inga* species collected in the southwestern Amazon are incompletely sorted in DNA barcode trees (Dexter, Pennington *et al.* 2010). No N_e estimates for *Inga* are available. We selected the data set from Dexter *et al.* (2010) who linked cpDNA *trnT–D* intron and nrDNA Internally Transcribed Spacer (ITS) sequences into a multi-locus DNA barcode of 1713–1771 nucleotides in total. The data set comprised 913 barcodes from 56 species.

Cypraeidae (Mollusca) are taxonomically one of the most extensively studied marine gastropods. Although Meyer & Paulay showed that subspecies rather than species best represent diversity in these DNA barcodes (Meyer and Paulay 2005) we adhered to species names, mainly because subspecies were generally less well sampled. No N_e estimates for Cypraeidae are available. The data set comprised 2008 mtDNA COI sequences of 211 species and had almost complete coverage of sister-species, some of which are reported to have diverged only recently (Meyer and Paulay 2005).

Only those species represented by 5 or more sequences were evaluated in the identification assessments. Their sequences were randomly distributed over a reference data set (80% per species) and a query data set (20% per species). Species represented by less than 5 sequences were kept in the reference data set, but not evaluated in the identification assessments (i.e. their sequences could therefore only contribute to the false positive rate of the query sequences that were evaluated).

3.2.3 Species 'barcode gap' and monophyly

To assess the existence of a 'barcode gap' in our data sets, we extracted within- and between- species K2P (Kimura 1980) distances from all 50 species in all 300 simulated reference data sets (100 of each N_{e}) and made comparisons between N_e's. We are aware that using K2P implies effective under-parameterization (Lemmon and Moriarty 2004) as we used HKY in the simulations, but we chose K2P as it is typically used in DNA barcode analyses (e.g. (Bucciarelli, Golani *et al.* 2002, Armstrong and Ball 2005, Neigel, Domingo *et al.* 2007, Ward, Costa *et al.* 2008, van Velzen, Larsen *et al.* 2009)). Repeating the analysis using HKY did not give different results (not shown). We evaluated the existence of 'barcode gaps' at species level by scoring a species as having a 'barcode gap' when the minimum between-species sequence distance exceeded the maximum within-species distance (Meyer and Paulay 2005, Meier, Zhang *et al.* 2008).

We assessed species-monophyly in DNA barcode trees of all 50 species in all 300 simulated reference data sets and subsequently compared results between N_e's. DNA barcode trees were reconstructed using NJ and parsimony using settings described below, and species were scored as either monophyletic or non-monophyletic based on the DNA barcode tree topologies.

3.2.4 Method performance

Neighbor joining (NJ). We used the neighbor joining algorithm (Saitou and Nei 1987) implemented in the R. package APE 2.5–3 (Paradis, Claude *et al.* 2004) and applied randomly shuffling of input order of sequences. We assessed tree topology in two ways, following Ross *et al.* (2008). 1. 'Strict assessment' meant that if the query was nested within a mono-specific cluster or clade it was identified as that species. Otherwise its identification was considered uncertain. This is equivalent to the 'Tree based identification, revised criteria' used by Meier *et al.* (2006) and is reported to have significantly lower false-positive rates (Ross, Murugan *et al.* 2008). 2. 'Liberal assessment' meant that if the query was sister to a mono-specific cluster it was identified as that species. Otherwise its identification was considered uncertain.

Parsimony (PAR). Maximum parsimony trees were estimated using TNT version 1.1 (Goloboff, Farris *et al.* 2008). Heuristic searches consisted of iterations of ratchet, sectorial searches, tree drift and tree fusing algorithms (Goloboff 1999)

through the TNT built-in function xmult, holding 1000 trees during search (hold 1000). Searches were stopped when four independent replicates found shortest trees of the same length (xmult= hits 4). Identical sequences were excluded before analysis and later restored to save computation time (riddup). Only one maximum parsimony tree was held after each analysis to make results comparable to NJ. We assessed tree topology in the same way as described for NJ above.

Nearest neighbor (NN). Nearest neighbors were calculated using the function dist.dna in the R. package APE version 2.5–3 (Paradis, Claude *et al.* 2004) based on the K2P model of sequence evolution (Kimura 1980). A query was identified as the species associated with its nearest neighbor (reference sequence with lowest distance to that query). In case nearest neighbors were from more than one species the query's identification was considered uncertain.

BLAST. Identification based on BLAST was performed using NCBI software version 2.2.25+ (Zhang, Schwartz *et al.* 2000). Reference data sets were stored in a BLAST database for subsequent matching with query sequences. Up to 100 hits with at least 80% identity were returned for each query, which was identified as the species associated with its best hit (highest bit score). In case more than one species were associated the query's identification was considered uncertain.

DNA-BAR. Reference data sets were converted to a matrix comprising presence/absence of distinguishers (sequence substrings) using the software 'degenbar' (DasGupta, Konwar et al.). Input parameters were as follows: distinguishers of length 5–50 nucleotides (l-min 5, l-max 50), up to 100 redundant distinguishers (Redundancy 100), GC content 0–100% (MinCandidGC 0, MaxCandidGC 100), annealing temperature 0–100°C (MinCandidTemp 0, MaxCandidTemp 100), salt and DNA concentration 50nM (SaltConc 50, DNAconc 50), and a maximum common substring weight of 100 (MaxCommSubstrWt 100) (note that degenbar was originally designed to pick DNA probes). In case of multi-locus DNA barcodes (i.e. *Inga* data set) loci in the reference alignment were separated by 50 'N' positions. The presence/absence matrix of distinguishers was then used as reference data set. Each query sequence was scored for presence/absence of distinguishers and identified as the species associated with the reference sequence with the greatest number of matching presence/absences. In case more than one reference sequence of the same species membership shared the greatest number of matches the query was identified as that species. In case reference sequences associated with different species shared the greatest number of matches identification was considered uncertain.

BLOG. Diagnostic logic mining analyses were performed with BLOG software version 2.4 (Bertolazzi, Felici *et al.* 2009) which is available online and on the Barcode Of Life Data Portal (Sarkar and Trizna 2011) (an off-line version is available from EW upon request). Input parameters for feature selection were as follows: a maximum number of 35 features chosen (BETA=35), a maximum of 200 iterations (GRASPITER =200), and a maximum time of 500 minutes for analysis (GRASPSECS=30000). Each query sequence was scanned to see if it satisfied any of

the logic formulas generated by BLOG and identified as the species associated with the matching logic formula. In case a query satisfied more than one logic formula the logic formula having lowest false positive rate on the reference data set was taken as the identification. In case error rates of logic formulas were equal identification was considered uncertain.

3.2.5 Statistical tests

We assessed relative performance of the six methods in terms of their identification success with simulated and empirical data. Identification success was defined in two ways: 1. 'Species identification success' was scored as the number of species for which all query sequences were correctly identified. 2. 'Sequence identification success' was scored as the number of correctly identified query sequences per data set, which is equivalent to sensitivity; i.e. true positives/(true positives + false negatives).

We evaluated the influence of (i) species divergence times (recently diverged versus old); (ii) method used; and (iii) N_e on species identification success, using Friedman tests (Friedman 1937) in which the sum of identification success measures per replicate was used as the observation. Significant differences between methods were revealed in post-hoc pairwise Wilcoxon signed rank tests based on paired observations (Wilcoxon 1947). To account for the large number of comparisons we applied Bonferroni correction (Bonferroni 1935) to all tests combined (i.e. multiplying p-values by total number of tests performed). A corrected value of $p<0.01$ was considered statistically significant.

3.3 Results

3.3.1 Data simulation

The 50 species in the simulated ultrametric species tree had divergence times between 98 and 553 116 generations (see figure 3.2). We classified half the species (with divergence times between 98 and 76621 generations) as 'recently diverged' and the other half (with divergence times between 76 621 and 553 116 generations) as 'old', see Figure 3.2.

3.3.2 Species 'barcode gap' and monophyly

Maximum within-species distance equals or exceeded minimum between species distance for a substantial proportion (37%) of the species in the simulated data sets, indicating absence of a barcode gap. This proportion positively correlates with effective population size (N_e), which is explained mainly by an increase of the within-species distances under larger N_e, see Figure 3.3. On the contrary, with 54% for old species and 20% for recently diverged species this proportion decreases with

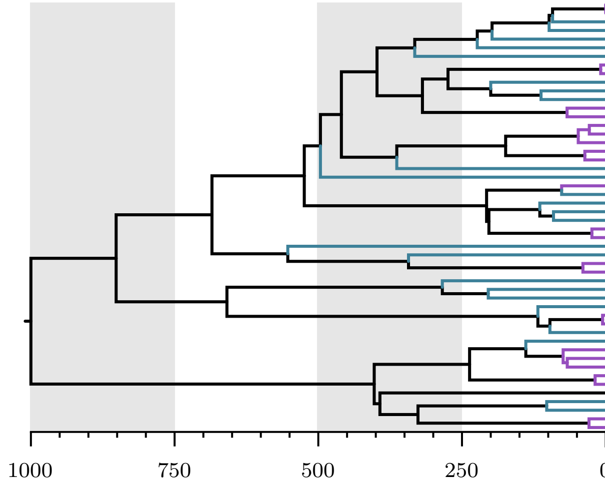


Figure 3.2: Simulated ultrametric species tree. Phylogenetic tree with 50 species simulated under the Yule model and with a total tree depth of 1 million generations. Terminal branches subtending species considered as 'recently diverged' are in purple, those subtending species considered as 'old' are in blue.

increasing divergence time (mostly dark dots fall below the 'barcode gap' line in Figure 3.3). As expected, percentage of species-monophyly was lower for species that had diverged more recently (Figure 3.4). While the oldest species (553 116 generations) was always monophyletic the two youngest species (98 generations) were never. Between these extremes, percentages increased more rapidly for data sets simulated under coalescence with smaller N_e (Figure 3.4).

3.3.3 Method performance

The comparative evaluation of methods shows, as expected, that species identification success generally decreased with increasing N_e, see figure 3.5 for results across all methods. Data sets that were simulated according to the smallest N_e (1000 individuals) had highest average success score with 89% (p<0.000 01). With an average success score of 81%, datasets that were simulated according to the largest N_e (50 000 individuals) were most challenging in terms of species identification (p<0.000 01). Similarly, species identification success rates of all methods are lower for species that have diverged more recently, see Figure 3.6 for results across all methods. On average, the 25 recently diverged species were correctly identified in 75% of cases, significantly less than 97% for the 25 old species (p<0.000 01). Query identification success showed the same pattern, where scores for old species were generally higher than for recently diverged species and showed less variation (data not shown). We therefore report relative performance of methods compared for recently diverged species only.

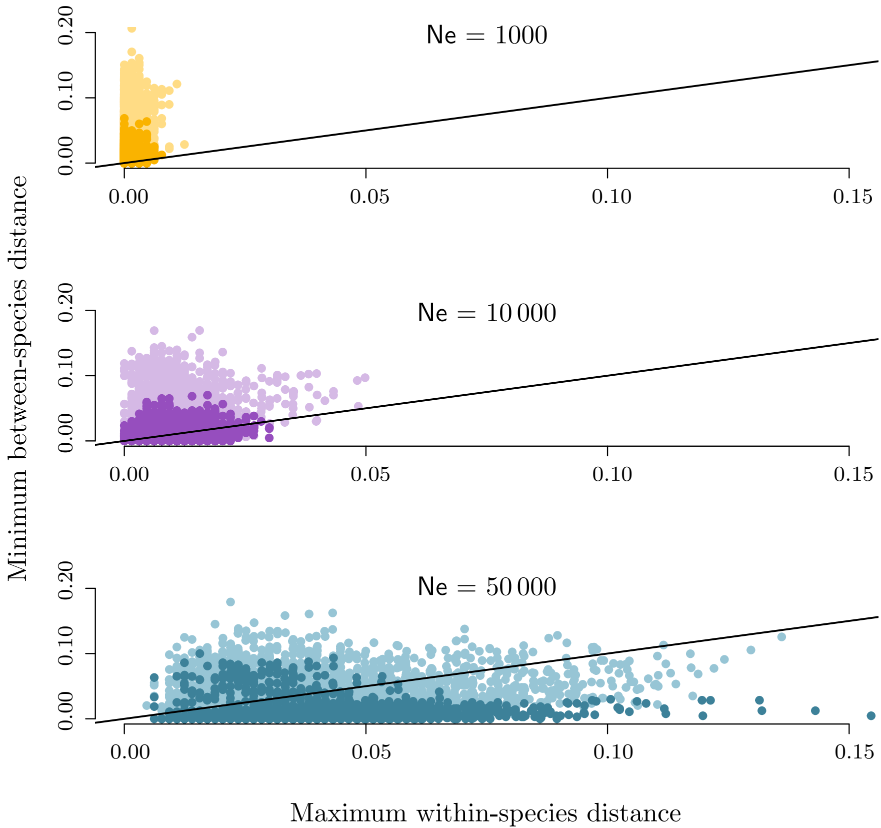


Figure 3.3: Species 'barcode gap'. Scatterplots of minimum between- over maximum within-species distance for 5000 simulated species in the reference data sets with 16 samples per species. Simulations under coalescence with effective population sizes (N_e) of 1000 (yellow, top), 10 000 (purple, middle) and 50 000 (blue, bottom) individuals. Brightness of the dots correlates with species divergence times, i.e. recently diverged species are dark and old species are light. Species plotted above the diagonal lines have a barcode gap.

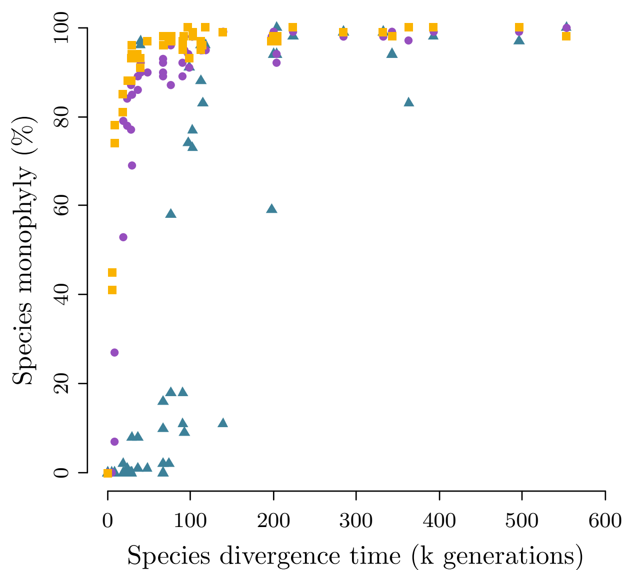


Figure 3.4: Species monophyly over time of divergence. Scatterplot of percentage species monophyly (N=100) based on NJ DNA barcode trees for 50 simulated species from the reference data sets (16 individuals per species) plotted against their divergence times. Simulations under coalescence with effective population sizes of 1000 (yellow squares), 10 000 (purple dots) and 50 000 (blue triangles) individuals.

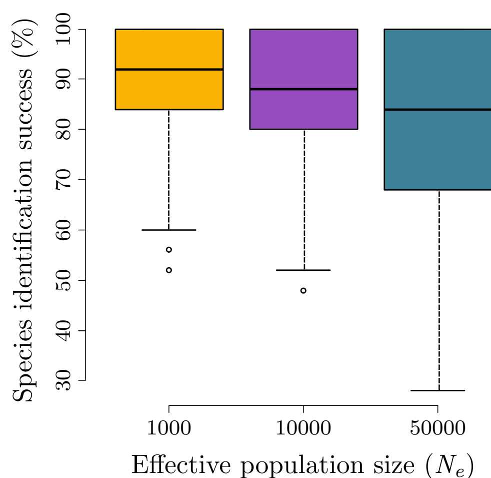


Figure 3.5: Influence of effective population size (N_e) on species identification success. Boxplots of percent species identification success (N=100) based on query data sets simulated under coalescence with effective population sizes of 1000 (yellow), 10 000 (purple) and 50 000 (blue) individuals.

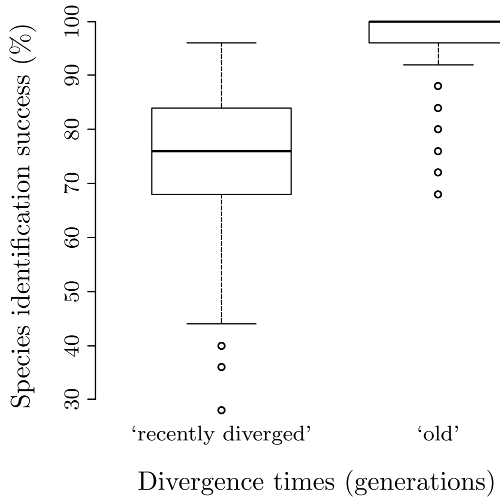


Figure 3.6: Influence of species divergence on species identification success. Boxplots of percent species identification success (N=300) based on query data sets for species that were either 'recently diverged' (divergence times between 98 and 76 621 generations) or 'old' (divergence times between 76 621 and 553 116 generations).

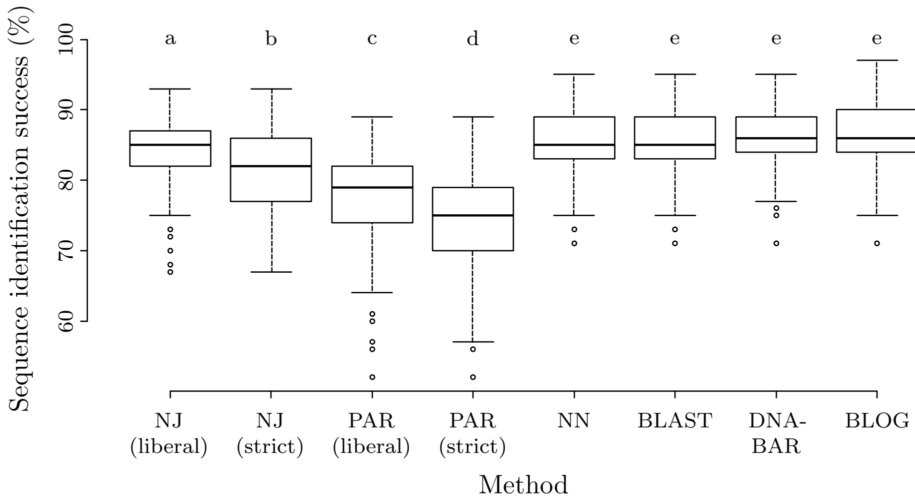


Figure 3.7: Method performance. Boxplots of sequence identification success (N=300) of six methods that were applied to recently diverged species in simulated query data sets. NJ = neighbor joining, PAR = parsimony, NN = nearest neighbor. Success scores not significantly different in post-hoc pairwise Wilcoxon tests are indicated by same superscripts.

Table 3.2: Relative method performance based on simulated data for recently diverged species.

Data set	NJ (liberal)	NJ (strict)	PAR (liberal)	PAR (strict)	NN	BLAST	DNA-BAR	BLOG
N_e = 1,000	83.69	83.58	73.31	73.14	86.18	86.18	**86.25**	85.96
N_e = 10,000	85.53	84.27	79.79	78.38	86.11	86.09	86.83	**88.15**
N_e = 50,000	84.20	77.35	79.53	72.32	84.76	84.56	**85.24**	84.58
overall	84.47^{a}	81.73^{b}	77.54^{c}	74.61^{d}	85.68^{e}	85.61^{e}	86.11^{e}	**86.23^{e}**

DNA barcode query identification success scores (%, N=100) of six methods applied to barcode sequence datasets simulated under three different effective population sizes (N_e). NJ = neighbor joining, PAR = parsimony, NN = nearest neighbor. Highest scores are in boldface. Overall success scores (%, N=300) not significantly different in post-hoc pairwise Wilcoxon tests are indicated by same superscripts.

3

Table 3.3: Relative method performance based on empirical data.

Data set	NJ (liberal)	NJ (strict)	PAR (liberal)	PAR (strict)	NN	BLAST	DNA-BAR	BLOG
Drosophila (118)	83.90	80.51	83.90	80.51	82.20	83.90	83.90	**96.61**
Inga (172)	91.28	90.12	81.40	80.23	88.37	82.56	**94.19**	90.12
Cypraeidae (354)	91.53	90.40	85.31	83.90	91.24	92.66	**93.22**	92.66
Overall	88.90	87.01	83.53	81.55	87.27	86.37	90.43	**93.13**

DNA barcode query identification success scores (%) of six methods applied to three empirical data sets. NJ = neighbor joining, PAR = parsimony, NN = nearest neighbor (liberal) = liberal assessment, (strict) = strict assessment. Number of query sequences in each data set is in brackets. Overall success scores (bottom line) are averaged over the three data sets. Highest scores are in boldface.

Diagnostic method BLOG performed best (86.2%) in terms of overall query identification success for recently diverged species based on simulated data (Table 3.2, Figure 3.7), although not significant (p=0.033). Diagnostic method DNA-BAR (86.1%) as well as similarity-based methods NN (85.7%) and BLAST (85.6%) performed only slightly worse than BLOG and significantly better than tree-based methods (p<0.000 01). Of the two tree-based methods NJ generally performed better than PAR and liberal assignment performed better than strict assignment for both methods (all p<0.000 01).

3.3.4 Empirical data sets

Based on empirical data diagnostic method BLOG performed best (93.1%) in terms of overall query identification success (see Table 3.3). Diagnostic method DNA-BAR performed only slightly worse (90.4%) and had the best score for two out of three empirical data sets (*Inga* and Cypraeidae).

Drosophila. The most divergent *Drosophila* sequences had 19.5% pairwise distance, and the largest within-species divergence was 17.5% for *D. angor*. Fifteen of 19 species had sufficient coverage (i.e. were represented by 5 or more sequences). Based on the reference data set comprising 497 sequences, 11 species were monophyletic in a NJ tree (73.3%) and 9 had a 'barcode gap' (60.0%). Based on the query data set (118 sequences) BLOG outperformed all other methods in terms of query identification success (114 query sequences correctly identified). DNA-BAR and BLAST identified 99 query sequences correctly as did NJ and PAR based on liberal assignment; NN identified 97 query sequences correctly; NJ and PAR identified 95 query sequences correctly based on strict assignment; see Table 3.3.

Inga. The two most divergent *Inga* sequences had 1.5% pairwise distance, and largest within-species divergence was 0.7% for *I. capitata*. Thirty five of 56 species had sufficient coverage (i.e. were represented by 5 or more sequences). Based on the reference data set (736 sequences) 25 species were monophyletic in a NJ tree (71.4%) and only 16 had a 'barcode gap' (45.7%). Based on the query data set (172 sequences) DNA-BAR outperformed all other methods in terms of query identification success (162 query sequences correctly identified). NJ identified 157 query sequences correctly based on liberal assignment; BLOG identified 155 query sequences correctly as did NJ based on strict assignment. NN identified 152 query sequences correctly; BLAST identified 142 query sequences correctly; PAR identified 140 query sequences correctly based on liberal assignment and 138 based on strict assignment.

Cypraeidae The most divergent Cypraeidae sequences had 28.5% pairwise distance, and largest within-species divergence was 17.1% for *Leporicypraea mappa*. Hundred twelve of 211 species had sufficient coverage (i.e. were represented by 5 or more sequences). Based on the reference data set (1654 sequences) only 81 species were monophyletic in a NJ tree (38.4%) and only 77 had a 'barcode gap' (36.5%). Based on the query data set (354 sequences) DNA-BAR outperformed all other methods in terms of query identification success (330 query sequences correctly identified). BLOG and BLAST identified 328 query sequences correctly; NJ identified 324 query sequences correctly based on strict assignment; NN identified 323 query sequences correctly; NJ identified 320 query sequences correctly based on strict assignment; PAR identified 302 query sequences correctly based on liberal assignment and 297 based on strict assignment.

3.4 Discussion

DNA barcoding works well for most species, although significant differences in population dynamics probably exist between, e.g. vertebrates, insects and plants. Indeed, DNA barcoding success rates have been estimated to be around 98% for animals and 70% for plants (Hollingsworth, Forrest *et al.* 2009, Hebert, deWaard *et al.* 2010, Dinca, Zakharov *et al.* 2011) with the relatively low success rate for the latter having been attributed to various causes such as high incident of hybrid species in angiosperms (Fazekas, Kesanakurti *et al.* 2009), long generation times or slow mutation rates of woody species (Hollingsworth, Graham *et al.* 2011) and limited dispersal of seeds (Petit and Excoffier 2009, Hollingsworth, Graham *et al.* 2011). Overall, the fact that DNA barcoding works so well is considered to be mainly due to conspecific sequences generally having their coalescent well after time of species divergence (Hebert, Cywinska *et al.* 2003).

3

Our results corroborate this notion in that, although our data sets contained incompletely-sorted species, identification success rates were generally high (>80%). Nevertheless, species that are recently diverged pose a consistent problem for identification based on DNA barcodes (Wallman and Donnellan 2001, Meyer and Paulay 2005, Kaila and Stahls 2006, Elias, Hill *et al.* 2007, Dexter, Pennington *et al.* 2010, Lou and Golding 2010, Yassin, Markow *et al.* 2010), as indicated by our findings in which methods proved not to be equally robust with regard to incomplete lineage sorting effects in recently diverged species (Figures 3.4 and 3.7). As such species are usually of special interest scientifically or from regulatory perspective (Armstrong, Cameron *et al.* 1997, Wallman and Donnellan 2001, Boykin, Shatters *et al.* 2006, Nolan, Carpenter *et al.* 2007, Aveskamp, Woudenberg *et al.* 2009, Paredes-Esquivel, Donnelly *et al.* 2009, Azpurua, De la Cruz *et al.* 2010, Dexter, Pennington *et al.* 2010, Skoracka and Dabert 2010), yet also difficult to identify using morphology (Kaila and Stahls 2006, Nolan, Carpenter *et al.* 2007, McBride, van Velzen *et al.* 2009, Newmaster and Ragupathy 2009, van Velzen, Larsen *et al.* 2009, Dexter, Pennington *et al.* 2010), finding robust analytical methods is warranted, and commonly used methods such as neighbor joining may not suffice.

3.4.1 Method performance

Tree-based methods

Our results based on simulated data of recently diverged species show that DNA barcode identification of recently diverged species can be significantly improved by applying methods that do not rely on tree representation. The two tree-based methods tested here, i.e. neighbor joining (NJ) and parsimony (PAR), perform worst in terms of query identification success, even with liberal assignment. This finding is in concordance with results from other studies comparing relative performance of

DNA barcoding methods (Meier, Shiyang *et al.* 2006, Little, Stevenson *et al.* 2008, Virgilio, Backeljau *et al.* 2010, Little 2011), as well as with the generally accepted notion that gene trees (i.e. DNA barcode trees) do not necessarily reflect organismal history (Nichols 2001).

PAR consistently and significantly achieved the lowest identification rates here. We see two possible explanations for this result: First, heuristic searches are not guaranteed to find the shortest (i.e. most parsimonious) tree(s) and our search settings may have been insufficiently thorough (Goloboff 1999). Further analysis of some data sets with more thorough search settings did not result in shorter trees being found, however (data not shown), indicating that settings were in fact adequate. Second, several equally parsimonious trees may exist of which only one was used for identification here. Having chosen randomly among equally parsimonious trees may therefore have affected results negatively. NJ will always find a single, fully resolved tree (Saitou and Nei 1987) which may have more biological relevance than a randomly chosen maximum parsimony tree, hence resulting in more correct identifications using NJ. We did not include barcode query identification based on a consensus of all most parsimonious trees, but because a consensus tree by definition has reduced resolution we do not expect this could increase performance of PAR.

For both tree-based methods (i.e. NJ and PAR) strict assignment (i.e. requiring a query to be nested within a monospecific clade for identification) significantly reduced identification success compared to liberal assignment (i.e. allowing identification of a query that is sister to a monospecific clade). This was as expected because when a query is sister to a monospecific clade strict assignment yields an uncertain identification whereas liberal assignment will assign it to the species associated with that clade (Ross, Murugan *et al.* 2008). Although identification can be wrong in some of these cases, even few correct identifications will result in a higher success rate for liberal assignment compared with strict assignment (Little, Stevenson *et al.* 2008, Ross, Murugan *et al.* 2008).

There are other tree-based methods for matching DNA barcodes available but we expect that these do not outperform NJ as tested here. For example, Bayesian methods for tree inference (Huelsenbeck and Ronquist 2001) do not find a single, fully resolved tree and will therefore share the drawbacks of PAR. The Statistical Assignment Package (SAP) (Munch, Boomsma *et al.* 2008) was already found to perform less well than NJ on a Gymnosperm multi-locus DNA barcode data set, even when using the 'constrained NJ' algorithm for tree estimation (Little 2011).

Similarity-based and diagnostic methods

These methods perform significantly better with 31% reduction of error rates compared to tree-based methods (26% when counting tree-based results using liberal assignment only), see Table 3.2 and Figure 3.7. Although not significant, diagnostic methods (i.e. BLOG and DNA-BAR) outperformed all other methods tested here.

This confirms their suspected superiority as they allow selecting differentiating characters whilst ignoring any obscuring within-species variation (DeSalle, Egan *et al.* 2005). Obviously, diagnostic methods are not guaranteed to have this advantage in all cases. For example, in another study (Little 2011) the diagnostic method CAOS (Sarkar, Planet *et al.* 2008) did not perform well; possibly because it is dependent upon tree topology for extracting diagnostic characters. The two similarity-based methods (i.e. NN and BLAST) performed only slightly worse compared to the diagnostic methods. This may seem surprising because of the large overlap of within- and between-species distances in our data sets (see Figure 3.3). But even when there is no 'barcode gap' for a particular species, the closest match for a query sequence can well be conspecific, resulting in correct identification (Meier, Shiyang *et al.* 2006). The two methods tested here either require (NN) or produce (BLAST) a sequence alignment, but reliable homology assessment and alignment can be problematic when sequences are variable in length (Little, Stevenson *et al.* 2008, Kuksa and Pavlovic 2009). Alternative similarity-based methods have been proposed that make a projection of sequences based on the decomposition of sequence strings and are therefore in effect alignment-free (Little, Stevenson *et al.* 2008, Austerlitz, David *et al.* 2009, Kuksa and Pavlovic 2009, Seo 2010, Little 2011). String decomposition can be performed in various ways, however, and optimal settings may differ between data sets. For example, preliminary tests of query identification using the recently proposed alignment-free method BRONX (Little 2011) showed high success rates for the multi-locus *Inga* data set (90.1%) but very low success rates for the *Drosophila* (53.4%) and Cypraeidae (74.6%) data sets, using the same (default) settings (data not shown).

Although diagnostic and similarity-based methods show similar performance in terms of correct query identification, they markedly differ in their computational cost. Similarity methods such as NN and BLAST are computationally relatively inexpensive because they only involve finding a query's closest match (Ross, Murugan *et al.* 2008). By contrast, diagnostic methods must select and extract diagnostic characters, which is computationally expensive (Bertolazzi, Felici *et al.* 2009). As an example, while the NN analysis of a simulated data set took only ~2 s on a 3GHz dual core desktop computer, analyzing the same data set with BLOG required ~7 min of computation (both analyses using one thread only). Nevertheless, a similarity analysis such as NN has to be repeated for every query sequence requiring identification, thus multiplying the computation time by the number of queries. Diagnostic characters, once they are extracted, can be used to identify any query sequence by simply matching it to these diagnostics – which is much faster than similarity matching in the case of BLOG.

An essential advantage of BLOG over all other methods tested here is that the diagnostic logic formulas extracted by BLOG contain additional information with regards to species identification (Bertolazzi, Felici *et al.* 2009). Such formulas list the nucleotide(s) by which a species can be differentiated from others and as such can

3

be compared with species descriptions in the traditional taxonomic sense (Goldstein and DeSalle 2011). Other methods can then be compared with trying to match an unknown specimen to all specimens in a collection. We envision that the logic formulas can provide valuable information for other applications. For example, the formulas can be included in species descriptions and taxonomic revisions (Damm, Schierwater *et al.* 2010), whereas relative similarities cannot. Obviously, diagnostic formulas exist only relative to a particular alignment but the same is true for morphological characteristics traditionally used for describing species, and in well-sampled clades this problem may well disappear. Diagnostic logic formulas can also be used for designing detection assays based on species-specific nucleotides (e.g. DNA chips and microarrays) and hence assist the development of tools for monitoring and regulation of species. For this purpose DNA-BAR is potentially even better suited than BLOG because it extracts diagnostics that are (combinations of) actual sequence strings that can be used as DNA probes (DasGupta, Konwar et al.). However, DNA-BAR does not incorporate species-level information in its analysis and selects diagnostics for sequences rather than for species (DasGupta, Konwar et al.). Moreover, diagnostics selected by DNA-BAR appear to be much more complex than the diagnostic logic formulas extracted by BLOG (personal observations), making DNA-BAR less suitable for extracting species-specific information.

The greatest challenge for diagnostic methods is scalability. Because diagnostic characters are dependent on their context, finding simple diagnostics becomes more difficult with increasing size of the reference database. For example, preliminary analysis of a large data set with 3000 DNA barcodes from over 600 bird species (data not shown) indicate that an alignment of such size is prohibitive for finding simple species-specific logic formulas using the current version of BLOG. Because datasets are ever increasing in size this is an important problem that can be in general tackled in different ways. With reference to this specific application, we see two solutions: 1. A similarity approach with some species groups flagged as 'problematic': Identification of a member of such group would then need to be confirmed with diagnostics specific for species in that group. 2. A combined similarity- and diagnostic approach where sequences are first binned into local alignments (e.g. at the level of families or genera) based on similarity; subsequently, diagnostics are applied only within these local alignments.

Statistical methods

We did not test any statistical methods for identification based on DNA barcodes. Nevertheless, when species identifications have economic or legal implications (e.g. in detection of quarantine organisms or forensics) there is an obvious need for probabilities associated with barcode matches. However, DNA barcode sequences are essentially short, meaning that they typically contain insufficient information to feed probabilistic models, especially when recently diverged species are concerned. We

therefore advocate confirmation of identifications based on DNA barcodes by other lines of evidence (e.g. multiple independent loci, serological tests or morphological expert opinion) rather than relying on DNA barcodes only in such cases.

Empirical data sets

Our results based on empirical data are largely consistent with results based on simulated data. Few differences in overall results exist, however: Where scores for tree-based NJ were suboptimal based on simulated data, they were comparable to at least some of the similarity-based and diagnostic methods when applied to the empirical data sets. For the *Drosophila* data set PAR performed equally well as NJ. It should be noted that with only three data sets assessing significance of differences in method performance is limited, underlining the advantage of using simulated data. In addition, DNA barcode identification success can depend on taxonomic sampling. In ‘regional’ data sets (i.e. samples from a particular geographic region only) within-species variation is usually underestimated because of un-sampled haplotypes, while between-species differences are usually overestimated because of un-sampled taxa (Moritz and Cicero 2004, Meyer and Paulay 2005, Meier, Shiyang *et al.* 2006, Elias, Hill *et al.* 2007, Wiemers and Fiedler 2007). Therefore, regional data sets such as *Inga* are expected to inflate DNA barcode identification success rates in contrast to ‘clade-based’ data sets (i.e. sampling all extant species across their entire distribution) such as Cypraeidae. Nevertheless, because the selected data sets comprise genetic markers from all three genomic compartments, result from different sampling efforts and represent broad phylogenetic diversity (i.e. insects, plants and gastropods) we interpret consistency in our findings as an indication that they will equally apply to other genetic markers and clades.

3.5 Conclusions

We found similarity-based (NN, BLAST) and diagnostic methods (BLOG, DNA-BAR) to significantly outperform tree-based methods (NJ, PAR), when applied to simulated DNA barcode data of recently diverged species. Diagnostic methods BLOG and DNA-BAR performed best on both simulated and empirical data and BLOG had the highest correct query identification rate overall. Although similarity-based methods have better scalability compared to BLOG they do not reveal any species-level information that can be used outside the realm of DNA barcoding. Diagnostic logic formulas extracted by BLOG provide information that can be used for e.g. taxonomy and species detection assays. Method choice therefore should depend on requirement of either computation speed or information content. In the end, recently diverged species remain difficult to identify, but we expect that our results contribute to alleviating this problem.

Acknowledgments

We thank the following people for kindly sharing empirical data sets for analysis: Brian Golding and Melanie Lou provided data of *Drosophila*, Kyle Dexter provided data of *Inga*, Chris Meyer provided data of Cypraeidae. Paola Bertolazzi and Guido Drovandi are acknowledged for their contribution to BLOG software engineering and scientific advice. Leandro Jones and Wayne Matten gave advice on configuration of TNT and BLAST, respectively. Marleen Botermans and Marc Sosef commented on early drafts of the manuscript and the editor and a reviewer provided valuable suggestions for further improvement.

Chapter 4

Species diversification in *Cymothoe* tropical forest butterflies (Lepidoptera, Nymphalidae)

Effects of changing climate and host plant association

Robin van Velzen[1,2], Niklas Wahlberg[3], Marc SM Sosef[1,2] & Freek T Bakker[1]

Biological Journal of the Linnean Society, 2013 108 (3), 546–564

[1] Biosystematics Group, Department of Plant Sciences, Wageningen University and Research Centre, Wageningen, The Netherlands
[2] Naturalis Biodiversity Centre (Section NHN), Wageningen University, Generaal Foulkesweg 37, 6703 BL Wageningen, The Netherlands
[3] Laboratory of Genetics, Department of Biology, University of Turku, Turku, Finland

Abstract

Evolution of novel host plant associations has been hypothesized to increase species diversification rates in butterflies because it allows for the colonization of new niches. Alternatively, such rates might be influenced by environmental factors such as geological events and changing climate. By generating a time-calibrated species-level molecular phylogenetic tree of the African forest butterfly genera *Harma* and *Cymothoe* we tested whether rates of net species diversification could best be explained by shifts to novel host plants or by palaeoclimatic factors.

Results show that, after the divergence of *Harma* and *Cymothoe* in the Miocene (15 Mya), net species diversification is low during the first 7 Myr. Coinciding with the onset of diversification of *Cymothoe* in the late Miocene (around 7.5 Mya) there is a sharp and significant increase in diversification rate, suggesting a rapid radiation. This increased rate did not correlate with host plant transition from Achariaceae to *Rinorea* (Violaceae) host plants, but rather with a period of global cooling and desiccation, indicating that tropical forest fragmentation may well have played a role in *Cymothoe* crown diversification.

4.1 Background

Evolution of a novel trait can promote species diversification because it allows entering new niches (e.g. Mayr 1963). This has been shown in animals (Clabaut, Bunje *et al.* 2007, Alfaro, Santini *et al.* 2009, Brakefield 2011), plants (Bakker, Culham *et al.* 2005, Merckx, Chatrou *et al.* 2008, Soltis, Albert *et al.* 2009, Johnson, FitzJohn *et al.* 2011) and Fungi (Berbee and Taylor 1993, Ahren and Tunlid 2003, Wang and Qiu 2006, Sung, Poinar Jr *et al.* 2008, Silva, Talhinhas *et al.* 2012). For butterflies, host plant association has been hypothesized to be such a trait (Ehrlich and Raven 1964) and indeed correlations between host plant shifts and increased rates of species diversification have been found in various butterfly clades (Weingartner, Wahlberg *et al.* 2006, Fordyce 2010). In addition to host plant association switches, species diversification rates in butterflies can be influenced by environmental factors such as geological events (Hall 2005, Mallarino, Bermingham *et al.* 2005, Wahlberg and Freitas 2007, Casner and Pyrcz 2010) and changing climate (Peña and Wahlberg 2008, Aduse-Poku, Vingerhoedt *et al.* 2009, Müller and Beheregaray 2010, Condamine, Sperling *et al.* 2012).

Here, we test whether rates of net species diversification (speciation minus extinction) in a clade of African forest butterflies can best be explained by shifts to novel host plants or by environmental factors. Although phylogenetic trees obviously provide at most correlation with and not evidence for causal effect, their power to estimate branching times and trait evolution in an absolute time-scale makes them ideal for studying species diversification (Rabosky 2006). Resolving the relative contribution of speciation and extinction to diversification on the basis of a phylogenetic tree is problematic, however, mainly because estimating rates of extinction is challenging without clear fossil evidence of extinct lineages (Rabosky 2010). Nevertheless, net species diversification can be estimated as long as significant phylogenetic sampling is achieved (Rabosky 2006). We selected the genera *Harma* Doubleday, 1848 and *Cymothoe* Hübner, 1819 (Nymphalidae, Limenitidinae) comprising a clade of butterflies confined to the forested regions of tropical Africa and Madagascar. Within this clade we see a sister relationship between monospecific *Harma* and *Cymothoe* comprising approximately 82 species (Ackery, Smith *et al.* 1995, Williams 2012). Apparently rates of net species diversification have differed between these genera.

Harma and *Cymothoe* also differ in larval host plant associations: within *Cymothoe* roughly half the species are highly specialized on particular species of *Rinorea* Aublet (Violaceae, Malpighiales). Most are even monophagous (Fontaine 1982, Amiet and Achoundong 1996, McBride, van Velzen *et al.* 2009), see Figure 4.1. The other species of *Cymothoe* feed exclusively on species of Achariaceae (also Malpighiales), which are also host to *Harma* (van Son 1979, Kielland 1990, Larsen 1991, Pringle, Henning *et al.* 1994, Amiet and Achoundong 1996). With respect to most other ecological traits, species of *Harma* and *Cymothoe* are highly similar: they are all

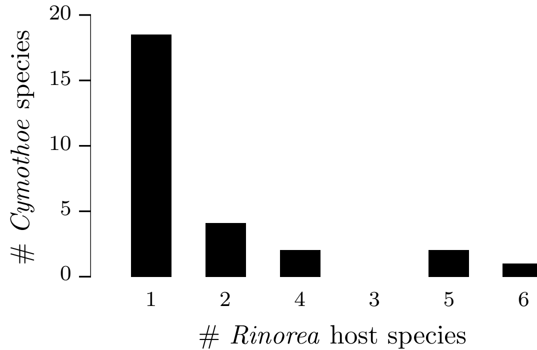


Figure 4.1: Host plant specialization of *Cymothoe* species feeding on *Rinorea*. Frequencies of number of *Rinorea* host plants per *Cymothoe* species.

forest butterflies, frugivorous and sexually dimorphic. This points to a high degree of niche conservatism and multiple species can usually be found in syntopy (i.e. living together at the same locality) if the relevant host plants are available (Amiet and Achoundong 1996, Larsen 2005). Within *Cymothoe*, egg clutch size is variable but correlated with host plant use: most species lay single eggs, possibly to minimize risk of predation, whereas some species associated with cyanogenic Achariaceae lay clutches with dozens of eggs on the same leaf and their larvae live gregariously (Amiet 2000). Given that most other life history and ecological traits are relatively constant within the *Harma*-*Cymothoe* clade, host plant shifts between Achariaceae and *Rinorea* would therefore *a priori* appear the most likely intrinsic trait affecting diversification rates in *Cymothoe*.

Alternatively, because nearly all species of *Harma* and *Cymothoe* are confined to wet forests (Larsen 2005), (historic) forest fragmentation could potentially lead to reproductive isolation with subsequent allopatric speciation in this clade. For example, sister species *C. egesta* Cramer, 1775 and *C. confusa* Aurivillius, 1887 occur nearly allopatric with only a small zone of overlap in Cameroon, where they were found to feed on distinct *Rinorea* host species (Amiet 1997), and population genetic analyses suggest that they indeed have allopatric origins (McBride, van Velzen *et al.* 2009). In another case, the closely related species *C. caenis* Drury, 1773 and *C. druryi* van Velzen & Larsen, 2009 are also geographically separated (van Velzen, Larsen *et al.* 2009), again suggesting allopatric speciation. Over geological time scales, Africa has experienced large fluctuations in climate and associated vegetation cover (Coetzee 1993, Jacobs 2004, Segalen, Lee-Thorp *et al.* 2007). The earliest evidence for angiosperm rainforest in Africa is from the Palaeocene (55–65 Mya), after which the lowland forest biome reached a peak in the late Eocene and Oligocene (23–40 Mya) (Jacobs 2004). The grass-dominated savannah biome began to expand in the middle Miocene (16 Mya) and became widespread in the late Miocene (8 Mya) at the expense of wet forest habitat (Morley and Richards 1993, Senut, Pickford *et*

al. 2009). Subsequent Pleistocene climatic fluctuations resulted in several cycles of fragmentation and expansion of the areas occupied by lowland rain forest (Dupont, Jahns *et al.* 2000, Cohen, Stone *et al.* 2007, Dupont 2011). Given the apparent niche conservatism of *Harma* and *Cymothoe* with respect to forest habitats, and given the approximate age of Limenitidinae at around 57 My (Wahlberg, Leneveu *et al.* 2009), climatic events are therefore likely candidates of environmental factors influencing their diversification.

Our aims were to: 1. Generate a species-level molecular phylogenetic tree for *Harma* and *Cymothoe*, calibrated in an absolute time scale; 2. Identify shifts in species diversification; 3. Reconstruct ancestral host plant associations based on contemporary associations; and 4. Assess correlations between species diversification and host use as well as climate and forest fragmentation.

4.2 Materials and methods

4.2.1 Taxon sampling

We included 52 species of *Cymothoe* (covering 63% of known extant species) and monospecific *Harma* in our study, mostly from newly collected specimens, either collected by RvV or kindly donated by a network of collectors (see acknowledgements), or from museum specimens obtained from the African Butterfy research institute (ABRI) (Nairobi), from RMCA (Tervuren), and from NHM (London). We effectively sampled all species that are morphologically divergent or represent major clades based on adult and larval morphology (Amiet 2000). Although relationships between the *Harma-Cymothoe* clade and other Limenitidinae are largely unknown, we chose representatives from three different Limenitidinae tribes as outgroup: *Neptis ida* Moore, 1858 (Neptini), *Lebadea martha* Fabricius, 1787 (incertae sedis) and *Limenitis reducta* Staudinger, 1901 (Limenitidini), resulting in a total of 56 taxa for which the accession, locality and other meta data are given in Table 4.1. Please note that some species names included here (e.g. *C. superba*, *C. butleri*) were not corroborated in our DNA barcoding study (see Chapter 2), and that *C. baylissii* ined. is unpublished.

4.2.2 Molecular methods

We extracted DNA from one or two legs, paper-dried or freshly preserved in 96% ethanol using the QIAgen DNeasy Blood & Tissue kit according to the manufacturer's instructions. We sequenced five genes that are known to be informative at the species and genus level (Peña and Wahlberg 2008, Wahlberg, Leneveu *et al.* 2009): Cytochrome *c* oxidase subunit 1 (COI) from the mitochondrial genome; *wingless* (*wgl*), ribosomal protein S5 (*RpS5*), glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*), and isocitrate dehydrogenase (*IDH*) from the nuclear genome. Primers

Table 4.1 Voucher information and GenBank accession numbers.

Species	Sex	Voucher	Collector	Collection date
C. adela	Male	SS_041	S. Szabolcs	17–26 March 2009
C. alcimeda	Female	SW_005	S. Woodhall	2006
C. althea	Male	TL_011	T. B. Larsen	April 2006
C. alticola				
C. altisidora	Male	JW_023	J. Wieringa	1 February 2008
C. amaniensis				
C. amenides	Male	RV_364	R. van Velzen	25 May 2006
C. angulifascia				
C. anitorgis	Female	OB_023	O. Brattström	11 February 2009
C. aramis	Male	OB_060	O. Brattström	28 April 2010
C. arcuata				
C. aubergeri	Female	CREO_110	R. Onstein	15 November 2009
C. aurivillii	Male	ABRI_098	M. Hassan	2006
C. baylissii ined.	Female	JB_001	J. Bayliss	November 2010
C. beckeri	Female	RV_386	R. van Velzen	2 June 2006
C. bouyeri	Male	ABRI_402	R. Ducarme	December 2008
C. butleri	Male	RV_392	R. van Velzen	27 July 2008
C. caenis		NW_10216	F. Molleman	
C. capella	Female	GW_4401	G. vandeWeghe	7 January 2004
C. caprina	Female	FM_180	F. Molleman	20 August 2004
C. coccinata	Female	RV_414	R. van Velzen	17 November 2010
C. collarti	Female	GW_13018	G. vandeWeghe	15 April 2007
C. collinsi				
C. colmanti	Male	TB_8923	T. Bouyer	June 2011
C. confusa	Female	RV_332	R. van Velzen	15 May 2006
C. consanguis	Male	RW_052	R. Warren	9 April 2009
C. coranus	Male	SW_13025	S. Woodhall	6 August 2006
C. cottrelli	Male	ABRI_087	S. C. Collins	July 2004
C. crocea	Male	FM_183	F. Molleman	26 August 2004
C. cyclades	Female	ABRI_330	P. Walwanda	February 1996
C. distincta	Male	FM_176	F. Molleman	26 August 2004
C. druryi	Male	TL_031	T. B. Larsen	April 2006
C. dujardini				
C. egesta	Female	OB_058	O. Brattström	1 April 2010
C. eris	Male	RMCA_242	Unknown	15 July 1989
C. euthalioides				
C. excelsa	Male	GW_14099	G. vandeWeghe	1 February 2008
C. fontainei	Female	GW_14237	G. vandeWeghe	4 February 2008
C. fumana	Male	CREO_119	R. Onstein	29 November 2009
C. haimodia	Male	GW_14221	G. vandeWeghe	3 February 2008
C. harmilla	Female	GW_12490	G. vandeWeghe	28 September 2007
C. hartigi	Male	SS_042	S. Szabolcs	17–26 March 2009
C. haynae	Male	GW_4404	G. vandeWeghe	7 January 2004
C. heliada	Female	GW_10623	G. vandeWeghe	31 March 2007
C. herminia	Female	RV_226	R. van Velzen	28 April 2006

Country	Location	COI	*wgl*	*GAPDH*	*RpS5*	*IDH*
Sierra Leone	Belebu	HE964949	HE964890	HE963090	HE964843	
South Africa	Port St. Johns	HE964951	HE964892	HE963092	HE964845	
Sierra Leone	Gola Forest	HE964953	HE964893	HE963093	HE964847	HE964799
Gabon	Alanga–Aboumi	HE964926	HE964871			
Cameroon	Londji 2	HE964938	HE964880	HE963082	HE964833	
Nigeria	Afi Mts.	HE964927		HE963072		
Nigeria	Afi Mts.	HE964930	HE964873	HE963075	HE964826	HE964791
Ghana	Kakum	HE964904	HE964852	HE963053	HE964804	HE964784
Tanzania	Kihansi Forest	HE964899				
Mozambique	Mt. Mabu	HE964924	HE964869	HE963069	HE964821	
Cameroon	Nkolo	HE964939	HE964881	HE963083	HE964834	HE964796
Congo DRC	Kasuo	BARCODE				
Kenya	Kakamega	HE964941	HE964883	HE963085	HE964836	
Uganda	Kibale Forest	GQ864754	GQ864442	GQ864952	GQ865420	GQ865083
Congo DRC	Bondo	HE964920	HE964866	HE963066	HE964818	
Cameroon	Doumo Pierre	HE964910		HE963058	HE964809	
Nigeria	Ologbo Forest	HE964942	HE964884	HE963086	HE964837	HE964798
Rwanda	Nyungwe Forest	HE964915	HE964861	HE963062	HE964813	
Congo DRC	Mamove	BARCODE				
Cameroon	Ducam-Duclair	HE964937	HE964879	HE963081	HE964832	HE964795
Nigeria	Rhoko	HE964946	HE964888		HE964841	
South Africa	Umdoni Parc	HE964952			HE964846	
Malawi	Nyika	BARCODE				
Cameroon	Doumo Pierre	HE964911	HE964857	HE963059	HE964810	HE964786
Congo DRC	Mt. Hoyo	BARCODE				
Cameroon	Doumo Pierre	HE964909	HE964856	HE963057	HE964808	
Sierra Leone	Gola Forest	HE964955	HE964895	HE963095		
Nigeria	Rhoko	HE964929	HE964872	HE963074	HE964825	
Cameroon	Moloundou	BARCODE				
Gabon	Lonmin	HE964916	HE964862	HE963063	HE964814	
Gabon	Lonmin	HE964918	HE964864	HE963064	HE964816	
Ghana	Kakum	HE964905	HE964853	HE963054	HE964805	HE964785
Gabon	Lonmin	HE964917	HE964863		HE964815	HE964788
Gabon	Waka	HE964914	HE964860	HE963061	HE964812	
Sierra Leone	Belebu	HE964950	HE964891	HE963091	HE964844	
Congo DRC	Bondo	HE964921			HE964819	
Gabon	Waka	HE964912	HE964858	HE963060	HE964811	HE964787
Cameroon	Mt. Kala	HE964934	HE964877	HE963079	HE964830	

Table 4.1 Voucher information and GenBank accession numbers, continued.

Species	Sex	Voucher	Collector	Collection date
C. hesiodina				
C. hesiodotus	Female	FM_008	F. Molleman	1 September 2004
C. hobarti	Female	RV_390	R. van Velzen	27 July 2008
C. howarthi				
C. hyarbita	Male	GW_9174	G. vandeWeghe	19 December 2006
C. hypatha	Female	GW_11340	G. vandeWeghe	22 April 2007
C. indamora	Male	RW_030	R. Warren	25 March 2008
C. isiro	Male	ABRI_060	R. Ducarme	December 2006
C. jodutta	Male	RV_060	R. van Velzen	13 April 2006
C. lambertoni	Male	ABRI_075	S. C. Collins	24 October 2010
C. lucasi	Female	GW_9483	G. vandeWeghe	21 December 2007
C. lurida	Male	TL_024	T. B. Larsen	April 2007
C. mabillei	Female	CREO_100	R. Onstein	26 October 2009
C. magambae	Male	ABRI_095	T.C.E. Congdon	16–30 March 2005
C. Magnus				
C. melanjae	Male	ABRI_083	Unknown	March/April 2008
C. meridionalis	Female	FM_167	F. Molleman	3 October 2004
C. nigeriensis	Male	RW_036	R. Warren	15 March 2008
C. ochreata	Female	ABRI_328	S. C. Collins	26 November 2004
C. oemilius	Female	RV_322	R. van Velzen	12 May 2006
C. ogova	Male	JW_002	J. Wieringa	21 March 2007
C. okomu	Female	RW_038	R. Warren	15 March 2008
C. orphnina	Female	TB_8118	T. Bouyer	May 2011
C. owassae	Male	ABRI_048	S. C. Collins	8–13 March 2007
C. preussii				
C. radialis				
C. reginae-elisabethae	Male	RD_077	R. Ducarme	3 September 2007
C. reinholdii	Male	RD_098	R. Ducarme	28 August 2007
C. sangaris	Male	RV_199	R. van Velzen	26 April 2006
C. sassiana				
C. serpentina				
C. superba	Male	GW_14794	G. Vande weghe	16 March 2008
C. teita	Female	DB_003	D. Bonte	17 March 2006
C. vumbui				
C. weymeri		SS_036	S. Szabolcs	29 March 2009
C. zenkeri	Male	ABRI_058	J. B. Ganiot	May 2007
C. zombana				
Harma theobene		NW102-8	F. Molleman	
Lebadea martha		NW100-13	T. B. Larsen	12 April 2004
Limenitis reducta		NW67-2	N. Wahlberg	26 April 2001
Neptis ida		NW98-3	C. Schulze	

Country	Location	COI	*wgl*	*GAPDH*	*RpS5*	*IDH*
Cameroon	Doumo Pierre	HE964908	HE964855	HE963056	HE964807	
Kenya	Kakamega	HE964940	HE964882	HE963084	HE964835	HE964797
Gabon	Tchimbélé	HE964922	HE964867	HE963067		
Gabon	Bateké	HE964913	HE964859			
Nigeria	Rhoko	HE964943	HE964885	HE963087	HE964838	
Congo DRC	Biakatu	BARCODE				
Cameroon	Mt. Kala	HE964932	HE964875	HE963077	HE964828	HE964792
Madagascar	Tsaratanana Mt.	HE964898	HE964848			
Gabon	Tchimbélé	HE964923	HE964868	HE963068	HE964820	HE964789
Ghana	Bobiri Forest	HE964954	HE964894	HE963094		
Ghana	Atewa	HE964903	HE964851	HE963052	HE964803	HE964783
Tanzania	South Pare Mts.	BARCODE				
Malawi	Mt. Mulanje	BARCODE				
Cameroon	Doumo Pierre	BARCODE				
Nigeria	Okomu	HE964944	HE964886	HE963088	HE964839	
Uganda	Budongo	HE964902	HE964850		HE964802	
Cameroon	Ducam-Duclair	HE964936	HE964878	HE963080	HE964831	HE964794
Gabon	Evouta	HE964925	HE964870	HE963070	HE964822	HE964790
Nigeria	Okomu	HE964945	HE964887	HE963089	HE964840	
Congo DRC	Mt. Hoyo	BARCODE				
Equatorial Guinea	Bioko; Moka	BARCODE				
Nigeria	Biakatu	BARCODE				
Congo DRC	Biakatu	HE964931	HE964874	HE963076	HE964827	
Cameroon	Mt. Kala	HE964933	HE964876	HE963078	HE964829	HE964793
Gabon	Lonmin	HE964919	HE964865	HE963065	HE964817	
Kenya	Ngangao	HE964906		HE963055	HE964806	
Ghana	Tano Ofin	HE964948	HE964889		HE964842	
CAR	Ndoloko	HE964896			HE964800	
Uganda	Kibale Forest	GQ864775	GQ864463	GQ864978	GQ865447	GQ865103
Bangladesh	Lowacherra Forest	GQ864784	GQ864472	GQ864991	GQ865460	GQ865116
France	Bagnoles, Aude	AY090217	AY090150	EU141509	EU141409	EU141568
Indonesia	Palolo Valley	EU141369	EU141250			

GenBank accession numbers starting with HE were newly generated for this study;
CAR = Central African Republic.

4

and laboratory protocols were taken from Wahlberg & Wheat (2008). Direct sequencing of polymerase chain reaction (PCR) products was performed on Applied Biosystems 3170xl Genetic Analyser at the University of Turku, or sent to Macrogen (South Korea) for sequencing. The resulting chromatograms were examined by eye in BioEdit (Hall 1999). All five genes are protein-coding, and thus alignment was trivial. GenBank accession numbers of the DNA sequences are given in Table 4.1.

4.2.3 Phylogenetic inference

Congruence tests

We tested the null hypothesis of congruence of phylogenetic signal between genes using the incongruence length difference (ILD) test (Farris, Kallersjo *et al.* 1994, Farris, Kallersjo *et al.* 1995, Cunningham 1997) as implemented by the partition homogeneity test in PAUP∗ 4.0b10 (Swofford 2003). For all ILD tests, uninformative (invariant and autapomorphic) characters were excluded and heuristic searches with random taxon sampling and tree bisection-reconstruction branch swapping were conducted. To establish a null distribution for each test, 1000 randomized data partitions of equal size to the originals were generated and ILDs were calculated for each replicate. The threshold for significance was a P-value of 0.01.

Model testing

We determined the relative fit of candidate models of nucleotide evolution for each gene and genomic compartment (mitochondrial versus nuclear) using JModelTest 0.1.1 (Posada 2008): three different substitution models (HKY, K80, and GTR) with or without estimated base frequencies, gamma-shaped distribution of rates (4 categories) and proportion of invariant sites – amounting to assessments of 24 different models. Models were optimised on maximum likelihood trees and best-fitting models of nucleotide evolution were selected based on the Akaike Information Criterion (AIC). Best-fitting models per partition are given in Table 4.2.

Inference

We estimated phylogenetic trees for each gene and genomic compartment using Bayesian Inference (BI) and maximum likelihood (ML) methods. We performed BI using MrBayes 3.1.2 (Ronquist and Huelsenbeck 2003), executing two independent Markov-chain Monte Carlo (MCMC) runs with four Metropolis-coupled chains each for 20 million generations and sampling every 1000 generations. Convergence of the two independent MCMC runs was assessed topologically (i.e. based on clade frequencies) using the online service AWTY (Nylander, Wilgenbusch *et al.* 2008) and based on model parameters using Tracer 1.5 (Rambaut and Drummond 2009). The

Table 4.2: Character partitions, their characteristics and models selected for phylogenetic inference.

Partition	#taxa	#characters	#informative	Model
mtDNA				
COI	56	1475	349	HKY+G
nDNA	56	2366	391	K80+I+G
wgl	50	363	363	SYM+I+G
GAPDH	46	692	104	GTR+I+G
rps5	49	597	77	GTR+I+G
IDH	21	714	116	GTR+G

#taxa = number of taxa; #characters = total number of characters; #informative = number of informative characters.

first 2 million generations (10%) were discarded as burn-in before calculating a 50-percent majority-rule consensus based on the posterior set of trees. We performed ML using Garli 2.0 (Zwickl 2006) with 16 independent search replicates, random starting trees and stopping each search when no better tree was found in 20 000 generations. Population and mutation settings for the genetic algorithm in Garli were left at their default values. In case only one best tree was found search replicates were incremented by eight until the best likelihood score was found multiple times independently. To estimate branch support we performed 100 bootstrap pseudo-replicates with a single search per pseudo-replicate. Bootstrapped trees were combined into a single file to calculate bootstrap values for all nodes. All BI and ML analyses were run on the online CIPRES science gateway (Miller, Pfeiffer *et al.* 2010).

Congruent data were combined and partitioned according to gene and genomic compartment. Phylogenetic trees based on the combined data were inferred using BI with 40 million generations per MCMC run and ML with 16 search replicates. Partitioning schemes were compared using Bayes factors between BI marginal likelihoods.

4.2.4 Timing of divergences

Time-calibrated phylogenetic trees were inferred using BEAST (Drummond and Rambaut 2007) at the online CIPRES science gateway (Miller, Pfeiffer *et al.* 2010). Analyses were based on combined data partitioned per genomic region and per gene. Because Limenitidinae fossils are unknown and hence unavailable for node calibration we used a putative secondary time calibration based on a recent study of the evolutionary history of Nymphalidae based on host-plant ages and 6 butterfly fossils (Wahlberg, Leneveu *et al.* 2009). This study had an estimate for the *Harma-Cymothoe* clade of 15.09 (95% HPD = 8.21–22.77) Mya. We set a prior distribution

for the age of the *Harma-Cymothoe* clade accordingly, which followed a lognormal distribution with log(mean) of 2.83 and log(standard deviation) of 0.26 Mya. Relaxed lognormal clocks were estimated for each genomic region separately to accommodate differences in mean substitution rates between mitochondrial and nuclear DNA. The *Harma-Cymothoe* clade was constrained to be monophyletic and we set a Yule prior on speciation with a uniform distribution between 0 and 5 for birth rate. To avoid problems associated with long branches, only *Neptis* was included as outgroup. Because uniform prior distributions for the mean substitution rates caused overestimation of age estimates in preliminary runs, we set a prior following a gamma distribution with a mean of 1 and a shape parameter of 0.001. All other priors were left at their default settings. We performed 4 independent MCMC runs with random starting trees, 40 million generations per run and sampling every 10 000 generations. The first 4 million generations (10%) of each run were discarded as burn-in.

4.2.5 Diversification analyses

Adding missing taxa

Diversification measures assume complete species sampling, whereas our phylogenetic data set includes only 63% of all extant species. Missing data are a common phenomenon in evolutionary studies and ignoring them can compromise analyses and produce incorrect results (Pybus and Harvey 2000, Nakagawa and Freckleton 2008, Garamszegi and Møller 2011). There are various techniques for correcting incomplete species sampling in diversification studies. Some deal with missing species directly, either by assuming that species sampling is random (FitzJohn, Maddison *et al.* 2009), or by considering clades with missing species as unresolved ("terminally unresolved trees"; Alfaro, Santini *et al.* 2009, FitzJohn, Maddison *et al.* 2009). Others generate a null distribution by randomly pruning taxa from simulated data with complete sampling (Harmon, Weir *et al.* 2008), or add missing species to phylogenetic trees before analysis (e.g. Purvis, Nee *et al.* 1995, Barraclough and Vogler 2002, Day, Cotton *et al.* 2008). A problem here is that estimates of speciation and extinction rates can be influenced by the way missing species are placed on the phylogenetic tree (Cusimano, Stadler *et al.* 2012). Recently, Cusimano *et al.* (2012) proposed a technique of simulating missing species under speciation/extinction models, thereby overcoming this problem. However, the missing species are simulated as branching times only and thus cannot be used for topology-based analyses.

We corrected for missing species by adding missing taxa as empty sequences at the tree inference stage in a Bayesian framework (Kuhn, Mooers *et al.* 2011). This has the advantage that the full suite of Bayesian phylogenetic tools (e.g. clock models, molecular evolutionary parameters, priors on tree topology) can be incorporated into the tree-building process along with the missing taxa (Kuhn, Mooers *et al.*

2011). In addition, it allows for retaining data from all sampled species, contrary to the terminally unresolved tree approach (e.g. Alfaro, Santini *et al.* 2009, FitzJohn, Maddison *et al.* 2009), thus taking all available phylogenetic data into account. We included 16 missing taxa in the data set as three 'N' codes and a short piece of COI (the DNA barcode) for an additional 13 species, amounting to a total of 82 species of *Cymothoe*.

Phylogenetic placement of missing species was controlled through monophyly constraints derived from morphological and taxonomic information (Amiet 2000, Larsen 2005) and implemented in BEAST. Hence, MCMC operators could move missing taxa at liberty but in accordance with monophyly constraints and the Yule prior on speciation (Kuhn, Mooers *et al.* 2011). All other settings were the same as for the divergence time analyses. Obviously, because of implementing a Yule prior, the resulting posterior set of trees is biased towards a constant rate of diversification. Because rate constancy is the typical null model for diversification rate analyses, the bias will be conservative, however (Kuhn, Mooers *et al.* 2011).

Temporal shifts in diversification

A lineage-through-time (LTT) plot based on the posterior set of trees with complete species sampling was generated using Tracer 1.5 (Rambaut and Drummond 2009), and compared with trends of global temperatures using oxygen isotope fractionation data in Benthic foraminifera from Zachos *et al.* (2001), which serve as a proxy for the total global mass of glacial ice sheets.

To test for a temporal shift in diversification rate we fitted a candidate set of rate-constant (Yule and birth-death) and rate-variable (DDX, DDl and Yule-2-rate) diversification models to the posterior set of trees using the R package LASER (Rabosky 2006). We recorded the decrease in AIC (ΔAIC) of the best fitting rate-variable model compared to the best-fitting rate-constant model as the test statistic (better-fitting models have lower AIC scores). In order to avoid Type I errors, the observed ΔAIC test statistics were compared with a null distribution of ΔAIC values based on fitting the same models on trees simulated under a constant-rate Yule model (Rabosky 2006).

Phylogenetic shifts in diversification

To test for branch-specific shifts in diversification rate we fitted a candidate set of nested diversification models with increasing complexity to the posterior set of trees using stepwise AIC in the R package MEDUSA (Alfaro, Santini *et al.* 2009). We set the maximum of fitted models to 5 (modelLimit=5) and selected the best fitting model that resulted in an improvement in AIC score above a threshold of 4.248. This corrected threshold ensures a significant increase ($p<0.05$) in model fit for trees with 82 tips, and was calculated automatically by MEDUSA.

Trait-dependent diversification

Host plant use (coded binary as Achariaceae or *Rinorea*) was optimised over the posterior set of 14,400 trees based on a reversible jump MCMC as implemented in BayesTraits (Barker, Meade *et al.* 2007). We fitted a continuous-time Markov model consisting of two parameters: the transition rates (q) from Achariaceae to *Rinorea* (qA→R) and from *Rinorea* to Achariaceae (qR→A). Posterior probabilities of the ancestral host plant association were estimated for *Harma* and *Cymothoe* combined, *Cymothoe* alone, and selected clades within *Cymothoe* using the 'Addnode' command. For the reversible jump model we set a hyperprior seeding an exponential distribution on the interval 0–30. The MCMC chain was run for 40 million generations and sampled every 10 000-th generation. The first sample was discarded as burn-in. In addition, we reconstructed ancestral host plant use at the tree inference stage using BEAST version 1.7.2 (Drummond, Suchard *et al.* 2012), allowing us to estimate the number of A→R and R→A transitions along each branch using Markov Jumps (Minin and Suchard 2008). We set a prior following a gamma distribution with shape 1 and scale parameter of 0.001 for the host plant transition rates to avoid overestimation of host plant transitions; all other settings were the same as for the complete species sampling analysis.

Correlation between host plant use and diversification rate was tested using binary-state speciation and extinction (BiSSE; Maddison, Midford *et al.* 2007) models implemented in the R package Diversitree version 0.7-2 (FitzJohn, Maddison *et al.* 2009). Because estimating extinction rates from phylogenetic trees is generally problematic (Rabosky 2010) and extinction rate estimates from preliminary analyses approached zero, we constrained the BiSSE model so that only net diversification rates were estimated (i.e. the single-parameter diversification model was realised by fixing the extinction parameter of the two-parameter BiSSE model at zero). Our BiSSE model therefore consisted of four parameters: the diversification rates of lineages associated with Achariaceae (λA) and with *Rinorea* (λR) inferred under the variable rate pure-birth model, and the transition rates from Achariaceae to *Rinorea* (qA→R) and *Rinorea* to Achariaceae (qR→A). Maximum likelihood estimates based on the BiSSE model were estimated for all 14 400 posterior trees to account for phylogenetic uncertainty. All trees were scaled to a total length of 1 before analysis in order to make diversification rate estimates based on different trees directly comparable. Species with unknown host plant use were re-coded to either *Rinorea* or Achariaceae feeders in concordance with results from BayesTraits and BEAST 1.7.2 to avoid overestimation of reversals. In addition, we used Bayesian methods to estimate posterior probability distributions for each of these four parameters using MCMC to account for uncertainty in both the phylogeny and parameter estimates. We used an exponential prior distribution for each parameter with a mean of two times ln(number of species) (Johnson, FitzJohn *et al.* 2011). We ran 5000 generations per tree with a step size of 2.5 for the speciation rates and 0.5

Table 4.3: P-values of ILD tests showing congruence between data partitions.

	mtDNA COI	*wgl*	*gapdh*	*rps5*	*idh*
wgl	0.9910				
gapdh	0.8040	0.0160			
rps5	1.0000	0.0540	0.4720		
idh	1.0000	0.9860	1.0000	0.9960	
nDNA	0.3680	0.2950	0.1740	0.2300	1.0000

for character transition rates, based on preliminary runs. The first 500 generations (10%) were discarded as burn-in and effective sample size of the remaining 4500 MCMC steps as calculated using the CODA package (Plummer, Best *et al.* 2006) was high (i.e. > 1095) for each parameter per tree separately. Because MCMC as implemented in Diversitree is computationally expensive, we analysed a subset of 1440 posterior trees (i.e. sampling every 100 000 generations), scaled to 1.

4.3 Results

4.3.1 Phylogenetic inference

Congruence tests

Although the incongruence length difference (ILD) test is generally susceptible to Type I errors (Darlu and Lecointre 2002) the null hypothesis of homogeneity between gene data sets was not rejected (P > 0.01, see Table 4.3). Therefore, the gene data sets were combined to maximise explanatory power (Bull, Huelsenbeck *et al.* 1993).

Inference

Trees based on different genes and genomic compartments were congruent and we found no conflict between trees inferred under ML and BI. All BI converged except for the analyses partitioned according to genes that, although they gave highest marginal likelihood values overall, experienced reduced convergence and mixing (measured as low effective parameter sampling sizes and exchange rates between chains). In order to improve exchange rates and mixing we re-ran the analyses with a heating temperature reduced from 0.20 to 0.10, but to no effect. Over-parameterization is known to impede convergence of Bayesian MCMC (Rannala 2002) and we suspect that this is also the case here. For these reasons we report BI results based on combined data partitioned according to genome, despite their smaller marginal likelihoods.

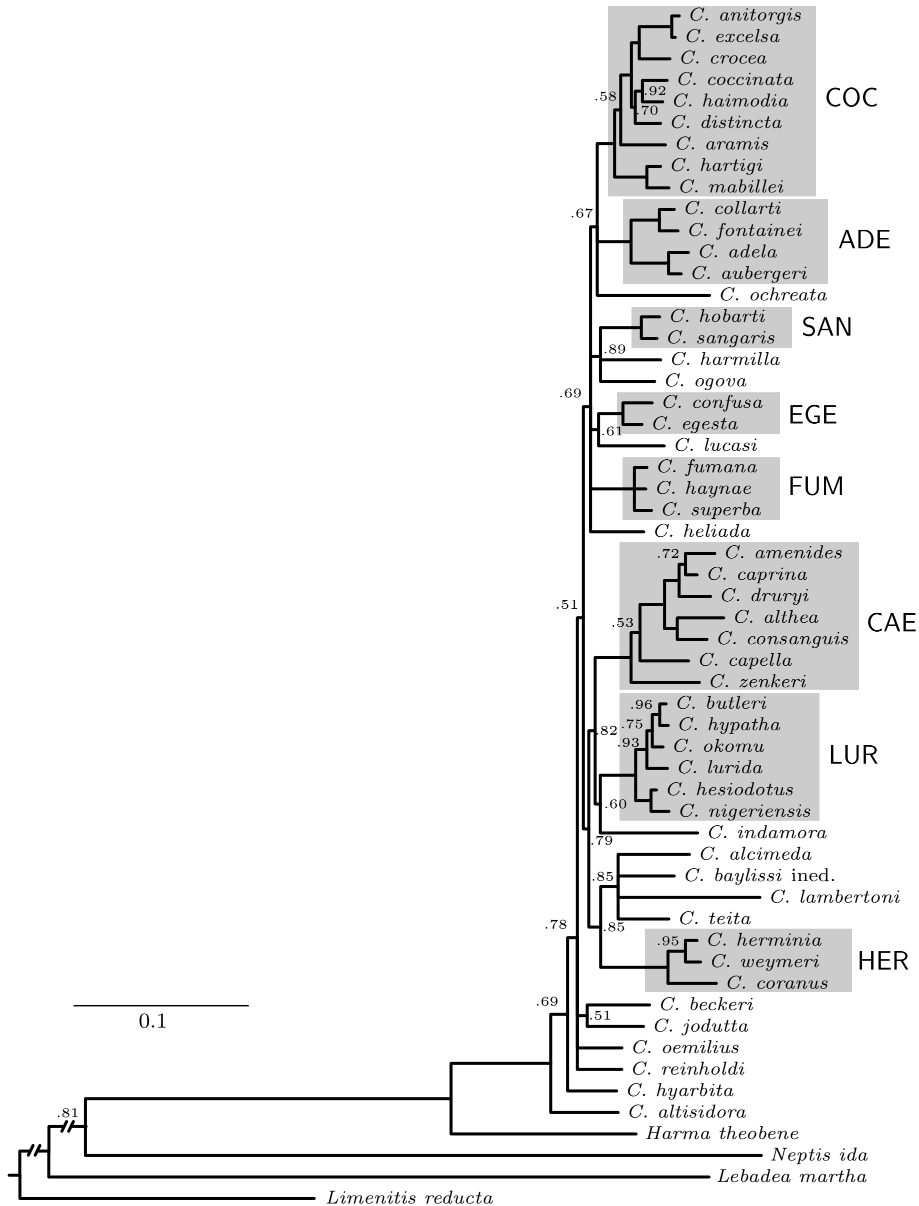


Figure 4.2: Phylogenetic tree of *Harma* and *Cymothoe*. Bayesian tree based on combined data partitioned according to genome. Scalebar indicates substitutions per site; branch labels indicate posterior clade probabilities below 1.00; highlighted clades have high support and are consistent with previous classifications based on morphology (Amiet 2000), see text.

Phylogenetic patterns

All inferred phylogenetic trees confirm that *Harma* and *Cymothoe* are sisters on a relatively long branch within Limenitidinae, with *Neptis ida* as the closest outgroup. The backbone of the clade comprising *Cymothoe* is largely unresolved (i.e its nodes have low or no support) with relatively short branches (see Figure 4.2). Below this backbone, eight clades are recovered with high support (i.e. posterior probability of 1.00) and correlate with previous classifications based on morphology (Amiet 2000).

The Coccinata clade (COC) consists of species characterized by small wing span and an orange to orangey-red ground colour of the males. The West African endemics *C. hartigi* Belcastro, 1990 and *C. mabillei* Overlaet, 1944 appear to be closely related and sister to all other species in the COC clade. The Adela clade (ADE; medium wing span, ochrous males) consists of two geographically separated pairs of sister species: a pair from West Africa with *C. adela* occurring in the Liberian and *C. aubergeri* Plantrou, 1977 occurring in the Ghana sub-region and a pair from Central Africa with *C. fontainei* Overlaet, 1952 occurring from Cameroon to central Congo DRC and *C. collarti* Overlaet, 1942 occurring in Kivu and Maniema provinces in eastern Congo DRC and in Rwanda. The Sangaris clade (SAN; small, blood-red) consists of species the males of which are highly similar but females are morphologically variable. The morphologically similar *C. ogova* Plötz, 1880 and the divergent *C. harmilla* Hewitson, 1874 appear to be related to this SAN clade. The Egesta clade (EGE; large, yellow) consists of the sister species *C. egesta* and *C. confusa* that are characterized by the males having an almost black transversal band over all wings. The morphologically similar but uncommon *C. orphnina* Karsch, 1894 (unsampled) is expected to be member of this clade too. The Fumana clade (FUM; large, yellow) consists of *C. fumana* Westwood, 1850, *C. haynae* Dewitz, 1887 and *C. superba* Aurivilius, 1898 which are characterized by the males having a mostly dark hindwing. Within the Caenis clade (CAE; small, creamy-white) the morphologically similar *C. consanguis* Aurivillius, 1896 and *C. althea* Cramer, 1776 appear to be sisters. The Lurida clade (LUR; large, ochrous-yellow or ochrous-orange) is characterized by an acutely angled forewing apex. Within this clade, the ochrous-orange *C. hesiodotus* Staudinger, 1890 and *C. nigeriensis* Overlaet, 1942 appear closely related and sister to the four ochrous-yellow species. The extremely rare *C. hesiodina* Schultze, 1908 (unsampled) is probably also related to this clade. The Herminia clade (HER; small, creamy-white) consists of the morphologically similar sister species *C. herminia* Grose-Smith, 1887 and *C. weymeri* Suffert, 1904, that both occur from West Africa to the Albertine Rift, and of the East African *C. coranus* Grose-Smith, 1889. Relationships of morphologically divergent *C. ochreata* Grose-Smith, 1890, *C. lucasi* Doumet, 1859, *C. heliada* Hewitson, 1874, *C. indamora* Hewitson, 1866, *C. beckeri* Herrich-Schaeffer, 1858, *C. jodutta* Westwood, 1850, *C. reinholdi* Plötz, 1880, *C. hyarbita* Hewitson, 1866, and *C. altisidora* Hewitson, 1869 remain unresolved.

4.3.2 Timing of divergences

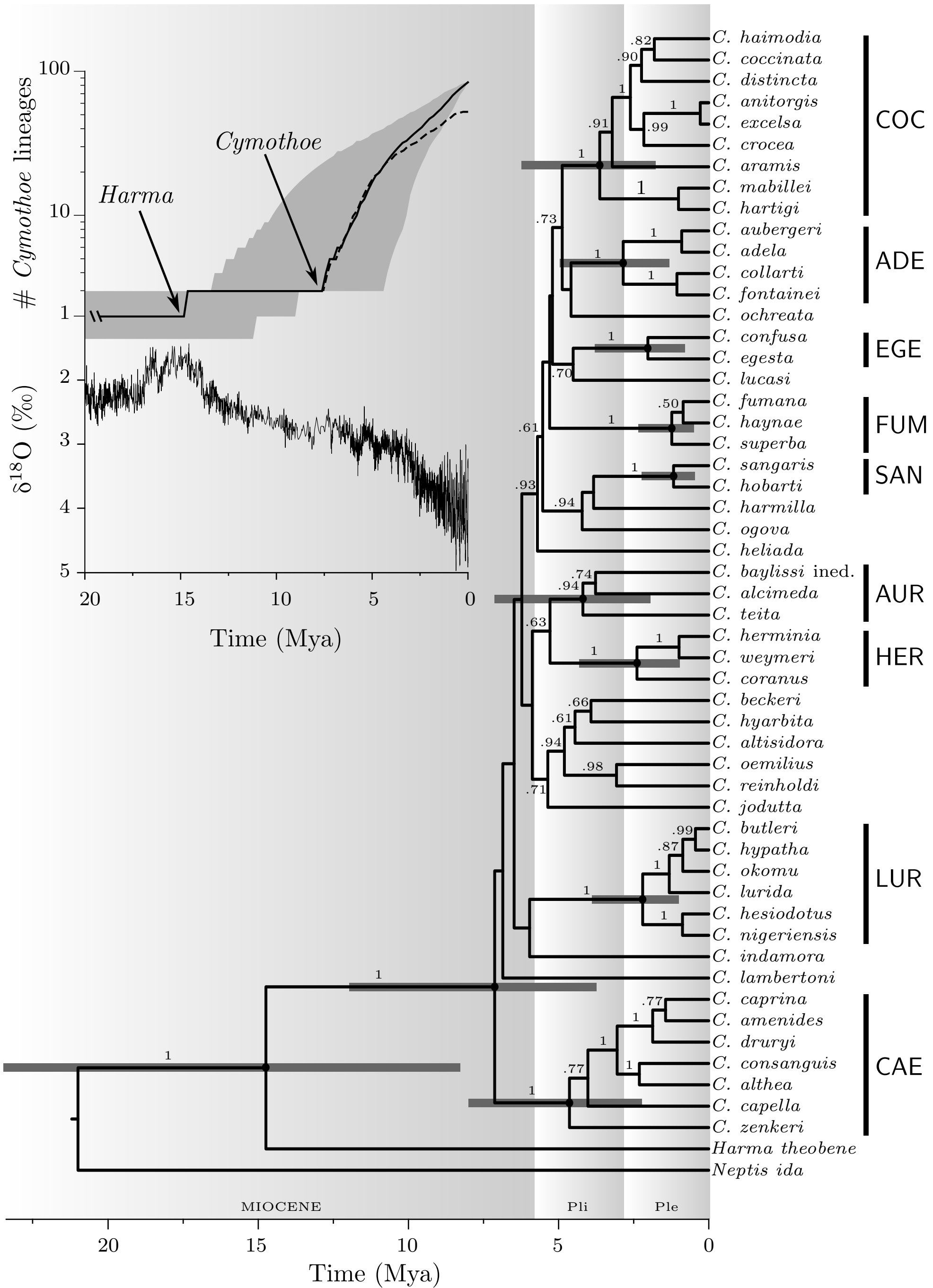
The most recent common ancestor of extent lineages of *Cymothoe* is estimated at 7.13 Mya (see Figure 4.3). Most crown clades described in the previous section originate not long after: CAE is the oldest with 4.65 Myr, followed by AUR with 4.19 Myr, COC with 3.63 Myr, ADE with 2.85, HER with 2.39 Myr, LUR with 2.2 Myr. The clades FUM (1.23 Myr) and SAN (1.17 Myr) are the youngest ones. Isolated species are also relatively old within the genus: the lineage leading to the Malagasy endemic *C. lambertoni* Oberthür, 1923 being most notable with an estimated age of 7.13 Myr, but this may be an artefact of its proportionally long branch compared to all other species, pushing it backwards in time in the relaxed clock analysis. The lineage leading to *Cymothoe indamora* is 5.96 Myr old; *C. heliada* is 5.7 Myr old; *C. jodutta* is 5.4 Myr old; *C. lucasii* is 4.5 Myr old. The other morphologically divergent species *C. beckeri, C. hyarbita, C. altisidora, C. oemilius* and *C. reinholdi* appear as a clade here (pp = 0.94, tmrca = 4.8 Mya), with the latter two species being sisters (pp = 0.98) of 3.1 Myr old. In addition, the East and South African endemics *C. teita* van Someren, 1939, *C. baylissii* ined., and *C. alcimeda* Godart, 1824 appear as a clade (pp = 0.94, tmrca = 4.2 Mya). Other East African montane endemics such as *C. aurivillii* Staudinger, 1899 (unsampled) are morphologically very similar to *C. teita* and are expected to be members of this Aurivillius (AUR) clade too.

4.3.3 Diversification analyses

Temporal shifts in diversification

Net species diversification is low during the first 7 Myr after the divergence of the *Harma* and *Cymothoe* lineages in the Miocene, followed by a sharp increase coinciding with the onset of diversification of extant lineages of *Cymothoe* in the late Miocene, around 7.5 (95%HPD = 3.68–11.97) Mya. The increased rate of net speciation correlates with a global trend of gradual cooling (Zachos, Pagani *et al.* 2001), see Figure 4.3.

Figure 4.3: Timing of divergences (next page). Time-calibrated maximum clade credibility tree. Pli = Pliocene; Ple = Pleistocene; branch labels indicate posterior clade probabilities above 0.50; vertical bars mark clades that have high support (pp = 1.00) and are consistent with previous classifications based on morphology (Amiet 2000). Inset shows species diversification rates in the *Cymothoe* and *Harma* clade and global temperature through time. Top: lineage-through-time (LTT) plot based on the posterior set of trees. Solid line shows number of lineages based on complete species sampling (shaded area indicates 95% HPD interval); dashed line shows number of lineages based on divergence time analysis with incomplete species sampling. Bottom: oxygen isotope fractionation data in Benthic foraminifera which serve as a proxy for the total global mass of glacial ice sheets and temperature; redrawn from Zachos et al. (2001)



4

Rate-variable models of diversification fitted our data best in 14 272 of 14 400 posterior trees (99.1%) with pure-birth (i.e. Yule) being the best fitting rate-constant model (97.8%) and Yule-2-rate being the best rate-variable model (99.7%). Mean ΔAIC was 5.94 indicating that rate-variable models have a much better fit to our data than rate-constant models ($p = 0.070$), see Figure 4.4. Fitting the Yule-2-rate model to our data revealed a temporal shift with a 5-fold increase in the mean diversification rate.

Phylogenetic shifts in diversification

Branch-specific shifts in diversification rates gave a significant better fit than a constant rate on all 14 400 posterior trees (100%). A single rate shift in diversification rate was optimized on 14 074 posterior trees (97.7%), two rate shifts on 324 (2.3%) trees and 3 rate shifts on the 2 remaining trees. The first (or only) shift constituted a rate increase coinciding with the *Cymothoe* clade (76.2%), or with the *Cymothoe* clade excluding the LAM clade (9.1%), *C. indamora* (5.4%), or both LAM and *C. indamora* (9.2%). When two rate shifts were optimized, the second shift generally constituted a rate decrease coinciding with the LAM clade (1.9%) or the LAM clade and *C. indamora* combined (0.2%) within *Cymothoe*. As mentioned earlier, both lineages are subtended by relatively long branches explaining their estimated low diversification rate compared with the rest of the *Cymothoe* clade.

Trait-dependent diversification

According to the BayesTraits optimizations, Achariaceae are the reconstructed ancestral host association for the *Harma-Cymothoe* clade ($pp = 0.985$) as well as for *Cymothoe* ($pp = 0.991$). Within *Cymothoe*, host plant association appears to be conservative with a reconstructed ancestral association exclusively with *Rinorea* for the COC, ADE, EGE, FUM, SAN, and LUR clades ($pp > 0.965$) and exclusively with Achariaceae for the AUR HER and CAE clades ($pp > 0.983$). Reversible jump MCMC favoured a single-parameter model with $qR \rightarrow A$ set to zero ($pp = 0.772$) and the mean estimated rate of $qA \rightarrow R$ was 6.7 times that of $qR \rightarrow A$. The BEAST ancestral host plant reconstruction corroborated Achariaceae as ancestral host association and optimized a single shift to *Rinorea* feeding ($pp = 0.88$), and no reversals ($pp = 0.98$), see Figure 4.4. We should point out, however, that the clade uniting all *Rinorea*-feeders received increased posterior probability compared with the same analysis without host data. This is because the optimized trait contributes to the tree likelihood in BEAST and consequently the outcome could be biased towards fewer shifts. The BiSSE analysis confirmed that transitions between Achariaceae and *Rinorea* association are (uni)directional: with $qA \rightarrow R$ being 2 times $qR \rightarrow A$ and the posterior density distribution of $qR \rightarrow A$ approaching zero. *Rinorea*-feeding lineages show higher rates of diversification ($\lambda R = 2.68$, sd

= 0.44) than those feeding on Achariaceae (λA = 2.17, sd = 0.36), see Figure 4.4. This difference was not significant, however (p = 0.157), and cannot explain the apparent large difference in diversification between *Harma* and *Cymothoe*.

4.4 Discussion

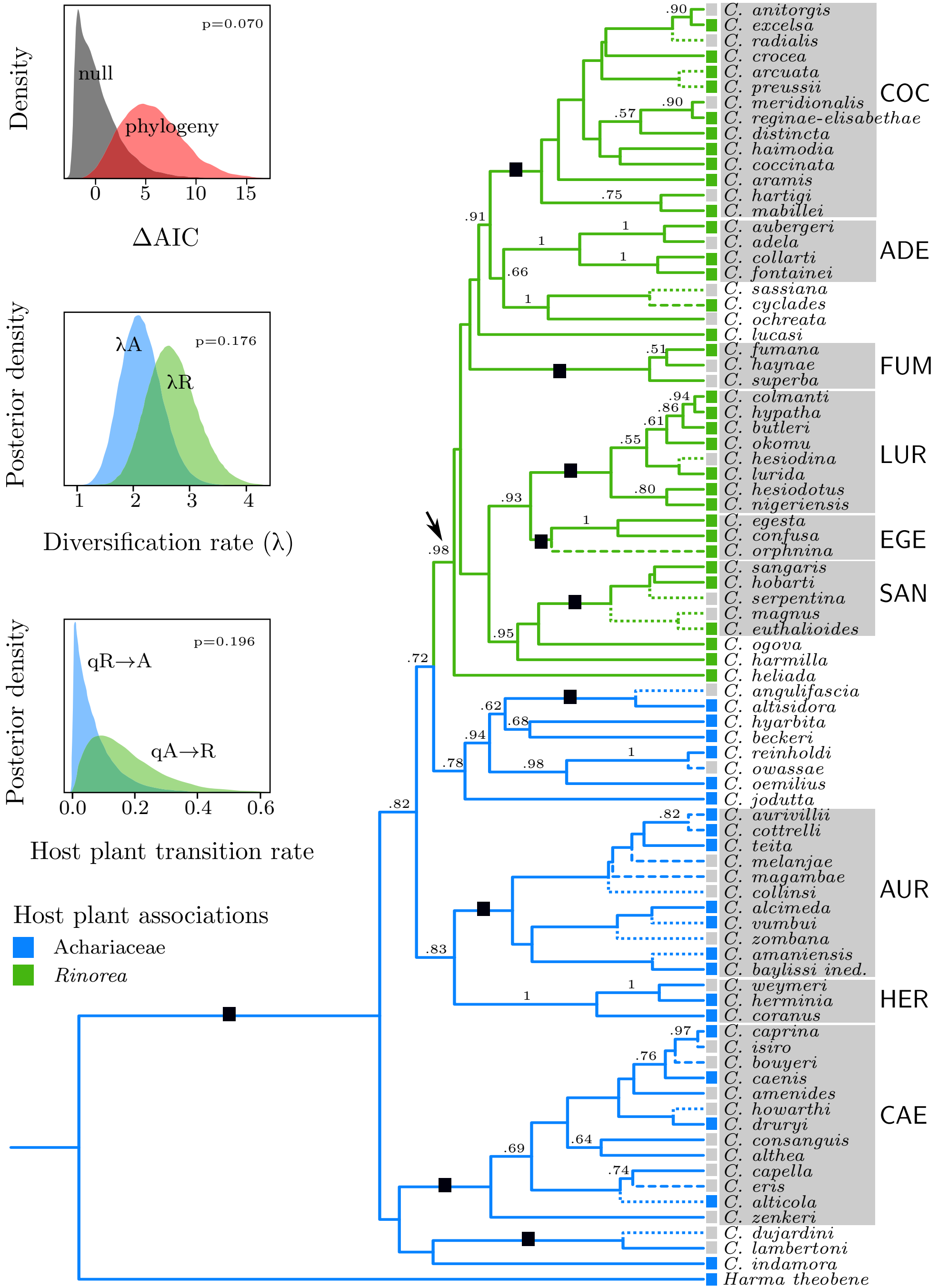
Based on our analyses we inferred a significant shift in species diversification rate of *Cymothoe* butterflies consistent with the late Miocene (7.5 Mya). To date, estimates of such diversifications in butterflies have been scarce and, compared with other species-level butterfly clades studied, patterns of diversification in *Cymothoe* and *Harma* are special in two ways. First, the estimated age is relatively young as, in general, most butterfly genera are thought to have diversified after the Cretaceous-Paleogene (K/Pg) boundary at 65 Mya (Wahlberg 2006, Heikkila, Kaila *et al.* 2012). This is echoed from various dating studies conducted over the last decade indicating Eocene to mid Miocene timeframes for most butterfly clades. Within Papilionidae, the genus *Papilio* (193 spp.) is thought to have diversified in the Oligocene 23–35 Mya and *Parnassius* (38 spp.) originated in the early Miocene 13–21 Mya (Nazari and Sperling 2007, Condamine, Sperling *et al.* 2012). Within Nymphalidae, genera have diversified to produce extant species from the mid Eocene (47 Mya) to present (Wahlberg, Leneveu *et al.* 2009). Satyrinae diversified in the Oligocene 23-36 Mya coinciding with the spread of their grass host plants (Peña and Wahlberg 2008) with the Satyrini tribe undergoing a quick radiation between 32 and 24 Mya (Peña, Nylin *et al.* 2011) and the Dirini tribe experiencing elevated diversification rates at 24-29 Mya (Price, Villet *et al.* 2011). Nymphalinae genera appear to have diversified during the early Miocene, with species diversification starting 23 Mya in the American subtribe Phyciodina (89 spp.) (Wahlberg and Freitas 2007), 21.7 Mya in *Melitaea* (Leneveu, Chichvarkhin *et al.* 2009), and 20 Mya in *Junonia* (Kodandaramaiah and Wahlberg 2007). Within Limenitidinae, the latitudinal gradient of species richness observed in *Adelpha* butterflies is the result of an increased diversification rate in the mid Miocene 10-15 Mya (Mullen, Savage *et al.* 2011). *Within Danainae, the genera Ithomia* (14.4 Mya) and *Napeogenes (*12.7 Mya*) started diversifying* in the Andes in the mid Miocene (Elias, Joron *et al.* 2009). In contrast, the relatively young age of *Cymothoe* diversification appears to be rare and only consistent with two other studies: the Asian tropical Lycaenid genus *Arhopala* (over 120 spp.) where large-scale climatic changes in the Miocene were hypothesized to have induced its initial diversification between 7 and 11 Mya (Megens, van Moorsel *et al.* 2004) and the Indo-Australian Pierinae genus *Delias* Hübner (165+ spp.) where species diversification showed an increase during the Pliocene–Pleistocene starting around 7 Mya (Braby and Pierce 2007). However, the mean diversification rate in *Cymothoe* is much higher than that of *Delias* as the latter already comprised around 25 species before the inferred rate increase (Braby and Pierce 2007).

Figure 4.4: Diversification analyses (next page). Maximum clade credibility tree from the BEAST trait reconstruction analysis based on complete species sampling (i.e. including missing taxa). Branch colours indicate reconstructed ancestral states and squares indicate current host plant associations with Achariaceae (blue), with *Rinorea* (green) or unknown (grey). Dashed lines subtend missing species for which only the DNA barcode was available; dotted lines subtend missing species for which no sequence data were available (empty sequences). Branch labels indicate posterior clade probabilities higher than 0.50; black squares indicate clade monophyly constraints restricting phylogenetic placement of missing species. Black arrow indicates clade receiving increased posterior probability compared to analyses without host plant data (see text). Insets show posterior densities of diversification models applied to trees with complete species sampling. Top panel shows ΔAIC of rate-variable models over rate-constant models indicating a temporal shift in diversification rate. Grey density shows results based on a null distribution of trees simulated under a Yule model; red density shows results based on our phylogenetic trees. Middle and bottom panels show posterior densities from the trait-dependent diversification (BiSSE) models. Blue densities indicate diversification rates of Achariaceae-feeding clades (λA) and transition rates from *Rinorea* to Achariaceae feeding (qR→A); green densities indicate diversification rates of *Rinorea*-feeding clades (λR) and transition rates from Achariaceae to *Rinorea* feeding (qR→A). Note that placement of empty sequences is not based on data and that nodes within constrained clades consequently have low posterior probabilities.

Secondly, the increase in *Cymothoe* diversification rate is abrupt. Not only is there a 10-fold increase in mean rates of species diversification, it also appears to be instantaneous on an evolutionary timescale, suggesting a rapid radiation (Rokas, Krüger *et al.* 2005). Most other species-level butterfly clades show a more gradual shift in diversification rates (e.g. Elias, Joron *et al.* 2009, Leneveu, Chichvarkhin *et al.* 2009) and although the rate of species diversification in the genus *Arhopala*, with 120 extant species in 11 Myr, has been higher than in *Cymothoe*, it is unclear if it constitutes a significant shift when compared with diversification rates in sister clades within Theclinae (Megens, van Moorsel *et al.* 2004).

4.4.1 Diversification and host association

Is the shift in species diversification due to an adaptive trait? We tested for a correlation between diversification and host plant association and found only slightly elevated rates of diversification in clades feeding on *Rinorea* compared with those feeding on the ancestral Achariaceae hosts. This minor difference therefore cannot explain the large rate difference. Moreover, even though the backbone nodes within *Cymothoe* are unresolved, lineages appear to have colonized *Rinorea* well after the increase in diversification rate (Figure 4.4) , rejecting a diversification-after-host-expansion hypothesis in this clade. Together, these results indicate that *Rinorea*-feeding has not contributed significantly to the increase in diversification rate within *Cymothoe*. Nevertheless, the shift is branch-specific, and we cannot rule out that some other *Cymothoe*-specific trait may have promoted species diversification in this clade. For instance, a past genetic catastrophic event such as chromosomal rearrangements can contribute to reproductive isolation and, hence, elevated rates of net diversification (Kandul, Lukhtanov *et al.* 2007).



p=0.070
null
phylogeny
Density
ΔAIC
p=0.176
λA
λR
Posterior density
Diversification rate (λ)
p=0.196
qR→A
qA→R
Posterior density
Host plant transition rate
Host plant associations
Achariaceae
Rinorea
COC
ADE
FUM
LUR
EGE
SAN
AUR
HER
CAE
C. anitorgis
C. excelsa
C. radialis
C. crocea
C. arcuata
C. preussii
C. meridionalis
C. reginae-elisabethae
C. distincta
C. haimodia
C. coccinata
C. aramis
C. hartigi
C. mabillei
C. aubergeri
C. adela
C. collarti
C. fontainei
C. sassiana
C. cyclades
C. ochreata
C. lucasi
C. fumana
C. haynae
C. superba
C. colmanti
C. hypatha
C. butleri
C. okomu
C. hesiodina
C. lurida
C. hesiodotus
C. nigeriensis
C. egesta
C. confusa
C. orphnina
C. sangaris
C. hobarti
C. serpentina
C. magnus
C. euthalioides
C. ogova
C. harmilla
C. heliada
C. angulifascia
C. altisidora
C. hyarbita
C. beckeri
C. reinholdi
C. owassae
C. oemilius
C. jodutta
C. aurivillii
C. cottrelli
C. teita
C. melanjae
C. magambae
C. collinsi
C. alcimeda
C. vumbui
C. zombana
C. amaniensis
C. baylissi ined.
C. weymeri
C. herminia
C. coranus
C. caprina
C. isiro
C. bouyeri
C. caenis
C. amenides
C. howarthi
C. druryi
C. consanguis
C. althea
C. capella
C. eris
C. alticola
C. zenkeri
C. dujardini
C. lambertoni
C. indamora
Harma theobene

4.4.2 Diversification and climate change

Can we explain *Cymothoe* diversification by considering environmental factors? The shift in species diversification correlates with a period of global cooling and desiccation in the late Miocene (Zachos, Pagani *et al.* 2001), see Figure 4.3. In Africa, these climatological changes led in part to the expansion of savannah at the expense of forest (Jacobs 2004, Segalen, Lee-Thorp *et al.* 2007, Senut, Pickford *et al.* 2009). Indeed, pollen records from Nigeria suggest that late Miocene dessication may have been responsible for the extinction of much of the West African humid tropical flora (Morley and Richards 1993) and the forest habitat of *Harma* and *Cymothoe* that must have been largely continuous throughout most of the Miocene thus became fragmented (Jacobs 2004). Given the niche conservatism within the *Harma*-*Cymothoe* clade we may assume that this habitat fragmentation led to reproductive isolation of populations. Indeed, many species of *Cymothoe* are currently confined to particular geographic regions (Larsen 2005), suggesting a predominantly allopatric mode of speciation (McBride, van Velzen *et al.* 2009). For *Cymothoe*, important regions of endemism are found in West, Central and East Africa and on Madagascar, with closely related species often occuring in different regions (McBride, van Velzen *et al.* 2009, van Velzen, Larsen *et al.* 2009). We note that our results do not support a Pleistocene climatic oscillation-based explanation involving forest fragmentation, but rather that *Cymothoe* diversification has been an ongoing process since the late Miocene desiccation (see above). Similar patterns of major splits in lineages in this geological period have been found in plants (Plana, Gascoigne *et al.* 2004, Couvreur, Chatrou *et al.* 2008).

The AUR clade consists of species that are endemic to particular montane forests in the eastern Arc mountains and its estimated age of 4.19 Myr correlates with the maximum of the eastern Arc mountain uplift (Sepulchre, Ramstein *et al.* 2006) suggesting that geology may have driven diversification in this clade. A similar case was found in the Andean butterfly genus *Lymanopoda* (Nymphalidae, Satyrinae), where species diversification was estimated to coincide with Andean orogeny (Casner and Pyrcz 2010).

The estimated age of 7.13 Mya for the lineage leading to the Malagasy endemic LAM clade suggests that it dispersed to Madagascar much later than most mammals (Yoder and Yang 2004, Poux, Madsen *et al.* 2005) and insects (Wirta, Viljanen *et al.* 2010, Sole, Wirta *et al.* 2011), including butterflies (Condamine, Sperling *et al.* 2012). This would mean that the lineage had less time to diversify and possibly encountered less unoccupied niches. In turn, this possibly explains why, where most other clades diversified dramatically on Madagascar (Vences, Wollenberg *et al.* 2009), the LAM clade consists of only two species (*C. lambertoni* and *C. dujardini* Viette, 1971).

Thus, changing climate as well as geological events are likely to have promoted species diversification in *Cymothoe*. Why *Harma* has not responded similarly to

habitat fragmentation remains unclear but, as it currently is the geographically most widespread species within the entire clade, it may be a better disperser and hence less prone to reproductive isolation, possibly explaining its lack of diversification in the last 15 million years.

4.4.3 Speciation versus extinction

Can the inferred shift in species diversification be explained by a difference in speciation rate or, alternatively, in extinction pruning lineages from the stem? Our estimates based on birth-death models suggest that extinction rates approach zero in *Harma* as well as in *Cymothoe*, suggesting that differences in diversification are mainly due to speciation. There is a general notion, however, that phylogenetic methods are not adequate for estimating extinction rates (Rabosky 2010), and this is probably even more true for stem groups. Nevertheless, elevated rates of extinction before the late Miocene seem less likely because the Miocene experienced a relatively stable hot and wet climate (Zachos, Pagani *et al.* 2001) promoting large habitats suitable for tropical forest butterflies. We therefore hypothesize that the diversification shift is mainly due to an elevated rate of speciation rather than extinction of stem lineages.

4.4.4 Diversification rate versus phylogenetic resolution

The first branches within the *Cymothoe* clade have low phylogenetic resolution, which can be attributed to lack of data, or to near simultaneous divergence of multiple lineages (hard polytomies). Coalescent theory predicts that short internal branches are prone to incomplete lineage sorting effects (Degnan and Rosenberg 2009). Because short internal branches are inherent to a high diversification rate, a negative correlation between diversification rate and phylogenetic resolution can be expected (Rokas, Krüger *et al.* 2005, Whitfield and Kjer 2008, Kodandaramaiah, Lees *et al.* 2010), which is confirmed by our data. We therefore hypothesize that lack of resolution within *Cymothoe* is inherent to its phylogenetic structure and that adding more sequence data will not solve the gene incongruencies because they might well be caused by incomplete lineage sorting. Methods have been developed that accomodate for incomplete lineage sorting by consolidating gene trees with a species tree (e.g. Liu and Pearl 2007, Heled and Drummond 2010). However, effective population size is an essential parameter for these methods and we expect that estimating ancestral population size for deeper internal branches is prohibitive. Therefore, we do not expect these methods to provide a solution to this problem. Resolving branches within a rapidly diversifying clade therefore remains a methodological challenge, even at recent timescales.

4.5 Conclusions

Whether the rapid recent diversification of *Cymothoe* is truly a unique phenomenon or is more general in butterflies than the current literature suggests, remains to be seen after more species-level phylogenetic trees become available. Even though shifts from Achariaceae to *Rinorea* host plants cannot explain the disparity in species diversification between *Harma* and *Cymothoe,* host plant transition seems to be directional from Achariaceae to *Rinorea*, possibly indicating an ultimate advantage of the *Rinorea*-association to *Cymothoe* butterflies. Indeed, *Rinorea* species are common understorey shrubs and trees and often occur gregariously in tropical African forests (Achoundong 1996), constituting a reliable and abundant food source. Possibly, shifts between different host plant species played a role in *Cymothoe* speciation within the highly specialized *Rinorea*-feeding clades. Reconstructing ancestral host associations at the species level requires a resolved (African) *Rinorea* phylogenetic tree and is the subject of Chapter 6.

Acknowledgements

We thank Pavel Matos for help in the lab. Specimens or samples used for this study were kindly donated by Julian Bayliss, Dries Bonte, Thierry Bouyer, Oskar Brattström, Steve Collins, Torben Larsen, Freerk Molleman, Renske Onstein, Sáfián Szabolcs, Robert Warren, Gael vandeWeghe, Steve Woodhall, and Jan Wieringa. Marc Suchard helped edit the BEAST ancestral host plant reconstruction xml files. Marleen Botermans gave valuable comments on early drafts of the manuscript. RVV acknowledges the Systematics Research Fund, Alberta Mennega Foundation, Hugo de Vries Foundation and the European Commission's Research Infrastructure Action (SYNTHESYS grant numbers BE-TAF-3810 and GB-TAF-4003) for financial support. NW acknowledges funding from the Academy of Finland (grant number 129 811).

Chapter 5

Phylogenetics and historical biogeography of *Rinorea* (Violaceae)

with emphasis on the African species

Robin van Velzen[1,2], Gregory Wahlert[3], Marc SM Sosef[1,2], Renske E. Onstein[4] & Freek T Bakker[1]

[1] Biosystematics Group, Department of Plant Sciences, Wageningen University and Research Centre, Wageningen, The Netherlands
[2] Naturalis Biodiversity Centre (Section NHN), Wageningen University, Generaal Foulkesweg 37, 6703 BL Wageningen, The Netherlands
[3] Department of Biology, University of Utah, Salt Lake City, Utah 84112, U.S.A.
[4] Institute of Systematic Botany, University of Zurich, Zollikerstrasse 107, Zurich 8008, Switzerland

Abstract

Rinorea Aubl. is a pantropical genus of shrubs and small trees. The genus is particularly diverse in Africa where species are ecologically important as they are often abundant or even dominant in particular forest types and act as larval host plants for highly specialized *Cymothoe* butterflies. Despite their importance, species identification of African *Rinorea* is difficult and a taxonomic revision of the African species is needed. Phylogenetic studies suggest that Neotropical taxa are sister to a Palaeotropical clade, with multiple independent dispersals to Madagascar, but these were based on plastid data only. Moreover, they lack an absolute timeframe needed to properly test biogeographic scenarios. We therefore present an updated phylogeny of *Rinorea* with increased taxonomic sampling, using plastid as well as nuclear DNA sequences, and estimate lineage divergence within an absolute timeframe.

Our results indicate that *Rinorea* originated in the Neotropics and reached Africa in the Eocene through trans-Atlantic dispersal. In Africa, the genus proliferated since the Oligocene into the high phylogenetic and morphological diversity that is seen today. From there, *Rinorea* dispersed to Asia in the Oligocene or early Miocene, probably after closing of the Tethys Sea. Madagascar has been colonized multiple times independently within a relatively recent time scale (Pliocene), suggesting that factors governing the independent colonizations of *Rinorea* to Madagascar may have been similar.

Phylogenetic relationships inferred from nuclear DNA data were generally congruent with those based on evidence from plastid haplotypes from earlier studies of *Rinorea* and helped resolve additional clades, some of which warrant further taxonomic study. It is clear that African *Rinorea* require comprehensive taxonomic revision; our contribution to understanding *Rinorea* infrageneric relationships will facilitate this task.

5.1 Background

Rinorea Aubl. is a pantropical genus of shrubs and small trees and with an estimated 225–275 species the second most species-rich genus in the family Violaceae (Ballard & Wahlert, unpubl. data). The diversity of species is well known in the Neotropics (49 spp.; Hekking 1988) and Madagascar (41 spp.; Wahlert 2010), but the numbers of species in Asia (ca. 30 spp.; Ballard and DeMuria, unpubl. data) and Africa are uncertain due to an outdated taxonomic framework. Estimates of the total number of African species range from 110 to 150 making Africa by far the richest continent in terms of species diversity (Achoundong 2000). Cameroon and Gabon are particularly species-rich (55 and 49 species, respectively; Bos 1989, Achoundong 1996, Bakker, van Gemerden *et al.* 2006, Sosef, Wieringa *et al.* 2006) and appear to be regions of endemism for the genus (Achoundong 1996, Achoundong 2000, Tchouto, de Wilde *et al.* 2009) coinciding well with other recognized hotspots of botanical diversity (Küper, Sommer *et al.* 2004).

African *Rinorea* mainly occur in the understory of humid or semi-deciduous forests (Achoundong 1996) where they can be abundant or even dominant in the lower forest strata (Guillet, Achoundong *et al.* 2001, Wagner 2001, Kenfack, Thomas *et al.* 2007, Githae, Chuah-Petiot *et al.* 2008, Djuikouo, Doucet *et al.* 2010, Chuyong, Kenfack *et al.* 2011). As they are usually restricted to specific environments (Achoundong 1996, Achoundong 2000, Adomou, Sinsin *et al.* 2006, Mwavu and Witkowski 2009, Tchouto, de Wilde *et al.* 2009, Djuikouo, Doucet *et al.* 2010), African *Rinorea* are considered useful bioindicators for forest typification (Achoundong 1996, Achoundong 2000, Tchouto, de Wilde *et al.* 2009). Their habitat specificity and common local endemism has been explained by low seed dispersal abilities because *Rinorea* fruits are dry dehiscent capsules that contain few and relatively heavy seeds that do not appear to be zoochorous (Hekking 1988, Achoundong 1996, Achoundong 2000, Bakker, van Gemerden *et al.* 2006, Tchouto, de Wilde *et al.* 2009). Local zones of high *Rinorea* diversity have been hypothesized to represent refuge areas during glacial maxima, possibly reflecting their low dispersal rate since then (Achoundong 2000, Tchouto, de Wilde *et al.* 2009). As various *Rinorea* species usually co-occur locally, they provide a case for future niche-partitioning studies (as in Wiens, Ackerly *et al.* 2010).

In addition, *Rinorea* species are ecologically significant as larval host plants for *Cymothoe* (Nymphalidae), a butterfly genus distributed throughout tropical Africa and Madagascar. At least 32 *Cymothoe* are highly specialized on particular species of *Rinorea* and most are even monophagous (Fontaine 1982, Amiet and Achoundong 1996, McBride, van Velzen *et al.* 2009). Several closely related (or even cryptic) *Cymothoe* species have been found to colonize different (usually related) species of *Rinorea* (Amiet 1997, Amiet 2000, McBride, van Velzen *et al.* 2009, van Velzen, Larsen *et al.* 2009) suggesting that *Rinorea* associations may have played a role in their diversification (McBride, van Velzen *et al.* 2009). Finally, species of *Rinorea*

are used as traditional medicinal plants in various parts of Africa (Bouquet 1969, Bouquet 1974, Iwu 1993, Neuwinger 2000, N'guessan, Soro *et al.* 2011), and have been found to contain weakly antioxidant essential oils (Agnaniet, Mounzeo *et al.* 2003) as well as an isoflavone that is considered to be an agent against schistosomiasis (Stewart, Bartholomew *et al.* 2000).

Despite their importance, species identification of African *Rinorea* is difficult and specimens are regularly unidentified in ecological (Kenfack, Thomas *et al.* 2007, e.g. Tchouto, de Wilde *et al.* 2009) and taxonomic studies (e.g. Robson 1960, Hawthorne and Jongkind 2006, Sosef, Wieringa *et al.* 2006). The most recent overview of all African *Rinorea* species was published by Brandt in 1914. He produced an infrageneric classification primarily based on characters of the androecium, which was adopted by De Wildeman (1920) and Melchior (1925). However, this classification includes some problematic groups that are 'transitional' (*R.* subsect. *Lobiferae*) or based on subtle differences such as the general shape of the anthers (*R.* subsects. *Kamerunenses* and *Brachypetalae*) or 'fleshyness' of the flowers (*R.* subsect. *Crassiflorae*) and it is therefore difficult to confidently assign species to their recognized infrageneric groups (Bos 1989, Wahlert 2010). Furthermore, since the last classification numerous new species have been published (Taton 1968, Achoundong and Cheek 2003, e.g. Achoundong and Bakker 2006) while others await their formal description (Achoundong 1996) further affirming the need for a revision of African *Rinorea.* Thus, although Dowsett-Lemaire & White (1990) stated that "*Rinorea* badly needs a critical pan-African revision" this has not yet been achieved today, rendering *Rinorea* an urgent case taxonomically. While a comprehensive revision is outside the scope of the present study, we contribute to a better understanding of the phylogenetic history of the genus, and therefore facilitate an improved infrageneric classification.

Recent phylogenetic studies within the Violaceae clade inferred the genus *Fusispermum* Cuatrec. as the first diverging extant lineage (Tokuoka 2008, Wurdack and Davis 2009). *Rinorea* diverges next, with the 'Apiculata' group (Hekking 1988; represented by R. apiculata) as sister to all other *Rinorea* (Tokuoka 2008). As the latter relationship received only low support we refer instead to the well-supported clade in that study comprising all *Rinorea* but excluding the morphologically divergent Apiculata group (i.e. *R. apiculata*, *R. crenata*, and *R. oraria*; Hekking 1988) as *Rinorea sensu stricto.*

After an initial phylogenetic reconstruction of African *Rinorea* indicated the presence of two main African haplotypes and two African/Malagasy sister group relationships (Bakker, van Gemerden *et al.* 2006), Wahlert & Ballard (Wahlert and Ballard 2012) produced a more inclusive phylogenetic hypothesis for the genus, based on plastid DNA only (99 accessions/68 taxa). They found that all Neotropical *Rinorea* s.s. are sister to a Palaeotropical clade which is subdivided into two clades: species from the first clade have an ovary with 1 ovule per locule, those from the second clade have 2 ovules per locule. Within the 1-ovule clade, Malagasy taxa appear sister to those from mainland Africa. Within the 2-ovule clade, again,

Malagasy taxa are sister to those from Africa, suggesting a total of 4 independent dispersals from mainland Africa to Madagascar (Bakker, van Gemerden *et al.* 2006, Wahlert and Ballard 2012), although this has never been tested within an absolute time-frame. Within the Palaeotropical *Rinorea* clade, Wahlert & Ballard (2012) found eleven African and Malagasy plastid haplotypes that appeared to be congruent with infrageneric groupings delimited in morphometric studies by Wahlert (2010), although not all haplotypes received high support.

In spite of the fact that these studies have collectively progressed our understanding of African *Rinorea* systematics greatly, two issues remain: (i) the lack of a nuclear DNA based phylogenetic perspective allowing reconstruction of actual clades instead of plastid haplotypes; (ii) some idea of the absolute time frame in which diversification in *Rinorea* has taken place. The latter would be important for testing historical biogeographic hypotheses. For instance, the disjunction between Neo- and Palaeotropical *Rinorea* may either be explained by a Gondwanan vicariance or by a more recent long-distance dispersal event (Queiroz 2005, Gheerbrant and Rage 2006). Likewise, it remains unknown if the independent dispersals from Africa to Madagascar were synchronous. Therefore, in order to address these questions, we present an updated phylogeny of *Rinorea* with increased taxonomic sampling, using plastid as well as nuclear DNA sequences, and estimate lineage divergences within an absolute time-frame.

5

5.2 Materials & methods

We selected *Fusispermum laxiflorum* and *Rinorea crenata* (Apiculata group) as outgroup taxa, based on Tokuoka (2008), Wahlert (2010), and Wahlert & Ballard (2012). The ingroup comprised 146 accessions covering 75 taxa, which include 4 Asian species (i.e.∼13% of Asian taxa), 5 Neotropical species (i.e.∼10% of Neotropical taxa; 2 from the informal group '*Rinorea*' and 3 from the group 'Pubiflora'; Hekking 1988). The majority though (128 accessions) comprised 60 African and/or Malagasy species (roughly 50% of known African taxa). Taxon sampling of African and Malagasy species represented each of the 11 infrageneric groups identified in earlier qualitative and morphometric studies of floral morphology by Wahlert (2010). Voucher information and GenBank accession numbers are given in Table 5.1.

DNA was extracted using the hot CTAB method (Doyle and Doyle 1987). We amplified the following three marker regions: 74 *trnL–F* sequences were Sanger-generated following standard protocols using primers C and F from Taberlet (1991); likewise, 116 nrDNA ITS sequences were generated using primers ITS1 + ITS4 (White, Bruns *et al.* 1990); 79 sequences of exon 12 of the low-copy nuclear-encoded gene *EMB2765* (Wurdack and Davis 2009) were generated using primers EMB2765ex12F and EMB2765ex12R and protocols by Horn *et al.* (2012). Another 59 trnL–F sequences were downloaded from GenBank (see Table 5.3).

Table 5.1 Voucher information and GenBank accession numbers.

Species	Collection	Herbarium	Country
R. albidiflora	Achoundong 2121	WAG	Cameroon
R. albidiflora	van Velzen 80	WAG	Cameroon
R. amietii ined.	van Velzen 87	WAG	Cameroon
R. amietii ined.	van Velzen 50	WAG	Cameroon
R. anguifera	Ballard 3192	WAG	Malaysia
R. angustifolia subsp. *ardisiiflora*	Luke 12745	WAG	Tanzania
R. angustifolia subsp. *engleriana*	Achoundong 2117	WAG	Cameroon
R. angustifolia subsp. *engleriana*	Onstein 41	WAG	Ghana
R. angustifolia subsp. *engleriana*	van Velzen 95	WAG	Cameroon
R. arborea	Burrows & Wahlert 10171	MO	Mozambique
R. arborea	Labat 3197	MO	Madagascar
R. bahiensis	Paula-Souza et al. 5679	SPF	Brazil
R. ballardii ined.	Wahlert 113	MO	Madagascar
R. batesii	van Velzen 54	WAG	Cameroon
R. bengalensis	Chase 2148	K	India
R. beniensis	ATBP 610	MO	Uganda
R. brachypetala	Jongkind 6729	WAG	Liberia
R. brachypetala	Koenen 165	WAG	Uganda
R. brachypetala	Sosef 2592	WAG	Gabon
R. brachypetala	van Velzen 100	WAG	Kenya
R. breviracemosa	Achoundong 2208	YA	Cameroon
R. breviracemosa	Jongkind 7011	WAG	Liberia
R. breviracemosa	Jongkind 6788	WAG	Liberia
R. breviracemosa	van Valkenburg 2549	WAG	Gabon
R. bullata	Skema 217	MO	Madagascar
R. callmanderi ined.	Callmander 582	MO	Madagascar
R. campoensis	Achoundong 2113	YA	Cameroon
R. campoensis	Bakker 12	WAG	Cameroon
R. campoensis	Shu 8294	YA	Cameroon
R. campoensis	Wieringa 4640	WAG	Gabon
R. campoensis	Shu 327	YA	Cameroon
R. caudata	Achoundong 2104	YA	Cameroon
R. caudata	van Velzen 49	WAG	Cameroon
R. chevalieri	Stévart 2859	MO	San Tome
R. claessensii	Onstein 31	WAG	Ghana
R. convallarioides subsp. *convallarioides*	Burrows 9908	BNRH	Mozambique
R. convallarioides subsp. *occidentalis*	Onstein 42	WAG	Ghana
R. crenata	Ballard 94-006	WIS	Costa Rica

Code	*trnL–F*	ITS	*EMB2765*
GA2121	AY739723		
RV80	t.b.s.	t.b.s.	t.b.s.
RV87	t.b.s.	t.b.s.	
RV50	t.b.s.	t.b.s.	t.b.s.
HB3192	AY739755		
QL12745	t.b.s.	t.b.s.	t.b.s.
GA2117	AY739724	t.b.s.	t.b.s.
RO41	t.b.s.	t.b.s.	t.b.s.
RV95	t.b.s.		
JB10171	JN714095	t.b.s.	t.b.s.
JL3197	JN714094		
JPS5679	JN714117	t.b.s.	t.b.s.
GW113	JN714093	t.b.s.	t.b.s.
RV54	t.b.s.	t.b.s.	t.b.s.
MC2148	AY739756		
ATBP610	JN714087	t.b.s.	t.b.s.
CJ6729	t.b.s.	t.b.s.	t.b.s.
EK165	t.b.s.	t.b.s.	t.b.s.
MS2592	t.b.s.	t.b.s.	t.b.s.
RV100	t.b.s.	t.b.s.	t.b.s.
GA2208	t.b.s.	t.b.s.	t.b.s.
CJ7011	t.b.s.	t.b.s.	
CJ6788		t.b.s.	
JV2549	t.b.s.	t.b.s.	
CS217	JN714069	t.b.s.	t.b.s.
MC582	JN714071		
GA2113	AY739725		
FB12	t.b.s.	t.b.s.	
GS8294	AY739742	t.b.s.	
JW4640	t.b.s.	t.b.s.	t.b.s.
GS327	AY739745	t.b.s.	
GA2104	AY739726		
RV49	t.b.s.	t.b.s.	
TS2859	JN714099	t.b.s.	t.b.s.
RO31	t.b.s.	t.b.s.	t.b.s.
JB9908	JN714088	t.b.s.	t.b.s.
RO42	t.b.s.	t.b.s.	t.b.s.
HB94-006	JN714119	t.b.s.	t.b.s.

Table 5.1 Voucher information and GenBank accession numbers, continued.

Species	Collection	Herbarium	Country
R. dentata	Bakker 6	WAG	Cameroon
R. dentata	Chatrou 578	WAG	Cameroon
R. dentata	Gereau 5541	MO?	Cameroon
R. dentata	van Velzen 119	WAG	Nigeria
R. dentata	van Velzen 75	WAG	Cameroon
R. dewildei ined.	van Velzen 43	WAG	Cameroon
R. dewitii	Achoundong 2123	YA	Cameroon
R. dewitii	van Velzen 42	WAG	Cameroon
R. dimakoensis ined.	Achoundong 2167	YA	Cameroon
R. dimakoensis ined.	Sonke 2591	YA	Cameroon
R. domatiosa	Crouch s.n.	n.a.	South Africa
R. domatiosa	Nowell 159	BHO	South Africa
R. cf. *ebolowensis*	van Andel 3620	WAG	Cameroon
R. elliptica	Burrows & Wahlert 10128	MO	Mozambique
R. cf.*exappendiculata*	Wieringa 4382	WAG	Gabon
R. gabunensis	Achoundong 2118	YA	Cameroon
R. gabunensis	Bakker 11	WAG	Cameroon
R. gabunensis	Bakker 9	WAG	Cameroon
R. gabunensis	van Velzen 32	WAG	Cameroon
R. gabunensis	Wieringa 4419	WAG	Gabon
R. gabunensis	Wieringa 4451	WAG	Gabon
R. gabunensis	Wieringa 4566	WAG	Gabon
R. greveana	Bolin & Razafindraibe M07-08	MO	Madagascar
R. guianensis	Paula-Souza et al. 9541	SPF	Brazil
R. horneri	Gentry & Lafrankie 66982	MO	Malaysia
R. hummelii	Maas 9505	WAG	Costa Rica
R. ilicifolia	Adam 30109	MO	Liberia
R. ilicifolia	Crouch s.n.	n.a.	South Africa
R. ilicifolia	Jongkind 6241	WAG	Liberia
R. ilicifolia	Onstein 29	WAG	Ghana
R. ilicifolia	Phillipson & Sitoni 4947	MO	Tanzania
R. ilicifolia	van Velzen 60	WAG	Cameroon
R. ilicifolia	Wieringa 5388	WAG	Ivory Coast
R. ilicifolia vaR. *amplexicaulis*	Abeid 1274	MO	Tanzania
R. ilicifolia vaR. *amplexicaulis*	Abeid 2282	MO	Tanzania
R. kamerunensis	Achoundong 2111	YA	Cameroon
R. kamerunensis	Chatrou 567	WAG	Cameroon
R. kamerunensis	van Velzen 72	WAG	Cameroon

Code	*trnL–F*	ITS	*EMB2765*
FB6	t.b.s.	t.b.s.	
LC578	t.b.s.	t.b.s.	t.b.s.
RG5541	JN714062		
RV119	t.b.s.		t.b.s.
RV75	t.b.s.		
RV43	t.b.s.	t.b.s.	t.b.s.
GA2123	AY739744		
RV42	t.b.s.	t.b.s.	
GA2167	t.b.s.	t.b.s.	t.b.s.
BS2591	AY739728	t.b.s.	t.b.s.
NCsn1	JN714098		t.b.s.
TN159	t.b.s.		
TA3620	t.b.s.	t.b.s.	t.b.s.
JB10128	JN714084	t.b.s.	t.b.s.
JW4382	AY739752	t.b.s.	t.b.s.
GA2118	AY739741	t.b.s.	t.b.s.
FB11	t.b.s.	t.b.s.	
FB9	t.b.s.	t.b.s.	
RV32	t.b.s.	t.b.s.	
JW4419	AY739739	t.b.s.	
JW4451	AY739740	t.b.s.	t.b.s.
JW4566		t.b.s.	
JB07-08	JN714083	t.b.s.	t.b.s.
JPS9541	JN714118	t.b.s.	t.b.s.
AG66982	JN714081	t.b.s.	
PM9505	t.b.s.	t.b.s.	
JA30109	JN714034	t.b.s.	
NCsn2	JN714032	t.b.s.	
CJ6241	t.b.s.	t.b.s.	
RO29	t.b.s.	t.b.s.	t.b.s.
PP4947	JN714033	t.b.s.	
RV60	t.b.s.	t.b.s.	
JW5388	t.b.s.	t.b.s.	t.b.s.
YA1274	JN714031	t.b.s.	
YA2282	JN714030	t.b.s.	
GA2111	AY739729	t.b.s.	t.b.s.
LC567	t.b.s.	t.b.s.	t.b.s.
RV72	t.b.s.		

Table 5.1 Voucher information and GenBank accession numbers, continued.

Species	Collection	Herbarium	Country
R. kemoensis	Sonke 2664	YA	Cameroon
R. lanceolata	Chase 2149	K	Malaysia
R. ledermannii	Achoundong 2116	YA	Cameroon
R. ledermannii	van Andel 3364	WAG	Cameroon
R. leiophylla	Achoundong 2115	YA	Cameroon
R. leiophylla	van Velzen 48	WAG	Cameroon
R. lepidobotrys	van Velzen 23	WAG	Cameroon
R. letouzeyi	Achoundong 2180	YA	Cameroon
R. liberica	Jongkind 6148	WAG	Liberia
R. longicuspis	Achoundong 2112	YA	Cameroon
R. longicuspis	Munzinger & Karamoko 38	WAG	Ivory Coast
R. longicuspis	Onstein 13	WAG	Ghana
R. longicuspis	van Velzen 21	WAG	Cameroon
R. longicuspis	van Velzen 62	WAG	Cameroon
R. longisepala	Achoundong 2109	YA	Cameroon
R. longisepala	van Velzen 39	WAG	Cameroon
R. macrocarpa	Paula-Souza et al. 9529	SPF	Brazil
R. mezilii	Achoundong 2119	YA	Cameroon
R. mezilii	Achoundong s.n.	n.a.	Cameroon
R. microdon	Jongkind 10470	WAG	Guinea
R. mutica	Wahlert & Rakotonasolo 12	MO	Madagascar
R. oblongifolia	Achoundong 2107	YA	Cameroon
R. oblongifolia	Bakker 13	WAG	Cameroon
R. oblongifolia	Jongkind 6717	WAG	Liberia
R. oblongifolia	Onstein 12	WAG	Ghana
R. oblongifolia	van Velzen 40	WAG	Cameroon
R. ovata	Achoundong 2122	YA	Cameroon
R. ovata	Shu 7751	WAG	Cameroon
R. prasina	Jongkind 2157	WAG	Ghana
R. preussii	van Velzen 15	WAG	Cameroon
R. pugionifera	Wahlert & Rakotonasolo 21	MO	Madagascar
R. racemosa	Paula-Souza et al. 9543	SPF	Brazil
R. rubrotincta	Achoundong 2124	YA	Cameroon
R. rubrotincta	Onstein 32	WAG	Ghana
R. rubrotincta	van Velzen 121	WAG	Nigeria
R. rubrotincta	van Velzen 126	WAG	Nigeria
R. rubrotincta	van Velzen 58	WAG	Cameroon
R. rubrotincta	Onstein 18	WAG	Ghana

Code	*trnL–F*	ITS	*EMB2765*
BS2664	AY739727	t.b.s.	t.b.s.
MC2149	JN714076		
GA2116	AY739730		
TA3364	t.b.s.	t.b.s.	t.b.s.
GA2115	AY739736	t.b.s.	t.b.s.
RV48	t.b.s.	t.b.s.	t.b.s.
RV23	t.b.s.	t.b.s.	t.b.s.
GA2180	AY739734	t.b.s.	
CJ6148		t.b.s.	
GA2112	AY739731	t.b.s.	t.b.s.
JM38	AY739757		
RO13	t.b.s.	t.b.s.	
RV21	t.b.s.	t.b.s.	t.b.s.
RV62	t.b.s.		
GA2109	AY739732		
RV39	t.b.s.	t.b.s.	t.b.s.
JPS9529	JN714108		
GA2119	AY739733	t.b.s.	t.b.s.
GAsn1	AY739752	t.b.s.	
CJ10470	t.b.s.		
GW12	JN714067		t.b.s.
GA2107	AY739735		
FB13	t.b.s.	t.b.s.	t.b.s.
CJ6717	t.b.s.		
RO12	t.b.s.	t.b.s.	t.b.s.
RV40	t.b.s.		t.b.s.
GA2122	AY739738	t.b.s.	t.b.s.
GS7751	AY739737	t.b.s.	
CJ2157		t.b.s.	
RV15	t.b.s.	t.b.s.	t.b.s.
GW21	JN713987	t.b.s.	t.b.s.
JPS9543	JN714110		
GA2124		t.b.s.	
RO32	t.b.s.		t.b.s.
RV121	t.b.s.	t.b.s.	t.b.s.
RV126		t.b.s.	t.b.s.
RV58	t.b.s.	t.b.s.	t.b.s.
RO18	t.b.s.		t.b.s.

5

Table 5.1 Voucher information and GenBank accession numbers, continued.

Species	Collection	Herbarium	Country
R. simoneae	Achoundong 2110	YA	Cameroon
R. sinuata	Achoundong 2178	YA	Cameroon
R. spinosa	Jongkind 3314	WAG	Madagascar
R. spinosa	Rouhan & Bernier 38	MO	Comoros Islands
R. spinosa	Wahlert & Rakotonasolo 4	MO	Madagascar
R. squamosa	Wahlert 95	MO	Madagascar
R. subauriculata	Wieringa 5155	WAG	Gabon
R. subintegrifolia	Achoundong 2108	YA	Cameroon
R. subintegrifolia	Jongkind 6054	WAG	Liberia
R. subintegrifolia	van Velzen 134	WAG	Nigeria
R. subintegrifolia	van Velzen 30	WAG	Cameroon
R. umbricola	Bakker 3	WAG	Cameroon
R. umbricola	van Velzen 83	WAG	Cameroon
R. verrucosa	Bakker 15	WAG	Cameroon
R. verrucosa	van Velzen 63	WAG	Cameroon
R. verrucosa	Wieringa 4474	WAG	Gabon
R. villiersii ined.	Achoundong 2174	YA	Cameroon
R. welwitschii	Jongkind 10522	WAG	Guinea
R. welwitschii	Luke 12962	EA	Tanzania
R. welwitschii	van Velzen 115	WAG	Nigeria
R. welwitschii	van Velzen 118	WAG	Nigeria
R. welwitschii	van Velzen 122	WAG	Nigeria
R. welwitschii	van Velzen 94	WAG	Cameroon
R. woermanniana	Wieringa 5172	WAG	Gabon
R. yaundensis	Harder 2964	MO	Ghana
R. yaundensis	Jongkind s.n.	n.a.	Ghana
R. yaundensis	van Velzen 55	WAG	Cameroon
R. yaundensis	van Velzen 56	WAG	Cameroon
R. zenkeri	van Velzen 59	WAG	Cameroon
R. sp.	Achoundong 2120	YA	Cameroon
R. sp.	Breteler 16000	WAG	Gabon
R. sp. nov near *ilicifolia*	Achoundong 2337	YA	Cameroon
R. sp. nov near *angustifolia engleriana*	Luke & Luke 13003	EA	Kenya
R. sp. nov near *keayii*	Luke 10365	EA	Tanzania

Code	*trnL–F*	ITS	*EMB2765*
GA2110	AY739743	t.b.s.	t.b.s.
GA2178	AY739750	t.b.s.	t.b.s.
CJ3314	AY739754	t.b.s.	
GR38	JN714027	t.b.s.	
GW4	JN714029	t.b.s.	
GW95	JN714046	t.b.s.	t.b.s.
JW5155	t.b.s.	t.b.s.	
GA2108	AY739746	t.b.s.	
CJ6054		t.b.s.	
RV134	t.b.s.	t.b.s.	t.b.s.
RV30	t.b.s.	t.b.s.	t.b.s.
FB3	t.b.s.	t.b.s.	
RV83	t.b.s.	t.b.s.	t.b.s.
FB15	t.b.s.	t.b.s.	
RV63	t.b.s.		
JW4474	AY739749	t.b.s.	t.b.s.
GA2174	t.b.s.	t.b.s.	t.b.s.
CJ10522	t.b.s.	t.b.s.	t.b.s.
QL12962	t.b.s.	t.b.s.	t.b.s.
RV115		t.b.s.	t.b.s.
RV118		t.b.s.	t.b.s.
RV122	t.b.s.	t.b.s.	t.b.s.
RV94	t.b.s.	t.b.s.	t.b.s.
JW5172	t.b.s.	t.b.s.	t.b.s.
DH2964	JN714063	t.b.s.	t.b.s.
CJsn1	t.b.s.		
RV55	t.b.s.	t.b.s.	t.b.s.
RV56		t.b.s.	
RV59	t.b.s.	t.b.s.	t.b.s.
GA2120	AY739748		
FJB16000	t.b.s.		
GA2337	t.b.s.	t.b.s.	
QL13003	t.b.s.	t.b.s.	t.b.s.
QL10365	JN714035	t.b.s.	

Herbarium acronyms follow Index Herbariorum (http://sweetgum.nybg.org/ih/); n.a. = voucher not available.

Table 5.2: Marker region alignments.

Marker region	#taxa	#characters	Version	Length	#informative
trnL-F	133	864	MAFFT	1120	133
			PRANK	1356	117
ITS	116	707	MAFFT	825	297
			PRANK	1022	286
EMB2765	79	942		942	142

#taxa = number of taxa; #characters = number of unaligned characters; #informative = number of informative sites within the alignment.

5.2.1 Phylogenetic analysis

Given the extensive length variation present in *trnL–F* and ITS, we adopted both MAFFT (Katoh and Toh 2008) and the 'phylogeny aware' progressive alignment approach implemented in PRANK (Loytynoja and Goldman 2009), as these approaches avoid subjective (i.e. manual) optimization of ensuing indels. Basically, we used both algorithms, conducted subsequent phylogenetic analyses on each alignment variant separately, and evaluated resulting clades accordingly.

MAFFT settings included the E-INS-i strategy and a gap opening cost of 1.0 ("mafft.bat --op 1 --ep 0 --maxiterate 1000 --retree 1 --genafpair --reorder input > output"). PRANK alignments were generated using webPRANK (Loytynoja and Goldman 2010) with default settings but without the use of clustalW for guide tree generation nor CHAOS alignment anchoring to increase alignment accuracy. The *EMB2765* marker contained no length variation and therefore alignment was not an issue.

We analysed each marker region separately, and *trnL–F* and ITS in both alignment versions (MAFFT and PRANK), amounting to a total of 5 unpartitioned matrices. In addition, in order to assess possible effects of missing data, we compiled four concatenated matrices (i.e. combining the marker regions; matrices 'A', 'B', 'C', 'D'), varying from 149 accessions (27% missing data; 'A') to 71 accessions (squared matrix, no missing data; 'D'), see Table 5.3. Each of A, B, C and D was concatenated twice, based on either MAFFT or PRANK alignment versions, amounting to 8 concatenated, and a grand total of 13 matrices to be analysed.

Maximum likelihood analyses were conducted using RAxML as implemented on the RAxML-HPC-Blackbox webserver (Stamatakis, Hoover *et al.* 2008), setting GTR as the model of sequence evolution and a gamma distribution (with 4 classes) to model rate variation over sites. Clade support was assessed by generating 100 replicates of rapid bootstrapping.

Bayesian inference was conducted using MrBayes 3.2.1, setting mixed models,

Table 5.3: Matrices and data coverage.

	Number of accessions			
Matrix	*trnL–F*	ITS	*EMB2765*	Total
Un-partitioned				
trnL–F	133			
ITS		116		
EMB2765			79	
Concatenated				
A	133	116	79	149
B	99	99	70	99
C	77	74	79	80
D	71	71	71	71

Concatenated matrices A,B,C and D have different levels of missing data, see text. Note that all matrices except *EMB2765* come in two alignment versions based on PRANK and MAFFT respectively, resulting in a total of 13 matrices (see Table 5.2).

allowing MrBayes to integrate over different GTR submodels using model jumps (nst = mixed), and gamma-distributed variation of rates (with 4 classes) to model rate variation over sites. Analyses consisted of two independent runs of 50 million generations, each with four metropolis coupled incrementally heated chains (temp = 0.05) and sampling every 10,000th generation. Convergence of the two independent MCMC runs was assessed topologically (i.e. based on clade frequencies) using the online service AWTY (Nylander, Wilgenbusch *et al.* 2008) and based on model parameters using Tracer 1.5 (Rambaut and Drummond 2009). The first 5 million generations (10%) were discarded as burn-in before calculating a 50-percent majority-rule consensus based on the posterior set of trees. We considered nodes with a posterior probability of > 0.90 as clades.

5.2.2 Divergence time estimation

Rinorea fossils are unknown and hence unavailable for node calibration. We used information from four fossils (see Table 5.4) from the so-called 'Parietal' clade (which includes. Achariaceae, Goupiaceae, Violaceae, Passifloraceae s.l., Lacistemataceae and Salicaceae s.l.; Wurdack and Davis 2009), that have been used in other studies (Davis, Webb *et al.* 2005, Hearn 2006, Wang, Moore *et al.* 2009, Bell, Soltis *et al.* 2010, Marcussen, Jakobsen *et al.* 2012). Because dating analyses are dependent on balanced taxonomic sampling to give credible age estimates (Linder, Hardy *et al.* 2005, Milne 2009, Pirie and Doyle 2012), and taxonomic sampling at the same level as for *Rinorea* was not possible within Malpighiales, we performed two successive analyses:

Analysis 1 aimed for balanced taxonomic sampling of Malpighiales and was calibrated using fossil information in combination with a maximum age estimate for Malpighiales. As the *trnL–F* and ITS regions were too variable for reliable alignment outside of *Rinorea*, character sampling comprised a 4-gene data set, *atpB*, *matK*, *rbcL*, and 18S (Tokuoka and Tobe 2006, Tokuoka 2008) for 159 taxa from Malpighiales, including 5 taxa representing major clades within *Rinorea*, compiled from GenBank. We chose the lognormal distribution for the fossil calibrations because it can assign the highest point probability for the nodal age to be somewhat older than the fossil (Ho and Phillips 2009), with an arbitrary value of 1 for both mean and scale and offset as indicated by the fossil age (applying the fossil calibrations as hard minimum bounds for nodes gave essentially the same results). In addition we imported a secondary calibration for the Malpighiales crown age of 102 (100–104) Myr from Wang *et al.* (2009) which is in agreement with older as well as newer estimates (Davis, Webb *et al.* 2005, Bell, Soltis *et al.* 2010). This calibration was set as a normal distribution with mean 102 and standard deviation of 2. Marginal prior distributions for calibrated nodes are not necessarily the same as the calibration distributions because they are also dependent on the prior distributions for the tree and eventual monophyly constraints (Heled and Drummond 2012). For our analysis the mismatch between the calibration and the actual marginal prior distributions is not expected to be large, however, because all our calibrated nodes were enforced to be monophyletic. This was confirmed by running the MCMC analysis without the sequence data to show the marginal prior distributions.

Analysis 2 aimed for comprehensive taxonomic sampling of *Rinorea* and was calibrated using a secondary calibration based on age estimates from Analysis 1. Taxonomic and character sampling was the same as for concatenated matrix A. We constrained the *Fusispermum* outgroup and *R. crenata* to be subsequent sisters to the *Rinorea* s.s. ingroup in accordance with previous studies (Tokuoka 2008, Wurdack and Davis 2009) and our own analyses (see Results). The age estimate of the *Rinorea* s.s. crown from Analysis 1 was imported and used as normally distributed secondary calibration. Because the same nodes and node-depths were targeted in both analyses, we expect that their age estimates are robust to the particular taxon and specimen sampling and can hence be translated from Analysis 1 to Analysis 2.

Both analyses used an uncorrelated lognormal relaxed clock (Drummond, Ho *et al.* 2006) as implemented in BEAST 1.7.2 (Drummond, Suchard *et al.* 2012) via the online CIPRES Science Gateway (Miller, Pfeiffer *et al.* 2010). We applied GTR + gamma as the model of DNA sequence evolution for each gene separately; a diffuse gamma (shape 0.001, scale 1000) as prior distribution for ucld.mean, and an exponential prior distribution for yule.birthRate (with mean 0.04 for analysis 1 and 0.076 for analysis 2). Each analysis consisted of five independent runs with 50 million generations, sampling every 10,000th generation. Effective sampling sizes of parameter values and convergence between runs was assessed using Tracer (Rambaut

Table 5.4: Fossil taxa used for calibrating nodes in the divergence time estimation analysis.

Extant taxon	Fossil taxon	Fossil type	Age (Ma)	Reference
Casearia	*Casearia* type	Pollen	37	Graham (1985)
Passiflora	*Passiflora kirchheimeri*	Seeds	37	Mai (1967)
Salix+Populus	*Pseudosalix handleyi*	Flowers, Fruits	48	Boucher *et al.* (2003)
Viola	*Viola* sp.	Seeds	18	Kovar-Eder *et al.* (2001)

and Drummond 2009). Topological convergence was assessed using the online system for graphical exploration of MCMC convergence AWTY (Nylander, Wilgenbusch *et al.* 2008). Mean clade ages as well as 95% highest posterior density (HPD) intervals were calculated. To estimates the age of biogeographical disjunctions (vicariance or dispersal) we considered stem (i.e. the latest outgroup split) as well as crown (i.e. earliest ingroup split) as maximum and minimum ages (Poux, Madsen *et al.* 2005).

5.3 Results

Our taxonomic sampling covered most of the known infrageneric variation in African *Rinorea*, as well as a fair representation of that known from the Neotropics, Madagascar, and Asia. For some widely distributed species, for instance *Rinorea ilicifolia* and *R. angustifolia*, our sampling covered a wide biogeographic distribution throughout West, Central and Southern Africa and Madagascar.

Sequences from *trnL–F* and ITS sequences were length-variable up to 864 and 707 base pairs, respectively. Sequences from *EMB2765* were 942 base pairs long, see Table 5.2. Because the PRANK algorithm separated characters with uncertain homology, ensuing alignments were longer than those based on MAFFT. Such 'thinning' of alignment also led to less informative sites, however, see Table 5.2. Consequently, analyses based on PRANK alignments required more CPU-time while resulting phylogenetic trees were less resolved. Results based on the different marker regions were congruent, as were those based on the four concatenated matrices. Maximum likelihood bootstrap support values showed a slight decrease with increasing taxon sampling (not shown). Posterior probabilities for clades did not decrease with increasing specimen sampling, however, indicating that missing data had no negative effect on the results from our Bayesian analyses. We therefore consider the Bayesian analysis of concatenated matrix 'A' based on MAFFT alignment, with 149 accessions and 2887 characters, to be the best representation of our data.

5.3.1 Phylogenetic results

In general, our results confirm the cpDNA haplotypes found by Wahlert & Ballard (2012), see Figure 5.1. The single accession from the *Rinorea* Apiculata group, *R. crenata*, is separate from all other *Rinorea* (1.00 pp) that are in turn divided into a Neotropical (0.98 pp) and a Palaeotropical (0.98 pp) clade. Within the Neotropical clade, opposite-leaved 'Pubiflora' group (1.00 pp) and alternate-leaved '*Rinorea*' group (0.98 pp) are sisters. Within the Palaeotropical clade, opposite-leaved *R.* subsect. *Verticillatae* (Engler 1902) from Madagascar (1.00 pp) are sister to the Asian species (0.92 pp), that together are sister to a predominantly African clade (0.97 pp). Peculiar is that, within this African clade, two specimens from Gabon (JW4382, JW5172; 1.00 pp) are sister to all others (clade A; 0.96 pp) in the separate marker region as well as in the combined analyses. Two separate insertions of 6 and 7 nucleotides in the *trnL–F* marker region are synapomorphies for this clade. Clade A consists of *Rinorea* subsect. *Dentatae* s.s. (1.00 pp), clade B (0.94 pp), and the 1-ovule clade (0.99 pp) uniting *Rinorea* sect. *Ardisianthus*, the *Rinorea arborea* group, *R.* sect. *Cycloglossae*, and *R.* subsect. *Choriandra*. Clade B consists of *Rinorea* subsect. *Subintegrifoliae* (1.00 pp), clade C (0.98 pp), and clade D (0.96 pp). Clade C consists of a grade of *R.* subsect. *Dentatae* s.l. and *R.* subsect. *Crassiflorae* (1.00 pp). Clade D consists of *Rinorea* subsect. *Ilicifoliae* (1.00 pp), the *R. squamosa* group sensu Wahlert (2010; 1.00 pp), and *R.* subsect. *Brachypetalae* s.l. (i.e. including *R. kamerunensis*; 0.98 pp).

5.3.2 Divergence timing results

The dating analyses suggest a stem age of 56 Myr and a crown age of 45 Myr for *Rinorea* s.s. coinciding with the split between Neotropical and Palaeotropical *Rinorea* (Figure 5.2; estimated ages of selected clades are given in Table 5.5). Within the Neotropical clade, informal groups '*Rinorea*' and 'Pubiflora' diverged much later, around 30 Mya. The Palaeotropical clade starts diverging 38 Mya into an Asian and Malagasy clade and an African clade, and *Rinorea* s.s. may have reached all three currently colonized continents (Africa including Madagascar) by 27 Mya. African *Rinorea* start diversifying 33 Mya, while the Asian (18 Mya) and Malagasy (4 Mya) clades start diversifying later, suggesting that the latter two are younger.

Within African *Rinorea*, clades A–D and the 1-ovule clade started diversifying 20–30 Mya. *R.* sect. *Ardisianthus* (17 Myr) and *R.* subsect. *Ilicifoliae* (16 Myr) are the oldest groups; *R.* sect. *Cycloglossae*, and *R.* subsects. *Brachypetalae*, *Choriandra*, and *Dentatae* s.s. (11–14 Myr) have intermediate ages. *R.* subsect. *Crassiflorae*, the *R. arborea* group, and the *R. squamosa* group (7–9 Myr) are relatively young.

5.4 Discussion

Based on increased taxonomic sampling, our results generally confirm previous results from Wahlert & Ballard (2012). Most haplotypes were recovered and some (i.e. the *Rinorea arborea* group, *R.* sect. *Cycloglossae*, and *R.* subsects. *Choriandra* and *Subintegrifoliae*) now have high posterior probabilities. The South African endemic *R. domatiosa* remained unresolved within the 1-ovule clade but appeared related to *R.* sect. *Ardisianthus* (0.70 pp). Given the morphological similarities between *R. domatiosa* and *R. angustifolia* (e.g. ciliate staminal tube, petals reflexed at anthesis, racemose inflorescence) we expect this relation to receive higher probabilities when more DNA sequence data become available.

In addition we observed several new phylogenetic patterns. First, the clade comprising Asian *Rinorea* and the Malagasy *R.* subsect. *Verticillatae* are sister to all others within the Palaeotropical clade. Consequently, the 1-ovule clade is nested within the 2-ovule clade rather than being its sister as suggested by Wahlert & Ballard (2012), which would imply that the ovary 'lost' one ovule per locule evolutionary (see below). Second, *R. woermanniana* and JW4382 (with uncertain identification as *R. exappendiculata*) appear as sister to all other African species (0.98 pp) suggesting that they require infrageneric recognition. The pistil shape of *Rinorea woermanniana* is distinctive in having a pear-shaped ovary bearing a structurally distinct fusiform style with protruding stigma. Further study is required to determine whether this is a synapomorphy for the clade, however (specimen JW4382 only has immature fruits). Third, *R.* subsect. *Dentatae* appears polyphyletic. The clade comprising *Rinorea dentata* and allied species *R. yaundensis* and *R. zenkeri* is unresolved but may be sister to the 1-ovule clade (0.81 pp). Surprisingly, all other species previously classified within *R.* subsect. *Dentatae* (not sampled by Wahlert and Ballard 2012) are not closely related to this *Dentatae* s.s. clade but rather appear as a grade closely related to *R.* subsect. *Crassiflorae* (0.99 pp). Finally, the African subsects. *Ilicifoliae* and *Brachypetalae* and the *R. squamosa* group form a clade (0.93 pp). Given the limited taxonomic sampling within African *Rinorea* we cannot exclude the possibility that we have missed additional clades. We therefore feel that formal taxonomic recognition of newly discovered clades would be premature.

Figure 5.1: Phylogenetic tree of *Rinorea* (next pages). Bayesian analysis based on concatenated matrix A. Branch labels indicate posterior clade probabilities; balloons show state changes of selected morphological characters; within section Ilicifoliae asterisks indicate taxa with spiny leaves; square brackets mark clades that have high support (pp = 1.00) and are consistent with haplotypes found by Wahlert and Ballard (2012); BRA = Brazil; CAM = Cameroon; CIV = Ivory Coast; COM = Comoros; CRI = Costa Rica; GAB = Gabon, GHA = Ghana; GIN = Guinea; IND = India; KEN = Kenya; LBR = Liberia; MDG = Madagascar; MOZ = Mozambique; MYS = Malaysia; NGA = Nigeria; STP = Sao Tome and Principe; TZA = Tanzania; UGA = Uganda; ZAF = South Africa.

D .93
Brachypetalae+Squamosae+Ilicifoliae

0.4

.74
RV50 CAM *R. amietii* ined.
RV87 CAM *R. amietii* ined.
RV39 CAM *R. longisepala*
.87
FB3 CAM *R. umbricola*
RV83 CAM *R. umbricola*
1
GA2109 CAM *R. longisepala*
GA2116 CAM *R. ledermanni*
GA2174 CAM *R. villiersii ined.*
JW5155 GAB *R. subauriculata*
TA3364 CAM *R. ledermannii*
Crassiflorae

B
.92
.62
1
FB15 CAM *R. verrucosa*
JW4474 GAB *R. verrucosa*
.99
RV63 CAM *R. verrucosa*
1
FJB16000 GAB *R. sp.*
.99
.99
GA2120 CAM *R. sp.*
RV43 CAM *R. dewildei ined.*
C
.73
CJ6148 LIB *R. liberica*
CJ6717 LBR *R. oblongifolia*
.99
1
RO12 GHA *R. oblongifolia*
CJ10470 GIN *R. microdon*
GA2178 CAM *R. sinuata*
Dentatae s.l. grade
1
1
FB13 CAM *R. oblongifolia*
RV40 CAM *R. oblongifolia*
.75
GA2107 CAM *R. oblongifolia*

FB12 CAM *R. campoensis*
GA2113 CAM *R. campoensis*
.95
GS327 CAM *R. campoensis*
A
GS8294 CAM *R. campoensis*
.97
1
JW4640 GAB *R. campoensis*
CJ6054 LBR *R. subintegrifolia*
1
GA2108 CAM *R. subintegrifolia*
RV134 NGA *R. subintegrifolia*
RV30 CAM *R. subintegrifolia*
Subintegrifoliae

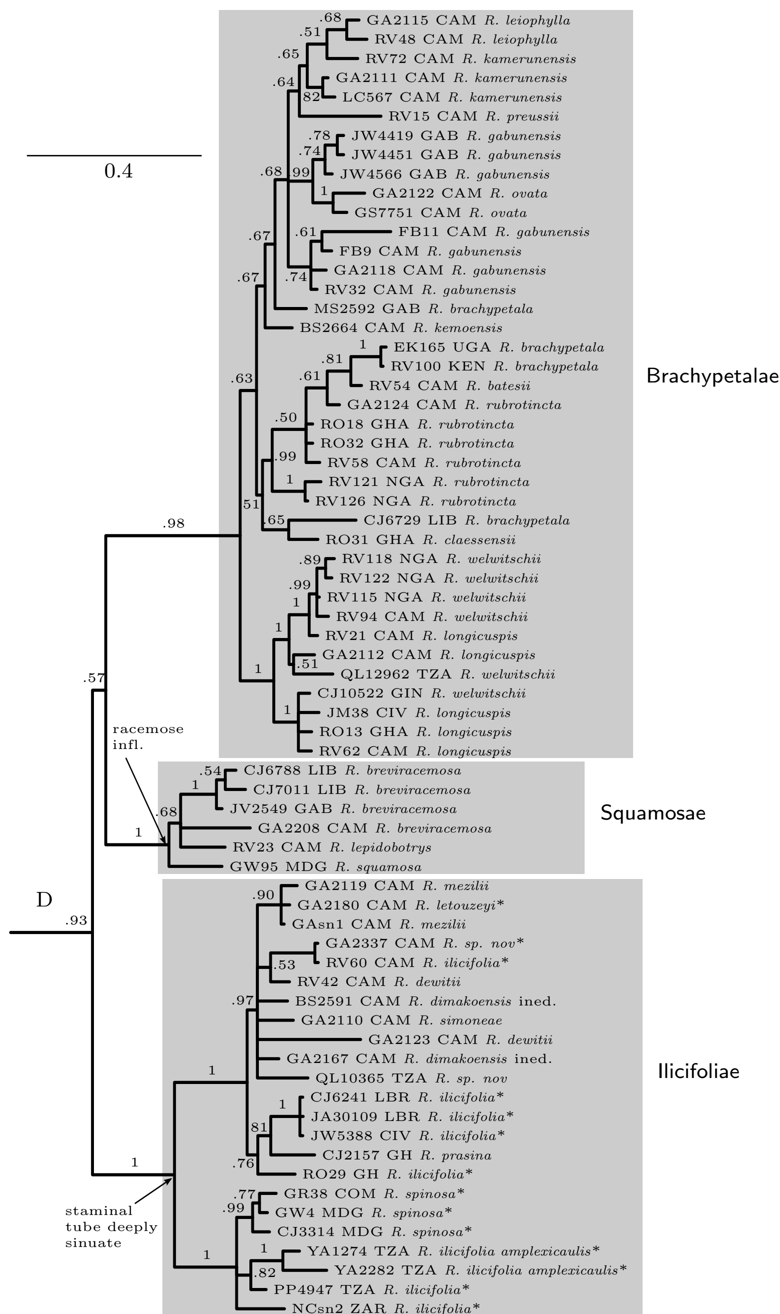
no staminal tube
.99
GA2104 CAM *R. caudata*
1
RV49 CAM *R. caudata*
.55
GA2121 CAM *R. albidiflora*
.99
RV80 CAM *R. albidiflora*
.96
GW21 MDG *R. pugionifera*
1
JB10128 MOZ *R. elliptica*
.61
JB07 MDG *R. greveana*
Choriandra

1
RO42 GHA *R. convallarioides occidentalis*
.99
TA3620 CAM *R. cf. ebolowensis*
.99
ATBP610 UGA *R. beniensis*
JB9908 MOZ *R. convallarioides*
Cycloglossae

race-mose infl.
1
RO41 GHA *R. angustifolia engleriana*
1
RV95 CAM *R. angustifolia engleriana*
African
TS2859 STP *R. chevalieri*
.97
1
GA2117 CAM *R. angustifolia engleriana*
.99
.91
QL13003 KEN *R. sp. nov*
QL12745 TZA *R. angustifolia ardisiiflora*
Ardisianthus
1 ovule per locule
.70
1
NCsn1 ZAF *R. domatiosa*
TN159 ZAF *R. domatiosa*
.86
GW113 MDG *R. ballardii ined.*
.93
JL3197 MDG *R. arborea*
JB10171 MOZ *R. arborea*
Arboreae

.81
CJsn1 GH *R. yaundensis*
.87
DH2964 GH *R. yaundensis*
RV55 CAM *R. yaundensis*
Palaeotropical
.62
RV56 CAM *R. yaundensis*
.98
RV59 CAM *R. zenkeri*
.95
RG5541 CAM *R. dentata*
RV119 NIG *R. dentata*
1
RV75 CAM *R. dentata*
1
FB6 CAM *R. dentata*
LC578 CAM *R. dentata*
Dentatae s.s.

1
JW4382 GAB *R. cf. exappendiculata*
Rinorea s.s.
JW5172 GAB *R. woermanniana*
1
.99
HB3192 MYS *R. anguifera*
Asian .92
MC2149 *R. lanceolata*
AG66982 MYS *R. horneri*
.95
MC2148 IND *R. bengalensis*
Verticillatae
opposite leaves
.50
CS217 MDG *R. bullata*
1
GW12 MDG *R. mutica*
1
MC582 MDG *R. callmanderi*
opposite leaves
1
JPS9529 BRA *R. macrocarpa*
Pubiflorae
JPS9543 BRA *R. racemosa*
Neotropical
.98
PM9505 CRI *R. hummelii*
.98
JPS5679 BRA *R. bahiensis*
Rinoreae
JPS9541 BRA *R. guianensis*
1
HB94 CRI *R. crenata*
Apiculata
Fusispermum laxiflorum

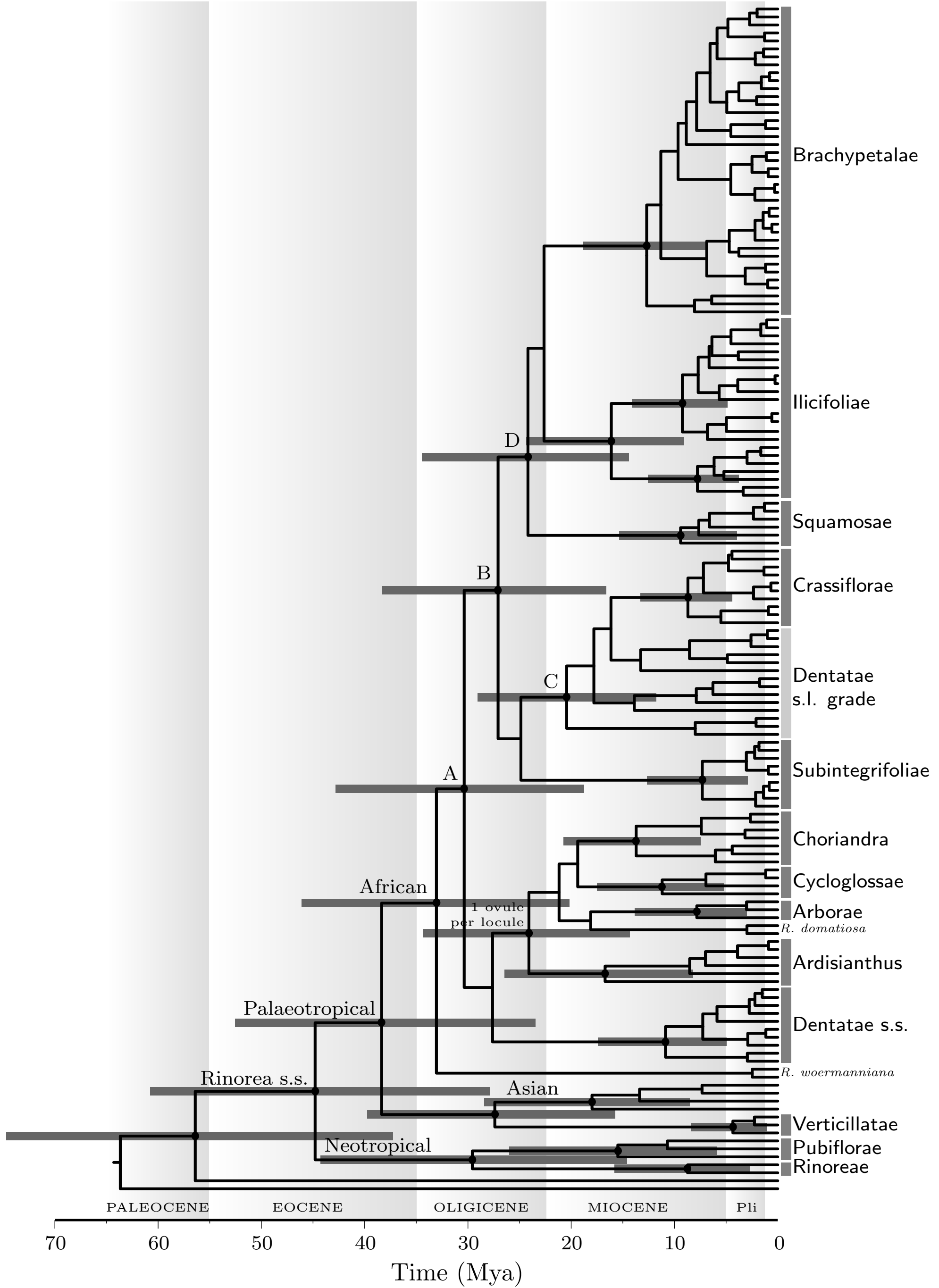

0.4
GA2115 CAM R. leiophylla
RV48 CAM R. leiophylla
RV72 CAM R. kamerunensis
GA2111 CAM R. kamerunensis
LC567 CAM R. kamerunensis
RV15 CAM R. preussii
JW4419 GAB R. gabunensis
JW4451 GAB R. gabunensis
JW4566 GAB R. gabunensis
GA2122 CAM R. ovata
GS7751 CAM R. ovata
FB11 CAM R. gabunensis
FB9 CAM R. gabunensis
GA2118 CAM R. gabunensis
RV32 CAM R. gabunensis
MS2592 GAB R. brachypetala
BS2664 CAM R. kemoensis
EK165 UGA R. brachypetala
RV100 KEN R. brachypetala
RV54 CAM R. batesii
GA2124 CAM R. rubrotincta
RO18 GHA R. rubrotincta
RO32 GHA R. rubrotincta
RV58 CAM R. rubrotincta
RV121 NGA R. rubrotincta
RV126 NGA R. rubrotincta
CJ6729 LIB R. brachypetala
RO31 GHA R. claessensii
RV118 NGA R. welwitschii
RV122 NGA R. welwitschii
RV115 NGA R. welwitschii
RV94 CAM R. welwitschii
RV21 CAM R. longicuspis
GA2112 CAM R. longicuspis
QL12962 TZA R. welwitschii
CJ10522 GIN R. welwitschii
JM38 CIV R. longicuspis
RO13 GHA R. longicuspis
RV62 CAM R. longicuspis
Brachypetalae
racemose infl.
CJ6788 LIB R. breviracemosa
CJ7011 LIB R. breviracemosa
JV2549 GAB R. breviracemosa
GA2208 CAM R. breviracemosa
RV23 CAM R. lepidobotrys
GW95 MDG R. squamosa
Squamosae
D
GA2119 CAM R. mezilii
GA2180 CAM R. letouzeyi*
GAsn1 CAM R. mezilii
GA2337 CAM R. sp. nov*
RV60 CAM R. ilicifolia*
RV42 CAM R. dewitii
BS2591 CAM R. dimakoensis ined.
GA2110 CAM R. simoneae
GA2123 CAM R. dewitii
GA2167 CAM R. dimakoensis ined.
QL10365 TZA R. sp. nov
CJ6241 LBR R. ilicifolia*
JA30109 LBR R. ilicifolia*
JW5388 CIV R. ilicifolia*
CJ2157 GH R. prasina
RO29 GH R. ilicifolia*
GR38 COM R. spinosa*
GW4 MDG R. spinosa*
CJ3314 MDG R. spinosa*
YA1274 TZA R. ilicifolia amplexicaulis*
YA2282 TZA R. ilicifolia amplexicaulis*
PP4947 TZA R. ilicifolia*
NCsn2 ZAR R. ilicifolia*
Ilicifoliae
staminal tube deeply sinuate

Diagnostic morphological characters traditionally used to delimit infrageneric groups within African *Rinorea* (i.e. leaf position, number of ovules, androecium structure; Brandt 1914), appeared to largely correlate with our phylogenetic tree (see Figure 5.1). However, some exceptions can be observed. For example, taxa with free stamens were traditionally classified separately from those with a staminal tube (Brandt 1914, De Wildeman 1920), and free filaments were considered primitive within *Rinorea* (Hekking 1988). Our results suggest that a staminal tube is a plesiomorphic character, however, and free filaments are a later reversal serving as an apomorphy within African *Rinorea*. Indeed, *Fusispermum* has connate filaments, similar to those in Palaeotropical *Rinorea* (Cuatrecasas 1950) corroborating that this is an ancestral rather than derived state. Our results thus confirm the hypothesis of Tokuoka (Tokuoka 2008) that free stamens are a synapomorphy for all Violaceae except *Fusispermum* and *Rinorea*. Likewise, sections with 1 ovule per locule were previously considered to be 'archaic' within African *Rinorea* (Achoundong 2000) but in our analyses appear to be derived. As the Malagasy *R.* subsect. *Verticillatae* and the Asian species have 2 ovules per locule, the Palaeotropical clade probably had an ancestor with 2 ovules per locule. The only other clades with single ovules are the '*Rinorea*' group and the Apiculata group suggesting that a single ovule per locule may have been the ancestral state for *Rinorea*. However, the Neotropical 'Pubiflorae' group consists of species with 1–3 ovule per locule and *Fusispermum* has many ovules that give rise to two different kinds of seed (Tokuoka 2008). Therefore, the ancestral number of ovules for *Rinorea* (and for Violaceae) could just as well be variable.

Although we have aimed for sampling all major infrageneric groups, our sampling (covering 50% of extant African species) was biased due to increased collection efforts especially in Cameroon. Therefore, testing historical biogeographic hypotheses within Africa or inferring evolutionary ecological trends in *Rinorea* requires denser sampling, especially in West and East Africa. Based on our taxonomic sampling for Neotropical (10%) and Asian (13%) *Rinorea*, theoretically a paraphyly or polyphyly hypothesis cannot be rejected for *Rinorea* on these continents. Nevertheless, as we included most infrageneric groups (Hekking, 1988; DeMuria & Ballard, unpublished data) and all available evidence points towards monophyly of Asian and Neotropical *Rinorea* s.s., we feel we can use this as a safe assumption throughout the rest of this discussion.

Figure 5.2: *Rinorea* Timing of divergences (next page). Time-calibrated maximum clade credibility tree based on concatenated matrix A. Horizontal bars show 95% HPD intervals of clade ages; grey vertical bars mark clades that have high support (pp = 1.00) and are consistent with haplotypes found by Wahlert and Ballard (2012); Pli = Pliocene.

5



Brachypetalae
Ilicifoliae
Squamosae
Crassiflorae
Dentatae s.l. grade
Subintegrifoliae
Choriandra
Cycloglossae
Arborae
R. domatiosa
Ardisianthus
Dentatae s.s.
R. woermanniana
Verticillatae
Pubiflorae
Rinoreae
D
B
C
A
African
1 ovule per locule
Palaeotropical
Rinorea s.s.
Asian
Neotropical
PALEOCENE
EOCENE
OLIGICENE
MIOCENE
Pli
70
60
50
40
30
20
10
0
Time (Mya)

Some species appear non-monophyletic in our phylogenetic tree, see Figure 5.1. Within the *R.* subsect. *Dentatae* s.s. clade two *R. dentata* specimens from southern Cameroon appear to be sister to all others, while three other specimens from Cameroon and Nigeria are more closely related to *R. zenkeri* and *R. yaundensis*. *Rinorea dentata* is distributed from Liberia to Congo (Tennant 1963, Hawthorne and Jongkind 2006) and can be variable in the field (Achoundong, pers. comm.) and indeed flowers are cream coloured in LC578 while they are yellow in RV119. Thus, the taxonomic status of this entity deserves closer attention which may require reinstatement of existing synonyms from southern Cameroon (i.e. *R. bipindensis* Engl. and *R. dinklagei* Engl.).

Rinorea angustifolia is a variable species distributed throughout much of tropical Africa (Tennant 1963, Grey-Wilson 1981). *R. angustifolia* subsp. *ardisiiflora* from Tanzania is separate from all others, while a specimen of a suspected new species from Kenya (Quentin Luke, pers. comm.) is closely related to *R. angustifolia* subsp. *engleriana* specimens from Cameroon and Ghana plus *R. chevallieri* from Sao Tomé. Pending better sampling of this group, the present results already show that possibly several more taxa are present and some synonyms may need to be reinstated.

Rinorea oblongifolia is another variable species (Chipp 1923) currently thought to be distributed across tropical Africa (Tennant 1963, Hawthorne and Jongkind 2006). Specimens from Cameroon appear as sister to all others in the *R.* subsect. *Dentatae* s.l. grade (0.99 pp), while those from Ghana and Liberia are closely related to *R. sinuata*, *R. microdon*, and *R. liberica* (1.00 pp). This separation is corroborated by morphological differences (e.g. flower colour and inflorescence structure), suggesting they are separate species (the type specimen is from Cameroon).

Rinorea ilicifolia is distributed throughout much of tropical Africa and easily recognized by its distinctive spiny leaf margins (Grey-Wilson 1981). Such spiny-leaved specimens occur in at least two separate clades, however. Those from Tanzania and South Africa cluster together with *R. spinosa* from Madagascar, while those from Liberia, Ivory Coast, Ghana, and Cameroon cluster together with non-spinose species from Ghana, Cameroon, and Tanzania. Again, this suggests the presence of at least two taxa.

R. longicuspis and *R. welwitschii* appear as sisters within *R.* subsect. *Brachypetalae* s.l., but the 11 specimens concerned do not nicely separate the two. The results suggest that some specimens might be misidentified (i.e. CJ10522, RV21, GA2112) which may mean that the most commonly used character for identification (i.e. leaf pubescence) is not reliable. Within *R.* subsect. *Brachypetalae* s.l., Gabonese specimens of *R. gabunensis* are sister to *R. leiophylla* (0.60 pp), while the Cameroonian specimens are separate (0.67 pp). Posterior probabilities are low, however. These results suggest that several species currently regarded as widespread throughout Africa (i.e. *R. angustifolia*, *R. dentata*, *R. oblongifolia*, and *R. ilicifolia*) may actually comprise species complexes containing species with a more local distribution.

With 11 Myr (*R. dentata*), 16 Myr (*R. ilicifolia*), 17 Myr (*R. angustifolia*), and even 27 Myr (*R. dentata*), our age estimates suggest that the coalescent within these widespread complexes is extremely deep, further corroborating specific status of the hypothesized 'local' species they contain. These observations suggest that *Rinorea* are even more restricted to certain environments than previously thought (Achoundong 1996, Achoundong 2000, Adomou, Sinsin *et al.* 2006, Githae, Chuah-Petiot *et al.* 2008, Mwavu and Witkowski 2009, Tchouto, de Wilde *et al.* 2009, Djuikouo, Doucet *et al.* 2010).

5.4.1 Trans-Atlantic dispersal

Our reconstructions of historical biogeographic patterns indicate that trans-Atlantic dispersal rather than Gondwanan vicariance explains the disjunction between Neotropical and Palaeotropical *Rinorea.* The disjunction is estimated at 45 Mya (95% HPD 28–61) in the Eocene epoch long after the tropical Atlantic had rifted in the Cretaceous roughly 119–105 Mya (McLoughlin 2001). Given that *Fusispermum* and *Rinorea* Apiculata outgroups are distributed in the Neotropics only, *Rinorea* s.s. is likely to have originated there and dispersed to the Palaeotropics. In addition, African lineages are the oldest within Palaeotropical *Rinorea*, lending further support to the hypothesis that dispersal was probably from America to Africa.

Dispersal of plants across the tropical Atlantic by ocean currents is well documented (Renner, Clausing *et al.* 2001, Renner 2004) and because in the Middle Eocene the Atlantic spanned only 1200 km between Liberia and Brazil (Houle 1998) this is the most probable route of dispersal. Such dispersals are usually inferred to be westward (i.e. out of Africa), in accordance with the prevalently westward-flowing sea currents, but eastward dispersal across the Atlantic has been described for tree species such as *Hernandia* rouglhy 3 Mya (Michalak, Zhang *et al.* 2010), *Annona senegalensis* 14–16 Mya (Richardson, Chatrou *et al.* 2004) and *Pradosia spinosa* 7–14 Mya (Bartish, Antonelli *et al.* 2011) as well as herbaceous species such as *Genlisea* (Fleischmann, Schäferhoff *et al.* 2010), *Pitcairnia feliciana* (Bromeliaceae) 12 Mya (Givnish, Millam *et al.* 2004), Melastomeae 12–14 Mya (Renner, Clausing *et al.* 2001), and *Maschalocephalus* (Rapateaceae) 7.3 Mya (Givnish, Millam *et al.* 2004).

The seasonal Atlantic North Equatorial Counter Current (NECC) flowing eastward from the Guyanas to the Gulf of Guinea (Fratantoni, Johns *et al.* 2000) may have been responsible for the dispersal of these taxa as well as of *Rinorea*, and indeed Lagrangian drifter data suggest that transatlantic travel from America to Africa is possible (Monzón-Argüello, López-Jurado *et al.* 2010). Dispersal of *Rinorea* seeds across the tropical Atlantic by aquatic birds is unlikely because *Rinorea* fruits and seeds are not known to be eaten by birds. In any case there are no bird migratory routes across the Atlantic (Berthold 2001). Dispersal by wind is equally unlikely because of the relatively large and heavy seeds.

Table 5.5: Estimated age of selected clades.

Clade	Age (Ma)	95% HPD
Violaceae crown	70.5	57.0–84.1
Rinorea s.s.	44.8	27.9–60.8
Neotropical	29.6	14.6–39.8
Palaeotropical	38.4	23.5–52.5
Asian+Verticillatae	27.4	15.8–39.8
Asian	18.0	8.5–28.4
R. subsect. *Verticillatae*	4.3	1.1–8.4
African	33.1	20.2–46.1
Clade A	30.4	18.8–42.8
R. subsect. *Dentatae* s.s.	10.9	5.0–17.4
Single ovule per locule clade	24.1	14.3–34.3
R. sect. *Ardisianthus*	16.7	8.2–26.5
R. arborea group	7.8	3.0–13.8
R. sect. *Cycloglossae*	11.2	5.2–17.5
R. subsect. *Choriandra*	13.7	3.3–12.2
Clade B	27.1	16.6–38.3
R. subsect. *Subintegrifoliae*	7.3	2.9–12.7
Clade C	20.4	11.6–29.0
R. subsect. *Crassiflorae*	8.5	4.4–13.3
Clade D	24.2	14.4–34.4
R. subsect. *Ilicifoliae*	16.1	9.1–24.3
East-African *Ilicifoliae*	7.8	3.8–12.5
West-African *Ilicifoliae*	9.2	4.9–14.1
R. squamosa group	9.4	4.0–15.4
R. subsect. *Brachypetalae* s.l.	12.7	7.0–18.9

An alternative explanation for the tropical disjunction is that *Rinorea* reached Africa through Laurasia. During the Eocene, tropical forests are thought to have extended further north, forming a continuous Boreotropical flora. Possibly, *Rinorea* reached Africa after range expansion in these Boreotropical forests, a scenario proposed for various other pantropical plant clades (Malcomber 2002, Richardson, Chatrou *et al.* 2004, Weeks, Daly *et al.* 2005, Couvreur, Pirie *et al.* 2011). We have two main arguments against such a scenario for *Rinorea*. First, fragmentation of a once continuous Boreotropical distribution predicts synchronous divergence between clades from all three tropical regions, whereas our phylogenetic results indicate a directional dispersal from America to Africa and no evidence for old Eurasian lineages. Second, given the poor dispersal abilities of *Rinorea*, long-distance dispersal via Laurasia seems to be less likely than transport over 1200 km by ocean currents, especially because South and North America were still separated by the Central American Seaway (Obando-Rodríguez, Bemis *et al.* 1996) and Africa and Eurasia by the Tethys Sea at that time (Savostin, Sibuet *et al.* 1986). Consequently, even though no experimental evidence exists for long-time survival of *Rinorea* seeds in seawater, we hypothesize that *Rinorea* seeds were transported by the Atlantic NECC from America to Africa.

5.4.2 Dispersal to Asia

Our analyses confirm a single dispersal of *Rinorea* to Asia, in accordance with previous analyses (Bakker, van Gemerden *et al.* 2006, Wahlert and Ballard 2012) as well as with patterns found in other plant clades (Renner and Meyer 2001, Malcomber 2002, Appelhans, Keßler *et al.* 2012). Time estimates suggest that the Asian clade diverged in the Oligocene or early Miocene (see Table 5.5), which is too recent to be explained by dispersal via the Indian tectonic plate (Ali and Aitchison 2008) or range expansion in Eocene Boreotropical forests (Malcomber 2002, Couvreur, Pirie *et al.* 2011). Our crown age estimate for Asian *Rinorea* (18 Myr) coincides with the collision between the Arabian and Eurasian plates in the Middle Miocene and closing of the Tethys sea, thus opening a novel dispersal route between Africa and Eurasia via the Arabian Peninsula and the Levant region. The same route has been hypothesized to facilitate range expansion from Africa into Asia by for example Anthemideae (Oberprieler 2005), Campanulaceae (Roquet, Sanmartín *et al.* 2009) and *Uvaria* (Zhou, Su *et al.* 2012). Possibly, tropical rain forests expanded at that same time, as an effect of the late Middle Miocene thermal maximum, allowing forest species to expand their ranges further North (Zhou, Su *et al.* 2012).

5.4.3 Multiple independent colonizations of Madagascar

The island of Madagascar is renowned for its exceptional biodiversity, with extraordinary levels of species diversity and endemism (Goodman and Benstead 2005). Its flora consists of over ten thousand Angiosperm species, 84% of which are endemic (Callmander, Phillipson *et al.* 2011) and probably resulted from a progressive differentiation of the autochthonous Gondwanan stock and natural introduction of taxa over time through long distance dispersal (Leroy 1978). Africa appears by far to be the most important source of floral dispersal to Madagascar (Yoder and Nowak 2006). Time-calibrated phylogenetic studies suggest that most introduced plants arrived on Madagascar in the Miocene (Renner 2004, Weeks and Simpson 2007, Bartish, Antonelli *et al.* 2011, Appelhans, Keßler *et al.* 2012, Zhou, Su *et al.* 2012) or later (Couvreur, Chatrou *et al.* 2008); dispersal of *Acridocarpus* (Malpighiaceae) in the Eocene being a notable exception (Davis, Bell *et al.* 2002).

Malagasy *Rinorea* taxa were recovered in five different clades, confirming previous claims that *Rinorea* colonized Madagascar multiple times independently from the African mainland (Bakker, van Gemerden *et al.* 2006, Wahlert and Ballard 2012). The ancestor of *R.* subsect. *Verticillatae* colonized Madagascar in the Miocene or late Oligocene between 4 and 27 Mya (95% HPD = 1–40); that of the *Rinorea arborea* group in the Pliocene or late Miocene between 3–8 Mya (95% HPD =0–14); that of *R.* subsect. *Choriandra* after Mid Miocene between 0 and 14 Mya (95% HPD = 0–21); *Rinorea squamosa* in the Pliocene or late Miocene between 0 and 9 Mya (95% HPD = 0–15); and *Rinorea spinosa* in the Pliocene between 2 and 6 Mya

(95% HPD = 0–10). As *R. angustifolia* also occurs on Madagascar, this constitutes a sixth independent colonization event (Wahlert and Ballard 2012), probably within the last 17 Myr. Considering diversification of the Palaeotropical clade since the late Eocene (38 Mya) these colonization events are all relatively recent. Intriguingly, the estimated times of colonization have considerable overlap between 4–6 Mya suggesting that factors governing these independent colonizations of *Rinorea* to Madagascar may have been similar. Human transport must be ruled out as an explanation for the arrival of *Rinorea* on Madagascar as the island was first colonized by settlers from the Sunda Islands only 2300 years ago (Burney, Burney *et al.* 2004). At its narrowest point, the Mozambique channel between Africa and Madagascar is 430 km and is relatively narrow, but prevailing currents have been directed westward towards Africa since the early Miocene (Ali and Huber 2010) possibly complicating an ocean journey to Madagascar at the time of *Rinorea* dispersal. However, recent data suggest a large circulation filling most of the northern part of the channel (Donguy and Piton 1991) and a series of large eddies propagating southward in southern part (Schouten, de Ruijter *et al.* 2003), possibly allowing plant parts to be transported by water in both directions.

5.5 Conclusions

Our data provide evidence that *Rinorea* originated in the Neotropics and reached Africa in the Eocene through trans-Atlantic dispersal. In Africa, the genus proliferated since the Oligocene into the large phylogenetic diversity that we see today. From there, *Rinorea* dispersed to Asia in the Oligocene or early Miocene probably after closing of the Tethys Sea. *Rinorea* has reached Madagascar multiple times independently within a relatively recent time scale (Pliocene), suggesting a possible common cause for these independent colonizations. The evidence from nuclear DNA sequences is generally congruent with plastid haplotypes found in earlier studies of *Rinorea*, and helped resolve additional clades, some of which warrant further taxonomic study (i.e. the *R. woermanniana* clade and Dentatae s.l. grade). It is clear that African *Rinorea* are in need of comprehensive taxonomic revision (Grey-Wilson 1981, Dowsett-Lemaire and White 1990, Wahlert 2010), and we hope that our contribution to understanding *Rinorea* infrageneric relationships will offer some handles to facilitate this task.

Acknowledgements

We are indebted to Ria Vrielink, Bram Lestrade and Sandrina Pardoel for help in the lab. Gaston Achoundong, Olivier Sene-Belinga, Safián Szabolcsz, Stephen Davey, Andrew Walker and Andrews Kankam Amankwah helped collect samples in the field. We thank Tinde van Andel, Lars Chatrou, Carel Jongkind, Erik Koenen, Quentin Luke, and Jan Wieringa for contributing silica samples. RvV acknowledges the Systematics Research Fund; RvV and REO acknowledge the Alberta Mennega Foundation and the Hugo de Vries Foundation for financial support.

Chapter 6

Asynchronous divergence of *Cymothoe* forest butterflies and their *Rinorea* host plants in tropical Africa

Robin van Velzen[1,2], Renske E. Onstein[3], Marc SM Sosef[1,2] & Freek T Bakker[1]

[1] Biosystematics Group, Department of Plant Sciences, Wageningen University and Research Centre, Wageningen, The Netherlands
[2] Naturalis Biodiversity Centre (Section NHN), Wageningen University, Generaal Foulkesweg 37, 6703 BL Wageningen, The Netherlands
[3] Institute of Systematic Botany, University of Zurich, Zollikerstrasse 107, Zurich 8008, Switzerland

Abstract

Insect herbivores are generally highly host specific and insect-plant associations seem phylogenetically conserved, with related insects usually feeding on related plants. However, the factors that generate, maintain, and constrain these associations remain unclear. We investigate the system of *Cymothoe* forest butterflies (Nymphalidae, Limenitidinae) and their *Rinorea* host plants (Violaceae) in tropical Africa, which is especially suitable for untangling processes shaping patterns of host plant associations because of its high level of specificity (mostly monophagous) and the high number of related species involved (33 herbivores and 32 hosts).

The aim of this chapter is to distinguish between alternative scenarios for the evolution of insect-host plant associations by comparing species-level phylogenetic trees of *Rinorea*-feeding *Cymothoe* with that of their hosts within an absolute timeframe. Our results show that: (i) Divergences among extant *Cymothoe* are more recent than those among their associated *Rinorea* hosts, suggesting asynchronous diversification of *Cymothoe* herbivores onto already diversified clades within African *Rinorea*; (ii) Phylogenetic trees of *Cymothoe* and their associated *Rinorea* host plants are discordant and current associations between *Cymothoe* herbivores and their *Rinorea* hosts have developed primarily through a process of host shifting rather than by cospeciation; and (iii) Related *Cymothoe* tend to feed on related *Rinorea* hosts.

Based on the available data, we propose a recent origin of *Rinorea*-feeding by *Cymothoe* butterflies with a single colonization of pre-existing lineages in the late Miocene. We find no support for reduced diversification of African *Rinorea* after the colonization. Current associations are best explained by a predominance of shifts among related plants, probably due to constraints in larval physiology and female oviposition behaviour. These findings are in agreement with a scenario of sequential evolution as a dominant pattern in insect-plant interactions.

6.1 Background

Given the extraordinary phylogenetic diversity of herbivorous insects and terrestrial plants, understanding their interactions is central to understanding global patterns in terrestrial biodiversity (Mitter, Farrell *et al.* 1988, Farrell and Mitter 1998, Price 2002, Futuyma and Agrawal 2009, Novotny, Miller *et al.* 2010). The fact that insect herbivores are generally highly host specific (Futuyma and Moreno 1988) and that those associations seem phylogenetically conserved (Ehrlich and Raven 1964, Mitter, Farrell *et al.* 1991) leads to numerous questions regarding the factors that generate, maintain, and constrain these interactions.

At least three main alternative scenarios of insect-host plant evolution exist: (i) *cospeciation*, in accordance with Fahrenholz' rule that parasites follow the speciation events of their hosts (Fahrenholz 1913), suggests that herbivores diversify in concert with their host plants (Farrell and Mitter 1990, Farrell and Mitter 1998). The assumption is that when host plant populations become reproductively isolated, so will the associated specialist insect herbivores (Janz 2011). This scenario predicts plant and herbivore phylogenies that are concordant both in topology and node age. (ii) *Escape and radiation,* envisioned by Ehrlich and Raven (1964), suggests that diversity of herbivorous insects and their hosts is the result of iterations of plant diversification after escape from herbivory due to novel defences, followed by colonization by insects that then diversify on the new and relatively underused resource. This scenario predicts discordant phylogenies with reciprocal asynchronous radiations of herbivores and plants due to their entering new adaptive zones (Simpson 1953). It is important to note that under this scenario insect diversification is promoted by colonization of novel host plants, while diversification in plants is promoted by the absence of associated insect herbivores (Janz 2011). (iii) The *sequential evolution* scenario is based on the assumption that interactions between insects and their host plants are asymmetric: while diversification of plants leads to increased resource heterogeneity and possibilities for speciation by herbivorous insects, herbivores have a negligible effect on speciation of their hosts (Jermy 1976). This scenario suggests that host use is labile, with colonization of pre-existing plant lineages (Jermy 1984, Bernays and Graham 1988) and that diversification is asynchronous as the plant lineages predate the insects that feed on them. Delay between plant and insect diversification is variable such that there is no correlation between the evolutionary age of plant groups and that of insect species living on them (Jermy 1984). This scenario also predicts discordant phylogenies; the main difference with the escape and radiation scenario is that insects have no effect on the diversification of their hosts.

Various molecular phylogenetic studies have explicitly addressed specific macro evolutionary hypotheses about insect-plant associations (Becerra and Venable 1999, Winkler and Mitter 2008, Futuyma and Agrawal 2009). Although specialized herbivorous insects can indeed remain conservative in terms of host plant use over

millions of years (Stone, Hernandez-Lopez *et al.* 2009), and some evidence has been found for significant synchronous co-cladogenesis at higher taxonomic levels (Farrell and Mitter 1990, Farrell and Mitter 1998, Becerra 2003), in most insect clades cospeciation does not explain observed patterns of host plant associations (Janz and Nylin 1998, Janz, Nyblom *et al.* 2001, Winkler and Mitter 2008). Even though the escape and radiation theory has been dominating theoretical and empirical research on coevolution (Jermy 1984, Agrawal 2007), its premise that herbivorous insects promote the development of novel defences in plants has remained controversial (Dicke 2000, Cornell and Hawkins 2003, Agrawal, Hastings *et al.* 2012). To our knowledge, there is no evidence for increased diversification in plants after escape from insect herbivores, or decreased diversification after becoming colonized. Contrarily, plants harbouring the most diverse herbivore communities are often ecologically dominant and widespread (Lewinsohn, Novotny *et al.* 2005, Nyman, Linder *et al.* 2012). In contrast, most phylogenetic studies support the hypothesis of sequential evolution (Janz and Nylin 1998, Ronquist and Liljeblad 2001, Lopez-Vaamonde, Wikström *et al.* 2006, Gómez-Zurita, Hunt *et al.* 2007, Smith, Godsoe *et al.* 2008, McKenna, Sequeira *et al.* 2009) and this scenario has become widely accepted (Winkler and Mitter 2008). Still, much about the evolution of associations between insect herbivores and their host plants is in need of unravelling (Dicke 2000, Janz 2011).

6.1.1 Host shifts

For example, even though host shifts in herbivorous insect lineages are commonplace, it is unclear whether they are an important driver for diversification, and if so, how. At higher taxonomic levels, shifts to a plant lineage distant from the ancestral host(s) could promote species diversification in herbivorous insects more than at lower taxonomic levels because it allows entering new 'larger', empty niches (Simpson 1953, Mayr 1963, Ehrlich and Raven 1964). For example, Pierid butterflies diversified after colonising Brassicaceae (Braby and Trueman 2006, Wheat, Vogel *et al.* 2007, Fordyce 2010). Indeed, correlations between host shifts involving distantly-related host plant lineages and increased rates of species diversification have been found in various insect clades (Braby and Trueman 2006, Janz, Nylin *et al.* 2006, Weingartner, Wahlberg *et al.* 2006, Winkler, Mitter *et al.* 2009, Fordyce 2010). Yet they may not account for the great diversity of herbivorous insects, however, as they are considered uncommon because of the inherent difficulty of colonizing distant hosts (Nyman 2010). At the species level, after a host shift, divergent selection on herbivores could ultimately lead to their reproductive isolation and, hence, ecological speciation (Emelianov, Dres *et al.* 2001, Berlocher and Feder 2002, Drès and Mallet 2002, Nosil, Crespi *et al.* 2002, Stireman, Nason *et al.* 2005, Singer and McBride 2010). Whether this process is common enough to generate the high levels of insect diversity remains controversial, however (Nyman, Vikberg *et al.* 2010), as many insect sister species

appear to have allopatric origins (e.g. McBride, van Velzen *et al.* 2009), and some clades of herbivorous insects appear to have diversified without any evidence for ecological speciation (e.g. Imada, Kawakita *et al.* 2011).

Another important question is whether related insect herbivores occur only on related hosts or on distantly related hosts with similar ecological characteristics. In general, specialized insects are expected to preferentially shift between hosts that are similar in terms of characteristics that influence herbivore fitness (Agosta 2006, Nyman 2010). As such, herbivores can be seen as tracking suitable resources, rather than host species (Janz, Nyblom *et al.* 2001). For example, phytochemistry plays an important role in host plant recognition as well as metabolic processing by specialized herbivores (Thompson and Pellmyr 1991, Dicke 2000, West and Cunningham 2002, Bossart 2003, Miles, del Campo *et al.* 2005), and plants that are chemically dissimilar from current hosts are expected to be difficult to colonize. Associations of herbivorous insects are therefore often thought to track patterns of secondary chemistry in their hosts (Becerra 1997, Agrawal 2007). Obviously, the distribution of such resources often reflects their evolutionary history so that phylogenetic patterns in host-parasite associations are to be expected (Ives and Godfray 2006), and indeed, related insects are generally found to feed on related plants (Ehrlich and Raven 1964, Jermy 1984, Mitter, Farrell *et al.* 1991, Winkler and Mitter 2008). However, some studies suggest that ecological factors explain current associations better than host phylogeny. For example, plant growth form was found to have a large effect on the probability of host shifts in butterflies in general (Janz and Nylin 1998). In addition, host shifts have been found to occur predominantly to common and/or widespread plants (Lewinsohn, Novotny *et al.* 2005, Nyman, Linder *et al.* 2012).

6.1.2 Synchronicity

In order to distinguish among various scenarios for insect-host plant evolution, it is important to compare relative divergence times of consumers and hosts (Percy, Page *et al.* 2004, Sorenson, Balakrishnan *et al.* 2004, De Vienne, Giraud *et al.* 2007). Studies of insect-host plant evolution based on accurately dated phylogenies are few (Nyman 2010), however, and relative timing of divergence in insects and their associated hosts is controversial (Wheat, Vogel *et al.* 2007, Kergoat, Le Ru *et al.* 2011). Evidence for synchronous diversification has been found in *Blepharida* beetles feeding on Burseraceae (Becerra 2003), Pieridae on Brassicaceae (Wheat, Vogel *et al.* 2007), and *Eois* butterflies on Neotropical *Piper* (Strutzenberger and Fiedler 2011), and weevils on various Angiosperm clades (McKenna, Sequeira *et al.* 2009), all thus supporting a cospeciation scenario. By contrast, asynchronous diversification, where diversification of herbivores occurs significantly later than that of their host plants, is in accordance with the escape and radiation and sequential evolution scenarios and has been found in Psyllid bugs feeding on Legumes (Percy, Page *et al.* 2004),

Tephritid flies on Asteraceae (Brändle, Knoll *et al.* 2005), and leaf-mining moths (Lopez-Vaamonde, Wikström *et al.* 2006), as well as leaf beetles (Gómez-Zurita, Hunt *et al.* 2007), weevils (McKenna, Sequeira *et al.* 2009), and leaf-mining sawflies (Leppänen, Altenhofer *et al.* 2012) on various Angiosperm clades. The lag between the time of divergence of plants and subsequent diversification of associated insects can range from tens of millions of years (Lopez-Vaamonde, Wikström *et al.* 2006) to only shortly after (Kergoat, Le Ru *et al.* 2011).

6.1.3 Species as fundamental units

Studies on insect diversification that focus on large clades allow comparing patterns across many independent lineages, which, when replicated among lineages, can reveal general processes. However, we have to keep in mind that generic-level phylogenetic studies cannot reveal population-specific patterns, which is the level at which processes underlying diversification operate (Agosta 2006). For instance, host shifts and extinctions can easily erode signs of cospeciation events over time (Nyman *et al.* 2010; Janz 2011), rendering studies over deep time-scales less suitable in this respect.

In addition, representing species-level associations by higher taxa complicates assessment of synchronous diversification because, when insect and plant species diverge in synchrony, divergence times of their respective higher taxa are not necessarily correlated. Likewise, if only the plants are represented by higher taxa in evolutionary association-studies, this logically introduces a bias towards finding older hosts.

In an ideal world evidence for the alternative scenarios for insect-host plant evolution should therefore be sought strictly at the species level in recently-diverged insect clades that are highly specialized in their host use (Janz and Nylin 1998, Nyman, Linder *et al.* 2012), and within an absolute time-frame (Percy, Page *et al.* 2004, Sorenson, Balakrishnan *et al.* 2004, Janz 2011).

6.1.4 *Cymothoe* butterflies and their *Rinorea* host plants

Here, we focus on the trophic interaction between *Cymothoe* butterflies (Nymphalidae, Limenitidinae) and their *Rinorea* host plants (Violaceae) (Amiet and Achoundong 1996). *Cymothoe* comprise a clade of 78 butterflies confined to the forested regions of tropical Africa and Madagascar. Species are highly sexually dimorphic with males exhibiting sometimes spectacular coloration and some females being subject to mimicry. Recent phylogenetic estimations indicated that the *Cymothoe* clade exhibits an elevated rate of species diversification compared with its sister genus *Harma*, correlating with climatological oscillations and global cooling (van Velzen *et al.* 2013).

Rinorea is a pantropical genus of shrubs and small trees. Africa accommodates the largest number of species; (110 to 150 spp.; Achoundong 2000); Cameroon and Gabon being particularly species-rich (55 and 49 species, respectively; Achoundong 1996, Bakker, van Gemerden *et al.* 2006, Sosef, Wieringa *et al.* 2006). African *Rinorea* are often abundant or even dominant in the understory of humid or semi-deciduous forests (Achoundong 1996, Kenfack, Thomas *et al.* 2007, Chuyong, Kenfack *et al.* 2011), possibly constituting a reliable resource for herbivorous insects. Phylogenetic analyses have shown that *Rinorea* is a relatively early diverging lineage within Violaceae (Tokuoka 2008, Wurdack and Davis 2009) and that African *Rinorea* (with the inclusion of some closely related Malagasy taxa) are monophyletic (see Chapter 5).

Based on a decade of field observations of *Cymothoe* female oviposition behaviour in Cameroon, Amiet & Achoundong (1996) found 13 species of *Cymothoe* feeding on various Achariaceae and 27 feeding on *Rinorea.* The latter showed a very high degree of trophic specialization: 18 species are monophagous, the other 9 stenophagous; feeding on 2 to 6 species of *Rinorea.* This species-level specificity was confirmed by larval choice assays showing that, as soon as they hatch, *Cymothoe* larvae are able to recognize their *Rinorea* host plant (Amiet and Achoundong 1996).

The *Cymothoe-Rinorea* system is highly suitable to investigate processes shaping evolutionary patterns of host plant associations and species diversity, because of the following reasons. First, if we would assume that *Cymothoe* have remained associated with the same *Rinorea* species over macroevolutionary time-scales, it seems plausible that they would diverge in concert (as observed in the high level of specificity typical for specialized ectoparasites that usually show a high degree of cospeciation (e.g. Hafner and Nadler 1988, Hughes, Kennedy *et al.* 2007). On the other hand, current monophagy obviously does not rule out a more dynamic history of associations. Second, the high number of related species involved (34 herbivores and 33 hosts) makes it possible to reconstruct the evolution of many associations at the species-level, as well as to quantify host range in terms of associated species instead of arbitrary higher taxa. Third, time-calibrations are available for both clades involved. Recent time-calibrated phylogenetic estimations suggest that extant *Cymothoe* diverged around 7.5 Mya in the late Miocene (van Velzen, Wahlberg *et al.* 2013) while the genus *Rinorea* is roughly 57 million years old (see Chapter 5), suggesting that *Cymothoe* may have colonized pre-existing *Rinorea* lineages. However, a comparison of associated species has never been performed and hence it remains unclear if *Rinorea*-feeding clades within *Cymothoe* diversified in synchrony with their hosts, or rather colonize pre-existing clades within African *Rinorea.*

Finally, it is a tropical system. Abundance of herbivorous insects and their plant hosts is highest in the tropics (Novotny, Drozd *et al.* 2006), but studies on insect host plant associations in tropical systems are scarce (Novotny, Miller *et al.* 2010), rendering our case study a valuable addition to current understanding of how insect-plant interactions shape tropical biodiversity.

6.1.5 Objectives

The aim of this paper is to distinguish between the alternative scenarios for the evolution of insect-host plant associations in *Cymothoe* and *Rinorea* by answering the following research questions: 1. Did *Cymothoe* lineages diverge in synchrony with their hosts, or rather colonize pre-existing clades within African *Rinorea*? 2. Are patterns of *Cymothoe* – *Rinorea* associations non-random so that related butterflies feed on related plants? 3. To what extent did plant hosts and herbivores cospeciate and to what extent have *Cymothoe* herbivores shifted between hosts?

In order to address these questions we recorded additional host plant associations in the field in Nigeria, Ghana, Gabon and Kenya. To gain insight into the degree and nature of (a)synchronicity in butterfly and host-plant diversification, we generated time-calibrated species-level phylogenetic trees based on compiled sequence data from nuclear, mitochondrial and plastid DNA, and compared divergence time estimates for associated clades. To see if closely related parasites use closely related hosts we compared pairwise phylogenetic distances among herbivores with those among their hosts. In order to assess the relative contribution of cospeciation versus host shifts to butterfly speciation, we reconstructed event-based historical scenarios of associations using cophylogenetic methods.

6.2 Materials and methods

6.2.1 Sampling

Host-plant associations

Data on associations between *Cymothoe* butterflies and their *Rinorea* host plants were based mostly on Amiet & Achoundong (1996) and Amiet (1997, 2000); with some additions described in McBride *et al.* (2009). To assess the consistency of these described associations, as well as to expand geographic sampling, we collected *Cymothoe*-*Rinorea* host plant association data in Cameroon, Kenya, Ghana and Nigeria (see Figure 6.1 for collecting localities).

Cameroon (visited by RvV in 2006) was selected to confirm associations described by Amiet & Achoundong (1996), and to collect from different forested areas in Central province (Eloundem, Kala), and Littoral province (Edéa). Kakamega forest in Kenya (RvV, 2008) is the easternmost remnant of the Guineo-Congolean rainforest (Larsen 1991) and was visited mainly to collect association data on East-African endemic species *C. hobarti* and *C. butleri*. Ghana (REO, 2009) was selected as representative of West African countries because its butterfly diversity is relatively well-documented

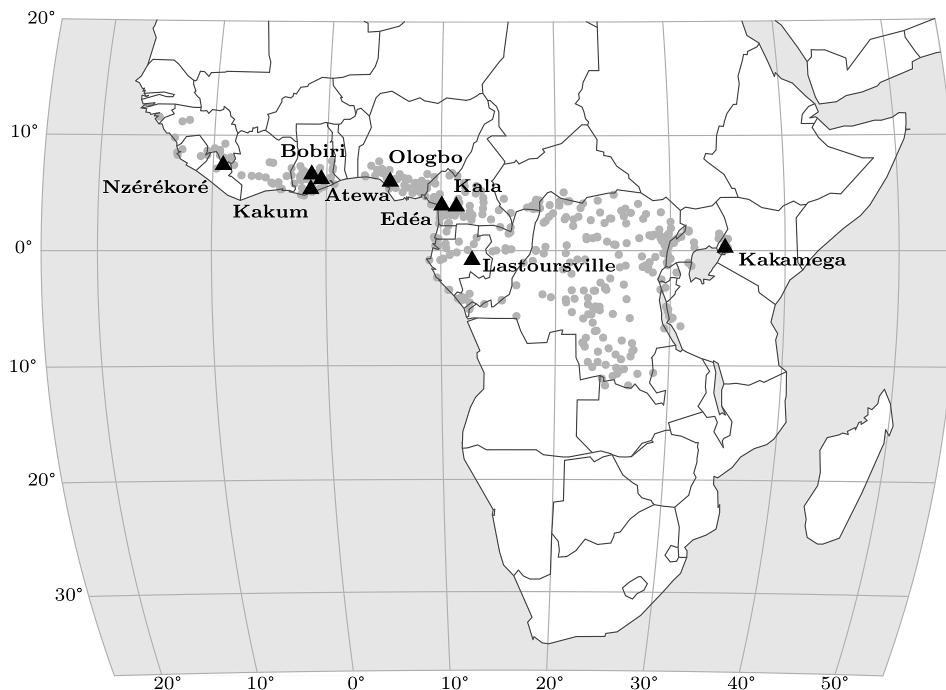


Figure 6.1: Geographical distribution of recorded *Cymothoe-Rinorea* host plant associations. Distribution of *Rinorea*-feeding *Cymothoe* species (grey dots) and localities where associations were observed in the field (black triangles). Distribution is based on specimens at the Royal Museum for Central Africa (Tervuren, Belgium), Natural History Museum (London, UK) and African Butterfly Research Centre (Nairobi, Kenya) as the research collection of RvV. Atewa includes Bunso Arboretum locality; Kala includes Eloundem.

(e.g. Larsen 2005, Bossart, Opuni-Frimpong *et al.* 2006). Associations were collected from different forested areas in the Eastern Region (Atewa Range forest reserve; Bunso Arboretum forest reserve), Ashanti Region (Bobiri butterfly sanctuary) and Central Region (Kakum National Park). Ologbo forest in Nigeria (RvV, 2010) was visited to collect association data on the local endemic species *C. okomu* and *C. nigeriensis* (Larsen 2005). In addition, single associations were observed and kindly shared by Jan Wieringa (Lastoursville Gabon, 2008) and Erik Koenen (Nzérékoré, Guinée, 2011). Observations consisted of monitoring female oviposition or larval feeding on *Rinorea* host plants; immature stages were identified using a DNA barcode library of >1000 sequences (van Velzen, Bakker *et al.* 2007, van Velzen, Larsen *et al.* 2009).

DNA sequence data

We aimed for complete species-level sampling of both *Cymothoe* and African *Rinorea*. For 59 *Cymothoe* taxa, we compiled DNA sequence data from cytochrome *c* oxidase subunit I (COI) from the mitochondrial genome, and *wingless* (*wgl*), ribosomal protein S5 (*RpS5*), glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*), and isocitrate dehydrogenase (*IDH*) from the nuclear genome (for more details see Chapter 4). Associated taxa for which phylogenetic sampling of multiple genes could not be achieved were represented by COI only. In order to reduce missing sequence data, we added COI sequences from conspecific individuals to four taxa: *C. coccinata* (RV_414+E028), *C. aubergeri* (CREO_110+116), *C. aramis* (OB_060+DK_001), and *C. mabillei* (CREO_100+101). We included *Harma theobene* for calibration purposes and *Neptis ida* as outgroup, based on Wahlberg *et al.* (2009). Please note that some species names included in our compilation (e.g. *C. superba*, *C. butleri*) were not corroborated in our DNA barcoding study (see Chapter 2), and that *C. baylissii* ined. is unpublished. For 73 *Rinorea* taxa (of which 56 African), we compiled DNA sequence data from the trnL (UAA) 5' exon–trnF (GAA) exon region from the plastid genome (*trnL–F*), nrDNA internally transcribed spacer (ITS), and exon 12 of the low-copy nuclear-encoded gene *EMB2765* from the nuclear genome (for more details see Chapter 5). We included *Fusispermum*, represented by concatenated sequences from *F. laxiflorum* (*trnL–F*) and *F. minutiflorum* (ITS), as outgroup, based on overall Violaceae phylogenetic studies by Tokuoka (2008), (Tokuoka 2008, Wurdack and Davis 2009) Our phylogenetic sampling included all currently documented associated species of *Cymothoe* and *Rinorea*, except two *Cymothoe* (*C. preussii* and *C. arcuata*), and two *Rinorea* (*R. subsessilis* and *R. keayi*) for which specimens were unavailable, amounting to a total of 31 *Rinorea*-feeding *Cymothoe* (94%) and 30 *Rinorea* hosts (91%), representing 54 of all 58 documented associations (93%). Information on voucher specimens and GenBank accession numbers is provided in Tables 4.1 and 5.1.

6.2.2 Phylogeny estimation and divergence timing

Data partitions and model testing

Data sets were partitioned per genomic region in *Cymothoe* and per marker for *Rinorea*. We determined the relative fit of candidate models of nucleotide evolution for each partition using JModelTest 0.1.1 (Posada 2008). Candidate models comprised three different substitution models (HKY, K80, and GTR) with or without estimated base frequencies, gamma-shaped distribution of rates (4 categories) and proportion of invariant sites – amounting to a total of 24 different models. Models were optimized on maximum likelihood trees and best-fitting models of nucleotide evolution were selected based on the corrected Akaike Information Criterion. Best-fitting candidate models per partition are given in Table 6.1.

Table 6.1: Character partitions, their characteristics and models selected for phylogenetic inference.

Partition	#accessions	#characters	#informative	Model
Cymothoe				
COI	61	1475	314	HKY + G
nDNA	54	2366	213	HKY + I + G
Rinorea				
trnL-F	71	1120	94	GTR + G
ITS	65	825	277	GTR + G
EMB2765	55	942	127	HKY + I + G

#accessions = number of accessions; #characters = total number of characters; #informative = number of informative characters.

Phylogenetic inference

To determine divergence dates, we used the Bayesian lognormal uncorrelated relaxed clock approach implemented in BEAST, version 1.7.2 (Drummond, Suchard *et al.* 2012). Because neither *Cymothoe* nor *Rinorea* fossils are known and hence unavailable for node calibration we used putative secondary time calibrations based on fossils from related clades. To calibrate the *Cymothoe* phylogenetic tree, we used the same secondary calibration as in van Velzen *et al.* (2013), based on six butterfly fossils (Wahlberg, Leneveu *et al.* 2009). This calibration comprised a lognormal prior with mean log 2.83 and stdev log 0.26 million years for the most recent common ancestor (mrca) of a monophyletic *Harma* and *Cymothoe*. To calibrate the *Rinorea* phylogenetic tree, we used the same secondary calibration as in Chapter 5, based on information from four fossils from the so-called 'Parietal' clade (which includes. Achariaceae, Goupiaceae, Violaceae, Passifloraceae s.l., Lacistemataceae and Salicaceae s.l.; Wurdack and Davis 2009). This calibration comprised a normal prior with mean 57 and a standard deviation of 9.7 million years for the mrca of *Rinorea* s.l. (i.e. including *R. crenata*). For details on these secondary calibrations see Chapters 4 and 5.

Both BEAST analyses applied the best-fitting models of nucleotide sequence evolution. In order to avoid overestimation of rates, ucld.mean received a diffuse gamma prior (shape 0.001, scale 1000), and the yule.birthRate parameter received an exponential prior with a mean corresponding to the average diversification rate. The average diversification rate was calculated as $(log(N/2))/T$, where N is the number of species and T is the calibrated age (Magallon and Sanderson 2001). Five independent Markov Chain Monte Carlo searches were run for 10 million generations each, sampling every 10 000-th generation. Searches achieved adequate mixing as assessed by large effective sample sizes of parameters (> 656) and convergence as assessed by repeatability of results over multiple independent searches. Results were

compared in terms of model parameters in Tracer 1.5 (Rambaut and Drummond 2009) and clade frequencies using the online service AWTY (Nylander, Wilgenbusch *et al.* 2008). After discarding the first million generations (10%) as burnin, results were pooled in logCombiner and maximum clade-credibility (MCC) trees were calculated using treeAnnotator (Drummond, Suchard *et al.* 2012). In order to avoid drawing conclusions based on clades with low posterior probability, nodes with a posterior probability below 0.90 were collapsed before all subsequent analyses.

Comparison of pairwise phylogenetic distances

To assess whether related *Cymothoe* feed on related *Rinorea*, we applied the Permutation test described by Hommola *et al.* (2009) to *Rinorea* and *Cymothoe* pairwise distance matrices in combination with our association data. The Permutation test is a modification of the Mantel test which requires one-to-one associations between parasites and hosts, which is usually not the case for most host parasite systems, including that of *Cymothoe* and *Rinorea.* In practice, therefore, certain host or parasite taxa can either be omitted or duplicated in order to achieve a one-to-one relationship before performing the Mantel test. Omitting data is obviously not desirable because it discards information, and taxon replication introduces bias. To alleviate this problem the Permutation test is applied to matrices of associations expressed as distances between hosts and parasites, respectively. As in the Mantel test, correlation between both matrices is measured with the Pearson's correlation coefficient (Hommola, Smith *et al.* 2009). To assess significance, a Null distribution of correlation coefficients is generated based on random permutations of the taxa; the associations and the original distance matrices are retained (this is consistent with the assumption of uncorrelated host and parasite phylogenies, because, if hosts and parasites have evolved independently, interaction between them does not depend on the position of the host and its parasite in their respective phylogenetic trees). The Permutation test thereby avoids simulation and hence assumptions about underlying processes. P-values are calculated simply as the percentage of permutations in which the correlation coefficient for permuted data is greater than or equal to the correlation coefficient based on the observed data (Hommola, Smith *et al.* 2009).

The Permutation test can be applied to patristic distances based on phylogenetic trees, or directly on sequence distances. Using sequence distances circumvents phylogenetic reconstruction, tree uncertainty and associated sampling effects (Hommola, Smith *et al.* 2009, Wilson, Forister *et al.* 2012). On the other hand, sequence distances are no proxy for shared ancestry (i.e. clades) because they do not distinguish between plesiomorphies and apomorphies. We therefore applied the test to sequence distances based on the Kimura 2-parameter model as well as to phylogenetic distances based on our calibrated phylogenetic MCC trees. Phylogenetic distances were calculated using the cophenetic function in the R package APE (Popescu, Huber *et al.* 2012). Null distributions were based on 10,000 permutations. We modified

the R source code distributed by Hommola *et al.* (2009) so that it accepts taxon labels and returns the calculated correlation coefficient in addition to the p-value; our code is given as supplementary data.

We also applied the more widely-used ParaFit test (Legendre, Desdevises *et al.* 2002) to the *Rinorea* and *Cymothoe* pairwise distance matrices in combination with the association data. This test transforms host and parasite distances derived from their phylogenetic trees or sequence alignments into matrices of principal coordinates. Dependence of the host and parasite distances is measured as the sum of squares of all values in matrix D, which is calculated as $C * t(A) * B$, where A is the matrix of associations, B is the matrix of principal coordinates computed from the distances between hosts, and C is the matrix of principal coordinates computed from the distances between parasites. To assess significance, a Null distribution of sums of squares is generated based on random permutations of the host-parasite associations (Legendre, Desdevises *et al.* 2002). This means that permuted association data are topologically not equivalent and may not reflect particularities of the observed associations, however. Results might therefore be affected by unusual features of the host–parasite graph rather than relate to relationships between host and parasite phylogenies (Hommola, Smith *et al.* 2009). In addition, the ParaFit test was found to have lower statistical power than the Permutation test (Hommola, Smith *et al.* 2009). We used the ParaFit function implemented in the R package APE (Popescu, Huber *et al.* 2012). The Null distribution was based on 10,000 permutations.

It is important to note that the Permutation and ParaFit tests were claimed to test host-parasite cospeciation and coevolution, respectively (Legendre, Desdevises *et al.* 2002, Hommola, Smith *et al.* 2009), even though they assess phylogenetic correlation only. The latter can be the result of various processes other than cospeciation or coevolution, including constrained host shifting or simple resource tracking (Menken 1996, Sorenson, Balakrishnan *et al.* 2004, De Vienne, Giraud *et al.* 2007). We therefore apply these tests exclusively in the sense of assessing phylogenetic correlation, i.e. whether related insects feed on related plants.

6.2.3 Cophylogeny estimation

In order to reconstruct historical scenarios of associations, *Cymothoe* and *Rinorea* phylogenetic trees were reconciled using cophylogenetic methods. Generally, this is done in a Parsimony framework, given a cost regime for the recoverable historical events. In our context, the recoverable historical events are cospeciation (parallel divergence of insect and host plant lineages), duplication (divergence of insects without host plant divergence), host shift (divergence of an insect onto an additional host plant lineage), and loss (absence of an insect on a host plant lineage where it would otherwise be expected). As the problem of reconstructing an optimal cophylogenetic history is computationally intractable (i.e. NP-complete; Ovadia, Fielder *et al.* 2011) heuristic search methods are applied. It is worth noting that

Table 6.2 List of currently known *Cymothoe-Rinorea* host plant associations.

Cymothoe	Rinorea	Source(s)
C. aramis	*R. dentata*	Amiet & Achoundong (1996)
C. arcuata	*R. subsessilis*	Amiet & Achoundong (1996)
C. aubergeri	*R. angustifolia engleriana*	REO pers. obs.
C. butleri	*R. brachypetala*	RVV pers. obs.
C. cf. *colmanti*	*R. batesii*	Amiet & Achoundong 1996 RVV pers. obs.
C. coccinata	*R. dentata*	Amiet & Achoundong 1996 RVV pers. obs.
C. coccinata	*R. yaundensis*	Amiet & Achoundong 1996 RVV pers. obs.
C. coccinata	*R. zenkeri*	Amiet & Achoundong (1996)
C. confusa	*R. simonae*	Amiet & Achoundong (1996)
C. confusa	*R. mezilii*	Amiet & Achoundong (1996)
C. confusa	*R. dimakoensis* ined.	Amiet & Achoundong (1996)
C. confusa	*R. letouzeyi*	Amiet & Achoundong (1996)
C. confusa	*R. dewitii*	Amiet & Achoundong 1996 RVV pers. obs.
C. confusa	*R. ilicifolia* CAM	Amiet & Achoundong 1996 McBride et al. (2009) RVV pers. obs.
C. crocea	*R. dentata*	Amiet & Achoundong (1996)
C. cyclades	*R. convallarioides occidentalis*	Amiet & Achoundong (1996)
C. distincta	*R. amietii* ined.	Amiet & Achoundong (1996)
C. distincta	*R. subsessilis*	Amiet & Achoundong (1996)
C. egesta	*R. breviracemosa*	McBride et al. (2009)
C. egesta	*R. ilicifolia* GH	REO pers. obs.
C. egesta	*R. lepidobotrys*	Amiet & Achoundong 1996 RVV pers. obs.
C. excelsa	*R. oblongifolia* CAM	Amiet & Achoundong 1996 RVV pers. obs.
C. fontainei	*R. caudata*	Amiet & Achoundong (1996)
C. fumana	*R. oblongifolia* CAM	Amiet & Achoundong 1996 RVV pers. obs.
C. fumana	*R. oblongifolia* GH	REO pers. obs.
C. fumana	*R. longisepala*	Amiet & Achoundong (1996)
C. fumana	*R. amietii ined.*	Amiet & Achoundong (1996)
C. fumana	*R. ledermannii*	Amiet & Achoundong (1996)
C. fumana	*R. subsessilis*	Amiet & Achoundong (1996)
C. haimodia	*R. longisepala*	Amiet & Achoundong (1996)
C. haimodia	*R. ledermannii*	Amiet & Achoundong (1996)

Guinea	Ghana	Nigeria	Cameroon	Gabon	Kenya
	Kakum (8)				
					Kakamega (2)
			Eloumden (1)		
		Ologbo (5)			
			Eloumden (1)		
			Edéa (5)		
			Eloumden (1)		
	Atewa (6) Kakum (22)				
			Kala (2)		
			Kala (1)		
			Kala (4)		
	Worobong (1)				

Table 6.2 List of currently known *Cymothoe-Rinorea* host plant associations, continued.

Cymothoe	Rinorea	Source(s)
C. harmilla	*R. angustifolia engleriana*	Amiet & Achoundong (1996)
C. heliada	*R. angustifolia engleriana*	Amiet & Achoundong (1996)
C. hesiodotus	*R. welwitschii*	Amiet & Achoundong (1996)
C. hobarti	*R. brachypetala*	RVV pers. obs.
C. hypatha	*R. welwitschii*	Amiet & Achoundong (1996)
C. hypatha	*R. rubrotincta*	Amiet & Achoundong (1996)
C. lucasi	*R. gabunensis*	Amiet & Achoundong 1996 Wieringa pers. obs.
C. lurida	*R. longicuspis*	Amiet & Achoundong (1996)
C. mabillei	*R. microdon*	Koenen pers. obs.
C. mabillei	*R. oblongifolia* GH	REO pers. obs.
C. nigeriensis	*R. rubrotincta*	RVV pers. obs.
C. ogova	*R. verrucosa*	Amiet & Achoundong (1996)
C. ogova	*R. dewildei* ined.	Amiet & Achoundong (1996)
C. okomu	*R. welwitschii*	RVV pers. obs.
C. orphnina	*R. keayi*	Amiet & Achoundong (1996)
C. preussii	*R. sinuata*	Amiet & Achoundong (1996)
C. preussii	*R. campoensis*	Amiet & Achoundong (1996)
C. reginae-elisabethae	*R. yaundensis*	Amiet & Achoundong (1996)
C. sangaris sp.1	*R. longicuspis*	REO pers. obs.
C. sangaris sp.2	*R. welwitschii*	Amiet & Achoundong (1996)
C. sangaris sp.2	*R. longicuspis*	Amiet & Achoundong (1996)
C. sangaris sp.2	*R. batesii*	Amiet & Achoundong 1996 RVV pers. obs.
C. sangaris sp.2	*R. rubrotincta*	Amiet & Achoundong (1996)
C. sangaris sp.3	*R. preussii*	Amiet & Achoundong 1996 RVV pers. obs.
C. sangaris sp.4	*R. welwitschii*	RVV pers. obs.
C. sangaris sp.4	*R. longicuspis*	REO pers. obs.
C. sangaris sp.4	*R. rubrotincta*	RVV pers. obs.

Guinea	Ghana	Nigeria	Cameroon	Gabon	Kenya
					Kakamega (1)
				Lastoursville (1)	
Nzérékoré (1)					
	Atewa (1) Bobiri (2) Bunso (4) Kakum (8) Worobong (1)				
		Ologbo (4)			
		Ologbo (1)			
	Atewa (5)				
			Eloundem (4)		
			Kala (1)		
		Ologbo (3)			
	Atewa (1)				
		Ologbo (1)			

Data based on literature and personal observations from six different countries; Bracketed numbers indicate number of independent observations in that particular forest. Note that *R. oblongifolia* and *R. ilicifolia* are not monophyletic species and that specimens from Cameroon (CAM) and Ghana (GH) represent distinct lineages (see Chapter 5)

cophylogenetic methods are based on the assumption that congruence between host and parasite phylogenies is the result of cospeciation (Charleston and Page 2009, Conow, Fielder *et al.* 2010). However, because preferential shifts to related hosts may also generate congruent trees, results must be interpreted with caution (De Vienne, Giraud *et al.* 2007).

Because we consider absences of associations between *Cymothoe* and *Rinorea* to be an artefact of sampling (i.e. we probably missed existing associations) rather than a true absence, reconstructions based on such species without recorded associations will thus overestimate the number of losses. Therefore, *Rinorea* species that are not known to be hosts of *Cymothoe* and *Cymothoe* species associated with Achariaceae or with unknown association were pruned from the phylogenetic trees.

Cophylogenetic histories of *Cymothoe* and associated *Rinorea* host plants were reconstructed using Jane, version 4 (Conow, Fielder *et al.* 2010). Jane uses a polynomial time dynamic programming algorithm in conjunction with a genetic algorithm to find optimal cophylogenetic scenarios. It allows for multi-host parasites and can calculate scenarios under a range of user-specified cost regimes for the historical events (Cruaud, Ronsted *et al.* 2012). We explored three arbitrary cost regimes, where cost for host shifts (C_{hs}) was always higher than that for cospeciation (C_{cs}): (i) $C_{cs} = 0$ and all other events = 1; (ii) $C_{cs} = 0$, $C_{hs} = 2$ and all other events = 1; and (iii) $C_{hs} = 2$ and all other events = 1. The genetic algorithm comprised 1000 generations and held a population of size 1000 and we let the algorithm return a maximum of 10,000 scenarios. To assess significance of the reconstructed scenarios, a null distribution of costs was generated based on 100 random permutations of the associations. We assessed support for specific host shifts by considering their frequencies in the different cost scenarios.

6.3 Results

6.3.1 Host plant observations

Our new *Cymothoe*-*Rinorea* host plant observations amount to a total of 23 associations of which 12 were new to science (see Table 6.2). Some associations were observed multiple times independently in different forests and countries. However, most are based on single observations and it is expected that extended sampling will reveal additional associations in the future. This implies that we can currently not assume any hard absences in *Cymothoe*-*Rinorea* association data; the common and widespread but seemingly uncolonized *R. subintegrifolia* being a possible exception (Amiet and Achoundong 1996). In any case, with a total of 58 associations between 33 *Cymothoe* butterflies and 34 *Rinorea* host plant species, 21 of which are monophagous (see Table 6.2), the data presented here constitute a significant improvement over previous biogeographical sampling (Amiet and Achoundong 1996,

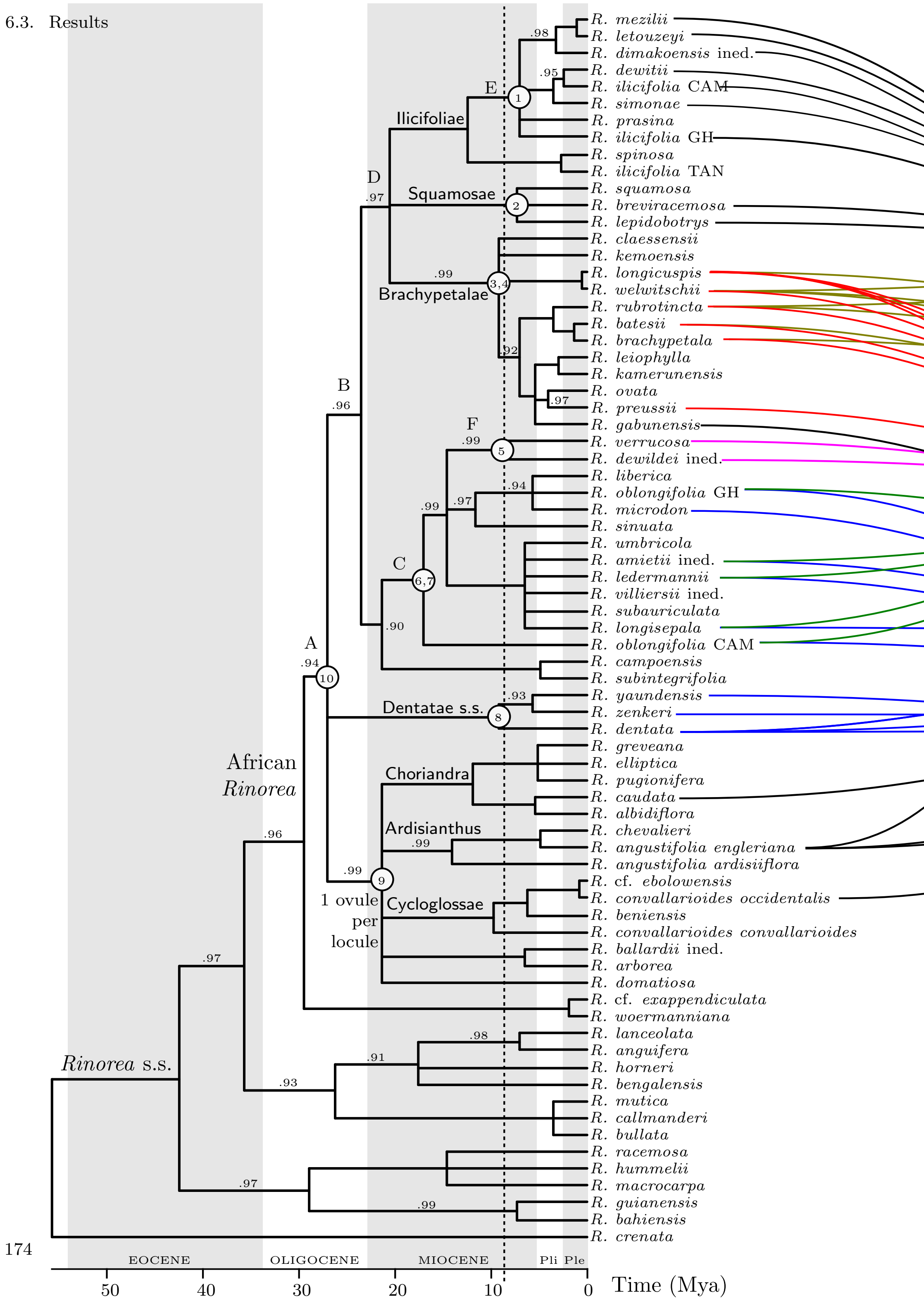
Amiet 1997, Amiet 2000, McBride, van Velzen *et al.* 2009) and represents most of the butterflies' geographical range.

6.3.2 Phylogeny estimation and distribution of associations

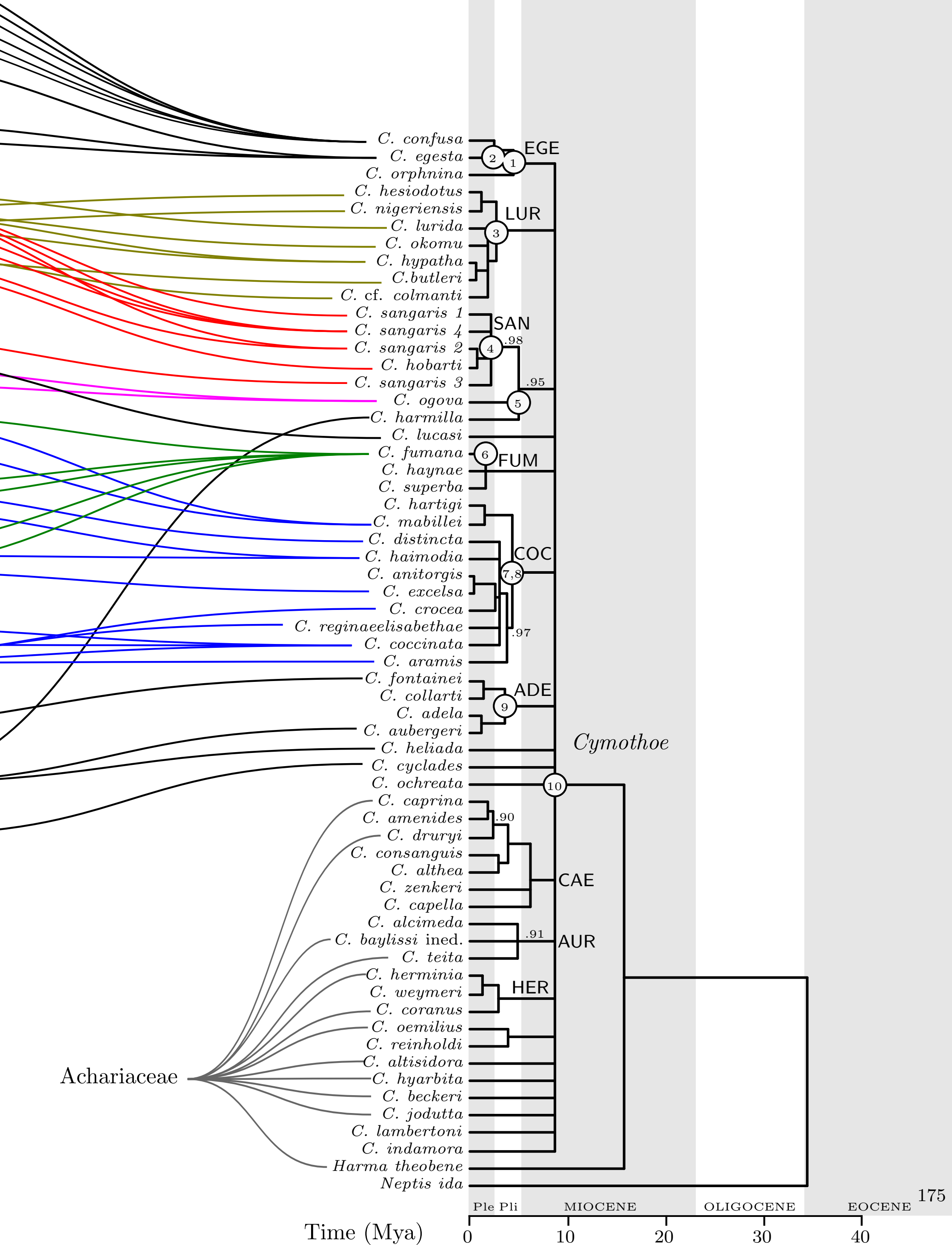
The *Rinorea* phylogenetic tree (Figure 6.2, left) had well supported clades and was congruent with previous estimations based on similar sequence data with higher population-level sampling (Chapter 5). We recovered African *Rinorea* (with the inclusion of some closely related Malagasy taxa) as monophyletic and African clades corresponding with formal or informal groupings based on morphology (Wahlert and Ballard 2012). *Cymothoe* host plants are distributed over most of the major African clades with multiple hosts occurring in *R.* subsect. *Ilicifoliae*, *R. squamosa* group, *R.* subsect. *Brachypetalae* s.l., *R.* clade C, *R.* subsect. *Dentatae, R.* subsect. *Choriandra*, *R.* sect. *Ardisianthus*, and *R.* sect. *Cycloglossae* (Wahlert and Ballard 2012; Chapter 5). The mrca of all *Cymothoe* hosts coincides with that of Clade A (i.e. all African *Rinorea* except *R. woermanniana* and *R.* cf. *exappendiculata*).

The *Cymothoe* phylogenetic tree (Figure 6.2, right) was congruent with previous estimations based on similar sequence data but with lower taxonomic sampling (van Velzen, Wahlberg *et al.* 2013). We recovered *Cymothoe* as monophyletic; comprising twenty unresolved lineages, some of which were single, morphologically distinct species. Clades within *Cymothoe* correspond with groupings based on morphological characters in adult and immature stages (Amiet 2000). Lineages containing species associated with *Rinorea* constituted roughly half of the phylogenetic diversity within the genus: clades EGE, LUR, SAN together with *C. ogova* and *C. harmilla*, FUM, COC, ADE and isolated species *C. heliada* and *C. cyclades.* Given morphological similarities between *C. cyclades* and *C. ochreata* we expect that the latter is also associated with *Rinorea.* We also note that *C. haynae*, member of the FUM clade is known to be associated with *Rinorea* in DRC, but the host species was unidentified (Fontaine 1982).

Figure 6.2: Time-calibrated tanglegram (next pages). Comparison of phylogenetic trees for *Rinorea* (left) and *Cymothoe* (right), with documented associations between *Rinorea* hosts and *Cymothoe* parasites indicated. Trees are drawn to the same scale with branch lengths proportional to divergence times (Mya = million years ago) as estimated in separate calibrated BEAST analyses. Ten associated clades corresponding to data points in Figure 6.3 are numbered. The vertical dashed line indicates the putative time of *Rinorea* colonization by *Cymothoe.* Nodes in both trees were rotated to maximize the appearance of congruence between the two trees.



R. mezilii
R. letouzeyi
R. dimakoensis ined.
R. dewitii
R. ilicifolia CAM
R. simonae
R. prasina
R. ilicifolia GH
R. spinosa
R. ilicifolia TAN
R. squamosa
R. breviracemosa
R. lepidobotrys
R. claessensii
R. kemoensis
R. longicuspis
R. welwitschii
R. rubrotincta
R. batesii
R. brachypetala
R. leiophylla
R. kamerunensis
R. ovata
R. preussii
R. gabunensis
R. verrucosa
R. dewildei ined.
R. liberica
R. oblongifolia GH
R. microdon
R. sinuata
R. umbricola
R. amietii ined.
R. ledermannii
R. villiersii ined.
R. subauriculata
R. longisepala
R. oblongifolia CAM
R. campoensis
R. subintegrifolia
R. yaundensis
R. zenkeri
R. dentata
R. greveana
R. elliptica
R. pugionifera
R. caudata
R. albidiflora
R. chevalieri
R. angustifolia engleriana
R. angustifolia ardisiiflora
R. cf. ebolowensis
R. convallarioides occidentalis
R. beniensis
R. convallarioides convallarioides
R. ballardii ined.
R. arborea
R. domatiosa
R. cf. exappendiculata
R. woermanniana
R. lanceolata
R. anguifera
R. horneri
R. bengalensis
R. mutica
R. callmanderi
R. bullata
R. racemosa
R. hummelii
R. macrocarpa
R. guianensis
R. bahiensis
R. crenata
Ilicifoliae
Squamosae
Brachypetalae
Dentatae s.s.
Choriandra
Ardisianthus
Cycloglossae
1 ovule per locule
African Rinorea
Rinorea s.s.
A
B
C
D
E
F
EOCENE
OLIGOCENE
MIOCENE
Pli
Ple
50
40
30
20
10
0
Time (Mya)



C. confusa
C. egesta
C. orphnina
C. hesiodotus
C. nigeriensis
C. lurida
C. okomu
C. hypatha
C.butleri
C. cf. colmanti
C. sangaris 1
C. sangaris 4
C. sangaris 2
C. hobarti
C. sangaris 3
C. ogova
C. harmilla
C. lucasi
C. fumana
C. haynae
C. superba
C. hartigi
C. mabillei
C. distincta
C. haimodia
C. anitorgis
C. excelsa
C. crocea
C. reginaeelisabethae
C. coccinata
C. aramis
C. fontainei
C. collarti
C. adela
C. aubergeri
C. heliada
C. cyclades
C. ochreata
C. caprina
C. amenides
C. druryi
C. consanguis
C. althea
C. zenkeri
C. capella
C. alcimeda
C. baylissi ined.
C. teita
C. herminia
C. weymeri
C. coranus
C. oemilius
C. reinholdi
C. altisidora
C. hyarbita
C. beckeri
C. jodutta
C. lambertoni
C. indamora
Harma theobene
Neptis ida
Achariaceae
EGE
LUR
SAN
FUM
COC
ADE
CAE
AUR
HER
Cymothoe
.98
.95
.97
.90
.91
1
2
3
4
5
6
7,8
9
10
Ple Pli
MIOCENE
OLIGOCENE
EOCENE
Time (Mya)
0
10
20
30
40

Visual inspection of the tanglegram plotting *Cymothoe-Rinorea* associations onto the phylogenetic trees (Figure 6.2) reveals that related *Cymothoe* butterflies feed on related *Rinorea* host plants. The *Cymothoe* EGE clade is strictly associated with members of the closely related *R.* subsect. *Ilicifoliae* and the *R. squamosa* group. *Cymothoe orphnina* is associated with *R. keayi* (unsampled) which also belongs to *R.* subsect. *Ilicifoliae*. The *Cymothoe* LUR and SAN clades are each associated exclusively with members of *R.* subsect. *Brachypetalae* s.l. The *Cymothoe* COC clade is associated with members of *R.* subsect. *Dentatae* and clade C. *Cymothoe fontainei* and *C. aubergeri* of the ADE clade are associated with *Rinorea* species sharing a single ovule per locule as a synapomorphy.

6.3.3 Comparative dating

Based on our time-calibrated analyses, divergence times of *Cymothoe* herbivores are substantially more recent than those of their *Rinorea* hosts, see Figures 6.2 and 6.3, and time estimates of selected associated *Cymothoe* and *Rinorea* nodes are not correlated (Pearson's correlation coefficient 0.544, $p = 0.10$). We estimated the origin of all *Cymothoe* (node 10) at 4.6–14.1 Mya; significantly younger than that of their associated *Rinorea* hosts (Clade A; 16.3–38.9 Mya). Likewise, *Cymothoe* clade ADE (node 9; 1.7–6.0 Mya) is much younger than the associated *Rinorea* clade uniting all African species with a single ovule per locule (12.6–31.5 Mya), as are *Cymothoe* clades SAN (node 4; 0.9–3.8 Mya) and LUR (node 3; 1.3–4.8 Mya) compared with the associated *R.* subsect. *Brachypetalae* (4.8–14.2 Mya), and *Cymothoe* clade COC (node 7; 2.1–7.2 Mya) compared with the associated *Rinorea* clade C (9.7–25.4 Mya), see left panel of Figure 6.3.

Mean divergence times of the remaining associated clades show the same overall pattern of young herbivores on older hosts, but we cannot exclude the possibility that their actual divergence times were synchronous, because their 95% HPD intervals overlap. This is the case for *Cymothoe* clade EGE (node 1; 2.0–7.7 Mya) and associated *Rinorea* clade E (3.5–11.7 Mya), for *Cymothoe* clade COC (node 8; 2.2–7.2 Mya) and associated *R.* subsect. *Dentatae* s.s. (3.9–16.3 Mya) as well as for *C. ogova* (node 5; 2.4–8.2 Mya) and associated *Rinorea* clade F (3.9–15.5 Mya) and for *C. egesta* (node 2; 1.0–4.5 Mya) and the associated *R. squamosa* group (2.3–11.4 Mya), see right panel of Figure 6.3.

6.3.4 Comparison of pairwise phylogenetic distances

In agreement with a pattern of dependence of host and parasite phylogenetic trees, phylogenetic (i.e. patristic) distances among *Cymothoe* and *Rinorea* were significantly correlated according to the Permutation test ($r = 0.47$, $p < 0.0001$). Raw sequence distances among *Cymothoe* and *Rinorea* were also significantly correlated, although less strongly so ($r = 0.25$, $p = 0.0007$), suggesting that reconstructed clades provide

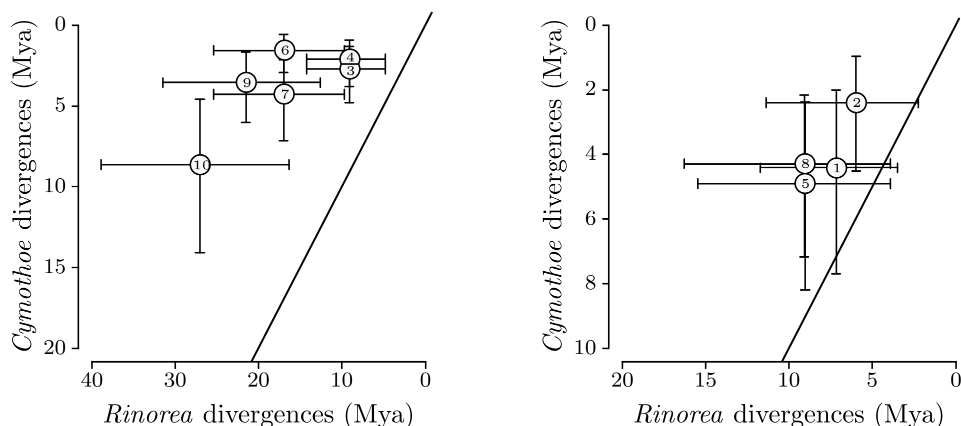


Figure 6.3: Asysnchronous divergences of *Cymothoe* butterflies and their *Rinorea* host plants. Comparison of estimated ages of corresponding nodes in the phylogenies of *Cymothoe* butterflies and their *Rinorea* host plants highlighted in Figure 6.2. Error bars show 95% highest posterior density (HPD) intervals; the diagonal line indicates equal times. Mean divergence times of *Cymothoe* herbivores are substantially more recent than those of their *Rinorea* hosts; 95% HPD intervals are significantly different for 6 selected clades (left panel) and overlapping for 4 lineages (right panel). Mya = million years ago.

a better proxy for *Cymothoe*-*Rinorea* associations than raw sequence distances alone. The ParaFit test did not reject the Null hypothesis that the two phylogenetic trees and the set of host-parasite association links are independent ($p = 0.433$), confirming previous claims that its statistical power is much less compared with the Permutation test (Hommola, Smith *et al.* 2009). In any case, our data support the hypothesis that pairs of closely related *Rinorea* are more often host for pairs of closely related *Cymothoe* herbivores than expected by chance.

6.3.5 Cophylogenetic analyses

As could be expected from the results above, recontructions of cophylogenetic histories indicate only few cospeciation events. According to the numerous optimal scenarios based on three different cost regimes, Jane consistently recovered more host shifts than cospeciation events, even though host shifts were given higher costs, see Table 6.3. Nevertheless, total cost of the reconstructed scenarios was always lower than those based on random permutations of associations, suggesting that the results are significant ($p < 0.01$). Only four specific host shifts received high support under all three cost regimes: *C. excelsa* to *R. oblongifolia* from Cameroon, *C. mabillei* to *R. oblongifolia* from Ghana and to *R. microdon*, and both *C. hobarti* and *C. butleri* to *R. brachypetala*, see Table 6.4.

Table 6.3: Cophylogenetic scenarios of *Cymothoe* and associated *Rinorea* host plants.

Cost regime	Cospeciations	Duplications	Host shifts	Losses	Failures to diverge	Total cost	Significance
01111	11–12	1–4	13–17	7–8	22	47	P<0.01
01211	13–14	4–5	10–12	9–12	22	59	P<0.01
11211	11–12	4–6	12–13	8–9	22	72	P<0.01

Table 6.4: Frequencies of well-supported host shifts.

Cost regime	*C. hobarti*→ *R. brachypetala*	*C. excelsa*→ *R. oblongifolia* CAM	*C. mabillei*→ *R. oblongifolia* GH and *R. microdon*	*C. butleri*→ *R. brachypetala*
01111	1.00	0.85	0.96	1.00
01211	1.00	1.00	0.94	0.67
11211	1.00	1.00	0.96	1.00

6.4 Discussion

Time-calibrated species-level phylogenetic trees provide strong evidence for asynchronous diversification of *Cymothoe* herbivores onto an already diversified group of *Rinorea* plants, and hence strong support for the *sequential evolution scenario*. Whereas African *Rinorea* diversified some 30 million years ago (van Velzen, Wahlberg *et al.* 2013), the *Cymothoe* feeding on them originated much later, about 7.5 million years ago (van Velzen, Wahlberg *et al.* 2013). Our results show that, also at a finer scale, divergence times for the herbivores are substantially more recent than those for their hosts, see Figures 6.2 and 6.3. Such asynchronous diversification is in accordance with a growing body of evidence that diversification times of herbivorous insects generally postdate those of their hosts (Janz and Nylin 1998, Ronquist and Liljeblad 2001, Lopez-Vaamonde, Wikström *et al.* 2006, Gómez-Zurita, Hunt *et al.* 2007, Smith, Godsoe *et al.* 2008, McKenna, Sequeira *et al.* 2009). Despite recent methodological improvements (Funk and Omland 2003, Phil-Eze and Okoro 2009, Heled and Drummond 2012), divergence time estimation remains an imprecise endeavour. In addition, as the *Cymothoe* and *Rinorea* phylogenetic trees are calibrated independently, their estimated ages are not directly comparable (Sorenson, Balakrishnan *et al.* 2004). Nevertheless, as we have given relatively wide confidence intervals to the calibration priors in both analyses, as well as used the 95% highest posterior density intervals when comparing associated clades, our conclusions are

conservative. Irrespective of absolute timing, relative divergence times of associated clades were uncorrelated ($p = 0.10$), also refuting a hypothesis of synchronicity.

Phylogenetic trees of *Cymothoe* and their associated *Rinorea* host plants are not topologically congruent. Cophylogenetic reconstructions indicate that current associations between *Cymothoe* herbivores and their *Rinorea* hosts have developed primarily through a process of host shifting rather than by cospeciation, refuting a scenario of long-term shared phylogenetic history (Farrell and Mitter 1998, Farrell 2001, Becerra 2003). Given the asynchronicity of associated *Rinorea* and *Cymothoe* lineages this is hardly surprising, and our results thus support the currently widely recognized idea that host shifts are the most important processes structuring plant–insect associations in many clades (Jermy 1984, Janz and Nylin 1998, Janz, Nyblom *et al.* 2001, Agosta 2006, Winkler and Mitter 2008). Given that preferential shifts to related hosts may also result in congruent trees (De Vienne, Giraud *et al.* 2007), rate estimates of host shifts based on cophylogeny estimations are probably underestimated (Janz 2011) and we assume that this is also the case here. In addition, ancestral host associations can be retained for long periods of time as part of the repertoire of herbivorous insects, allowing shifts to go back and forth between related host species (Janz, Nyblom *et al.* 2001). In any case, most reconstructed cospeciation events involve *Cymothoe* and *Rinorea* lineages that did not diverge in synchrony, rendering a cospeciation scenario untenable.

Nevertheless, comparison of pairwise phylogenetic distances among *Cymothoe* with those among *Rinorea* show that related *Cymothoe* feed on related *Rinorea* hosts. Indeed, while the distribution of current *Cymothoe* host-plant associations appears opportunistic at smaller scales, it is surprisingly conservative at larger scales (see Figure 6.2). Again, this pattern is in accordance with other studies comparing insect and host plant phylogenetic trees (Mitter, Farrell *et al.* 1991, Janz and Nylin 1998, Ronquist and Liljeblad 2001, Braby and Trueman 2006, Lopez-Vaamonde, Wikström *et al.* 2006, Wilson, Forister *et al.* 2012) and can best be explained by a process where host shifts most often involve related hosts and where shifts between major host clades have been infrequent (Janz, Nyblom *et al.* 2001, Nyman 2010). Such limited colonization is a reasonable expectation for herbivores as plants from the same clade may share characteristics that influence herbivore fitness (Agosta 2006), such as secondary phytochemistry (Ehrlich and Raven 1964, Menken 1996, Becerra 1997, Agrawal 2007). In *Cymothoe*, successful host colonization may therefore be influenced by the degree of mismatch between larval physiology in combination with female oviposition cues and the chemical makeup of a novel *Rinorea* host.

Species not feeding on *Rinorea* use Achariaceae as host plants (see Figure 6.2). Achariaceae are known to be cyanogenic, and *Cymothoe* host plants *Rawsonia lucida* and *Lindackeria dentata* (Achoundong 1996) contain the cyanogenic precursor cyclopentenyl glycine (Andersen, Clausen *et al.* 2001, Jaroszewski, Ekpe *et al.* 2004). Although *Rinorea* is non-cyanogenic, a novel cyclopentenyl glycine was encountered in *Rinorea ilicifolia*, suggesting that these compounds act as a common chemical

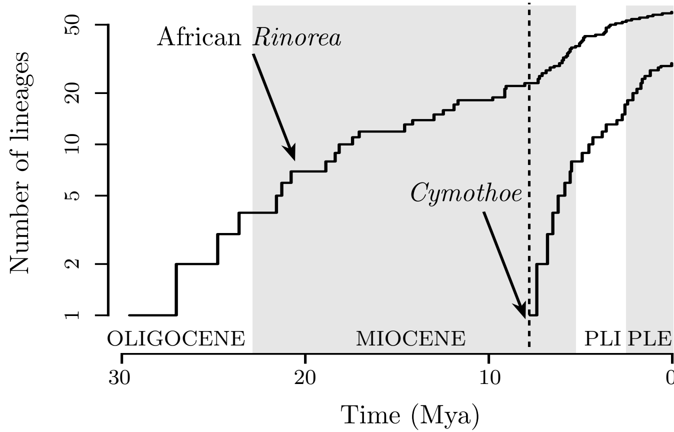


Figure 6.4: Lineages of *Cymothoe* and *Rinorea* through time. Lineage-through-time plots for *Rinorea*-feeding *Cymothoe* (lower line) and African *Rinorea* (upper line). Number of lineages are shown on logarithmic scale. The vertical dashed line indicates the putative time of *Rinorea* colonization by *Cymothoe*. Note that even though *Rinorea*-feeding *Cymothoe* are unresolved in our phylogenetic reconstructions, we hypothesize that *Rinorea* were colonized only once, see discussion. Mya = million years ago.

recognition template for ovipositing *Cymothoe* females (Clausen, Frydenvang *et al.* 2002). Alternatively, hosts shifts may be influenced by insect carnivores. Butterfly eggs and larvae are prone to predation by various other insect clades such as parasitoid wasps that use chemical cues from the host plant to locate their prey (Dicke 2000, Lill, Marquis *et al.* 2002). Shifting to novel host plants outside their typical host range can therefore provide an 'enemy free space' (Bernays and Graham 1988, Ishihara and Ohgushi 2008, Diamond and Kingsolver 2010). *Cymothoe* females generally lay single eggs on suitable *Rinorea* hosts (Amiet and Achoundong 1996), possibly to minimize parasitation risk for their offspring, suggesting that predation has sufficient impact to influence their behaviour. It appears reasonable to assume that when parasites locate *Cymothoe* eggs or larvae by specific host plant cues, shifting to another host could significantly increase fitness of *Cymothoe* individuals. Whether such selection is consistent enough through time and space to have an influence at the species level remains controversial, however (Heard, Stireman *et al.* 2006).

Lineage-through-time (LTT) plots of *Cymothoe* and African *Rinorea* based on our phylogenetic trees are shown in Figure 6.4. Both plots show a steady increase in lineages indicating that *Cymothoe* diversification is correlated with that of its hosts. Perhaps surprisingly, diversification of extant African *Rinorea* seems to be slightly increased since the time of their colonization by *Cymothoe*. Instead of attributing this pattern to reciprocal effects between *Rinorea* and *Cymothoe* diversification, we consider global cooling and climatic oscillations since the late Miocene (Zachos,

Pagani *et al.* 2001) as a likely explanation. It is important to note, however, that whereas we achieved near-complete species sampling of *Cymothoe*, taxonomic sampling of African *Rinorea* was limited to roughly 50% of all extant taxa. This means that our associated lineage-through-time plot may be biased (Cusimano and Renner 2010) and can therefore only act as a preliminary proxy for patterns of *Rinorea* diversification. In any case, there is no support for reduced diversification in *Rinorea* after colonization as would be predicted under the *escape and radiation scenario* (Ehrlich and Raven 1964).

6.4.1 An evolutionary scenario for *Cymothoe* host plant associations

Based on current data, can we generate a scenario for the evolutionary history of associations between *Cymothoe* and *Rinorea*? Major lineages within *Cymothoe* originated within a very short timeframe, around 7.5 Mya (van Velzen, Wahlberg *et al.* 2013). Based on the data presented here, it is clear that roughly half of these lineages are specialized on particular *Rinorea* clades. These lineages are unresolved (Figure 6.2), probably due to incomplete lineage sorting effects within a rapidly diversifying *Cymothoe* clade (van Velzen, Wahlberg *et al.* 2013) and we therefore cannot confirm whether *Rinorea*-feeding *Cymothoe* are monophyletic or not.

Nevertheless, we consider such monophyly likely for the following reasons. First, it is parsimonious to assume a single shift from ancestral Achariaceae to *Rinorea* rather than parallel independent shifts by multiple species. Second, theory predicts that shifts between unrelated host plants are less common than those between related ones. However, it is important to note that 'relatedness' in this case is relative to the level of specialism of the herbivore as hosts can be similar from the perspective of a generalist but at the same time distant from that of a specialist (Nyman 2010). By consequence, insect lineages with narrow tolerances are likely to speciate by shifting between host within young, species-rich plant groups, whereas more generalist lineages could speciate by shifting among higher plant taxa (Nyman 2010). Violaceae and Achariaceae are related families within Malpighiales as they are both members of a clade of largely unresolved lineages sharing parietal placentation (Wurdack and Davis 2009). Nevertheless, these families diverged roughly 70 My before *Rinorea* was colonized (Bell, Soltis *et al.* 2010). Given this divergence, in combination with the family-level specialism of ancestral Achariaceae-feeders, we assume that the shift from Achariaceae to *Rinorea* constitutes a rare event. Finally, we have found no evidence for reverse colonization of Achariaceae by *Rinorea*-feeding clades, further corroborating our hypothesis of a single host shift by one ancestral *Cymothoe*.

The dashed vertical line in Figures 6.2 and 6.4 corresponds with the putative time of the *Cymothoe* colonization event, 7.5 Mya. At that time, all major *Rinorea* host plant lineages were already present. Nevertheless, when ignoring extinction,

African *Rinorea* comprised only about 20 species by then, 11 of which are ancestors of current host plants, see Figure 6.2 and 6.4. This suggests that African *Rinorea* constituted a relatively homogeneous resource when they were colonized. Given this historical setting, we hypothesize that *Cymothoe* colonized multiple sympatric species of *Rinorea.* Subsequent host shifts were possibly mediated by changing geographical distributions of the butterflies and/or the ecologically fitting *Rinorea* hosts, in accordance with results found for other herbivorous insects (Agosta 2006, Brändle, Kühn *et al.* 2008, Jahner, Bonilla *et al.* 2011, Slove and Janz 2011, Nyman, Linder *et al.* 2012, Wilson, Forister *et al.* 2012).

Based on our estimations it is clear that current patterns of associations are the result of host shifts between related plants within a gradually diversifying African *Rinorea.* Reconstructing specific shifts from one particular host to another is difficult, however. This is because of methodological problems, as under parsimony reconstructions, opportunistic association patterns can be explained by a vast number of equally likely host shift histories. But, it is also inherent to the fact that under a scenario of recurrent host shifts history tends to erase its own tracks (Nyman 2010, Janz 2011).

Nevertheless, four well-supported shifts could be identified. First, current associations of the only two East-African endemics *C. butleri* and *C. hobarti* with *R. brachypetala* are explained by assuming two parallel host shifts in 96% of Jane reconstructions (see Table 6.4). As *R. brachypetala* is the only *Rinorea* species occurring in the remnants of Guineo-Congolean rainforest in Kenya (Fischer, Rembold *et al.* 2010) these shifts may have facilitated the dispersal of *Cymothoe* species belonging to two separate clades (LUR and SAN) from Central to East Africa. Both *C. hobarti* (0.7 Mya) and *C. butleri* (0.6 Mya) diverged only recently from their respective sisters. This means that they possibly have retained much of the patterns caused by the processes underlying the shift. This makes them ideal candidates for studying genetic, biogeographic, ecological, physiological, and behavioural factors influencing host shifts and dispersal in tropical butterflies. Likewise, current association of West African endemic *C. mabillei* with West African *R. oblongifolia* and *R. microdon* is explained by a host shift in 93% of Jane reconstructions, again possibly facilitating its dispersal.

Finally, given that *Rinorea* subsect. *Ilicifoliae* and the *R. squamosa* group had already diverged long before *Cymothoe* clade EGE, associations between *C. egesta* and its *R. lepidobotrys* and *R. breviracemosa* host plants in Cameroon can only be explained by assuming a host shift from ancestral *R.* subsect. *Ilicifoliae* hosts, see Figure 6.2. Because *C. egesta* is also feeding on *R. ilicifolia* in Ghana, this shift probably represents local host-range expansion. Given the physiological and presumably phytochemical disparity between *R ilicifolia* and the derived *R. breviracemosa* and *R. lepidobotrys* hosts it can be expected that they exert differential selective pressures on *Cymothoe* populations. As such, *C. egesta* may represent a case of host race formation and, perhaps, incipient speciation (Menken 1981, Drès

and Mallet 2002, Nosil, Crespi *et al.* 2002, Singer and McBride 2010, Borer, van Noort *et al.* 2011). Larval differentiation of the populations feeding on *R. squamosa* group (see Figure 2.16) appear to substantiate such a hypothesis. Unravelling processes underlying host shifts requires comprehensive population-level sampling of *Cymothoe* and *Rinorea* hosts, characterization of host phytochemistry, and accurate biogeographic reconstructions, in combination with behavioural, physiological and genetic experiments.

Available host association data suggest that, for *Rinorea*-feeding *Cymothoe*, current host ranges vary from 1–6 closely related *Rinorea* species, with the majority of associations being monophagous. Host ranges expanded from mono- to stenophagy multiple times independently, as stenophagous species occur in multiple clades within *Cymothoe*. In analogy to problems associated with identifying host shifts, ancestral host ranges are inherently difficult to reconstruct (Janz, Nylin *et al.* 2006). It seems clear, however, that host range in *Cymothoe* is dynamic with alternations between range expansions and retractions at recent time-scales.

There seems to be a correlation between the size of the distribution area of *Cymothoe* species and their host range, in accordance with patterns found in other Lepidoptera (e.g. Jahner, Bonilla *et al.* 2011). For example, within the COC clade *C. coccinata* (3 hosts within *R.* subsect. *Dentatae* s.s.) is the most widespread species as well as the one with the highest number of hosts. Similarly, *C. confusa* (6 hosts within *R.* subsect. *Ilicifoliae*) and *C. fumana* (6 hosts within *Rinorea* clade C) have the broadest host ranges within *Rinorea*-feeding *Cymothoe* and are amongst the geographically most widespread species. This observation could very well be a sampling effect, however, as common species with large distribution areas are encountered more often in the field, and so íf they are stenophagous the have a higher chance that this is recorded. On the other hand a species might well be common and widespread *because* it is associated with multiple hosts. Sampling effects could be alleviated by specifically targeting rarer species (both of *Rinorea* and *Cymothoe*) during field work, although probably never be completely excluded.

Compared with other tropical butterfly-hostplant associations, the *Cymothoe-Rinorea* system represents a special case in terms of the large number of congeneric species involved in both herbivores and host plants (Amiet and Achoundong 1996). At the same time, phylogenetic estimations have indicated that *Cymothoe* experienced increased rates of diversification, a pattern that we associated with climatic change earlier (see Chapter 4). The question could therefore be asked whether the two patterns might be correlated, i.e. if climatic change and resulting habitat fragmentation could have led to increased incidence of shifts between *Rinorea* hosts. Answering this question requires better understanding of the general principles that govern the evolution of insect-host plant associations.

6.5 Conclusion

Based on the available data, we propose a recent origin of *Rinorea*-feeding by *Cymothoe* butterflies with a single colonization of pre-existing lineages in the late Miocene. Even though there is a significant trend of related *Cymothoe* feeding on related *Rinorea* hosts, insect and host phylogenetic trees are discordant, and their reconciliation requires assuming many host shifts, some of which may have facilitated the dispersal of *Cymothoe* lineages. Current host-plant associations within *Cymothoe* have evidently been produced by a combination of host conservatism and shifts among closely related *Rinorea*, probably mediated through constraints in larval physiology and female oviposition behaviour. Host range in *Rinorea*-feeding *Cymothoe* is dynamic, with specialization as well as host range expansions at recent time-scales. These findings are in close agreement with the scenario of sequential colonization as a dominant pattern in insect-plant interactions.

Acknowledgements

We are indebted to Gaston Achoundong, Oskar Brattström, Steve Collins, Stephen Davey, Andrews Kankam Amankwah, Gerco Niezing, Andrew Walker, Olivier Sene-Belinga, and Safián Szabolcsz for their help in organising and performing fieldwork. We thank Tinde van Andel, Lars Chatrou, Carel Jongkind, Erik Koenen, Quentin Luke, Greg Wahlert and Jan Wieringa for contributing *Rinorea* silica samples. Julian Bayliss, Dries Bonte, Thierry Bouyer, Oskar Brattström, Steve Collins, Torben Larsen, Freerk Molleman, Robert Warren, Gael vandeWeghe, and Steve Woodhall kindly donated *Cymothoe* samples. Jan Wierina and Erik Koenen collected additional *Cymothoe*-*Rinorea* host plant associations in Gabon and Guinea, respectively. Marleen Botermans provided valuable comments on an early draft of the manuscript. The Alberta Mennega Foundation, Hugo de Vries Foundation (RvV and REO), and the Systematics Research Fund (awarded jointly by the Systematics Association and the Linnean Society to RvV) are acknowledged for financial support.

Chapter 7

General discussion

Abstract

This thesis aimed to elucidate the evolutionary history of the highly specialized associations between *Cymothoe* tropical forest butterflies and their *Rinorea* host plants. Using current methods for phylogenetic reconstruction and divergence time estimation, we assessed the role of host association and switching in *Cymothoe* species diversification; applying a statistical framework whenever possible. In this chapter I summarize and synthesize the answers to the research questions posed in Chapter 1 and provide a chronological overview of the reconstructed evolutionary events. Implications of our results to other scientific fields such as biological regulation and agriculture are discussed. Challenges and possibilities for future research are identified.

7.1 Summarizing results

Systematics of *Cymothoe*

Reliable classification and taxonomy of species is of the essence for any biological study. Generally, some species groups are notoriously problematic, however. The Afrotropical butterfly genus *Cymothoe* is such a group, as species are highly dimorphic sexually and associating males with females can be problematic because either males or females can be morphologically similar between species. Another issue is that, while species of *Cymothoe* are of special interest due to their highly specialized associations with *Rinorea* host plants, morphology-based identification of immature stages found on host plants in the field is problematic. We therefore presented an analysis of an extensive dataset of 1204 DNA barcode sequences and assessed whether it facilitates species delimitation and identification in *Cymothoe* (see Chapter 2).

We applied a novel practical decision pipeline for integrating DNA barcodes, morphology and biogeography within a taxonomic framework, which proved instrumental for solving taxonomic problems in *Cymothoe*. In addition, our DNA barcode data set allowed for the identification of 42 immature specimens from six different countries, significantly increasing the data on *Cymothoe* host plant associations. Nevertheless, we concluded that some sections within *Cymothoe* remain difficult to interpret taxonomically even with DNA barcodes, mainly because of lack of interspecific divergence and incomplete lineage sorting of DNA barcode haplotypes. These patterns are probably due to species being relatively young, i.e. they diverged only recently.

DNA barcoding of recently diverged species

Recently diverged species are frequently of special interest scientifically as well as from a regulatory perspective, but they are also frequently difficult to identify. DNA barcoding has proven instrumental in species identification, especially in insects and vertebrates, but for the identification of recently diverged species it has been reported to be problematic in some cases.

Therefore, in order to assess what are the best methods for matching DNA barcodes from recently diverged species, we compared six methods in their ability to correctly match DNA barcodes from selected published empirical data sets as well as simulated data (see Chapter 3). Our results showed that, even though recently diverged species pose a significant problem for effective DNA barcoding, sensitive similarity-based and diagnostic (non-tree based) methods can significantly improve identification performance compared with the commonly used tree-based methods such as neighbor joining.

Cymothoe species diversification

In order to to better understand the dynamics underlying species diversification in *Cymothoe* we set up a series of studies based on our best possible estimate of phylogenetic relationships in this clade. *Harma* and *Cymothoe* constitute sister genera of African specialist herbivores with a striking difference in species diversification: *Harma* being monospecific and *Cymothoe* comprising ~78 species. As *Cymothoe* has an expanded host range due to shift(s) from the ancestral Achariaceae to *Rinorea*, we asked whether its increased diversification was the result of these shifts or rather with –presumed- climate-driven habitat fragmentation (and allopatric speciation), i.e. independent of host plant shifts.

In order to answer this question we generated a species-level molecular phylogenetic tree of *Cymothoe* and *Harma*, and calibrated it within an absolute time frame. We then identified significant shifts in species diversification rates and assessed correlations of estimated diversification with reconstructed host plant associations and with trends in global temperature variation at geological scales (see Chapter 4). Based on these analyses we inferred a significant shift in species diversification rate of *Cymothoe* butterflies in the late Miocene. Compared with other butterfly clades, this shift was interpreted to be abrupt as well as recent.

Collectively, our reconstructions indicated that shift(s) to *Rinorea* had only marginal effect on diversification in the *Harma Cymothoe* clade. Rather, elevated net diversification rate of *Cymothoe* correlates with putative climate-driven habitat fragmentation since the late Miocene. We conclude that forest fragmentation caused by changing climate in the late Miocene, as well as Eastern Arc Mountain uplift are likely to have promoted species diversification in *Cymothoe*.

Systematics and historical biogeography of *Rinorea*

Rinorea (Violaceae) is a pantropical genus of understory trees. The majority of species are in Africa, some of which are host plants of *Cymothoe*. Although phylogenetic studies based on plastid data have collectively progressed our understanding of *Rinorea* systematics (Bakker, van Gemerden *et al.* 2006, Wahlert and Ballard 2012), several issues remain. First, taxonomic sampling was relatively low and most *Cymothoe* host plants were lacking in these studies. We therefore increased taxonomic sampling to include 50% of African *Rinorea* and nearly all known *Cymothoe* host plants. Secondly, the lack of a nuclear DNA based phylogenetic perspective precludes reconstruction of actual clades instead of plastid haplotypes. We therefore based our analyses on plastid as well as nuclear DNA sequences. Finally, the absolute time frame in which diversification in *Rinorea* has taken place has remained unknown so far. An absolute time frame is required when testing scenarios of *Cymothoe*-*Rinorea* host plant associations (see Chapter 6), but also for answering open questions about the historical biogeography of *Rinorea*. For example, there appears to be an early

disjunction between Neo- and Palaeotropical *Rinorea*. But without absolute time estimates it is unclear whether this disjunction should be explained by Gondwanan vicariance or by a more recent long-distance dispersal event (Queiroz, 2005). Also, within the Palaeotropical clade, at least 4 Malagasy taxa appear within separate clades from mainland Africa, suggesting independent dispersals from mainland Africa to Madagascar(Bakker, van Gemerden *et al.* 2006, Wahlert and Ballard 2012). It remains unknown, however if these putative independent dispersals from Africa to Madagascar were synchronous.

We therefore estimated *Rinorea* lineage divergences within an absolute time frame (chapter 4). Our evidence derived from nuclear DNA sequences was generally congruent with that from plastid haplotypes suggesting that the latter represent actual clades. Moreover, it helped resolving additional phylogenetic relationships, some of which warrant further taxonomic study. Based on divergence time estimates we concluded that Violaceae originated in the Neotropics long after the breakup of Gondwana and that *Rinorea* reached Africa in the Eocene through trans-Atlantic dispersal. In Africa, the genus proliferated since the Oligocene resulting in the phylogenetic and morphological diversity that is seen today. *Rinorea* subsequently dispersed from Africa to Asia in the Oligocene or early Miocene, probably due to closing of the Tethys Sea.

Malagasy *Rinorea* taxa were recovered as sister to African lineages in five different clades, confirming previous claims that *Rinorea* colonized Madagascar multiple times independently from the African mainland (Bakker, van Gemerden *et al.* 2006, Wahlert and Ballard 2012). These independent colonizations happened within a relatively recent time scale (Pliocene), suggesting that factors governing the independent colonizations of *Rinorea* to Madagascar may have been similar, or simultaneous.

Evolution of *Cymothoe*-*Rinorea* host plant associations

The insect-plant system of *Cymothoe* butterflies feeding on *Rinorea* host plants is highly suitable to investigate processes shaping evolutionary patterns of host plant associations, mainly because of both the high level of specialization and the large number of species involved. Assuming that *Cymothoe* have remained specialized on the same *Rinorea* species over macroevolutionary time-scales, it seems plausible that they would diverge in synchrony (i.e. cospeciate). On the other hand, current monophagy obviously does not rule out a more dynamic history of associations, with *Cymothoe* colonizing already existing *Rinorea* lineages. In order to distinguish between alternative scenarios for the evolution of insect-host plant associations we integrated the time-calibrated phylogenetic evidence from *Cymothoe* and *Rinorea* presented in Chapters 4 and 5 with updated host association records from the field.

To gain insight in the relative levels of cospeciation versus host shifts, we reconstructed event-based historical scenarios of associations using phylogenetic rec-

onciliation methods. We found only low levels of congruence between *Cymothoe* and *Rinorea* phylogenetic trees. Indeed, event-based cophylogenetic analyses suggested that current associations between *Cymothoe* herbivores and their *Rinorea* hosts have developed primarily through a process of host shifting rather than by cospeciation. In order to assess the degree of (a)synchronicity in butterfly and host-plant diversification, we compared divergence time estimates for associated clades and found strong support for asynchronous diversification of *Cymothoe* herbivores onto already diversified clades within African *Rinorea*. In shiort, whereas African *Rinorea* diversified some 30 million years ago, the *Cymothoe* feeding on them originated much later, about 7.5 million years ago. Also at a finer scale, divergence times for the herbivores are substantially more recent than those for their hosts, also supporting a scenario of sequential evolution rather than cospeciation.

Despite such asynchronicity and prevalence of host shifts, comparisons of pairwise phylogenetic distances among *Cymothoe* with those among *Rinorea* species showed that related *Cymothoe* do feed on related *Rinorea*. Indeed, while the distribution of current *Cymothoe* host-plant associations appears random at the species level, it is surprisingly conservative at 'deeper' phylogenetic scales. This suggests that, even though host shifts are common in *Cymothoe*, they occur predominantly between closely related *Rinorea* species. Possibly, this reflects relative conservation of phytochemical factors in the evolution of *Rinorea* clades.

7.2 A chronology of reconstructed historical events in *Cymothoe* and *Rinorea*

Absolute time frames were imperative for distinguishing among competing hypotheses and historical scenarios in this thesis. In this paragraph I provide a chronological overview of the evolutionary events that shaped current biogeographic distributions of and host plant associations between *Cymothoe* and *Rinorea*, within a paleogeographic context (see Figure 7.1).

The overview starts about 65 million years ago, when a massive comet or asteroid ∼10 km in diameter is considered to have collided with Earth a few miles from the present-day town of Chicxulub in Mexico (Alvarez, Alvarez *et al.* 1980, Hildebrand, Penfield *et al.* 1991). The proposed impact had catastrophic effects on the global environment, including a lingering impact winter that made it impossible for plants and plankton to carry out photosynthesis (Pope, Baines *et al.* 1997). This resulted in a mass extinction event known as the Cretaceous-Paleogene (K-Pg) extinction event. With some three-quarters of plant and animal species on Earth (including all non-avian dinosaurs) going extinct over a geologically short period of time, it was one of the three largest mass extinctions in the past 500 million years (Schulte, Alegret *et al.* 2010, Renne, Deino *et al.* 2013). At that time, the continents formed

by the break-up of Gondwana (that started about 184 Mya) were well separated (see Figure 7.1). Madagascar had reached its current position with respect to Africa by 130–118 Ma, and most of its biota were removed by the K-Pg extinction event (Yoder and Nowak 2006). The rift of the tropical Atlantic had separated South America and Africa since 119–105 Mya (McLoughlin 2001). Consequently, Africa was already deprived of direct contact with other continents for at least 40 Ma before the K-Pg extinction event took place (Gheerbrant and Rage 2006).

The Violaceae crown clade estimate (70 Mya) predates the K-Pg mass extinction by only few million years, suggesting that all extant Violaceae descend from only few surviving lineages (*Fusispermum* representing the oldest descendent lineage) (Tokuoka 2008). Lowland rainforest expanded after the K-Pg extinction event on all continents, but because all early diverging Violaceae lineages are Neotropical, the family was at that time probably restricted to the Neotropics. It therefore seems likely that Violaceae is of Neotropical origin. The Apiculata group diverged from *Rinorea* s.s. roughly 56 Mya. At that time, global temperatures were relatively high (Eocene optimum) permitting a belt of tropical vegetation around the Northern Hemisphere that may have facilitated dispersal of tropical taxa across continents. The maximum divergence time estimate for Neotropical and Palaeotropical *Rinorea* (∼45 Mya) postdates this thermal optimum, however, and is therefore better explained by a transatlantic dispersal hypothesis (chapter five). When *Rinorea* colonized Africa in the late Eocene, the lowland forest biome was already at its peak: even though global temperatures dropped significantly in the Oligocene, pollen records indicating a shift to more forested conditions in Africa (Jacobs 2004).

In the early Miocene, isolation of the African continent finally ended, as it became definitively connected with Eurasia through the Middle East, and the Tethys seaway closed 16–20 Mya (Gheerbrant and Rage 2006). This resulted in changes in tropical ocean currents and ensuing global cooling in the middle Miocene (Rommerskirchen, Condon *et al.* 2011, Zhang, Nisancioglu *et al.* 2011). By consequence, grass-dominated savannah began to expand in Africa. At this time, the lineages comprising *Harma* and *Cymothoe* diverged and were associated with Achariaceae (see Chapter 4). In the late Miocene (8 Mya) savannah had become widespread at the expense of wet forest habitat (Morley and Richards 1993, Senut, Pickford *et al.* 2009), and *Cymothoe* abruptly began to diversify and colonized African *Rinorea* (see Chapter 4). Given that Achariaceae-feeding and *Rinorea*-feeding *Cymothoe* comprise an equal number of lineages (and species), it seems likely that the *Rinorea*-colonization event happened at the onset of the *Cymothoe* radiation. The resulting *Cymothoe-Rinorea* associations can therefore perhaps be seen as originating from an instantaneous adaptive' event.diversification.

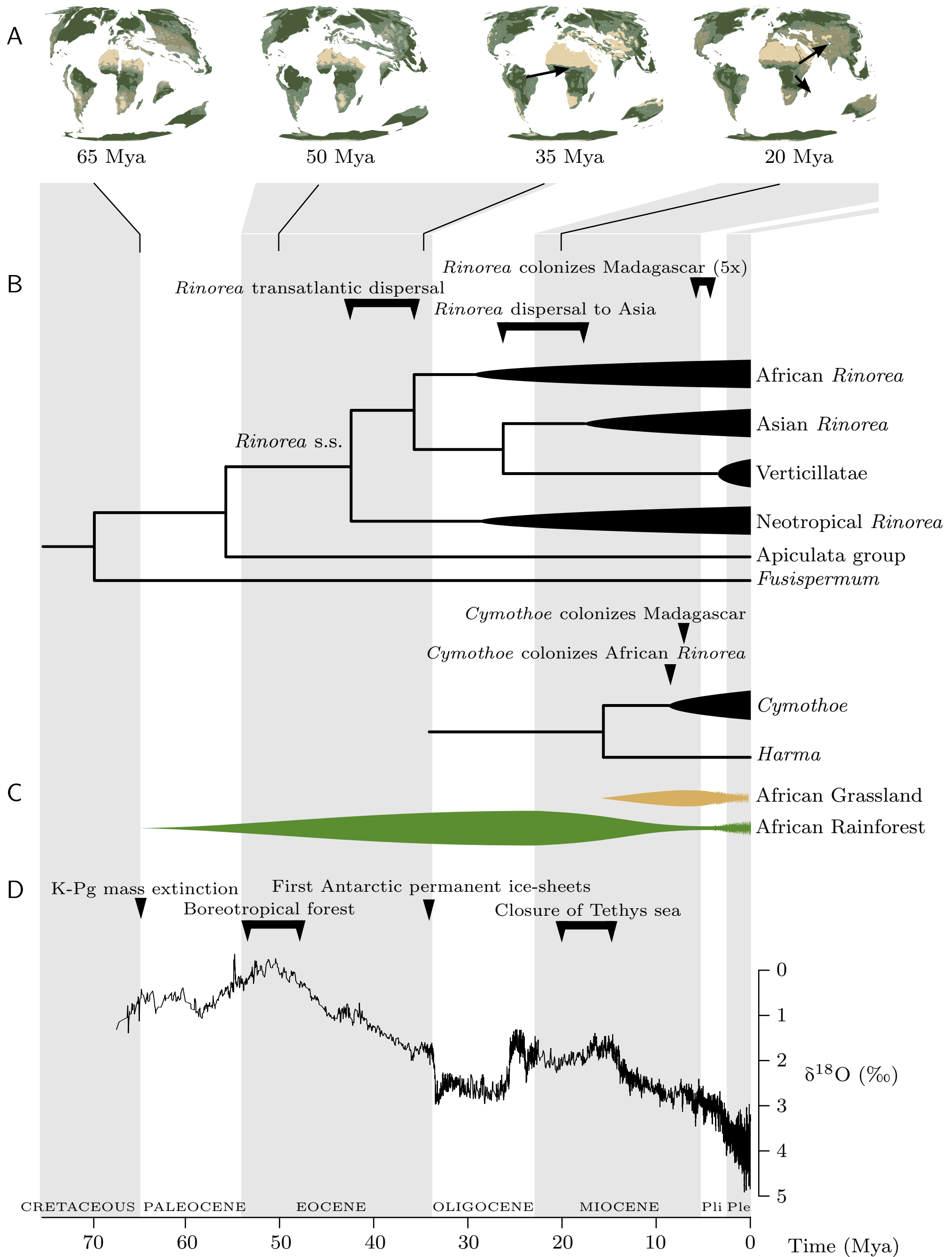
Possibly soon after (<7 Mya), a single lineage of *Cymothoe* colonized Madagascar (see Chapter 4). Intriguingly, time estimates of five independent colonizations of Madagascar by different *Rinorea* lineages show considerable overlap around the same time period (4–6 Mya; Chapter 4). Given the time after the K-Pg mass extinction

on Madagascar 65 Mya, colonization from mainland Africa by both *Cymothoe* and *Rinorea* is therefore surprisingly recent, perhaps facilitated by (a change in) circular currents in the Mozambique Channel. The rapid radiation of *Cymothoe* continues until recent times and some species groups appear to be of Pleistocene origin.

7.3 Unruliness of rapid radiations

In this thesis the *Cymothoe* clade emerged as a case of strikingly rapid diversification. Such radiations are of special interest because they allow us to disentangle alternative factors that may influence macro-evolutionary processes (Rokas, Krüger *et al.* 2005, Gavrilets and Losos 2009, Fordyce 2010), as we have done in Chapter 4. Rapidly diversifying lineages also pose specific analytical challenges, however, mainly because speciation events happen at an elevated rate. By consequence, haplotypes that were present in ancestral species may not experience enough time to get sorted between descendant species, (i.e. incomplete lineage sorting), before subsequent speciation events. Because the increase in diversification of *Cymothoe* coincided with the onset divergence of its extant lineages and continued until the present, effects of incomplete lineage sorting can be seen at various levels. At a shallow phylogenetic level, many species within the clade are relatively young, and our DNA barcode data suggest extensive haplotype sharing between recently diverged species. Obviously, this complicates delineation and identification based on such data (chapter one). At a deeper level, despite our significant sequencing efforts, phylogenetic relationships between the main *Cymothoe* lineages remained unresolved and we attributed this also to incomplete lineage sorting effects (chapter four). Given our DNA barcoding results from extant species one can indeed imagine that the ancestors of these main lineages at their time likewise constituted a group of species with low levels of divergence and similarly experiencing incomplete lineage sorting, occasional hybridization and introgression. Obviously, in such cases reconstructing the exact phylogenetic history millions of years later is extremely difficult.

Figure 7.1: Chronology of reconstructed historical events in *Cymothoe* and *Rinorea* (next page). (A) Palaeographic reconstructions showing relative position of continents in Mollweide projection, redrawn with permission from Ron Blakey (2008; cpgeosystems.com). Arrows indicate postulated *Rinorea* long-distance dispersal routes. (B) Summary of time-calibrated phylogenetic events in *Rinorea* and *Cymothoe*. (C) Schematic representation of extent of savannah grassland and lowland rainforest in Africa over time, based on Jacobs (2004). (D) Selected key tectonic and biotic events and global temperature through time, as oxygen isotope fractionation patterns in Benthic foraminifera which serve as a proxy for the total global mass of glacial ice sheets and temperature; redrawn from Zachos *et al.* (2001); Mya = million years ago; Pli = Pliocene, Ple = Pleistocene.



7

7.4 The end of cospeciation theory?

Comprising highly specialized associations between 33 species of a single clade of butterflies and 34 host species from a single clade of plants, the *Cymothoe-Rinorea* system may seem like an ideal candidate for finding patterns of cospeciation. However, our results clearly reject a scenario of cospeciation of *Cymothoe* and *Rinorea*. Instead, *Cymothoe* was found to have colonized *Rinorea* lineages that originated millions of years earlier, and current associations appear to be shaped through host shifts between related *Rinorea* (see Chapter 6). Such asynchronous diversification and prevalence of host shifts is in accordance with studies in many other clades of herbivorous insects (Jermy 1984, Janz and Nylin 1998, Janz, Nyblom *et al.* 2001, Ronquist and Liljeblad 2001, Brändle, Knoll *et al.* 2005, Agosta 2006, Lopez-Vaamonde, Wikström *et al.* 2006, Gómez-Zurita, Hunt *et al.* 2007, Winkler and Mitter 2008, McKenna, Sequeira *et al.* 2009, Leppänen, Altenhofer *et al.* 2012). And similar results were also found in other host-symbiont systems (Smith, Godsoe *et al.* 2008, Won, Jones *et al.* 2008, Pagán, Firth *et al.* 2010, McLeish and van Noort 2012). Indeed, based on literature, a recent review concluded that convincing examples of cospeciation between host and symbionts seem to be the exception rather than the rule (de Vienne, Refrégier *et al.* 2013). The few supported cases of cospeciation mostly involve vertically transmitted endosymbionts such the sulfur-oxidizing bacteria in Vesicomyid clams (Peek, Feldman *et al.* 1998), and *Buchnera* bacteria in *Brachycaudus* aphids (Jousselin, Desdevises *et al.* 2009). The well-known case of cospeciation between pocket gophers (Rodentia, Geomyidae) and their chewing lice (Hafner et al., 1994, 2003) is probably linked to their unusual life history and ecology: Species of pocket gophers are mostly allopatric and spend most of their time in tunnels, decreasing the likelihood of their parasites shifting to other hosts (Hafner, Spradling *et al.* 2004). The chewing lice are obligate parasites that spend their entire life on their host and have no dispersal stage, also disfavouring host shifts (Clayton and Johnson 2003). Overall, our results are therefore in agreement with a growing body of substantial evidence to suggest that evolutionary dynamics of hosts and parasites do not favour cospeciation (de Vienne, Refrégier *et al.* 2013).

Phylogenetic reconstruction of interspecies associations is problematic, however, especially when host switches are common. Cophylogenetic methods can only make accurate predictions when host and parasite phylogenetic trees are congruent. With increasing incongruence, predictions become more and more stochastic and dependent on the costs set for the various recoverable historical events (e.g. cospeciation, duplication, host shift, and loss). Indeed, even though our results clearly indicated a dynamic history of associations between *Cymothoe* and *Rinorea*, we were unable to accurately reconstruct ancestral host ranges or specific shifts from one host plant to another (see Chapter 6). This is logical because there are endless possible historical explanations for phylogenetic incongruence. Also, when evolutionary changes (in this context: host shifts) are frequent, historical events become obscured by subsequent

ones, leading to saturation of phylogenetic signal. Another issue is that currently available analytical tools are based on minimizing costs in a arsimony framework and consequently cannot deliver probabilities. In phylogenetic analysis of DNA sequence data, the development of models has alleviated (at least in part) some of the above issues. As opposed to Parsimony, models can accommodate rates of change and therefore take into account iterative events. Perhaps more importantly, they allow calculation of likelihoods or, in a Bayesian framework, posterior probabilities. Although many models for host-parasite population dynamics exist (e.g. Herbert and Isham 2000, Restif and Koella 2003, Briggs and Hoopes 2004), adequate models for host-parasite association on a macro-evolutionary scale are wanting. A Bayesian approach implementing a simple stochastic model with a single rate for host switching exists (Huelsenbeck, Rannala *et al.* 2000), but to my knowledge is rarely used in empirical studies. Also, host associations of herbivorous insects may be governed by factors other than those in other host-symbiont systems where e.g. the mode and timing of transmission can be important. Therefore, development of a macro-evolutionary model specific for insect-host plant seems necessary.

One useful way of looking at the evolution of insect host plant associations is by applying an island biogeographic perspective. For an herbivorous insect, resource suitability varies between plant tissues, individuals, and species. Nevertheless, when considering a single tissue type (e.g. leaves in the context of *Cymothoe*), the overall variation among individuals within species will tend to be smaller than the variation between species. Consequently, corresponding tissues of conspecific individuals together can be seen as constituting a distinct island in resource space, and related species form resource archipelagos (Janzen 1968) , and associations can perhaps be regarded stable versus instable equilibria. These islands and archipelagoes create the adaptive landscape (Wright 1932) and hence the context for host-plant use evolution in insects (Nyman 2010). Plants can become colonized by different species at different times. On the other hand, plants can lose herbivores when the latter switch to another host or go extinct. Consequently, the number of species associated with a given host is determined by a colonization-extinction balance, as on islands (MacArthur and Wilson 1967, Janzen 1968). Assuming an island biogeographic model, common and widespread plants equal large and relatively well-connected islands where extinction is predicted to be less likely and occasional 'loss' of associated herbivores is counterbalanced by an increased rate of recolonization from related plant species (i.e. nearby islands) (Nyman 2010). Indeed, common and widespread plants with many relatives tend to have more associated insect herbivore species than rare, phylogenetically isolated ones (e.g. Marquis, Price *et al.* 1991, Kelly and Southwood 1999, Lewinsohn, Novotny *et al.* 2005).

A problem with the island biogeographic perspective on insect-plant associations is that host plants are not static. Over evolutionary time, plant lineages diverge phylogenetically but also in resource space, due to selection and drift (Webb 2000, Agrawal 2007). As such, plants can be seen as "moving targets" (Nyman 2010)

and a macro-evolutionary model of insect-host plant associations needs to take into account these evolutionary changes. In addition, for a model of insect-host plant associations to be useful, in most cases it would need to include geographic species distributions. This is because, logically, host shifts cannot occur between species that have allopatric distributions. Incorporating allopatry makes a model much more complex, however. Especially because reconstructing historical species distributions is already challenging and requires additional models of climatic variables and species ecological envelopes (Elith and Leathwick 2009, Morin and Thuiller 2009). On the other hand, for relatively uniformly distributed systems such as *Cymothoe*-*Rinorea* (restricted to forested regions in Africa), allowing host shifts regardless of species distributions seems a reasonable simplification of reality.

7.5 Chemical ecology

Chemistry is likely to play a key role in shaping the associations between *Cymothoe* and their *Rinorea* hosts. In general, insect-host plant associations are determined by adult females who choose a suitable host for their offspring. Oviposition behaviour of butterflies is induced by recognition of plant metabolites via receptors in the tarsus of the foreleg (Honda 1995, Tsuchihara, Hisatomi *et al.* 2009). Determining which metabolites act as oviposition stimulant for butterflies is difficult, however.

Cymothoe species not feeding on *Rinorea* (Violaceae) use members of Achariaceae as host plants. Achariaceae are known to be cyanogenic, and *Cymothoe* host plant genera *Caloncoba, Kiggelaria, Lindackeria* and *Rawsonia* contain cyclopentanoid glycosides that act as precursor for cyanide production (Cramer, Rehfeldt *et al.* 1980, Raubenheimer and Elsworth 1988, Achoundong 1996, Andersen, Clausen *et al.* 2001, Jaroszewski, Ekpe *et al.* 2004). Although *Rinorea* is non-cyanogenic, Clausen *et al.* (2002) discovered a novel cyclopentanoid glycoside in *Rinorea ilicifolia*, which led them to hypothesize that these metabolites act as a common chemical recognition template for ovipositing *Cymothoe* females. As was outlined in Chapter 1, our preliminary untargeted comparison of metabolites in *Rinorea ilicifolia* suggested that cyclopentanoid glycosides do not play a role in oviposition site recognition its specialist herbivore *Cymothoe egesta*, however. In retrospect, this finding could have been expected because, although the discovery of a chemical similarity between Achariaceae, Passifloraceae, Turneraceae and Violaceae was originally surprising (Clausen, Frydenvang *et al.* 2002), recent results have shown that these families are closely related within Malpighiales (Meier, Shiyang *et al.* 2006). Given these close phylogenetic relationships, many chemical similarities can be expected, each an equally valid candidate common oviposition stimulant in these butterflies. Moreover, the high level of specialization of *Rinorea*-feeding *Cymothoe* indicates that oviposition is determined by phytochemical differences rather than similarities. Indeed, it seems unlikely that each *Rinorea* host contains a unique form of cyclopentanoid glycoside.

Probably, disentangling *Cymothoe* chemical oviposition cues therefore requires tightly controlled behavioural and electrophysiological experiments in different species.

Another interesting avenue could be to perform a comparative untargeted metabolomics study in *Rinorea*. This could potentially identify metabolites that are associated with oviposition by *Cymothoe* on particular host plant species. Ethanol-soluble leaf surface compounds of multiple *Rinorea* species are available for analysis, which may shed light on currently puzzling host plant association patterns such as *C. egesta* that is associated with species from distantly related *Rinorea* from subsection *Ilicifoliae* and the *R. squamosa* group. Another intriguing question is whether host-specific metabolites are sequestered by *Cymothoe* and act as mate recognition signal, because this would facilitate assortative mating and, ultimately, ecological speciation.

Obviously, plant metabolites evolve under selection for traits other than recognition by ovipositing insects. Over the course of the hundreds of millions of years that plants have been parasitized by insects, they have developed many mechanisms to avoid or reduce their impact. The highly toxic metabolites such as alkaloids, furanocoumarins, and glucosinolates are an obvious example, although they may also act as recognition in specialists that have developed ways to tolerate their effects, or even sequester them for their own defence (Nishida 2002). Plants perceive herbivory through molecular signals that may be produced by the plant in response to tissue damage, or by the herbivore (e.g. insect secretions or plant compounds that are modified by the insect while feeding) (Pieterse, van der Does *et al.* 2012). These signals trigger pathways that mount defensive responses and usually involve interplay of different plant hormones. Herbivorous insects (and other attackers), on the other hand, can rewire plant immune signalling circuitry for their own benefit (Zarate, Kempema *et al.* 2007, Weech, Chapleau *et al.* 2008, Bruessow, Gouhier-Darimont *et al.* 2010, Giron, Frago *et al.* 2013, Savchenko, Pearse *et al.* 2013).

One of the defensive responses by plants is to emit volatiles either at the site of damage or systemically from undamaged parts of affected plants (Heil and Ton 2008). These induced volatiles alter the interactions of the plant with its environment, amongst others by attracting insect predators and parasitoids (De Moraes, Lewis *et al.* 1998, Dicke and van Loon 2000). Butterfly eggs and larvae are prone to predation by various other insect clades such as parasitoid wasps, and we found this to be true also for *Cymothoe*. Because insect carnivores use chemical cues from the host plant to locate their prey (Dicke 2000, Lill, Marquis *et al.* 2002), insect herbivores such as *Cymothoe* may enter an 'enemy free space' by shifting to a novel host (Bernays and Graham 1988, Ohsaki and Sato 1994, Ishihara and Ohgushi 2008, Diamond and Kingsolver 2010). Where other butterflies usually lay clutches with multiple eggs, *Cymothoe* females associated with *Rinorea* generally lay single eggs on suitable hosts (Amiet and Achoundong 1996), possibly to minimize predation risk for their offspring. This suggests that parasitization causes sufficient selection pressure to influence their behaviour. Whether such selection is consistent enough through time

and space to have an influence at the species level remains controversial, however (Heard, Stireman *et al.* 2006).

7.6 *Wolbachia* and insect speciation

Insect plant associations can also be mediated through symbiotic microorganisms that enable their insect host to effectively colonize a plant, either through aiding in metabolism of plant tissue or through manipulating the plant immune signalling circuitry (Hosokawa, Kikuchi *et al.* 2007, Barr, Hearne *et al.* 2010, Casteel, Hansen *et al.* 2012, Frago, Dicke *et al.* 2012). There is no information about *Cymothoe* symbiotic microorganisms, except for *Wolbachia* alpha-proteobacterial endosymbionts, the presence of which was confirmed by our DNA sequence data. In fact, our extensive DNA barcode data set revealed putative introgression between closely related but morphologically well-defined species and we hypothesized that this may have been mediated by *Wolbachia* endosymbionts (see Chapter 2). In addition to promoting introgression of mitochondrial haplotypes, *Wolbachia* can also cause cytoplasmic incompatibility between individuals that are infected with different strains. Consequently, when a species is infected with two different strains, this can lead to reproductive isolation and (sympatric) speciation (Bordenstein, O'Hara *et al.* 2001). Given the rapid radiation of *Cymothoe*, we can hypothesize that *Wolbachia* has promoted speciation in this clade (and perhaps has even played a role in the forest-fragmentation setting at which it happened). Based on this hypothesis, two predictions can be tested: First, in order to have played a significant role in their diversification, the majority of *Cymothoe* species are expected to be infected by *Wolbachia*. Second, sister species are expected to host different *Wolbachia* strains. Confirming (or rejecting) these predictions requires sequencing of *Wolbachia*-specific genetic markers (e.g. *WSP* gene sequences). Experimental confirmation of cytoplasmic incompatibility would require experimental crosses between individuals that are infected or treated with antibiotics.

7.7 Herbivorous insects and food security

Plant-insect interactions are receiving increasing attention because of their importance in crop production (Schoonhoven, van Loon *et al.* 2005). Global demand for food is ever increasing due to continuing population and consumption growth. Increased productivity is an essential part of the solution and reducing crop loss due to pests such as insect herbivores is important. As an example, actual losses due to insect pests in worldwide production of wheat, maize and cotton are around 10%, suggesting major economic impact (Oerke 2006). Herbivorous insects can have detrimental impact on crops in two ways: the first is through direct damage

done to the plant by the feeding insect, which eats leaves or burrows in stems, fruit, or roots. There are many pest species of this type, both in larvae and adults, among orthoptera, homoptera, heteroptera, coleoptera, lepidoptera, and diptera (Schoonhoven, van Loon *et al.* 2005). The second is through indirect damage in which the insect transmits e.g. bacterial, viral, or fungal infection into a crop. Examples include viral diseases transmitted by whiteflies or thrips, and bacterial infections causing Asian citrus greening transmitted by psyllids (Varma and Malathi 2003, Bove 2006).

At the same time, an increasing number of insects and plants that have evolved in separation are currently coming into contact. This is because of increased global trade and associated intentional and accidental transportation of species across continents (Work, McCullough *et al.* 2005). In addition, due to global climate change, species ranges are in flux and ranges of previously separated species and populations may start to overlap, providing novel opportunities for selection and host shifts (Bale, Masters *et al.* 2002). It is therefore tempting to find implications of our results on insect-host plant associations for agricultural systems. We found that *Cymothoe* host plant associations are labile but that host shifts are ultimately phylogenetically constrained. This seems to be a general pattern because similar results have been found for various other insect clades (Mitter, Farrell *et al.* 1991, Janz and Nylin 1998, Ronquist and Liljeblad 2001, Braby and Trueman 2006, Lopez-Vaamonde, Wikström *et al.* 2006, Wilson, Forister *et al.* 2012). Given these constraints, one might expect that predicting host shifts by specialized insects onto crops is trivial. When an insect is specialized on a plant that is closely related to a crop, it is likely that it can become a pest. Similarly, when a crop is distantly related to the current host, successful colonization is unlikely. Such extrapolation from natural to agricultural ecosystems is unwarranted, however, because these systems are very different (Gassmann, Onstad *et al.* 2009). For efficiency reasons, agricultural landscapes are homogeneous landscapes comprising monocultures. As a consequence, selection pressures exerted in agricultural systems are more intense than in natural systems (Macfadyen and Bohan 2010). This is because crops represent an unlimited and locally concentrated availability of plant food, allowing herbivores to build up large populations (Oerke 2006, Balmer, Pfiffner *et al.* 2013). In addition, diversity of natural enemies is relatively low in agricultural systems (Letourneau, Jedlicka *et al.* 2009, Haddad, Crutsinger *et al.* 2011). Therefore, selection for a trait allowing effective colonization of a crop is rarely balanced by selection from predators (Boyer, Baker *et al.* 2007, Gardiner, Landis *et al.* 2009, Balmer, Pfiffner *et al.* 2013). Last, human agricultural practices such as field rotation and pest management impose particularly strong selection, and crop pests may go through dramatic population bottlenecks, favouring fixation of resistance traits (Hendry, Kinnison *et al.* 2011, Curzi, Zavala *et al.* 2012). As an example, Western corn rootworm (*Diabrotica virgifera*) depends on the continuous availability of corn. Exploiting this specialism, farmers broadly adopted a practice of annual crop rotation between corn and the

distantly related non-host soybean. Although initially providing excellent control of Western corn rootworm, this practice dramatically reduced landscape heterogeneity in East-central Illinois and imposed intense selection pressure. This selection pressure resulted in behavioural changes and "rotation-resistant" adults, that are able to survive on soybean (Sammons, Edwards *et al.* 1997).

A solution to the above issues of rising selection pressures on insect herbivores (as well as other pests) should probably be sought in more integrated approaches to agriculture. Integrated pest management aims for pest damage reduction by the most economical means, and with the least possible hazard to people and the environment. It comprises a suite of complementary technologies such as crop- and region-specific protection solutions and takes pest life history and population dynamics into account (Kogan 1998, Birch, Begg *et al.* 2011).

7.8 A plea for integration of taxonomic data

In the current information age, data management and integration is the key to successful execution and dissemination of scientific work. The advent of new high-throughput sequencing technologies has enabled automated generation of high volumes of genetic sequence data. In Chapter 2, for example, we have demonstrated that high-throughput DNA barcoding can generate large, population-level sequence data sets. DNA sequence data are made publicly available through digital repositories such as GenBank and BOLD (Ratnasingham and Hebert 2007).

At the same time, access to crucial taxonomic information is being facilitated by an increasing number of digital databases of e.g. species names, digitized original descriptions, and online imagery of type specimens and historical literature (Jorger, Norenburg *et al.* 2012). Certainly, online resources such as the International Plant Names Index (www.ipni.org; Croft, Cross *et al.* 1999), Biodiversity Heritage Library (www.biodiversitylibrary.org), and Aluka (www.aluka.org) have greatly facilitated the systematic work presented in this thesis. The TimeTree public knowledge database aggregates divergence time estimates from the literature to facilitate retrieval by the greater community of scientists (Hedges, Dudley *et al.* 2006). The iPlant Collaborative has created a Taxonomic Name Resolution Service that can help correct incorrect (i.e. misspelled or synonymous) taxonomic names (Boyle, Hopkins *et al.* 2013). Nevertheless, there is room for improvement in some areas. Some data repositories are poorly curated (e.g. GenBank contains many DNA sequences referencing misidentified specimens or misspelled species names). Providing an interface for moderation by the user community could help to efficiently improve data integrity of such large databases. In addition, libraries should better coordinate their digitization efforts. Repositories of taxonomic literature such as the Biodiversity Library currently contain many digital duplicates of the same historical literature, while most of the more specialized works are still digitally unavailable

and consequently remain obscure in the digital age.

The most crucial issue in my view, however, is that data still remain fragmented over different repositories. The main challenge therefore currently lies in data integration. For example, systematic and taxonomic workflows could be greatly enhanced when a species name could provide cross-reference linking to the original publication and description, (type) specimen data, images, and DNA sequences. Because computers cannot readily interpret species names other than as a string of text, it would be helpful if all taxonomic names had associated digital identifiers, similar to the digital object identifier indexing system already in use to facilitate retrieval of e.g. scientific literature. The advantage of such identifiers is that they are permanent, whereas the location of metadata about that species may change. A project with such goals has been presented in the form of a Global Names Architecture (Patterson, Cooper *et al.* 2010) but unfortunately does not appear to have taken off. Notable online services that aggregate data based on species names are the Encyclopedia of Life (eol.org) and Tropicos (www.tropicos.org). However, these services are not cross-referenced and each uses their own name referencing system.

Species names have been effective identifiers and entries to biological information for centuries. But the digital age requires that they can be unambiguously referenced. With global digital identifiers for taxonomic names, a future can be envisioned where updated taxonomic information is accessible for almost everybody from everywhere, in a truly integrated way and with interlinked purpose-specific web interfaces.

References

Abdo, Z. and Golding, G.B. (2007). A step toward barcoding life: A model-based, decision-theoretic method to assign genes to preexisting species groups. *Systematic Biology* **56**(1): 44–56.

d'Abrera, B. (2004). Butterflies of the Afrotropical region part II. Nymphalidae, Libytheidae. Melbourne, Hill House Publishers.

Achoundong, G. (1996). Les *Rinorea* comme indicateurs des grands types forestiers du Cameroun. in: *The Biodiversity of African Plants.* L. J. G. Van der Maesen, X. M. Van der Burgt and J. M. Van Medenbach-de Rooy. Dordrecht, Kluwer Academic Publishers: 536–544.

Achoundong, G. (2000). Les *Rinorea* et l'étude des refuges forestiers en Afrique. in: *Dynamique à long terme des écosystèmes forestiers intertropicaux.* M. Servant and S. Servant-Vildary. Paris, Mémoire UNESCO: 19–29.

Achoundong, G. and Bakker, F.T. (2006). Deux nouvelles espèces de *Rinorea*, série Ilicifoliae (Violaceae) du Cameroun. *Adansonia* **28**(1): 129–136.

Achoundong, G. and Cheek, M. (2003). Two new species of *Rinorea* (Violaceae) from western Cameroon. *Kew Bulletin* **58**(4): 957–964.

Ackery, P.R. (1988). Hostplants and classification - a review of Nymphalid butterflies. *Biological Journal of the Linnean Society* **33**(2): 95–203.

Ackery, P.R., Smith, C.R. and Vane-Wright, R.I. (1995). Carcasson'terights African butterflies: A catalogue of the Rhopalocera of the Afrotropical region. East Melbourne, CSIRO.

Adomou, A.C., Sinsin, B. and van der Maesen, L.J.G. (2006). Phytosociological and chorological approaches to phytogeography: A meso-scale study in Benin. *Systematics and Geography of Plants* **76**(2): 155–178.

Aduse-Poku, K., Vingerhoedt, E. and Wahlberg, N. (2009). Out-of-Africa again: A phylogenetic hypothesis of the genus Charaxes (Lepidoptera: Nymphalidae) based on five gene regions. *Molecular Phylogenetics and Evolution* **53**(2): 463–478.

Agnaniet, H., Mounzeo, H., Menut, C., Bessiere, J.-M. and Criton, M. (2003). The Essential oils of *Rinorea subintegrifolia* O. Ktze and *Drypetes gossweileri* S. Moore occurring in Gabon. *Flavour and Fragrance Journal* **18**(3): 207–210.

Agosta, S.J. (2006). On ecological fitting, plant–insect associations, herbivore host shifts, and host plant selection. *Oikos* **114**(3): 556–565.

Agrawal, A.A. (2007). Macroevolution of plant defense strategies. *Trends in Ecology & Evolution* **22**(2): 103–109.

Agrawal, A.A., Hastings, A.P., Johnson, M.T.J., Maron, J.L. and Salminen, J.-P. (2012). Insect herbivores drive real-time ecological and evolutionary change in plant populations. *Science* **338**(6103): 113–116.

Ahren, D. and Tunlid, A. (2003). Evolution of parasitism in nematode-trapping fungi. *Journal of Nematology* **35**(2): 194–197.

Alfaro, M.E., Santini, F., Brock, C., Alamillo, H., Dornburg, A., Rabosky, D.L., Carnevale, G. and Harmon, L.J. (2009). Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proceedings of the National Academy of Sciences of the United States of America* **106**(32): 13410–13414.

Ali, J.R. and Aitchison, J.C. (2008). Gondwana to Asia: Plate tectonics, paleogeography and the biological connectivity of the Indian sub-continent from the Middle Jurassic through latest Eocene (166–35 Ma). *Earth-Science Reviews* **88**(3–4): 145–166.

Ali, J.R. and Huber, M. (2010). Mammalian biodiversity on Madagascar controlled by ocean currents. *Nature* **463**(7281): 653–656.

Allmann, S., Spathe, A., Bisch-Knaden, S., Kallenbach, M., Reinecke, A., Sachse, S., Baldwin, I.T. and Hansson, B.S. (2013). Feeding-induced rearrangement of green leaf volatiles reduces moth oviposition. *eLife* **2**: e00421.

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

Alvarez, L.W., Alvarez, W., Asaro, F. and Michel, H.V. (1980). Extraterrestrial cause for the Cretaceous-Tertiary extinction. *Science* **208**(4448): 1095–1108.

Amiet, J.-L. (1997). Spécialisation trophique et premières états chez les *Cymothoe*: implications taxonomiques (Lepidoptera, Nymphalidae). *Bulletin de la Société entomologique de France* **102**(1): 15–29.

Amiet, J.-L. (2000). Les premiers états des *Cymothoe*: Morphologie et intérêt phylogénique (Lepidoptera, Nymphalidae). *Bulletin de la Société entomologique de France* **106**(4): 349–390.

Amiet, J.-L. and Achoundong, G. (1996). Un exemple de spécialisation trophique chez les Lépidoptères: les *Cymothoe* camerounaises inféodées aux *Rinorea* (Violacées) (Lepidoptera, Nymphalidae). *Bulletin de la Société entomologique de France* **101**(5): 449–466.

Andersen, L., Clausen, V., Oketch-Rabah, H.A., Lechtenberg, M., Adsersen, A., Nahrstedt, A. and Jaroszewski, J.W. (2001). Gynocardin and cyclopentenylglycine in *Rawsonia lucida*. *Biochemical Systematics and Ecology* **29**(2): 219–222.

Appelhans, M.S., Keßler, P.J.A., Smets, E., Razafimandimbison, S.G. and Janssens, S.B. (2012). Age and historical biogeography of the pantropically distributed Spathelioideae (Rutaceae, Sapindales). *Journal of Biogeography* **39**(7): 1235–1250.

Aris-Brosou, S. and Yang, Z. (2002). Effects of models of rate evolution on estimation of divergence dates with special reference to the Metazoan 18S ribosomal RNA phylogeny. *Systematic Biology* **51**(5): 703–714.

Armstrong, K.F. and Ball, S.L. (2005). DNA barcodes for biosecurity: invasive species identification. *Philosophical Transactions of the Royal Society B-Biological Sciences* **360**(1462): 1813–1823.

Armstrong, K.F., Cameron, C.M. and Frampton, E.R. (1997). Fruit fly (Diptera: Tephritidae) species identification: A rapid molecular diagnostic technique for quarantine application. *Bulletin of Entomological Research* **87**(2): 111–118.

Austerlitz, F., David, O., Schaeffer, B., Bleakley, K., Olteanu, M., Leblois, R., Veuille, M. and Laredo, C. (2009). DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. *BMC Bioinformatics* **10(Suppl 14)**: S10.

Aveskamp, M.M., Woudenberg, J.H.C., de Gruyter, J., Turco, E., Groenewald, J.Z. and Crous, P.W. (2009). Development of taxon-specific sequence characterized amplified region (SCAR) markers based on actin sequences and DNA amplification fingerprinting (DAF): a case study in the *Phoma exigua* species complex. *Molecular Plant Pathology* **10**(3): 403–414.

Azpurua, J., De la Cruz, D., Valderama, A. and Windsor, D. (2010). *Lutzomyia* sand fly diversity and rates of infection by *Wolbachia* and an exotic *Leishmania* species on Barro Colorado Island, Panama. *PLoS Neglected Tropical Diseases* **4**(3): e627.

Bakker, F.T., Culham, A., Marais, E.M. and Gibby, M. (2005). Nested radiation in Cape *Pelargonium*. in: *Plant species-level systematics: new perspectives on pattern & process*. F. T. Bakker, L. W. Chatrou, B. Gravendeel and P. Pelser. Liechtenstein, Gantner Verlag.

Bakker, F.T., van Gemerden, B.S. and Achoundong, G. (2006). Molecular systematics of African *Rinorea* Aublet. (Violaceae). in: *Taxonomy and ecology of African plants, their conservation and sustainable use*. S. A. Ghazanfar and H. J. Beentje. Richmond, Kew Publishing.

Bale, J.S., Masters, G.J., Hodkinson, I.D., Awmack, C., Bezemer, T.M., Brown, V.K., Butterfield, J., Buse, A., Coulson, J.C., Farrar, J., Good, J.E.G., Harrington, R., Hartley, S., Jones, T.H., Lindroth, R.L., Press, M.C., Symrnioudis, I., Watt, A.D. and Whittaker, J.B. (2002). Herbivory in global climate change research: direct effects of rising temperature on insect herbivores. *Global Change Biology* **8**(1): 1–16.

Balmer, O., Pfiffner, L., Schied, J., Willareth, M., Leimgruber, A., Luka, H. and Traugott, M. (2013). Noncrop flowering plants restore top-down herbivore control in agricultural fields. *Ecology and Evolution*: 2634–2646.

Barker, D., Meade, A. and Pagel, M. (2007). Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. *Bioinformatics* **23**(1): 14–20.

Barr, K.L., Hearne, L.B., Briesacher, S., Clark, T.L. and Davis, G.E. (2010). Microbial symbionts in insects influence down-regulation of defense genes in maize. *PLoS ONE* **5**(6): e11339.

Barraclough, T.G. and Vogler, A.P. (2002). Recent diversification rates in North American tiger beetles estimated from a dated mtDNA phylogenetic tree. *Molecular Biology and Evolution* **19**(10): 1706–1716.

Bartish, I.V., Antonelli, A., Richardson, J.E. and Swenson, U. (2011). Vicariance or long-distance dispersal: historical biogeography of the pantropical subfamily Chrysophylloideae (Sapotaceae). *Journal of Biogeography* **38**(1): 177–190.

Bastos, A., Nair, D., Taylor, P., Brettschneider, H., Kirsten, F., Mostert, E., von Maltitz, E., Lamb, J., van Hooft, P., Belmain, S., Contrafatto, G., Downs, S. and Chimimba, C. (2011). Genetic monitoring detects an overlooked cryptic species and reveals the diversity and distribution of three invasive *Rattus* congeners in south Africa. *BMC Genetics* **12**(1): 26.

Beaurain, M. (1988). Note sur *Cymothoe coranus* Grose-Smith et description de trois nouvelles sous-espèces (Lepidoptera, Nymphalidae). *Lambillionea* **88**: 39–47.

Becerra, J.X. (1997). Insects on plants: Macroevolutionary chemical trends in host use. *Science* **276**(5310): 253–256.

Becerra, J.X. (2003). Synchronous coadaptation in an ancient case of herbivory. *Proceedings of the National Academy of Sciences of the United States of America* **100**(22): 12804–12807.

Becerra, J.X. and Venable, D.L. (1999). Macroevolution of insect-plant associations: The relevance of host biogeography to host affiliation. *Proceedings of the National Academy of Sciences of the United States of America* **96**(22): 12626–12631.

Bell, C.D., Soltis, D.E. and Soltis, P.S. (2010). The Age and diversification of the angiosperms re-revisited. *American Journal of Botany* **97**(8): 1296–1303.

Berbee, M.L. and Taylor, J.W. (1993). Dating the evolutionary radiations of the true fungi. *Canadian Journal of Botany* **71**(8): 1114–1127.

Berenbaum, M.R. and Zangerl, A.R. (1998). Chemical phenotype matching between a plant and its insect herbivore. *Proceedings of the National Academy of Sciences of the United States of America* **95**(23): 13743–13748.

Berger, L.A. (1981). Les Papillons du Zaïre. Brussels, Weissenbruch.

Berlocher, S.H. and Feder, J.L. (2002). Sympatric speciation in phytophagous insects: Moving beyond controversy? *Annual Review of Entomology* **47**: 773–815.

Bernays, E. and Graham, M. (1988). On the evolution of host specificity in phytophagous arthropods. *Ecology* **69**(4): 886–892.

Berthold, P. (2001). Bird migration: A general survey, 2nd edition. New York, Oxford University Press.

Bertolazzi, P., Felici, G., Festa, P. and Lancia, G. (2008). Logic classification and feature selection for biomedical data. *Computers and Mathematics With Applications* **55**(5): 889–899.

Bertolazzi, P., Felici, G. and Lancia, G. (2010). Application of feature selection and classification to computational molecular biology. in: *Biological Data Mining.* J. Y. Chen and S. Lonardi. Boca Raton, Chapman & Hall/CRC: 257–294.

Bertolazzi, P., Felici, G. and Weitschek, E. (2009). Learning to classify species with barcodes. *Bmc Bioinformatics* **10**.

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

Birch, A.N.E., Begg, G.S. and Squire, G.R. (2011). How agro-ecological research helps to address food security issues under new IPM and pesticide reduction policies for global crop production systems. *Journal of Experimental Botany* **62**(10): 3251–3261.

Blakey, R.C. (2008). Gondwana paleogeography from assembly to breakup-A 500 m.y. odyssey. in: *Resolving the Late Paleozoic Ice Age in Time and Space.* C. R. Fielding, T. D. Frank and J. L. Isbell. **441**: 1–28.

Bonferroni, C. (1935). Il calcolo delle assicurazioni su gruppi di teste. in: *Studi in Onore del Professore Salvatore Ortu Carboni.* Rome, Tipografi del Senato: 13–60.

Bordenstein, S.R., O'Hara, F.P. and Werren, J.H. (2001). *Wolbachia*-induced incompatibility precedes other hybrid incompatibilities in *Nasonia. Nature* **409**(6821): 707–710.

Borer, M., van Noort, T., Arrigo, N., Buerki, S. and Alvarez, N. (2011). Does a shift in host plants trigger speciation in the Alpine leaf beetle *Oreina speciosissima* (Coleoptera, Chrysomelidae)? *BMC Evolutionary Biology* **11**(1): 310.

Bos, J.J. (1989). Clef des Violaceae ligneuses pour la partie occidentale de l'Afrique centrale. *Adansonia* **4**: 461–468.

Bossart, J.L. (2003). Covariance of preference and performance on normal and novel hosts in a locally monophagous and locally polyphagous butterfly population. *Oecologia* **135**(3): 477–486.

Bossart, J.L., Opuni-Frimpong, E., Kuudaar, S. and Nkrumah, E. (2006). Richness, abundance, and complementarity of fruit-feeding butterfly species in relict sacred forests and forest reserves of Ghana. *Biodiversity and Conservation* **15**(1): 333–359.

Boucher, L.D., Manchester, S.R. and Judd, W.S. (2003). An extinct genus of Salicaceae based on twigs with attached flowers, fruits, and foliage from the Eocene Green River Formation of Utah and Colorado, USA. *American Journal of Botany* **90**(9): 1389–1399.

Bouquet, A. (1969). Féticheurs et Médecines Traditionnelles du Congo (Brazzaville). Paris, Mémoires ORSTOM.

Bouquet, A. (1974). Plantes Médicinales de la Côte D'Ivoire. Paris, ORSTOM.

Bouyer, T. and Ducarme, R. (2013). Description de la femelle de *Cymothoe bouyeri* Vande weghe, 2011 (Lepidoptera, Nymphalidae) *Entomologia Africana* **18**(1): 1–2.

Bouyer, T. and Joly, E. (1995). Notes sur *Cymothoe harmilla* avec description d'une nouvelle sous-espèce et des premiers stades. *Lambillionea* **95**(4): 526–532.

Bove, J.M. (2006). Huanglongbing: A destructive, newly-emerging, century-old disease of citrus. *Journal of Plant Pathology* **88**(1): 7–37.

Boyer, S.L., Baker, J.M. and Giribet, G. (2007). Deep genetic divergences in *Aoraki denticulata* (Arachnida, Opiliones, Cyphophthalmi): a widespread 'mite harvestman' defies DNA taxonomy. *Molecular Ecology* **16**(23): 4999–5016.

Boykin, L.M., Shatters, R.G., Hall, D.G., Burns, R.E. and Franqui, R.A. (2006). Analysis of host preference and geographical distribution of *Anastrepha suspensa* (Diptera: Tephritidae) using phylogenetic analyses of mitochondrial cytochrome oxidase I DNA sequence data. *Bulletin of Entomological Research* **96**(5): 457–469.

Boyle, B., Hopkins, N., Lu, Z., Raygoza Garay, J.A., Mozzherin, D., Rees, T., Matasci, N., Narro, M., Piel, W., Mckay, S., Lowry, S., Freeland, C., Peet, R. and Enquist, B. (2013). The Taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC Bioinformatics* **14**(1): 16.

Braby, M.F. and Pierce, N.E. (2007). Systematics, biogeography and diversification of the Indo-Australian genus *Delias* Hübner (Lepidoptera: Pieridae): phylogenetic evidence supports an 'out-of-Australia' origin. *Systematic Entomology* **32**(1): 2–25.

Braby, M.F. and Trueman, J.W.H. (2006). Evolution of larval host plant associations and adaptive radiation in pierid butterflies. *Journal of Evolutionary Biology* **19**(5): 1677–1690.

Brakefield, P.M. (2011). Evo-devo and accounting for Darwin's endless forms. *Philosophical Transactions of the Royal Society B-Biological Sciences* **366**(1574): 2069–2075.

Brändle, M., Knoll, S., Eber, S., Stadler, J. and Brandl, R. (2005). Flies on thistles: support for synchronous speciation? *Biological Journal of the Linnean Society* **84**(4): 775–783.

Brändle, M., Kühn, I., Klotz, S., Belle, C. and Brandl, R. (2008). Species richness of herbivores on exotic host plants increases with time since introduction of the host. *Diversity and Distributions* **14**(6): 905–912.

Brandt, M. (1914). Übersicht über die afrikanischen Arten der Gattung *Rinorea* Aubl. *Botanische Jahrbücher für Systematik, Pflanzengeschichte und Pflanzengeographie* **50 suppl.**: 405–418.

Briggs, C.J. and Hoopes, M.F. (2004). Stabilizing effects in spatial parasitoid–host and predator–prey models: a review. *Theoretical Population Biology* **65**(3): 299–315.

Britten, R. (1986). Rates of DNA sequence evolution differ between taxonomic groups. *Science* **231**(4744): 1393–1398.

Bruessow, F., Gouhier-Darimont, C., Buchala, A., Metraux, J.-P. and Reymond, P. (2010). Insect eggs suppress plant defence against chewing herbivores. *The Plant Journal* **62**(5): 876–885.

Bucciarelli, G., Golani, D. and Bernardi, G. (2002). Genetic cryptic species as biological invaders: the case of a Lessepsian fish migrant, the hardyhead silverside *Atherinomorus lacunosus*. *Journal of Experimental Marine Biology and Ecology* **273**(2): 143–149.

Bull, J.J., Huelsenbeck, J.P., Cunningham, C.W., Swofford, D.L. and Waddell, P.J. (1993). Partitioning and combining data in phylogenetic analysis. *Systematic Biology* **42**(3): 384–397.

Burney, D.A., Burney, L.P., Godfrey, L.R., Jungers, W.L., Goodman, S.M., Wright, H.T. and Jull, A.J.T. (2004). A chronology for late prehistoric Madagascar. *Journal of Human Evolution* **47**(1–2): 25–63.

Butlin, R., Debelle, A., Kerth, C., Snook, R.R., Beukeboom, L.W., Cajas, R.F.C., Diao, W., Maan, M.E., Paolucci, S., Weissing, F.J., van de Zande, L., Hoikkala, A., Geuverink, E., Jennings, J., Kankare, M., Knott, K.E., Tyukmaeva, V.I., Zoumadakis, C., Ritchie, M.G., Barker, D., Immonen, E., Kirkpatrick, M., Noor, M., Garcia, C.M., Schmitt, T., Schilthuizen, M. and Marie Curie, S.N. (2012). What do we need to know about speciation? *Trends in Ecology & Evolution* **27**(1): 27–39.

Callmander, M.W., Phillipson, P.B., Schatz, G.E., Andriambololonera, S., Rabarimanarivo, M., Rakotonirina, N., Raharimampionona, J., Chatelain, C., Gautier, L. and Lowry, P.P. (2011). The Endemic and non-endemic vascular flora of Madagascar updated. *Plant Ecology and Evolution* **144**(2): 121–125.

del Campo, M.L., Miles, C.I., Schroeder, F.C., Mueller, C., Booker, R. and Renwick, J.A. (2001). Host recognition by the tobacco hornworm is mediated by a host plant compound. *Nature* **411**(6834): 186–189.

Casner, K.L. and Pyrcz, T.W. (2010). Patterns and timing of diversification in a tropical montane butterfly genus, *Lymanopoda* (Nymphalidae, Satyrinae). *Ecography* **33**(2): 251–259.

Casteel, C.L., Hansen, A.K., Walling, L.L. and Paine, T.D. (2012). Manipulation of plant defense responses by the tomato Psyllid (*Bactericerca cockerelli*) and its associated endosymbiont *Candidatus Liberibacter psyllaurous. PLoS ONE* **7**(4): e35191.

Castillo, D., Mell, J., Box, K. and Blumenstiel, J. (2011). Molecular evolution under increasing transposable element burden in *Drosophila*: A speed limit on the evolutionary arms race. *BMC Evolutionary Biology* **11**(1): 258.

Charlat, S., Duplouy, A., Hornett, E., Dyson, E., Davies, N., Roderick, G., Wedell, N. and Hurst, G. (2009). The Joint evolutionary histories of *Wolbachia* and mitochondria in *Hypolimnas bolina. BMC Evolutionary Biology* **9**(1): 64.

Chipp, T.F. (1923). New species of *Rinorea* from West Africa. *Bulletin of miscellaneous information, Royal Botanic Gardens, Kew* **1923**: 289–299.

Chuyong, G.B., Kenfack, D., Harms, K.E., Thomas, D.W., Condit, R. and Comita, L.S. (2011). Habitat specificity and diversity of tree species in an African wet tropical forest. *Plant Ecology* **212**(8): 1363–1374.

Clabaut, C., Bunje, P.M.E., Salzburger, W. and Meyer, A. (2007). Geometric morphometric analyses provide evidence for the adaptive character of the tanganyikan cichlid fish radiations. *Evolution* **61**(3): 560–578.

Clausen, V., Frydenvang, K., Koopmann, R., Jorgensen, L.B., Abbiw, D.K., Ekpe, P. and Jaroszewski, J.W. (2002). Plant analysis by butterflies: Occurrence of cyclopentenylglycines in Passifloraceae, Flacourtiaceae, and Turneraceae and discovery of the novel nonproteinogenic amino acid 2-(3 '-cyclopentenyl)glycine in *Rinorea. Journal of Natural Products* **65**(4): 542–547.

Clayton, D.H. and Johnson, K.P. (2003). Linking coevolutionary history to ecological process: Doves and lice. *Evolution* **57**(10): 2335–2341.

Coetzee, J.A. (1993). African flora since the terminal Jurassic. New Haven, Yale Univ Press.

Cohen, A.S., Stone, J.R., Beuning, K.R.M., Park, L.E., Reinthal, P.N., Dettman, D., Scholz, C.A., Johnson, T.C., King, J.W., Talbot, M.R., Brown, E.T. and Ivory, S.J. (2007). Ecological consequences of early Late Pleistocene megadroughts in tropical Africa. *Proceedings of the National Academy of Sciences* **104**(42): 16422–16427.

Coley, P.D. and Barone, J.A. (1996). Herbivory and plant defenses in tropical forests. *Annual Review of Ecology and Systematics* **27**: 305–335.

Condamine, F.L., Sperling, F.A.H., Wahlberg, N., Rasplus, J.-Y. and Kergoat, G.J. (2012). What causes latitudinal gradients in species diversity? Evolutionary processes and ecological constraints on swallowtail biodiversity. *Ecology Letters* **15**(3): 267–277.

Conow, C., Fielder, D., Ovadia, Y. and Libeskind-Hadas, R. (2010). Jane: a new tool for the cophylogeny reconstruction problem. *Algorithms for Molecular Biology* **5**.

Cornell, H.V. and Hawkins, B.A. (2003). Herbivore responses to plant secondary compounds: A test of phytochemical coevolution theory. *American Naturalist* **161**(4): 507–522.

Couvreur, T.L.P., Chatrou, L.W., Sosef, M.S.M. and Richardson, J.E. (2008). Molecular phylogenetics reveal multiple tertiary vicariance origins of the African rain forest trees. *BMC Biology* **6**: 54.

Couvreur, T.L.P., Pirie, M.D., Chatrou, L.W., Saunders, R.M.K., Su, Y.C.F., Richardson, J.E. and Erkens, R.H.J. (2011). Early evolutionary history of the flowering plant family Annonaceae: steady diversification and boreotropical geodispersal. *Journal of Biogeography* **38**(4): 664–680.

Cramer, U., Rehfeldt, A.G. and Spener, F. (1980). Isolation and biosynthesis of cyclopentenylglycine, a novel non-proteinogenic amino-acid in Flacourtiaceae. *Biochemistry* **19**(13): 3074–3080.

Crisp, M.D. and Cook, L.G. (2005). Do early branching lineages signify ancestral traits? *Trends in Ecology & Evolution* **20**(3): 122–128.

Croft, J., Cross, N., Hinchcliffe, S., Lughadha, E.N., Stevens, P.F., West, J.G. and Whitbread, G. (1999). Plant names for the 21st century: the International Plant Names Index, a distributed data source of general accessibility. *Taxon* **48**(2): 317–324.

Cruaud, A., Ronsted, N., Chantarasuwan, B., Chou, L.S., Clement, W.L., Couloux, A., Cousins, B., Genson, G., Harrison, R.D., Hanson, P.E., Hossaert-Mckey, M., Jabbour-Zahab, R., Jousselin, E., Kerdelhue, C., Kjellberg, F., Lopez-Vaamonde, C., Peebles, J., Peng, Y.Q., Pereira, R.A.S., Schramm, T., Ubaidillah, R., van Noort, S., Weiblen, G.D., Yang, D.R., Yodpinyanee, A., Libeskind-Hadas, R., Cook, J.M., Rasplus, J.Y. and Savolainen, V. (2012). An Extreme Case of Plant-Insect Codiversification: Figs and Fig-Pollinating Wasps. *Systematic Biology* **61**(6): 1029–1047.

Cuatrecasas, J. (1950). Fieldiana. *Studies in South American plants II* **27**(1): 55–113.

Cunningham, C.W. (1997). Can three incongruence tests predict when data should be combined? *Molecular Biology and Evolution* **14**(7): 733–740.

Curzi, M.J., Zavala, J.A., Spencer, J.L. and Seufferheld, M.J. (2012). Abnormally high digestive enzyme activity and gene expression explain the contemporary evolution of a *Diabrotica* biotype able to feed on soybeans. *Ecology and Evolution* **2**(8): 2005–2017.

Cusimano, N. and Renner, S.S. (2010). Slowdowns in diversification rates from real phylogenies may not be real. *Systematic Biology* **59**(4): 458–464.

Cusimano, N., Stadler, T. and Renner, S.S. (2012). A new method for handling missing species in diversification analysis applicable to randomly or non-randomly sampled phylogenies. *Systematic Biology* **61**(5): 785–792.

Damm, S., Schierwater, B. and Hadrys, H. (2010). An integrative approach to species discovery in odonates: from character-based DNA barcoding to ecology. *Molecular Ecology* **19**(18): 3881–3893.

Darlu, P. and Lecointre, G. (2002). When does the incongruence length difference test fail? *Molecular Biology and Evolution* **19**(4): 432–437.

DasGupta, B., Konwar, K.M., Măndoiu, I.I. and Shvartsman, A.A. (2005). DNA-BAR: distinguisher selection for DNA barcoding. *Bioinformatics* **21**(16): 3424–3426.

Davis, C.C., Bell, C.D., Fritsch, P.W. and Mathews, S. (2002). Phylogeny of *Acridocarpus-Brachylophon* (Malpighiaceae): Implications for tertiary tropical floras and Afroasian biogeography. *Evolution* **56**(12): 2395–2405.

Davis, C.C., Webb, C.O., Wurdack, K.J., Jaramillo, C.A. and Donoghue, M.J. (2005). Explosive radiation of Malpighiales supports a mid-Cretaceous origin of modern tropical rain forests. *American Naturalist* **165**(3): E36–E65.

Day, J.J., Cotton, J.A. and Barraclough, T.G. (2008). Tempo and mode of diversification of Lake Tanganyika Cichlid fishes. *PLoS ONE* **3**(3): e1730.

Dayrat, B. (2005). Towards integrative taxonomy. *Biological Journal of the Linnean Society* **85**(3): 407–415.

De Moraes, C.M., Lewis, W.J., Pare, P.W., Alborn, H.T. and Tumlinson, J.H. (1998). Herbivore-infested plants selectively attract parasitoids. *Nature* **393**(6685): 570–573.

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

De Vienne, D.M., Giraud, T. and Shykoff, J.A. (2007). When can host shifts produce congruent host and parasite phylogenies? A simulation approach. *Journal of Evolutionary Biology* **20**(4): 1428–1438.

De Vienne, D.M., Refrégier, G., López-Villavicencio, M., Tellier, A., Hood, M.E. and Giraud, T. (2013). Cospeciation vs host-shift speciation: methods for testing, evidence from natural associations and relation to coevolution. *New Phytologist* **198**(2): 347–385.

De Wildeman, E. (1920). Notes sur le genre *Rinorea* Aubl. *Bulletin du Jardin botanique de l'Etat à Bruxelles* **6**: 131–194.

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

DeSalle, R., Egan, M.G. and Siddall, M. (2005). The Unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philosophical Transactions of the Royal Society B-Biological Sciences* **360**(1462): 1905–1916.

Devictor, V., Mouillot, D., Meynard, C., Jiguet, F., Thuiller, W. and Mouquet, N. (2010). Spatial mismatch and congruence between taxonomic, phylogenetic and functional diversity: the need for integrative conservation strategies in a changing world. *Ecology Letters* **13**(8): 1030–1040.

Dexter, K.G., Pennington, T.D. and Cunningham, C.W. (2010). Using DNA to assess errors in tropical tree identifications: How often are ecologists wrong and when does it matter? *Ecological Monographs* **80**(2): 267–286.

Diamond, S.E. and Kingsolver, J.G. (2010). Fitness consequences of host plant choice: a field experiment. *Oikos* **119**(3): 542–550.

Dicke, M. (2000). Chemical ecology of host-plant selection by herbivorous arthropods: a multitrophic perspective. *Biochemical Systematics and Ecology* **28**(7): 601–617.

Dicke, M. and van Loon, J.J.A. (2000). Multitrophic effects of herbivore-induced plant volatiles in an evolutionary context. *Entomologia Experimentalis et Applicata* **97**(3): 237–249.

Dinca, V., Zakharov, E.V., Hebert, P.D.N. and Vila, R. (2011). Complete DNA barcode reference library for a country's butterfly fauna reveals high performance for temperate Europe. *Proceedings of the Royal Society B-Biological Sciences* **278**(1704): 347–355.

Djuikouo, M.N.K., Doucet, J.-L., Nguembou, C.K., Lewis, S.L. and Sonké, B. (2010). Diversity and aboveground biomass in three tropical forest types in the Dja Biosphere Reserve, Cameroon. *African Journal of Ecology* **48**(4): 1053–1063.

Donguy, J.R. and Piton, B. (1991). The Mozambique channel revisited. *Oceanologica Acta* **14**(6): 549–558.

Dowsett-Lemaire, F. and White, F. (1990). New and noteworthy plants from the evergreen forests of Malawi. *Bulletin du Jardin botanique national de Belgique* **60**(1/2): 73–110.

Doyle, J.J. and Doyle, J.L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* **19**: 11–15.

Drès, M. and Mallet, J. (2002). Host races in plant–feeding insects and their importance in sympatric speciation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **357**(1420): 471–492.

Drummond, A.J., Ho, S.Y.W., Phillips, M.J. and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biol* **4**(5): e88.

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

Drummond, A.J., Suchard, M.A., Xie, D. and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution.*

Duffey, S.S. and Stout, M.J. (1996). Antinutritive and toxic components of plant defense against insects. *Archives of Insect Biochemistry and Physiology* **32**(1): 3–37.

Dupont, L. (2011). Orbital scale vegetation change in Africa. *Quaternary Science Reviews* **30**(25–26): 3589–3602.

Dupont, L.M., Jahns, S., Marret, F. and Ning, S. (2000). Vegetation change in equatorial West Africa: time-slices for the last 150 ka. *Palaeogeography, Palaeoclimatology, Palaeoecology* **155**(1): 95–122.

Edwards, A. and Cavalli-Sforza, L. (1963). The Reconstruction of evolution. *Annals of Human Genetics* **27**: 105–106.

Ehrlich, P.R. and Raven, P.H. (1964). Butterflies and plants - a study in coevolution. *Evolution* **18**(4): 586–608.

Elias, M., Hill, R.I., Willmott, K.R., Dasmahapatra, K.K., Brower, A.V.Z., Mallet, J. and Jiggins, C.D. (2007). Limited performance of DNA barcoding in a diverse community of tropical butterflies. *Proceedings of the Royal Society B-Biological Sciences* **274**(1627): 2881–2889.

Elias, M., Joron, M., Willmott, K., Silva-Brandao, K.L., Kaiser, V., Arias, C.F., Pinerez, L.M.G., Uribe, S., Brower, A.V.Z., Freitas, A.V.L. and Jiggins, C.D. (2009). Out of the Andes: patterns of diversification in clearwing butterflies. *Molecular Ecology* **18**(8): 1716–1729.

Elith, J. and Leathwick, J.R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics* **40**(1): 677–697.

Emelianov, I., Dres, M., Baltensweiler, W. and Mallet, J. (2001). Host-induced assortative mating in host races of the larch budmoth. *Evolution* **55**(10): 2002–2010.

Engler, A. (1902). Violaceae africanae. *Botanische Jahrbücher der Systematik* **XXXIII**: 132–147.

Fahrenholz, H. (1913). Ectoparasiten und Abstammungslehre. *Zoologischer Anzeiger* **41**: 371–374.

Farrell, B. and Mitter, C. (1990). Phylogenesis of insect plant interactions - Have *Phyllobrotica* leaf beetles (Chrysomelidae) and the Lamiales diversified in parallel? *Evolution* **44**(6): 1389–1403.

Farrell, B.D. (2001). Evolutionary assembly of the milkweed fauna: Cytochrome oxidase I and the age of Tetraopes beetles. *Molecular Phylogenetics and Evolution* **18**(3): 467–478.

Farrell, B.D. and Mitter, C. (1998). The Timing of insect/plant diversification: might *Tetraopes* (Coleoptera : Cerambycidae) and *Asclepias* (Asclepiadaceae) have co-evolved? *Biological Journal of the Linnean Society* **63**(4): 553–577.

Farris, J.S., Kallersjo, M., Kluge, A.G. and Bult, C. (1994). Testing Significance of Incongruence. *Cladistics-the International Journal of the Willi Hennig Society* **10**(3): 315–319.

Farris, J.S., Kallersjo, M., Kluge, A.G. and Bult, C. (1995). Constructing a significance test for incongruence. *Systematic Biology* **44**(4): 570–572.

Fazekas, A.J., Kesanakurti, P.R., Burgess, K.S., Percy, D.M., Graham, S.W., Barrett, S.C.H., Newmaster, S.G., Hajibabaei, M. and Husband, B.C. (2009). Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? *Molecular Ecology Resources* **9**: 130–139.

Federal Ministry of Environment, United Nations Development Programme and Global Environment Facility (2011). Niger Delta Biodiversity Project**:** UNDP GEF PIMS no.: 2047.

Felici, G. and Truemper, K. (2002). A MINSAT approach for learning in logic domains. *INFORMS Journal on Computing* **14**(1): 20–36.

R

Fischer, E., Rembold, K., Althof, A., Obholzer, J., Malombe, I., Mwachala, G., Onyango, J.C., Dumbo, B. and Theisen, I. (2010). Annotated checklist of the vascular plants of Kakamega forest, Western Province, Kenya. *Journal of East African Natural History* **99**(2): 129–226.

Fisher, B.L. and Smith, M.A. (2008). A revision of Malagasy species of *Anochetus* Mayr and *Odontomachus* Latreille (Hymenoptera: Formicidae). *PLoS ONE* **3**(5): e1787.

FitzJohn, R.G., Maddison, W.P. and Otto, S.P. (2009). Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Systematic Biology* **58**(6): 595–611.

Fitzpatrick, B.M., Fordyce, J.A. and Gavrilets, S. (2009). Pattern, process and geographic modes of speciation. *Journal of Evolutionary Biology* **22**(11): 2342–2347.

Fleischmann, A., Schäferhoff, B., Heubl, G., Rivadavia, F., Barthlott, W. and Müller, K.F. (2010). Phylogenetics and character evolution in the carnivorous plant genus *Genlisea* A. St.-Hil. (Lentibulariaceae). *Molecular Phylogenetics and Evolution* **56**(2): 768–783.

Floyd, R., Abebe, E., Papert, A. and Blaxter, M. (2002). Molecular barcodes for soil nematode identification. *Molecular Ecology* **11**(4): 839–850.

Fontaine, M. (1982). Genre *Cymothoe* Hübner. Lep. Nymphalidae - S. fam. Nymphalinae. Note sur les premiers états. *Lambillionea* **82**: 63–64, 67–72, 95–98.

Fordyce, J.A. (2010). Host shifts and evolutionary radiations of butterflies. *Proceedings of the Royal Society B* **277**(1701): 3735–3743.

Forister, M.L., Nice, C.C., Fordyce, J.A. and Gompert, Z. (2009). Host range evolution is not driven by the optimization of larval performance: the case of Lycaeides melissa (Lepidoptera: Lycaenidae) and the colonization of alfalfa. **160**(3): 551–561.

Frago, E., Dicke, M. and Godfray, H.C.J. (2012). Insect symbionts as hidden players in insect–plant interactions. *Trends in Ecology & Evolution* **27**(12): 705–711.

Fratantoni, D.M., Johns, W.E., Townsend, T.L. and Hurlburt, H.E. (2000). Low-latitude circulation and mass transport pathways in a model of the Tropical Atlantic Ocean. *Journal of Physical Oceanography* **30**(8): 1944–1966.

Friedman, M. (1937). The Use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* **32**(200): 675–701.

Funk, D.J., Filchak, K.E. and Feder, J.L. (2002). Herbivorous insects: model systems for the comparative study of speciation ecology. *Genetica* **116**(2-3): 251–267.

Funk, D.J. and Omland, K.E. (2003). Species-level paraphyly and polyphyly: Frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology Evolution and Systematics* **34**: 397–423.

Futuyma, D.J. and Agrawal, A.A. (2009). Macroevolution and the biological diversity of plants and herbivores. *Proceedings of the National Academy of Sciences of the United States of America* **106**(43): 18054–18061.

Futuyma, D.J. and Moreno, G. (1988). The Evolution of ecological specialization. *Annual Review of Ecology and Systematics* **19**: 207–233.

Garamszegi, L.Z. and Møller, A.P. (2011). Nonrandom variation in within-species sample size and missing data in phylogenetic comparative studies. *Systematic Biology* **60**(6): 876–880.

Gardiner, M.M., Landis, D.A., Gratton, C., DiFonzo, C.D., O'Neal, M., Chacon, J.M., Wayo, M.T., Schmidt, N.P., Mueller, E.E. and Heimpel, G.E. (2009). Landscape diversity enhances biological control of an introduced crop pest in the north-central USA. *Ecological Applications* **19**(1): 143–154.

Gassmann, A.J., Onstad, D.W. and Pittendrigh, B.R. (2009). Evolutionary analysis of herbivorous insects in natural and agricultural environments. *Pest Management Science* **65**(11): 1174–1181.

Gavrilets, S. and Losos, J.B. (2009). Adaptive radiation: Contrasting theory with data. *Science* **323**(5915): 732–737.

Gheerbrant, E. and Rage, J.-C. (2006). Paleobiogeography of Africa: How distinct from Gondwana and Laurasia? *Palaeogeography, Palaeoclimatology, Palaeoecology* **241**(2): 224–246.

Giron, D., Frago, E., Glevarec, G., Pieterse, C.M.J. and Dicke, M. (2013). Cytokinins as key regulators in plant–microbe–insect interactions: connecting plant growth and defence. *Functional Ecology* **27**(3): 599–609.

Githae, E.W., Chuah-Petiot, M., Mworia, J.K. and Odee, D.W. (2008). A botanical inventory and diversity assessment of Mt. Marsabit forest, a sub-humid montane forest in the arid lands of northern Kenya. *African Journal of Ecology* **46**(1): 39–45.

Givnish, T.J., Millam, K.C., Evans, T.M., Hall, J.C., Pires, J.C., Berry, P.E. and Sytsma, K.J. (2004). Ancient vicariance or recent long-distance dispersal? Inferences about phylogeny and South American-African disjunctions in Rapateaceae and Bromeliaceae based on ndhF sequence data. *International Journal of Plant Sciences* **165**(4): S35–S54.

Goldstein, P.Z. and DeSalle, R. (2011). Integrating DNA barcode data and taxonomic practice: Determination, discovery, and description. *BioEssays* **33**(2): 135–147.

Goloboff, P.A. (1999). Analyzing large data sets in reasonable times: solutions for composite optima. *Cladistics* **15**(4): 415–428.

Goloboff, P.A., Farris, J.S. and Nixon, K.C. (2008). TNT, a free program for phylogenetic analysis. *Cladistics* **24**(5): 774–786.

Gómez-Zurita, J., Hunt, T., Kopliku, F. and Vogler, A.P. (2007). Recalibrated tree of leaf beetles (Chrysomelidae) indicates independent diversification of Angiosperms and their insect herbivores. *PLoS ONE* **2**(4): e360.

Gompert, Z., Forister, M.L., Fordyce, J.A. and Nice, C.C. (2008). Widespread mito-nuclear discordance with evidence for introgressive hybridization and selective sweeps in Lycaeides. *Molecular Ecology* **17**(24): 5231–5244.

Goodman, S.M. and Benstead, J.P. (2005). Updated estimates of biotic diversity and endemism for Madagascar. *Oryx* **39**(1): 73–77.

Graham, A. (1985). Studies in Neotropical paleobotany. IV. The Eocene communities of Panama. *Annals of the Missouri Botanical Garden* **72**: 504–534.

Grey-Wilson, C. (1981). Notes on African Violaceae. *Kew Bulletin* **36**(1): 103–126.

Griffiths, A.M., Sims, D.W., Cotterell, S.P., El Nagar, A., Ellis, J.R., Lynghammar, A., McHugh, M., Neat, F.C., Pade, N.G., Queiroz, N., Serra-Pereira, B., Rapp, T., Wearmouth, V.J. and Genner, M.J. (2010). Molecular markers reveal spatially segregated cryptic species in a critically endangered fish, the common skate (*Dipturus batis*). *Proceedings of the Royal Society B-Biological Sciences* **277**(1687): 1497–1503.

Guillet, B., Achoundong, G., Happi, J.Y., Beyala, V.K.K., Bonvallot, J., Riera, B., Mariotti, A. and Schwartz, D. (2001). Agreement between floristic and soil organic carbon isotope (13C/12C, 14C) indicators of forest invasion of savannas during the last century in Cameroon. *Journal of Tropical Ecology* **17**(6): 809–832.

Haddad, N.M., Crutsinger, G.M., Gross, K., Haarstad, J. and Tilman, D. (2011). Plant diversity and the stability of foodwebs. *Ecology Letters* **14**(1): 42–46.

Hafner, M.S. and Nadler, S.A. (1988). Phylogenetic trees support the coevolution of parasites and their hosts. *Nature* **332**(6161): 258–259.

Hafner, M.S., Spradling, T.A., Light, J.E., Hafner, D.J. and Demboski, J.R. (2004). Systematic revision of pocket gophers of the *Cratogeomys gymnurus* species group. *Journal of Mammalogy* **85**(6): 1170–1183.

Hajibabaei, M., DeWaard, J.R., Ivanova, N.V., Ratnasingham, S., Dooh, R.T., Kirk, S.L., Mackie, P.M. and Hebert, P.D.N. (2005). Critical factors for assembling a high volume of DNA barcodes. *Philosophical Transactions of the Royal Society B-Biological Sciences* **360**(1462): 1959–1967.

Hajibabaei, M., Janzen, D.H., Burns, J.M., Hallwachs, W. and Hebert, P.D.N. (2006). DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences of the United States of America* **103**(4): 968–971.

Hajibabaei, M., Smith, M.A., Janzen, D.H., Rodriguez, J.J., Whitfield, J.B. and Hebert, P.D.N. (2006). A minimalist barcode can identify a specimen whose DNA is degraded. *Molecular Ecology Notes* **6**(4): 959–964.

Hall, J.P.W. (2005). Montane speciation patterns in *Ithomiola* butterflies (Lepidoptera : Riodinidae): are they consistently moving up in the world? *Proceedings of the Royal Society B-Biological Sciences* **272**(1580): 2457–2466.

Hall, T.A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium* **41**: 95–98.

Harmon, L.J., Weir, J.T., Brock, C.D., Glor, R.E. and Challenger, W. (2008). GEIGER: investigating evolutionary radiations. *Bioinformatics* **24**(1): 129–131.

Hasegawa, M., Kishino, H. and Yano, T.A. (1985). Dating of the human ape splitting by a molecular clock of mitochondrial-DNA. *Journal of Molecular Evolution* **22**(2): 160–174.

Hawthorne, W.D. and Jongkind, C.C.H. (2006). Woody plants of western African forests: A guide to the forest trees, shrubs and lianes from Senegal to Ghana. Richmond, Royal Botanical Gardens Kew.

Heard, S.B., Stireman, J.O., Nason, J.D., Cox, G.H., Kolacz, C.R. and Brown, J.M. (2006). On the elusiveness of enemy-free space: spatial, temporal, and host-plant-related variation in parasitoid attack rates on three gallmakers of goldenrods. *Oecologia* **150**(3): 421–434.

Hearn, D.J. (2006). *Adenia* (Passifloraceae) and its adaptive radiation: phylogeny and growth form diversification. *Systematic Botany* **31**(4): 805–821.

Hebert, P.D.N., Cywinska, A., Ball, S.L. and DeWaard, J.R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B-Biological Sciences* **270**(1512): 313–321.

Hebert, P.D.N., deWaard, J.R. and Landry, J.F. (2010). DNA barcodes for 1/1000 of the animal kingdom. *Biology Letters* **6**(3): 359–362.

Hebert, P.D.N. and Gregory, T.R. (2005). The Promise of DNA barcoding for taxonomy. *Systematic Biology* **54**(5): 852–859.

Hebert, P.D.N., Penton, E.H., Burns, J.M., Janzen, D.H. and Hallwachs, W. (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences of the United States of America* **101**(41): 14812–14817.

Hebert, P.D.N., Stoeckle, M.Y., Zemlak, T.S. and Francis, C.M. (2004). Identification of birds through DNA barcodes. *PLoS Biology* **2**(10): 1657–1663.

Hecq, J. and Larsen, T.B. (1997). New butterflies from the Okomu Nature Sanctuary in Nigeria (Lepidoptera: Nymphalidae). *Lambillionea* **97**: 101–104.

Hedges, S.B., Dudley, J. and Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**(23): 2971–2972.

Heikkila, M., Kaila, L., Mutanen, M., Peña, C. and Wahlberg, N. (2012). Cretaceous origin and repeated tertiary diversification of the redefined butterflies. *Proceedings of the Royal Society B-Biological Sciences* **279**(1731): 1093–1099.

Heil, M. and Ton, J. (2008). Long-distance signalling in plant defence. *Trends in plant science* **13**(6): 264–272.

Hekking, W.H.A. (1988). Violaceae part I. *Rinorea* and *Rinoreocarpus*. New York, The New York Botanical Garden.

Heled, J. and Drummond, A.J. (2010). Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution* **27**(3): 570–580.

Heled, J. and Drummond, A.J. (2012). Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Systematic Biology* **61**(1): 138–149.

Hendry, A.P., Kinnison, M.T., Heino, M., Day, T., Smith, T.B., Fitt, G., Bergstrom, C.T., Oakeshott, J., Jørgensen, P.S., Zalucki, M.P., Gilchrist, G., Southerton, S., Sih, A., Strauss, S., Denison, R.F. and Carroll, S.P. (2011). Evolutionary principles and their practical application. *Evolutionary Applications* **4**(2): 159–183.

Herbert, J. and Isham, V. (2000). Stochastic host-parasite interaction models. *Journal of Mathematical Biology* **40**(4): 343–371.

Hildebrand, A.R., Penfield, G.T., Kring, D.A., Pilkington, M., Camargo Z., A., Jacobsen, S.B. and Boynton, W.V. (1991). Chicxulub Crater: A possible Cretaceous/Tertiary boundary impact crater on the Yucatán Peninsula, Mexico. *Geology* **19**(9): 867–871.

Ho, S.Y.W. and Phillips, M.J. (2009). Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Systematic Biology* **58**(3): 367–380.

Hollingsworth, P.M., Forrest, L.L., Spouge, J.L., Hajibabaei, M., Ratnasingham, S., van der Bank, M., Chase, M.W., Cowan, R.S., Erickson, D.L., Fazekas, A.J., Graham, S.W., James, K.E., Kim, K.J., Kress, W.J., Schneider, H., van AlphenStahl, J., Barrett, S.C.H., van den Berg, C., Bogarin, D., Burgess, K.S., Cameron, K.M., Carine, M., Chacon, J., Clark, A., Clarkson, J.J., Conrad, F., Devey, D.S., Ford, C.S., Hedderson, T.A.J., Hollingsworth, M.L., Husband, B.C., Kelly, L.J., Kesanakurti, P.R., Kim, J.S., Kim, Y.D., Lahaye, R., Lee, H.L., Long, D.G., Madrinan, S., Maurin, O., Meusnier, I., Newmaster, S.G., Park, C.W., Percy, D.M., Petersen, G., Richardson, J.E., Salazar, G.A., Savolainen, V., Seberg, O., Wilkinson, M.J., Yi, D.K., Little, D.P. and CBOL Plant Working Group (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America* **106**(31): 12794–12797.

Hollingsworth, P.M., Graham, S.W. and Little, D.P. (2011). Choosing and using a plant DNA barcode. *PLoS ONE* **6**(5): e19254.

Hommola, K., Smith, J.E., Qiu, Y. and Gilks, W.R. (2009). A permutation test of host-parasite cospeciation. *Molecular Biology and Evolution* **26**(7): 1457–1468.

Honda, K. (1995). Chemical basis of differential oviposition by Lepidopterous insects. *Archives of Insect Biochemistry and Physiology* **30**(1): 1–23.

Horn, J.W., van Ee, B.W., Morawetz, J.J., Riina, R., Steinmann, V.W., Berry, P.E. and Wurdack, K.J. (2012). Phylogenetics and the evolution of major structural characters in the giant genus *Euphorbia* L. (Euphorbiaceae). *Molecular Phylogenetics and Evolution* **63**(2): 305–326.

Hosokawa, T., Kikuchi, Y., Shimada, M. and Fukatsu, T. (2007). Obligate symbiont involved in pest status of host insect. *Proceedings of the Royal Society B: Biological Sciences* **274**(1621): 1979–1984.

Houle, A. (1998). Floating islands: a mode of long-distance dispersal for small and medium- sized terrestrial vertebrates. *Diversity and Distributions* **4**(5/6): 201–216.

Hsieh, C.H., Wang, C.H. and Ko, C.C. (2007). Evidence from molecular markers and population genetic analyses suggests recent invasions of the western north pacific region by biotypes B and Q of *Bemisia tabaci* (Gennadius). *Environmental Entomology* **36**(4): 952–961.

Hubert, N., Hanner, R., Holm, E., Mandrak, N.E., Taylor, E., Burridge, M., Watkinson, D., Dumont, P., Curry, A., Bentzen, P., Zhang, J.B., April, J. and Bernatchez, L. (2008). Identifying Canadian freshwater fishes through DNA barcodes. *PLoS ONE* **3**(6): e2490.

Hudson, R., Futuyama, D. and Antonovics, J. (1990). Gene genealogies and the coalescent process. in: *Oxford Surveys in Evolutionary Biology vol. 7*, Oxford University Press. **7**.

Huelsenbeck, J.P., Rannala, B. and Larget, B. (2000). A Bayesian framework for the analysis of cospeciation. *Evolution* **54**(2): 352–364.

Huelsenbeck, J.P. and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**(8): 754–755.

Hughes, J., Kennedy, M., Johnson, K.P., Palma, R.L. and Page, R.D.M. (2007). Multiple cophylogenetic analyses reveal frequent cospeciation between pelecaniform birds and *Pectinopygus* lice. *Systematic Biology* **56**(2): 232–251.

Hurst, G.D.D. and Jiggins, F.M. (2005). Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. *Proceedings of the Royal Society B: Biological Sciences* **272**(1572): 1525–1534.

Imada, Y., Kawakita, A. and Kato, M. (2011). Allopatric distribution and diversification without niche shift in a bryophyte-feeding basal moth lineage (Lepidoptera: Micropterigidae). *Proceedings of the Royal Society B-Biological Sciences* **278**(1721): 3026–3033.

Ishihara, M. and Ohgushi, T. (2008). Enemy-free space? Host preference and larval performance of a willow leaf beetle. **50**(1): 35–43.

Ivanova, N.V., DeWaard, J.R. and Hebert, P.D.N. (2006). An inexpensive, automation-friendly protocol for recovering high-quality DNA. *Molecular Ecology Notes* **6**(4): 998–1002.

Ives, A.R. and Godfray, H.C.J. (2006). Phylogenetic analysis of trophic associations. *The American Naturalist* **168**(1): E1–E14.

Iwu, M.M. (1993). Handbook of African medicinal plants. Boca Raton, CRC Press.

Jacobs, B.F. (2004). Palaeobotanical studies from tropical Africa: relevance to the evolution of forest, woodland and savannah biomes. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **359**(1450): 1573–1583.

Jahner, J.P., Bonilla, M.M., Badik, K.J., Shapiro, A.M. and Forister, M.L. (2011). Use of exotic hosts by lepidoptera: widespread species colonize more novel hosts. *Evolution* **65**(9): 2719–2724.

Janz, N. (2005). The relationship between habitat selection and preference for adult and larval food resources in the polyphagous butterfly Vanessa cardui (Lepidoptera : Nymphalidae). *Journal of Insect Behavior* **18**(6): 767–780.

Janz, N. (2011). Ehrlich and Raven revisited: Mechanisms underlying codiversification of plants and enemies. in: *Annual Review of Ecology, Evolution, and Systematics, Vol 42*. D. J. Futuyma, H. B. Shaffer and D. Simberloff. Palo Alto, Annual Reviews. **42**: 71–89.

Janz, N., Nyblom, K. and Nylin, S. (2001). Evolutionary dynamics of host-plant specialization: A case study of the tribe Nymphalini. *Evolution* **55**(4): 783–796.

Janz, N. and Nylin, S. (1998). Butterflies and plants: A phylogenetic study. *Evolution* **52**(2): 486–502.

Janz, N., Nylin, S. and Wahlberg, N. (2006). Diversity begets diversity: host expansions and the diversification of plant-feeding insects. *BMC Evolutionary Biology* **6**: 4.

Janzen, D.H. (1968). Host plants as islands in evolutionary and contemporary time. *The American Naturalist* **102**(928): 592–595.

Janzen, D.H. (1980). When is it coevolution. *Evolution* **34**(3): 611–612.

Jaroszewski, J.W., Ekpe, P. and Witt, M. (2004). Cyclopentanoid cyanohydrin glucosides and amides of a *Lindackeria dentata*. *Planta Medica* **70**(10): 1001–1003.

Jermy, T. (1976). Insect-host-plant relationship-co-evolution or sequential evolution? *Symposia biologica Hungarica* **16**: 109–113.

Jermy, T. (1984). Evolution of insect host plant relationships. *American Naturalist* **124**(5): 609–630.

Jermy, T. and Szentesi, Á. (2003). Evolutionary aspects of host plant specialisation – a study on bruchids (Coleoptera: Bruchidae). *Oikos* **101**(1): 196–204.

Jiggins, F.M. (2003). Male-killing *Wolbachia* and mitochondrial DNA: Selective sweeps, hybrid introgression and parasite population dynamics. *Genetics* **164**(1): 5–12.

Johnson, M.T.J., FitzJohn, R.G., Smith, S.D., Rausher, M.D. and Otto, S.P. (2011). Loss of sexual recombination and segregation is associated with increased diversification in evening primroses. *Evolution* **65**(11): 3230–3240.

Jorger, K., Norenburg, J., Wilson, N. and Schrodl, M. (2012). Barcoding against a paradox? Combined molecular species delineations reveal multiple cryptic lineages in elusive meiofaunal sea slugs. *BMC Evolutionary Biology* **12**(1): 245.

Jousselin, E., Desdevises, Y. and Coeur d'acier, A. (2009). Fine-scale cospeciation between *Brachycaudus* and *Buchnera aphidicola*: bacterial genome helps define species and evolutionary relationships in aphids. *Proceedings of the Royal Society B: Biological Sciences* **276**(1654): 187–196.

Kaartinen, R., Stone, G.N., Hearn, J., Lohse, K. and Roslin, T. (2010). Revealing secret liaisons: DNA barcoding changes our understanding of food webs. *Ecological Entomology* **35**(5): 623–638.

Kaila, L. and Stahls, G. (2006). DNA barcodes: Evaluating the potential of COI to diffentiate closely related species of *Elachista* (Lepidoptera : Gelechioidea : Elachistidae) from Australia. *Zootaxa* **1170**: 1–26.

Kandul, N.P., Lukhtanov, V.A., Pierce, N.E. and Mooers, A. (2007). Karyotypic diversity and speciation in *Agrodiaetus* butterflies. *Evolution* **61**(3): 546–559.

Katoh, K. and Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics* **9**(4): 286–298.

Kelley, S.T. and Farrell, B.D. (1998). Is specialization a dead end? The phylogeny of host use in *Dendroctonus* bark beetles (Scolytidae). *Evolution* **52**(6): 1731–1743.

Kelly, C.K. and Southwood, T.R.E. (1999). Species richness and resource availability: A phylogenetic analysis of insects associated with trees. *Proceedings of the National Academy of Sciences* **96**(14): 8013–8016.

Kenfack, D., Thomas, D.W., Chuyong, G. and Condit, R. (2007). Rarity and abundance in a diverse African forest. *Biodiversity and Conservation* **16**(7): 2045–2074.

Kergoat, G.J., Le Ru, B.P., Genson, G., Cruaud, C., Couloux, A. and Delobel, A. (2011). Phylogenetics, species boundaries and timing of resource tracking in a highly specialized group of seed beetles (Coleoptera: Chrysomelidae: Bruchinae). *Molecular Phylogenetics and Evolution* **59**(3): 746–760.

Kielland, J. (1990). Butterflies of Tanzania. Melbourne, Hill House.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide-sequences. *Journal of Molecular Evolution* **16**(2): 111–120.

Kingman, J.F.C. (1982). On the Genealogy of Large Populations. *Journal of Applied Probability* **19A**: 27–43.

Knowles, L.L. and Carstens, B.C. (2007). Delimiting species without monophyletic gene trees. *Systematic Biology* **56**(6): 887–895.

Kodandaramaiah, U., Lees, D.C., Muller, C.J., Torres, E., Karanth, K.P. and Wahlberg, N. (2010). Phylogenetics and biogeography of a spectacular Old World radiation of butterflies: the subtribe Mycalesina (Lepidoptera: Nymphalidae: Satyrini). *BMC Evolutionary Biology* **10**: 172.

Kodandaramaiah, U. and Wahlberg, N. (2007). Out-of-Africa origin and dispersal-mediated diversification of the butterfly genus *Junonia* (Nymphalidae : Nymphalinae). *Journal of evolutionary biology* **20**(6): 2181–2191.

Kogan, M. (1998). Integrated pest management: historical perspectives and contemporary developments. *Annual Review of Entomology* **43**(1): 243–270.

Koski, L.B. and Golding, G.B. (2001). The Closest BLAST hit is often not the nearest neighbor. *Journal of Molecular Evolution* **52**(6): 540–542.

Kovar-Eder, J., Kvacek, Z. and Meller, B. (2001). Comparing Early to Middle Miocene floras and probable vegetation types of Oberdorf N Voitsberg (Austria), Bohemia (Czech Republic), and Wackersdorf (Germany). *Review of Palaeobotany and Palynology* **114**(1–2): 83–125.

Kuhn, T.S., Mooers, A.Ø. and Thomas, G.H. (2011). A simple polytomy resolver for dated phylogenies. *Methods in Ecology and Evolution* **2**(5): 427–436.

Kuksa, P. and Pavlovic, V. (2009). Efficient alignment-free DNA barcode analytics. *BMC Bioinformatics* **10**.

Küper, W., Sommer, J.H., Jon, C.L., Mutke, J., Linder, H.P., Beentje, H.J., Rompaey, R.S.A.R.V., Chatelain, C., Sosef, M. and Barthlott, W. (2004). Africa's hotspots of biodiversity redefined. *Annals of the Missouri Botanical Garden* **91**(4): 525–535.

Kvie, K.S., Hogner, S., Aarvik, L., Lifjeld, J.T. and Johnsen, A. (2013). Deep sympatric mtDNA divergence in the autumnal moth (*Epirrita autumnata*). *Ecology and Evolution* **3**(1): 126–144.

Lahaye, R., Van der Bank, M., Bogarin, D., Warner, J., Pupulin, F., Gigot, G., Maurin, O., Duthoit, S., Barraclough, T.G. and Savolainen, V. (2008). DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences of the United States of America* **105**(8): 2923–2928.

Larsen, T.B. (1991). The Butterflies of Kenya and their natural history. Oxford, Oxford University Press.

Larsen, T.B. (2005). Butterflies of West Africa. Stenstrup, Apollo Books.

Legendre, P., Desdevises, Y. and Bazin, E. (2002). A statistical test for host-parasite coevolution. *Systematic Biology* **51**(2): 217–234.

Lemmon, A.R. and Moriarty, E.C. (2004). The Importance of proper model assumption in Bayesian phylogenetics. *Systematic Biology* **53**(2): 265–277.

Leneveu, J., Chichvarkhin, A. and Wahlberg, N. (2009). Varying rates of diversification in the genus *Melitaea* (Lepidoptera: Nymphalidae) during the past 20 million years. *Biological Journal of the Linnean Society* **97**(2): 346–361.

Leppänen, S.A., Altenhofer, E., Liston, A.D. and Nyman, T. (2012). Phylogenetics and evolution of host-plant use in leaf-mining sawflies (Hymenoptera: Tenthredinidae: Heterarthrinae). *Molecular Phylogenetics and Evolution* **64**(2): 331–341.

Leroy, J.F. (1978). Composition, origin, and affinities of the Madagascan vascular flora. *Annals of the Missouri Botanical Garden* **65**(2): 535–589.

Letourneau, D.K., Jedlicka, J.A., Bothwell, S.G. and Moreno, C.R. (2009). Effects of natural enemy biodiversity on the suppression of arthropod herbivores in terrestrial ecosystems. *Annual Review of Ecology, Evolution, and Systematics* **40**(1): 573–592.

Lewinsohn, T.M., Novotny, V. and Basset, Y. (2005). Insects on plants: Diversity of herbivore assemblages revisited. *Annual Review of Ecology Evolution and Systematics* **36**: 597–620.

Lewinsohn, T.M. and Roslin, T. (2008). Four ways towards tropical herbivore megadiversity. *Ecology Letters* **11**(4): 398–416.

Lill, J.T., Marquis, R.J. and Ricklefs, R.E. (2002). Host plants influence parasitism of forest caterpillars. **417**(6885): 170–173.

Linder, H.P., Hardy, C.R. and Rutschmann, F. (2005). Taxon sampling effects in molecular clock dating: An example from the African Restionaceae. *Molecular Phylogenetics and Evolution* **35**(3): 569–582.

Little, D.P. (2011). DNA barcode sequence identification incorporating taxonomic hierarchy and within taxon variability. *PLoS ONE* **6**(8): e26619.

Little, D.P., Stevenson, D.W., Cullman, L.B. and Cullman, D. (2008). A comparison of algorithms for identification of specimens using DNA barcodes: examples from gymnosperms. *Cladistics* **24**(1): 97–97.

Liu, L. and Pearl, D.K. (2007). Species trees from gene trees: Reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Systematic Biology* **56**(3): 504–514.

Lopez-Vaamonde, C., Wikström, N., Labandeira, C., Godfray, H.C.J., Goodman, S.J. and Cook, J.M. (2006). Fossil-calibrated molecular phylogenies reveal that leaf-mining moths radiated millions of years after their host plants. *Journal of Evolutionary Biology* **19**(4): 1314–1326.

Lou, M. and Golding, G.B. (2010). Assigning sequences to species in the absence of large interspecific differences. *Molecular Phylogenetics and Evolution* **56**(1): 187–194.

Loytynoja, A. and Goldman, N. (2009). Uniting alignments and trees. *Science* **324**(5934): 1528–1529.

Loytynoja, A. and Goldman, N. (2010). webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics* **11**: 579.

MacArthur, R.H. and Wilson, E.O. (1967). The Theory of island biogeography. Princeton, Princeton University Press.

Macfadyen, S. and Bohan, D.A. (2010). Crop domestication and the disruption of species interactions. *Basic and Applied Ecology* **11**(2): 116–125.

Maddison, W.P., Midford, P.E. and Otto, S.P. (2007). Estimating a binary character's effect on speciation and extinction. *Systematic Biology* **56**(5): 701–710.

Magallon, S. and Sanderson, M.J. (2001). Absolute diversification rates in Angiosperm clades. *Evolution* **55**(9): 1762–1780.

Mai, D. (1967). Die Florenzonen, der Florenwechsel und die Vorstellungen über den. Klimaablauf im Jungtertiär der Deutschen Demokratischen Republik. *Abhandlungen des Zentralen Geologischen Instituts* **10**: 55–81.

Makarenkov, V. and Lapointe, F.-J. (2004). A weighted least-squares approach for inferring phylogenies from incomplete distance matrices. *Bioinformatics* **20**(13): 2113–2121.

Malcomber, S.T. (2002). Phylogeny of *Gaertnera* Lam. (Rubiaceae) based on multiple DNA markers: evidence of a rapid radiation in a widespread, morphologically diverse genus. *Evolution* **56**(1): 42–57.

Mallarino, R., Bermingham, E., Willmott, K.R., Whinnett, A. and Jiggins, C.D. (2005). Molecular systematics of the butterfly genus *Ithomia* (Lepidoptera : Ithomiinae): a composite phylogenetic hypothesis based on seven genes. *Molecular Phylogenetics and Evolution* **34**(3): 625–644.

Marcussen, T., Jakobsen, K.S., Danihelka, J., Ballard, H.E., Blaxland, K., Brysting, A.K. and Oxelman, B. (2012). Inferring species networks from gene trees in high-polyploid North American and Hawaiian Violets (*Viola*, Violaceae). *Systematic Biology* **61**(1): 107–126.

Marquis, R.J., Price, P.W., ed., Lewinsohn, T.M., ed., Fernandes, G.W., ed. and Benson, W.W., ed. (1991). Herbivore fauna of *Piper* (Piperaceae) in a Costa Rican wet forest: diversity, specificity, and impact. in: *Plant-Animal Interactions, Evolutionary Ecology in Tropical and Temperate Regions*. P. W. Price, T. M. Lewinsohn, G. W. Fernandes and W. W. Benson. New York, Wiley: 179–208.

Matz, M.V. and Nielsen, R. (2005). A likelihood ratio test for species membership based on DNA sequence data. *Philosophical Transactions of the Royal Society B-Biological Sciences* **360**(1462): 1969–1974.

May, G.E., Gelembiuk, G.W., Panov, V.E., Orlova, M.I. and Lee, C.E. (2006). Molecular ecology of zebra mussel invasions. *Molecular Ecology* **15**(4): 1021–1031.

Mayr, E. (1963). Animal species and evolution. Cambridge, Harvard University press.

McBride, C.S., van Velzen, R. and Larsen, T.B. (2009). Allopatric origin of cryptic butterfly species that were discovered feeding on distinct host plants in sympatry. *Molecular Ecology* **18**(17): 3639–3651.

McFadden, C.S., Benayahu, Y., Pante, E., Thoma, J.N., Nevarez, P.A. and France, S.C. (2011). Limitations of mitochondrial gene barcoding in Octocorallia. *Molecular Ecology Resources* **11**(1): 19–31.

McKenna, D.D. and Farrell, B.D. (2006). Tropical forests are both evolutionary cradles and museums of leaf beetle diversity. *Proceedings of the National Academy of Sciences of the United States of America* **103**(29): 10947–10951.

McKenna, D.D., Sequeira, A.S., Marvaldi, A.E. and Farrell, B.D. (2009). Temporal lags and overlap in the diversification of weevils and flowering plants. *Proceedings of the National Academy of Sciences* **106**(17): 7083–7088.

McKeon, S., Lehr, M., Wilkerson, R., Ruiz, J., Sallum, M., Lima, J., Povoa, M. and Conn, J. (2010). Lineage divergence detected in the malaria vector *Anopheles marajoara* (Diptera: Culicidae) in Amazonian Brazil. *Malaria Journal* **9**(1): 271.

McLeish, M. and van Noort, S. (2012). Codivergence and multiple host species use by fig wasp populations of the *Ficus* pollination mutualism. *Bmc Evolutionary Biology* **12**(1): 1.

McLoughlin, S. (2001). The Breakup history of Gondwana and its impact on pre-Cenozoic floristic provincialism. *Australian Journal of Botany* **49**(3): 271–300.

McNeill, J. (2000). Naming the groups: developing a stable and efficient nomenclature. *Taxon* **49**(4): 705–720.

Megens, H.J., van Moorsel, C.H.M., Piel, W.H., Pierce, N.E. and de Jong, R. (2004). Tempo of speciation in a butterfly genus from the Southeast Asian tropics, inferred from mitochondrial and nuclear DNA sequence data. *Molecular Phylogenetics and Evolution* **31**(3): 1181–1196.

Meier, R., Shiyang, K., Vaidya, G. and Ng, P.K.L. (2006). DNA barcoding and taxonomy in diptera: A tale of high intraspecific variability and low identification success. *Systematic Biology* **55**(5): 715–728.

Meier, R., Zhang, G. and Ali, F. (2008). The Use of mean instead of smallest interspecific distances exaggerates the size of the 'barcoding gap' and leads to misidentification. *Systematic Biology* **57**(5): 809–813.

Melchior, H. (1925). Violaceae. in: *Die Natürlichen Pflanzenfamilien.* A. Engler. Leipzig, Verlag von Wilhelm Engelmann. **21**: 359.

Menken, S.B.J. (1981). Host races and sympatric speciation in small Ermine moths, Yponomeutidae. *Entomologia Experimentalis Et Applicata* **30**(3): 280–292.

Menken, S.B.J. (1996). Pattern and process in the evolution of insect-plant associations: *Yponomeuta* as an example. *Entomologia Experimentalis Et Applicata* **80**(1): 297–305.

Merckx, V., Chatrou, L., Lemaire, B., Sainge, M., Huysmans, S. and Smets, E. (2008). Diversification of myco-heterotrophic angiosperms: Evidence from Burmanniaceae. *BMC Evolutionary Biology* **8**(1): 178.

Meusnier, I., Singer, G., Landry, J.-F., Hickey, D., Hebert, P. and Hajibabaei, M. (2008). A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics* **9**(1): 214.

Meyer, C.P. and Paulay, G. (2005). DNA barcoding: Error rates based on comprehensive sampling. *PLoS Biology* **3**(12): 2229–2238.

Michalak, I., Zhang, L.-B. and Renner, S.S. (2010). Trans-Atlantic, trans-Pacific and trans-Indian Ocean dispersal in the small Gondwanan Laurales family Hernandiaceae. *Journal of Biogeography* **37**(7): 1214–1226.

Miles, C.I., del Campo, M.L. and Renwick, J.A.A. (2005). Behavioral and chemosensory responses to a host recognition cue by larvae of *Pieris rapae. Journal of Comparative Physiology A* **191**(2): 147–155.

Miller, M.A., Pfeiffer, W. and Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *Gateway Computing Environments Workshop (GCE), 2010*: 1–8.

Milne, R.I. (2009). Effects of taxon sampling on molecular dating for within-genus divergence events, when deep fossils are used for calibration. *Journal of Systematics and Evolution* **47**(5): 383–401.

Minin, V.N. and Suchard, M.A. (2008). Counting labeled transitions in continuous-time Markov models of evolution. *Journal of Mathematical Biology* **56**(3): 391–412.

Mitter, C., Farrell, B. and Futuyma, D.J. (1991). Phylogenetic studies of insect plant interactions - insights into the genesis of diversity. *Trends in Ecology & Evolution* **6**(9): 290–293.

Mitter, C., Farrell, B. and Wiegmann, B. (1988). The Phylogenetic study of adaptive zones - Has phytophagy promoted insect diversification? *American Naturalist* **132**(1): 107–128.

Monaghan, M.T., Wild, R., Elliot, M., Fujisawa, T., Balke, M., Inward, D.J.G., Lees, D.C., Ranaivosolo, R., Eggleton, P., Barraclough, T.G. and Vogler, A.P. (2009). Accelerated species inventory on Madagascar using coalescent-based models of species delineation. *Systematic Biology* **58**(3): 298–311.

Monzón-Argüello, C., López-Jurado, L.F., Rico, C., Marco, A., López, P., Hays, G.C. and Lee, P.L.M. (2010). Evidence from genetic and Lagrangian drifter data for transatlantic transport of small juvenile green turtles. *Journal of Biogeography* **37**(9): 1752–1766.

Morando, M., Avila, L.J. and Sites, J.W. (2003). Sampling strategies for delimiting species: Genes, individuals, and populations in the *Liolaemus elongatus-kriegi* complex (Squamata: Liolaemidae) in Andean–Patagonian South America. *Systematic Biology* **52**(2): 159–185.

Morin, X. and Thuiller, W. (2009). Comparing niche- and process-based models to reduce prediction uncertainty in species range shifts under climate change. *Ecology* **90**(5): 1301–1313.

Moritz, C. and Cicero, C. (2004). DNA barcoding: Promise and pitfalls. *PLoS Biology* **2**(10): 1529–1531.

Morley, R.J. and Richards, K. (1993). Gramineae cuticle: a key indicator of Late Cenozoic climatic change in the Niger Delta. *Review of Palaeobotany and Palynology* **77**(1–2): 119–127.

Mullen, S.P., Savage, W.K., Wahlberg, N. and Willmott, K.R. (2011). Rapid diversification and not clade age explains high diversity in neotropical *Adelpha* butterflies. *Proceedings of the Royal Society B-Biological Sciences* **278**(1713): 1777–1785.

Müller, C.J. and Beheregaray, L.B. (2010). Palaeo island-affinities revisited - Biogeography and systematics of the Indo-Pacific genus *Cethosia* Fabricius (Lepidoptera: Nymphalidae). *Molecular Phylogenetics and Evolution* **57**(1): 314–326.

Munch, K., Boomsma, W., Willerslev, E. and Nielsen, R. (2008). Fast phylogenetic DNA barcoding. *Philosophical Transactions of the Royal Society B-Biological Sciences* **363**(1512): 3997–4002.

Mwavu, E.N. and Witkowski, E.T.F. (2009). Seedling regeneration, environment and management in a semi-deciduous African tropical rain forest. *Journal of Vegetation Science* **20**(5): 791–804.

N'guessan, K., Soro, D. and Amon, A. (2011). Plantes utilisées en médecine traditionnelle dans le traitement des maladies cardiovasculaires, en pays Abbey et Krobou, dans le Sud de la Côte-d'Ivoire. *Phytothérapie* **9**(4): 199–208.

Nakagawa, S. and Freckleton, R.P. (2008). Missing inaction: the dangers of ignoring missing data. *Trends in Ecology & Evolution* **23**(11): 592–596.

Narita, S., Nomura, M., Kato, Y. and Fukatsu, T. (2006). Genetic structure of sibling butterfly species affected by *Wolbachia* infection sweep: evolutionary and biogeographical implications. *Molecular Ecology* **15**(4): 1095–1108.

Nazari, V. and Sperling, F.A.H. (2007). Mitochondrial DNA divergence and phylogeography in western Palaearctic Parnassiinae (Lepidoptera : Papilionidae): How many species are there? *Insect Systematics & Evolution* **38**(2): 121–138.

Neigel, J., Domingo, A. and Stake, J. (2007). DNA barcoding as a tool for coral reef conservation. *Coral Reefs* **26**(3): 487–499.

Neuwinger, H.D. (2000). African traditional medicine: a dictionary of plant use and applications. Stuttgart, Medpharm Scientific Publishers.

Newmaster, S.G. and Ragupathy, S. (2009). Testing plant barcoding in a sister species complex of pantropical *Acacia* (Mimosoideae, Fabaceae). *Molecular Ecology Resources* **9**: 172–180.

Nichols, R. (2001). Gene trees and species trees are not the same. *Trends in Ecology & Evolution* **16**(7): 358–364.

Nielsen, R. and Matz, M. (2006). Statistical approaches for DNA barcoding. *Systematic Biology* **55**(1): 162–169.

Nishida, R. (2002). Sequestration of defensive substances from plants by Lepidoptera. *Annual Review of Entomology* **47**: 57–92.

Nishida, R. (2005). Chemosensory basis of host recognition in butterflies - Multi-component system of oviposition stimulants and deterrents. *Chemical Senses* **30**(Suppl 1): i293–i294.

Nolan, D.V., Carpenter, S., Barber, J., Mellor, P.S., Dallas, J.F., Mordue, A.J. and Piertney, S.B. (2007). Rapid diagnostic PCR assays for members of the *Culicoides obsoletus* and *Culicoides pulicaris* species complexes, implicated vectors of bluetongue virus in Europe. *Veterinary Microbiology* **124**(1-2): 82–94.

Nosil, P., Crespi, B.J. and Sandoval, C.P. (2002). Host-plant adaptation drives the parallel evolution of reproductive isolation. *Nature* **417**(6887): 440–443.

Novotny, V., Drozd, P., Miller, S.E., Kulfan, M., Janda, M., Basset, Y. and Weiblen, G.D. (2006). Why are there so many species of herbivorous insects in tropical rainforests? *Science* **313**(5790): 1115–1118.

Novotny, V., Miller, S.E., Baje, L., Balagawi, S., Basset, Y., Cizek, L., Craft, K.J., Dem, F., Drew, R.A.I., Hulcr, J., Leps, J., Lewis, O.T., Pokon, R., Stewart, A.J.A., Allan Samuelson, G. and Weiblen, G.D. (2010). Guild-specific patterns of species richness and host specialization in plant–herbivore food webs from a tropical forest. *Journal of Animal Ecology* **79**(6): 1193–1203.

Nylander, J.A.A., Wilgenbusch, J.C., Warren, D.L. and Swofford, D.L. (2008). AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics* **24**(4): 581–583.

Nyman, T. (2010). To speciate, or not to speciate? Resource heterogeneity, the subjectivity of similarity, and the macroevolutionary consequences of niche-width shifts in plant-feeding insects. *Biological Reviews* **85**(2): 393–411.

Nyman, T., Linder, H.P., Pena, C., Malm, T. and Wahlberg, N. (2012). Climate-driven diversity dynamics in plants and plant-feeding insects. *Ecology Letters* **15**(8): 889–898.

Nyman, T., Vikberg, V., Smith, D.R. and Boeve, J.L. (2010). How common is ecological speciation in plant-feeding insects? A 'higher' Nematinae perspective. *BMC Evolutionary Biology* **10**.

Oates John, F., Bergl Richard, A. and Linder Joshua, M. (2004). Africa's Gulf of Guinea forests: Biodiversity patterns and conservation priorities. Washington, Conservation International.

Obando-Rodríguez, J.A., Bemis, E.D., Jackson, B.C., ed., Budd, A.F., ed. and Coates, A.G., ed. (1996). The Geologic evolution of the Central American isthmus, Chicago, IL, The University of Chicago Press, US.

Oberprieler, C. (2005). Temporal and spatial diversification of circum-Mediterranean Compositae-Anthemideae. *Taxon* **54**(4): 951–966.

Ødegaard, F. (2000). How many species of arthropods? Erwin's estimate revised. *Biological Journal of the Linnean Society* **71**(4): 583–597.

Oerke, E.C. (2006). Crop losses to pests. *The Journal of Agricultural Science* **144**(01): 31–43.

Ohsaki, N. and Sato, Y. (1994). Food plant choice of *Pieris* butterflies as a trade-off between parasitoid avoidance and quality of plants. *Ecology* **75**(1): 59–68.

Ovadia, Y., Fielder, D., Conow, C. and Libeskind-Hadas, R. (2011). The Cophylogeny reconstruction problem is NP-complete. *Journal of Computational Biology* **18**(1): 59–65.

Overlaet, F.G. (1952). Formes nouvelles ou peu connues de Nymphalides africains. *Annales du Musée Royal du Congo Belge* **ser. 8**(14): 1–50.

Padial, J., Miralles, A., De la Riva, I. and Vences, M. (2010). The Integrative future of taxonomy. *Frontiers in Zoology* **7**(1): 16.

Pagán, I., Firth, C. and Holmes, E.C. (2010). Phylogenetic analysis reveals rapid evolutionary dynamics in the plant RNA virus genus *Tobamovirus*. **71**(4): 298–307.

Page, R.D.M. (2003). Tangled trees: phylogeny, cospeciation, and coevolution. Chicago, University of Chicago press.

Paradis, E., Claude, J. and Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**(2): 289–290.

Paredes-Esquivel, C., Donnelly, M.J., Harbach, R.E. and Townson, H. (2009). A molecular phylogeny of mosquitoes in the *Anopheles barbirostris* subgroup reveals cryptic species: Implications for identification of disease vectors. *Molecular Phylogenetics and Evolution* **50**(1): 141–151.

Patterson, D.J., Cooper, J., Kirk, P.M., Pyle, R.L. and Remsen, D.P. (2010). Names are key to the big new biology. *Trends in Ecology & Evolution* **25**(12): 686–691.

Peek, A.S., Feldman, R.A., Lutz, R.A. and Vrijenhoek, R.C. (1998). Cospeciation of chemoautotrophic bacteria and deep sea clams. *Proceedings of the National Academy of Sciences* **95**(17): 9962–9966.

Peña, C., Nylin, S. and Wahlberg, N. (2011). The Radiation of Satyrini butterflies (Nymphalidae: Satyrinae): a challenge for phylogenetic methods. *Zoological Journal of the Linnean Society* **161**(1): 64–87.

Peña, C. and Wahlberg, N. (2008). Prehistorical climate change increased diversification of a group of butterflies. *Biology Letters* **4**(3): 274–278.

Percy, D.M., Page, R.D.M. and Cronk, Q.C.B. (2004). Plant-insect interactions: Double-dating associated insect and plant lineages reveals asynchronous radiations. *Systematic Biology* **53**(1): 120–127.

Petit, N. and Barbadilla, A. (2009). Selection efficiency and effective population size in *Drosophila* species. *Journal of Evolutionary Biology* **22**(3): 515 – 526.

Petit, R.J. and Excoffier, L. (2009). Gene flow and species delimitation. *Trends in Ecology & Evolution* **24**(7): 386–393.

Phil-Eze, P.O. and Okoro, I.C. (2009). Sustainable biodiversity conservation in the Niger Delta: a practical approach to conservation site selection. **18**(5): 1247–1257.

Pieterse, C.M.J., van der Does, D., Zamioudis, C., Leon-Reyes, A. and Van Wees, S.C.M. (2012). Hormonal modulation of plant immunity. *Annual Review of Cell and Developmental Biology* **28**(1): 489–521.

Pillon, Y., Fay, M.F., Shipunov, A.B. and Chase, M.W. (2006). Species diversity versus phylogenetic diversity: A practical study in the taxonomically difficult genus *Dactylorhiza* (Orchidaceae). *Biological Conservation* **129**(1): 4–13.

Pirie, M.D. and Doyle, J.A. (2012). Dating clades with fossils and molecules: the case of Annonaceae. *Botanical Journal of the Linnean Society* **169**(1): 84–116.

Plana, V., Gascoigne, A., Forrest, L.L., Harris, D. and Pennington, R.T. (2004). Pleistocene and pre-pleistocene *Begonia* speciation in Africa. *Molecular Phylogenetics and Evolution* **31**(2): 449–461.

Plummer, M., Best, N., Cowles, K. and Vines, K. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R News* **6**(1): 7–11.

Pons, J., Barraclough, T.G., Gomez-Zurita, J., Cardoso, A., Duran, D.P., Hazell, S., Kamoun, S., Sumlin, W.D. and Vogler, A.P. (2006). Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology* **55**(4): 595–609.

Pope, K.O., Baines, K.H., Ocampo, A.C. and Ivanov, B.A. (1997). Energy, volatile production, and climatic effects of the Chicxulub Cretaceous/Tertiary impact. *Journal of Geophysical Research-Planets* **102**(E9): 21645–21664.

Popescu, A.-A., Huber, K.T. and Paradis, E. (2012). Ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. *Bioinformatics* **28**(11): 1536–1537.

Popper, K.R. (1963). Conjectures and refutations: The growth of scientific knowledge. London, Routledge.

Posada, D. (2008). jModelTest: Phylogenetic model averaging. *Molecular Biology and Evolution* **25**(7): 1253–1256.

Poux, C., Madsen, O., Marquard, E., Vieites, D.R., de Jong, W.W. and Vences, M. (2005). Asynchronous colonization of Madagascar by the four endemic clades of primates, tenrecs, carnivores, and rodents as inferred from nuclear genes. *Systematic Biology* **54**(5): 719–730.

Powell, C.B. (1997). Discoveries and priorities for mammals in the freshwater forests of the Niger Delta. *Oryx* **31**(2): 83–85.

Price, B.W., Villet, M.H., Walton, S.M. and Barker, N.P. (2011). Using molecules and morphology to infer the phylogenetic relationships and evolutionary history of the Dirini (Nymphalidae: Satyrinae), a tribe of butterflies endemic to Southern Africa. *Systematic Entomology* **36**(2): 300–316.

Price, P.W. (2002). Resource-driven terrestrial interaction webs. *Ecological Research* **17**(2): 241–247.

Pringle, E., Henning, G. and Ball, J. (1994). Pennington's butterflies of southern Africa. Cape Town, Struik Winchester.

Puillandre, N., Lambert, A., Brouillet, S. and Achaz, G. (2012). ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology* **21**(8): 1864–1877.

Puillandre, N., Modica, M.V., Zhang, Y., Sirovich, L., Boisselier, M.C., Cruaud, C., Holford, M. and Samadi, S. (2012). Large-scale species delimitation method for hyperdiverse groups. *Molecular Ecology* **21**(11): 2671–2691.

Purvis, A., Nee, S. and Harvey, P.H. (1995). Macroevolutionary inferences from primate phylogeny. *Proceedings of the Royal Society of London Series B-Biological Sciences* **260**(1359): 329–333.

Pybus, O.G. and Harvey, P.H. (2000). Testing macro-evolutionary models using incomplete molecular phylogenies. *Proceedings of the Royal Society of London Series B-Biological Sciences* **267**(1459): 2267–2272.

Queiroz, A.d. (2005). The Resurrection of oceanic dispersal in historical biogeography. *Trends in Ecology & Evolution* **20**(2): 68–73.

Rabosky, D.L. (2006). LASER: A maximum likelihood toolkit for detecting temporal shifts in diversification rates from molecular phylogenies. *Evolutionary Bioinformatics* **2**: 247–250.

Rabosky, D.L. (2006). Likelihood methods for detecting temporal shifts in diversification rates. *Evolution* **60**(6): 1152–1164.

Rabosky, D.L. (2010). Extinction rates should not be estimated from molecular phylogenies. *Evolution* **64**(6): 1816–1824.

Rabosky, D.L., Slater, G.J. and Alfaro, M.E. (2012). Clade age and species richness are decoupled across the Eukaryotic tree of life. *PLoS Biol* **10**(8): e1001381.

Rach, J., DeSalle, R., Sarkar, I.N., Schierwater, B. and Hadrys, H. (2008). Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. *Proceedings of the Royal Society B* **275**(1632): 237–247.

Rannala, B. (2002). Identifiability of parameters in MCMC Bayesian inference of phylogeny. *Systematic Biology* **51**(5): 754–760.

Ratnasingham, S. and Hebert, P.D.N. (2007). BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes* **7**(3): 355–364.

Raubenheimer, D. and Elsworth, J.F. (1988). Gynocardin from the leaves of *Kiggelaria africana*. *Journal of Natural Products* **51**(4): 779–779.

Raychoudhury, R., Baldo, L., Oliveira, D.C.S.G. and Werren, J.H. (2009). Modes of acquisition of Wolbachia: horizontal transfer, hybrid introgression, and codivergence in the *Nasonia* species complex. *Evolution* **63**(1): 165–183.

Reid, B.N., Le, M., McCord, W.P., Iverson, J.B., Georges, A., Bergmann, T., Amato, G., Desalle, R. and Naro-Maciel, E. (2011). Comparing and combining distance-based and character-based approaches for barcoding turtles. *Molecular Ecology Resources* **11**(6): 956–967.

Renne, P.R., Deino, A.L., Hilgen, F.J., Kuiper, K.F., Mark, D.F., Mitchell, W.S., Morgan, L.E., Mundil, R. and Smit, J. (2013). Time scales of critical events around the Cretaceous-Paleogene boundary. *Science* **339**(6120): 684–687.

Renner, S.S. (2004). Plant dispersal across the tropical Atlantic by wind and sea currents. *International Journal of Plant Sciences* **165**(4): S23–S33.

Renner, S.S., Clausing, G. and Meyer, K. (2001). Historical biogeography of Melastomataceae: the roles of Tertiary migration and long-distance dispersal. *American Journal of Botany* **88**(7): 1290–1300.

Renner, S.S. and Meyer, K. (2001). Melastomeae come full circle: biogeographic reconstruction and molecular clock dating. *Evolution* **55**(7): 1315–1324.

Renwick, J.A.A. and Chew, F.S. (1994). Oviposition behavior in Lepidoptera. *Annual Review of Entomology* **39**: 377–400.

Restif, O. and Koella, J.C. (2003). Shared control of epidemiological traits in a coevolutionary model of host-parasite interactions. *American Naturalist* **161**(6): 827–836.

Richardson, J.E., Chatrou, L.W., Mols, J.B., Erkens, R.H.J. and Pirie, M.D. (2004). Historical biogeography of two cosmopolitan families of flowering plants: Annonaceae and Rhamnaceae. *Philosophical Transactions of the Royal Society B-Biological Sciences* **359**(1450): 1495–1508.

Robe, L., Cordeiro, J., Loreto, E. and Valente, V. (2010). Taxonomic boundaries, phylogenetic relationships and biogeography of the *Drosophila willistoni* subgroup (Diptera: Drosophilidae). *Genetica* **138**(6): 601–617.

Robson, N.K.B. (1960). Violaceae. *Flora Zambesiaca* **1**(1): 246–260.

Rokas, A., Krüger, D. and Carroll, S.B. (2005). Animal evolution and the molecular signature of radiations compressed in time. *Science* **310**(5756): 1933–1938.

Rommerskirchen, F., Condon, T., Mollenhauer, G., Dupont, L. and Schefuss, E. (2011). Miocene to Pliocene development of surface and subsurface temperatures in the Benguela Current system. *Paleoceanography* **26**(3): PA3216.

Ronquist, F. and Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**(12): 1572–1574.

Ronquist, F. and Liljeblad, J. (2001). Evolution of the gall wasp-host plant association. *Evolution* **55**(12): 2503–2522.

Roquet, C., Sanmartín, I., Garcia-Jacas, N., Sáez, L., Susanna, A., Wikström, N. and Aldasoro, J.J. (2009). Reconstructing the history of Campanulaceae with a Bayesian approach to molecular dating and dispersal–vicariance analyses. *Molecular Phylogenetics and Evolution* **52**(3): 575–587.

Ross, H.A., Murugan, S. and Li, W.L.S. (2008). Testing the reliability of genetic methods of species identification via simulation. *Systematic Biology* **57**(2): 216–230.

Saitou, N. and Nei, M. (1987). The Neighbor joining method - a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**(4): 406–425.

Sammons, A.E., Edwards, C.R., Bledsoe, L.W., Boeve, P.J. and Stuart, J.J. (1997). Behavioral and feeding assays reveal a western corn rootworm (Coleoptera : Chrysomelidae) variant that is attracted to soybean. *Environmental Entomology* **26**(6): 1336–1342.

Sarkar, I.N., Planet, P.J., Bael, T.E., Stanley, S.E., Siddall, M., DeSalle, R. and Figurski, D.H. (2002). Characteristic attributes in cancer microarrays. *Journal of Biomedical Informatics* **35**(2): 111–122.

Sarkar, I.N., Planet, P.J. and Desalle, R. (2008). CAOS software for use in character-based DNA barcoding. *Molecular Ecology Resources* **8**(6): 1256–1259.

Sarkar, I.N., Thornton, J.W., Planet, P.J., Figurski, D.H., Schierwater, B. and DeSalle, R. (2002). An automated phylogenetic key for classifying homeoboxes. *Molecular Phylogenetics and Evolution* **24**(3): 388–399.

Sarkar, I.N. and Trizna, M. (2011). The Barcode of life data portal: bridging the biodiversity informatics divide for DNA barcoding. *PLoS ONE* **6**(7): e14689.

Savchenko, T., Pearse, I.S., Ignatia, L., Karban, R. and Dehesh, K. (2013). Insect herbivores selectively suppress the HPL branch of the oxylipin pathway in host plants. *The Plant Journal* **73**(4): 653–662.

Savostin, L.A., Sibuet, J.C., Zonenshain, L.P., Lepichon, X. and Roulet, M.J. (1986). Kinematic evolution of the Tethys Belt from the Atlantic Ocean to the Pamirs since the Triassic. *Tectonophysics* **123**(1–4): 1–35.

Schindel, D.E. and Miller, S.E. (2005). DNA barcoding a useful tool for taxonomists. *Nature* **435**(7038): 17–17.

Schlick-Steiner, B.C., Steiner, F.M., Seifert, B., Stauffer, C., Christian, E. and Crozier, R.H. (2010). Integrative taxonomy: A multisource approach to exploring biodiversity. *Annual Review of Entomology* **55**(1): 421–438.

Schmidt, B.C. and Sperling, F.A.H. (2008). Widespread decoupling of mtDNA variation and species integrity in *Grammia* tiger moths (Lepidoptera: Noctuidae). *Systematic Entomology* **33**(4): 613–634.

Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., Chen, W. and Consortium, F.B. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences* **109**(16): 6241–6246.

Schoonhoven, L.M., van Loon, J.J.A. and Dicke, M. (2005). Insect-plant biology. Oxford, Oxford University Press.

Schouten, M.W., de Ruijter, W.P.M., van Leeuwen, P.J. and Ridderinkhof, H. (2003). Eddies and variability in the Mozambique Channel. *Deep-Sea Research Part II: Topical Studies in Oceanography* **50**(12–13): 1987–2003.

Schulte, P., Alegret, L., Arenillas, I., Arz, J.A., Barton, P.J., Bown, P.R., Bralower, T.J., Christeson, G.L., Claeys, P., Cockell, C.S., Collins, G.S., Deutsch, A., Goldin, T.J., Goto, K., Grajales-Nishimura, J.M., Grieve, R.A.F., Gulick, S.P.S., Johnson, K.R., Kiessling, W., Koeberl, C., Kring, D.A., MacLeod, K.G., Matsui, T., Melosh, J., Montanari, A., Morgan, J.V., Neal, C.R., Nichols, D.J., Norris, R.D., Pierazzo, E., Ravizza, G., Rebolledo-Vieyra, M., Reimold, W.U., Robin, E., Salge, T., Speijer, R.P., Sweet, A.R., Urrutia-Fucugauchi, J., Vajda, V., Whalen, M.T. and Willumsen, P.S. (2010). The Chicxulub asteroid impact and mass extinction at the Cretaceous-Paleogene boundary. *Science* **327**(5970): 1214–1218.

Segalen, L., Lee-Thorp, J.A. and Cerling, T. (2007). Timing of C-4 grass expansion across sub-Saharan Africa. *Journal of Human Evolution* **53**(5): 549–559.

Senut, B., Pickford, M. and Segalen, L. (2009). Neogene desertification of Africa. *Comptes Rendus Geosciences* **341**(8–9): 591–602.

Seo, T.-K. (2010). Classification of nucleotide sequences using support vector machines. *Journal of Molecular Evolution* **71**(4): 250–267.

Sepulchre, P., Ramstein, G., Fluteau, F., Schuster, M., Tiercelin, J.J. and Brunet, M. (2006). Tectonic uplift and Eastern Africa aridification. *Science* **313**(5792): 1419–1423.

Shaffer, H.B. and Thomson, R.C. (2007). Delimiting species in recent radiations. *Systematic Biology* **56**(6): 896–906.

Shapiro, E., Biezuner, T. and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature reviews. Genetics* **14**(9).

Silva, D.N., Talhinhas, P., Cai, L.E.I., Manuel, L., Gichuru, E.K., Loureiro, A., Várzea, V., Paulo, O.S. and Batista, D. (2012). Host-jump drives rapid and recent ecological speciation of the emergent fungal pathogen *Colletotrichum kahawae. Molecular Ecology* **21**(11): 2655–2670.

Simpson, G.G. (1953). The Major features of evolution. New York, Colombia University Press.

Singer, M.C. and McBride, C.S. (2010). Multitrait, host-associated divergence among sets of butterfly populations: implications for reproductive isolation and ecological speciation. *Evolution* **64**(4): 921–933.

Skoracka, A. and Dabert, M. (2010). The Cereal rust mite *Abacarus hystrix* (Acari: Eriophyoidea) is a complex of species: evidence from mitochondrial and nuclear DNA sequences. *Bulletin of Entomological Research* **100**(3): 263–272.

Slove, J. and Janz, N. (2011). The Relationship between diet breadth and geographic range size in the butterfly subfamily Nymphalinae – A study of global scale. *PLoS ONE* **6**(1): e16057.

Smith, C.I., Godsoe, W.K.W., Tank, S., Yoder, J.B. and Pellmyr, O. (2008). Distinguishing coevolution from covariance in an obligate pollination mutualism: asynchronous divergence in Joshua tree and its pollinators. *Evolution* **62**(10): 2676–2687.

Smith, M.A., Bertrand, C., Crosby, K., Eveleigh, E.S., Fernandez-Triana, J., Fisher, B.L., Gibbs, J., Hajibabaei, M., Hallwachs, W., Hind, K., Hrcek, J., Huang, D.-W., Janda, M., Janzen, D.H., Li, Y., Miller, S.E., Packer, L., Quicke, D., Ratnasingham, S., Rodriguez, J., Rougerie, R., Shaw, M.R., Sheffield, C., Stahlhut, J.K., Steinke, D., Whitfield, J., Wood, M. and Zhou, X. (2012). *Wolbachia* and DNA barcoding insects: Patterns, potential, and problems. *PLoS ONE* **7**(5): e36514.

Sole, C.L., Wirta, H., Forgie, S.A. and Scholtz, C.H. (2011). Origin of Madagascan Scarabaeini dung beetles (Coleoptera: Scarabaeidae): dispersal from Africa. *Insect Systematics & Evolution* **42**(1): 29–40.

Soltis, D.E., Albert, V.A., Leebens-Mack, J., Bell, C.D., Paterson, A.H., Zheng, C.F., Sankoff, D., dePamphilis, C.W., Wall, P.K. and Soltis, P.S. (2009). Polyploidy and Angiosperm diversification. *American Journal of Botany* **96**(1): 336–348.

Sorenson, M.D., Balakrishnan, C.N. and Payne, R.B. (2004). Clade-limited colonization in brood parasitic finches (Vidua spp.). *Systematic Biology* **53**(1): 140–153.

Sosef, M.S.M., Wieringa, J.J., Jongkind, C.C.H., Achoundong, G., Azizet Issembé, Y., Bedigian, D., van den Berg, R.G., Breteler, F.J., Cheek, M., Degreef, J., Faden, R.B., Goldblatt, P., van der Maesen, L.J.G., Ngok Banak, L., Niangadouma, R., Nzabi, T., Nziengui, B., Rogers, Z.S., Stévart, T., van Valkenburg, J.L.C.H., Walters, G. and de Wilde, J.J.F.E. (2006). Checklist of Gabonese vascular plants. *Scripta Botanica Belgica* **35**: 438.

Spencer, K.C. and Seigler, D.S. (1985). Cyanogenic glycosides and the systematics of the Flacourtiaceae. *Biochemical Systematics and Ecology* **13**(4): 421–431.

Stamatakis, A., Hoover, P. and Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML web servers. *Systematic Biology* **57**(5): 758–771.

Steel, M. and McKenzie, A. (2001). Properties of phylogenetic trees generated by Yule-type speciation models. *Mathematical Biosciences* **170**(1): 91–112.

Stewart, M., Bartholomew, B., Currie, F., Abbiw, D.K., Latif, Z., Sarker, S.D. and Nash, R.J. (2000). Pyranoisoflavones from *Rinorea welwitschii*. *Fitoterapia* **71**(5): 595–597.

Stireman, J.O., Nason, J.D. and Heard, S.B. (2005). Host-associated genetic differentiation in phytophagous insects: General phenomenon or isolated exceptions? Evidence from a goldenrod-insect community. *Evolution* **59**(12): 2573–2587.

Stone, G.N., Hernandez-Lopez, A., Nicholls, J.A., Di Pierro, E., Pujade-Villar, J., Melika, G. and Cook, J.M. (2009). Extreme host plant conservatism during at least 20 million years of host plant pursuit by oak gallwasps. *Evolution* **63**(4): 854–869.

Stouthamer, R., Breeuwer, J.A.J. and Hurst, G.D.D. (1999). *Wolbachia pipientis*: Microbial manipulator of arthropod reproduction. *Annual Review of Microbiology* **53**: 71–102.

Strutzenberger, P. and Fiedler, K. (2011). Temporal patterns of diversification in Andean Eois, a species-rich clade of moths (Lepidoptera, Geometridae). *Journal of Evolutionary Biology* **24**(4): 919–925.

Stur, E. and Ekrem, T. (2011). Exploring unknown life stages of Arctic Tanytarsini (Diptera: Chironomidae) with DNA barcoding. *Zootaxa*(2743): 27–39.

Sung, G.-H., Poinar Jr, G.O. and Spatafora, J.W. (2008). The Oldest fossil evidence of animal parasitism by fungi supports a Cretaceous diversification of fungal–arthropod symbioses. *Molecular Phylogenetics and Evolution* **49**(2): 495–502.

Taberlet, P. (1991). Universal primers for amplification of 3 noncoding regions of chloroplast DNA. *Plant Molecular Biology* **17**(5): 1105–1109.

Taton, A. (1968). *Rinorea* nouveaux (Violaceae) de Congo-Kinshasa. *Bulletin du Jardin Botanique National de Belgique* **38**: 381–391.

Tchouto, M.G.P., de Wilde, J.J.F.E., de Boer, W.F., van der Maesen, L.J.G. and Cleef, A.M. (2009). Bio-indicator species and Central African rain forest refuges in the Campo-Ma'an area, Cameroon. *Systematics and Biodiversity* **7**(1): 21–31.

Templeton, A.R., Crandall, K.A. and Sing, C.F. (1992). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* **132**(2): 619–633.

Tennant, J.R. (1963). Notes on tropical African Violaceae. *Kew Bulletin* **16**: 409–435.

Thompson, J.N. and Pellmyr, O. (1991). Evolution of oviposition behavior and host preference in Lepidoptera. *Annual Review of Entomology* **36**: 65–89.

Tokuoka, T. (2008). Molecular phylogenetic analysis of Violaceae (Malpighiales) based on plastid and nuclear DNA sequences. *Journal of Plant Research* **121**(3): 253–260.

Tokuoka, T. and Tobe, H. (2006). Phylogenetic analyses of Malpighiales using plastid and nuclear DNA sequences, with particular reference to the embryology of Euphorbiaceae sens. str. *Journal of Plant Research* **119**(6): 599–616.

Trewick, S.A. (2008). DNA Barcoding is not enough: mismatch of taxonomy and genealogy in New Zealand grasshoppers (Orthoptera: Acrididae). *Cladistics* **24**(2): 240–254.

Tsuchihara, K., Hisatomi, O., Tokunaga, F. and Asaoka, K. (2009). An oviposition stimulant binding protein in a butterfly: Immunohistochemical localization and electrophysiological responses to plant compounds. *Communicative & Integrative Biology* **2**(4): 356–358.

van Someren, V.G.L. (1974). List of foodplants of some east african rhopalocera, with notes on the early stages of some Lycaenidae. *Journal of the Lepidopterists' Society* **28**(4): 315–333.

van Son, G. (1979). The Butterflies of Southern Africa, part 4, Nymphalidae: Nymphalinae. Pretoria, Transvaal Museum.

van Velzen, R., Bakker, F.T. and van Loon, J.J.A. (2007). DNA barcoding reveals hidden species diversity in *Cymothoe* (Nymphalidae) *Proceedings of the Netherlands Entomological Society Meeting* **18**: 95–103.

van Velzen, R., Larsen, T.B. and Bakker, F.T. (2009). A new hidden species of the *Cymothoe caenis*-complex (Lepidoptera: Nymphalidae) from western Africa. *Zootaxa* **2197**: 53–63.

van Velzen, R., Wahlberg, N., Sosef, M.S.M. and Bakker, F.T. (2013). Effects of changing climate on species diversification in tropical forest butterflies of the genus *Cymothoe* (Lepidoptera: Nymphalidae). *Biological Journal of the Linnean Society* **108**(3): 546–564.

van Velzen, R., Weitschek, E., Felici, G. and Bakker, F.T. (2012). DNA barcoding of recently diverged species: relative performance of matching methods. *PLoS ONE* **7**(1): e30490.

Vande weghe, G.R. (2010). Les papillons du Gabon. Libreville, Wildlife Conservation Society.

Vande weghe, G.R. (2011). Description d'une nouvelle espèce de *Cymothoe* Hübner, 1819 de République démocratique du Congo (Lepidoptera, Nymphalidae, Limenitinae). *Lambillionea* **110**(3): 305–306.

Varma, A. and Malathi, V.G. (2003). Emerging geminivirus problems: A serious threat to crop production. *Annals of Applied Biology* **142**(2): 145–164.

Vences, M., Wollenberg, K.C., Vieites, D.R. and Lees, D.C. (2009). Madagascar as a model region of species diversification. *Trends in Ecology & Evolution* **24**(8): 456–465.

Vet, L.E.M. and Dicke, M. (1992). Ecology of infochemical use by natural enemies in a tritrophic context. *Annual Review of Entomology* **37**(1): 141–172.

Virgilio, M., Backeljau, T., Nevado, B. and De Meyer, M. (2010). Comparative performances of DNA barcoding across insect orders. *BMC Bioinformatics* **11**: 206.

Wagner, T. (2001). Seasonal changes in the canopy arthropod fauna in *Rinorea beniensis* in Budongo Forest, Uganda. *Plant Ecology* **153**(1): 169–178.

Wahlberg, N. (2006). That awkward age for butterflies: Insights from the age of the butterfly subfamily Nymphalinae (Lepidoptera : Nymphalidae). *Systematic Biology* **55**(5): 703–714.

Wahlberg, N. and Freitas, A.V.L. (2007). Colonization of and radiation in South America by butterflies in the subtribe Phyciodina (Lepidoptera : Nymphalidae). *Molecular Phylogenetics and Evolution* **44**(3): 1257–1272.

Wahlberg, N., Leneveu, J., Kodandaramaiah, U., Peña, C., Nylin, S., Freitas, A.V.L. and Brower, A.V.Z. (2009). Nymphalid butterflies diversify following near demise at the Cretaceous/Tertiary boundary. *Proceedings of the Royal Society B* **276**(1677): 4295–4302.

Wahlberg, N. and Wheat, C.W. (2008). Genomic outposts serve the phylogenomic pioneers: Designing novel nuclear markers for genomic DNA extractions of lepidoptera. *Systematic Biology* **57**(2): 231–242.

Wahlert, G.E. (2010). Phylogeny, biogeography, and a taxonomic revision of *Rinorea* (Violaceae) from Madagascar and the Comoro Islands PhD, Ohio University.

Wahlert, G.E. and Ballard, H.E., Jr. (2012). A phylogeny of *Rinorea* (Violaceae) inferred from plastid DNA sequences with an emphasis on the African and Malagasy species. *Systematic Botany* **37**(4): 964–973.

Wallman, J.F. and Donnellan, S.C. (2001). The Utility of mitochondrial DNA sequences for the identification of forensically important blowflies (Diptera: Calliphoridae) in southeastern Australia. *Forensic Science International* **120**(1–2): 60–67.

Wang, B. and Qiu, Y.L. (2006). Phylogenetic distribution and evolution of mycorrhizas in land plants. *Mycorrhiza* **16**(5): 299–363.

Wang, H., Moore, M.J., Soltis, P.S., Bell, C.D., Brockington, S.F., Alexandre, R., Davis, C.C., Latvis, M., Manchester, S.R. and Soltis, D.E. (2009). Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proceedings of the National Academy of Sciences* **106**(10): 3853–3858.

Ward, R.D., Costa, F.O., Holmes, B.H. and Steinke, D. (2008). DNA barcoding of shared fish species from the North Atlantic and Australasia: minimal divergence for most taxa, but *Zeus faber* and *Lepidopus caudatus* each probably constitute two species. *Aquatic Biology* **3**(1): 71–78.

Ward, R.D., Hanner, R. and Hebert, P.D.N. (2009). The campaign to DNA barcode all fishes, FISH-BOL. *Journal of Fish Biology* **74**(2): 329–356.

Ward, R.D., Zemlak, T.S., Innes, B.H., Last, P.R. and Hebert, P.D.N. (2005). DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society B-Biological Sciences* **360**(1462): 1847–1857.

Webb, C.O. (2000). Exploring the phylogenetic structure of ecological communities: An example for rain forest trees. *Am Nat* **156**(2): 145–155.

Weech, M.-H., Chapleau, M., Pan, L., Ide, C. and Bede, J.C. (2008). Caterpillar saliva interferes with induced *Arabidopsis thaliana* defence responses via the systemic acquired resistance pathway. *Journal of Experimental Botany* **59**(9): 2437–2448.

Weeks, A., Daly, D.C. and Simpson, B.B. (2005). The Phylogenetic history and biogeography of the frankincense and myrrh family (Burseraceae) based on nuclear and chloroplast sequence data. *Molecular Phylogenetics and Evolution* **35**(1): 85–101.

Weeks, A. and Simpson, B.B. (2007). Molecular phylogenetic analysis of *Commiphora* (Burseraceae) yields insight on the evolution and historical biogeography of an "impossible" genus. *Molecular Phylogenetics and Evolution* **42**(1): 62–79.

Weiblen, G.D., Webb, C.O., Novotny, V., Basset, Y. and Miller, S.E. (2006). Phylogenetic dispersion of host use in a tropical insect herbivore community. *Ecology* **87**(7): S62–S75.

Weingartner, E., Wahlberg, N. and Nylin, S. (2006). Dynamics of host plant use and species diversity in *Polygonia* butterflies (Nymphalidae). *Journal of Evolutionary Biology* **19**(2): 483–491.

Weitschek, E., van Velzen, R. and Felici, G. (2011). Species classification using DNA Barcode sequences: A comparative analysis. IASI-CNR Research Report. Roma, IASI-CNR.

Werren, J.H., Zhang, W. and Guo, L.R. (1995). Evolution and phylogeny of *Wolbachia*: Reproductive parasites of arthropods. *Proceedings: Biological Sciences* **261**(1360): 55–63.

West, S.A. and Cunningham, J.P. (2002). A general model for host plant selection in phytophagous insects. *Journal of Theoretical Biology* **214**(3): 499–513.

Wheat, C.W., Vogel, H., Wittstock, U., Braby, M.F., Underwood, D. and Mitchell-Olds, T. (2007). The Genetic basis of a plant–insect coevolutionary key innovation. *Proceedings of the National Academy of Sciences* **104**(51): 20427–20431.

White, T., Bruns, T., Lee, S., Taylor, J., Innis, M., Gelfand, D. and Shinsky, J. (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. in: *PCR Protocols: A Guide to Methods and Applications*, Academic Press: 315–322.

Whitfield, J.B. and Kjer, K.M. (2008). Ancient rapid radiations of insects: Challenges for phylogenetic analysis. *Annual Review of Entomology* **53**: 449–472.

Whitworth, T.L., Dawson, R.D., Magalon, H. and Baudry, E. (2007). DNA barcoding cannot reliably identify species of the blowfly genus *Protocalliphora* (Diptera: Calliphoridae). *Proceedings of the Royal Society B: Biological Sciences* **274**(1619): 1731–1739.

Wiemers, M. and Fiedler, K. (2007). Does the DNA barcoding gap exist? - a case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers in Zoology* **4**: 8.

Wiens, J.J., Ackerly, D.D., Allen, A.P., Anacker, B.L., Buckley, L.B., Cornell, H.V., Damschen, E.I., Jonathan Davies, T., Grytnes, J.-A., Harrison, S.P., Hawkins, B.A., Holt, R.D., McCain, C.M. and Stephens, P.R. (2010). Niche conservatism as an emerging principle in ecology and conservation biology. *Ecology Letters* **13**(10): 1310–1324.

Wilcoxon, F. (1947). Probability tables for individual comparisons by ranking methods. *Biometrics* **3**(3): 119–122.

Wiley, E.O. (1978). The Evolutionary species concept reconsidered. *Systematic Biology* **27**(1): 17–26.

Williams, M.C. (2012). Butterflies and skippers of the Afrotropical Region. A digital encyclopaedia. from http://www.atbutterflies.com.

Wilson, J.S., Forister, M.L., Dyer, L.A., O'Connor, J.M., Burls, K., Feldman, C.R., Jaramillo, M.A., Miller, J.S., Rodríguez-Castañeda, G., Tepe, E.J., Whitfield, J.B. and Young, B. (2012). Host conservatism, host shifts and diversification across three trophic levels in two Neotropical forests. *Journal of Evolutionary Biology* **25**(3): 532–546.

Winkler, I.S. and Mitter, C. (2008). The Phylogenetic dimension of insect-plant interactions: a review of recent evidence. in: *The Evolutionary biology of herbivorous insects: specialization, speciation, and radiation.* K. Tilmon. Berkeley, University of California Press: 240–263.

Winkler, I.S., Mitter, C. and Scheffer, S.J. (2009). Repeated climate-linked host shifts have promoted diversification in a temperate clade of leaf-mining flies. *Proceedings of the National Academy of Sciences of the United States of America* **106**(43): 18103–18108.

Wirta, H., Viljanen, H., Orsini, L., Montreuil, O. and Hanski, I. (2010). Three parallel radiations of Canthonini dung beetles in Madagascar. *Molecular Phylogenetics and Evolution* **57**(2): 710–727.

Won, Y.-J., Jones, W.J. and Vrijenhoek, R.C. (2008). Absence of cospeciation between deep-sea Mytilids and their thiotrophic endosymbionts. *Journal of Shellfish Research* **27**(1): 129–138.

Work, T.T., McCullough, D.G., Cavey, J.F. and Komsa, R. (2005). Arrival rate of nonindigenous insect species into the United States through foreign trade. *Biological Invasions* **7**(2): 323–332.

Wright, S. (1932). The Roles of mutation, inbreeding, crossbreeding and selection in evolution. *Sixth International Congress of Genetics*, Ithaca. **1**: 356–366.

Wright, S. (1943). Isolation by distance. *Genetics* **28**(2): 114–138.

Wurdack, K.J. and Davis, C.C. (2009). Malpighiales phylogenetics: gaining ground on one of the most recalcitrant clades in the Angiosperm tree of life. *American Journal of Botany* **96**(8): 1551–1570.

Xi, Z., Ruhfel, B.R., Schaefer, H., Amorim, A.M., Sugumaran, M., Wurdack, K.J., Endress, P.K., Matthews, M.L., Stevens, P.F., Mathews, S. and Davis, C.C. (2012). Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proceedings of the National Academy of Sciences* **109**(43): 17519–17524.

Yang, Z. and Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences* **107**(20): 9264–9269.

Yassin, A., Capy, P., Madi-Ravazzi, L., Ogereau, D. and David, J.R. (2008). DNA barcode discovers two cryptic species and two geographical radiations in the invasive drosophilid *Zaprionus indianus. Molecular Ecology Resources* **8**(3): 491–501.

Yassin, A., Markow, T.A., Narechania, A., O'Grady, P.M. and DeSalle, R. (2010). The Genus *Drosophila* as a model for testing tree- and character-based methods of species identification using DNA barcoding. *Molecular Phylogenetics and Evolution* **57**(2): 509–517.

Yeates, D.K., Seago, A., Nelson, L., Cameron, S.L., Joseph, L.E.O. and Trueman, J.W.H. (2011). Integrative taxonomy, or iterative taxonomy? *Systematic Entomology* **36**(2): 209–217.

Yoder, A.D. and Nowak, M.D. (2006). Has vicariance or dispersal been the predominant biogeographic force in Madagascar? Only time will tell. *Annual Review of Ecology, Evolution, and Systematics* **37**(1): 405–431.

Yoder, A.D. and Yang, Z. (2000). Estimation of primate speciation dates using local molecular clocks. *Molecular Biology and Evolution* **17**(7): 1081–1090.

Yoder, A.D. and Yang, Z. (2004). Divergence dates for Malagasy lemurs estimated from multiple gene loci: geological and evolutionary context. *Molecular Ecology* **13**(4): 757–773.

Zachos, J., Pagani, M., Sloan, L., Thomas, E. and Billups, K. (2001). Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* **292**(5517): 686–693.

Zakharov, E.V., Lobo, N.F., Nowak, C. and Hellmann, J.J. (2009). Introgression as a likely cause of mtDNA paraphyly in two allopatric skippers (Lepidoptera: Hesperiidae). *Heredity* **102**(6): 590–599.

Zarate, S.I., Kempema, L.A. and Walling, L.L. (2007). Silverleaf whitefly induces salicylic acid defenses and suppresses effectual jasmonic acid defenses. *Plant Physiology* **143**(2): 866–875.

Zemlak, T.S., Ward, R.D., Connell, A.D., Holmes, B.H. and Hebert, P.D.N. (2009). DNA barcoding reveals overlooked marine fishes. *Molecular Ecology Resources* **9**: 237–242.

Zhang, A.B., He, L.J., Crozier, R.H., Muster, C. and Zhu, C.D. (2010). Estimating sample sizes for DNA barcoding. *Molecular Phylogenetics and Evolution* **54**(3): 1035–1039.

Zhang, C., Zhang, D.-X., Zhu, T. and Yang, Z. (2011). Evaluation of a Bayesian coalescent method of species delimitation. *Systematic Biology* **60**(6): 747–761.

Zhang, Z., Nisancioglu, K.H., Flatoy, F., Bentsen, M., Bethke, I. and Wang, H. (2011). Tropical seaways played a more important role than high latitude seaways in Cenozoic cooling. *Climate of the Past* **7**(3): 801–813.

Zhang, Z., Schwartz, S., Wagner, L. and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology* **7**(1–2): 203–214.

Zhou, L., Su, Y.C.F., Thomas, D.C. and Saunders, R.M.K. (2012). 'Out-of-Africa' dispersal of tropical floras during the Miocene climatic optimum: evidence from *Uvaria* (Annonaceae). *Journal of Biogeography* **39**(2): 322–335.

Zou, S., Li, Q., Kong, L., Yu, H. and Zheng, X. (2011). Comparing the usefulness of distance, monophyly and character-based DNA barcoding methods in species identification: A case study of Neogastropoda. *PLoS ONE* **6**(10): e26619.

Zuckerkandl, E. and Pauling, L.B. (1962). Molecular disease, evolution, and genetic heterogeneity. in: *Horizons in Biochemistry.* M. Kasha and B. Pullman. New York, Academic Press: 189–225.

Zwickl, D.J. (2006). Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. PhD thesis, University of Texas at Austin.

Summary

This thesis aimed to elucidate the evolutionary history of the associations between *Cymothoe* forest butterflies (Nymphalidae, Limenitidinae) and their *Rinorea* host plants (Violaceae) in tropical Africa. Insects are by far the most diverse group of multicellular organisms on earth. Because most insect species are herbivores, understanding the evolution of interactions between herbivorous insects and their host plants is therefore crucial to comprehend global patterns in terrestrial biodiversity. The *Cymothoe-Rinorea* system is especially suitable for untangling processes shaping patterns of insect-host plant associations because of its high level of specificity (mostly monophagous) and the large number of related species involved (33 insect herbivores and 32 hosts). Obviously, any evolutionary study relies on a solid classification and taxonomy of the organisms under study. Unfortunately, however, in *Cymothoe* as well as *Rinorea*, taxonomy and classification is still partly unresolved.

To improve taxonomy of *Cymothoe* and facilitate efficient identification of immature specimens found on *Rinorea* host plants, we generated an extensive dataset of 1204 DNA barcode sequences (Chapter 2). Application of a novel taxonomic decision pipeline for integrating DNA barcodes with morphology and biogeography proved instrumental for solving taxonomic problems in *Cymothoe* and five taxa within *Cymothoe* could be confidently raised to species level. In addition, our DNA barcode data set allowed for the identification of 42 immature specimens from six different countries, significantly increasing the data on *Cymothoe* host plant associations. Nevertheless, our results also demonstrated that not all species of *Cymothoe* can be confidently delimited or identified. We hypothesize that this is probably due to incomplete lineage sorting and introgression (the latter possibly mediated through *Wolbachia* endosymbionts) between recently diverged *Cymothoe* species. In order to assess what are the best methods for matching DNA barcodes from recently diverged species, we compared six methods in their ability to correctly match DNA barcodes from selected published empirical data sets as well as simulated data (Chapter 3). Our results showed that, even though recently diverged species pose a significant problem for effective DNA barcoding, sensitive similarity-based and diagnostic methods can significantly improve identification performance compared with the commonly used tree-based methods.

To improve classification and clarify the biogeographic history of *Rinorea*, we presented an updated phylogenetic tree of *Rinorea* with increased taxonomic sampling, using plastid as well as nuclear DNA sequences (Chapter 5). Phylogenetic relationships inferred from nuclear DNA data were generally congruent with those based on evidence from plastid haplotypes from earlier studies of *Rinorea* and helped resolve additional clades, some of which warrant further taxonomic study.

Divergence time estimations indicated that *Rinorea* originated in the Neotropics and reached Africa in the Eocene through trans-Atlantic dispersal. From Africa, *Rinorea* subsequently dispersed into Asia in the Oligocene or early Miocene, and colonized Madagascar multiple times independently within a relatively recent time scale (Pliocene), suggesting that factors governing the independent colonizations of *Rinorea* to Madagascar may have been similar.

In Chapter 4 we assessed whether differential rates of net species diversification in the African butterfly sister genera *Harma* (1 species) and *Cymothoe* (approximately 82 species) could best be explained by shifts to novel host plants (from Achariaceae to *Rinorea*) or by environmental factors such as changing climate. We generated the first time-calibrated species-level molecular phylogenetic tree of *Harma* and *Cymothoe* and found that, after their divergence in the Miocene (15 Mya), net species diversification was low during the first 7 Myr. Coinciding with the onset of diversification of *Cymothoe* in the late Miocene (around 7.5 Mya) there was a sharp and significant increase in diversification rate, suggesting a rapid radiation. This increased rate did not correlate with host plant transition from Achariaceae to *Rinorea* host plants, but rather with a period of global cooling and desiccation, indicating that tropical forest fragmentation may well have driven the elevated diversification rates in *Cymothoe*.

Finally, in Chapter 6 we integrated the time-calibrated phylogenetic evidence from *Cymothoe* and *Rinorea* presented in chapters 4 and 5 with updated host association records from the field, with the aim to distinguish between alternative scenarios for the evolution of insect-host plant associations. Our results showed that: (i) divergences among extant *Cymothoe* are more recent than those among their associated *Rinorea* hosts, suggesting asynchronous diversification of *Cymothoe* herbivores onto already diversified clades within African *Rinorea*; (ii) phylogenetic trees of *Cymothoe* and their associated *Rinorea* host plants are discordant and current associations between *Cymothoe* herbivores and their *Rinorea* hosts have developed primarily through a process of host shifting rather than by cospeciation; and (iii) related *Cymothoe* tend to feed on related *Rinorea* hosts. Based on the available data, we propose a recent origin of *Rinorea*-feeding by *Cymothoe* butterflies with a single colonization of pre-existing lineages in the late Miocene. Current associations are best explained by a predominance of shifts among related plants, probably due to constraints in larval physiology and oviposition behaviour. Overall, these findings are in agreement with a growing body of substantial evidence to suggest that divergences of herbivorous insects and their host plants are asynchronous, and that evolutionary dynamics of hosts and parasites do not favour cospeciation.

Insect-plant interactions are receiving increasing attention because of their importance in crop production and protection. At the same time, an increasing number of insects and plants that have evolved in separation are currently coming into contact due to human activities and climatic changes. It is therefore tempting to find implications of our findings for insect-host plant associations for agricultural

systems (Chapter 7). Based on our results, one might predict that insects will only become pests of crops that are closely related to their natural host. Extrapolating our findings to an agricultural setting is difficult, however, because of the difference in selective pressures between natural and agricultural ecosystems.

Samenvatting

Het doel van dit proefschrift was om de evolutionaire geschiedenis te ontrafelen van de associaties tussen *Cymothoe* vlinders (Nymphalidae, Limenitidinae) en hun *Rinorea* waardplanten in het tropisch bos van Afrika. Insekten zijn van alle meercellige organismen verreweg de meest diverse groep. Aangezien de meeste insecten herbivoor zijn, is inzicht in de evolutie van interacties tussen insecten en hun waardplanten van groot belang voor een beter begrip van de wereldwijde patronen in terrestrische biodiversiteit. Het systeem *Cymothoe*-*Rinorea* wordt gekenmerkt door een hoge mate van specificiteit (voornamelijk monofaag) en grote aantallen betrokken verwante soorten (33 insekten en 32 waardplanten) en is daardoor bijzonder goed geschikt om de processen die de patronen in insekt-waardplant associaties te ontrafelen. Evolutionair onderzoek staat of valt met een solide classificatie en taxonomie van de te bestuderen soorten. Maar de taxonomie en classificatie van zowel *Cymothoe* als *Rinorea* zijn echter nog gedeeltelijk onduidelijk.

Om de taxonomie van *Cymothoe* te verbeteren en efficiënte identificatie van onvolwassen exemplaren mogelijk te maken hebben we een uitgebreide data set van 1204 DNA barcode sequenties gegenereerd (Hoofdstuk 2). Bij het oplossen van taxonomische problemen in *Cymothoe* speelde een nieuw taxonomisch beslissingsprotocol voor het integreren van DNA barcodes met morfologie en biogeografie een grote rol; en vijf taxa binnen *Cymothoe* konden met zekerheid tot soort verheven worden. Daarnaast maakte onze DNA barcode data set de identificatie van 42 onvolwassen exemplaren uit zes verschillende landen mogelijk, waarmee de gegevens over *Cymothoe*-waardplant associaties significant werden uitgebreid. Desalniettemin lieten onze resultaten ook zien dat niet alle *Cymothoe* soorten met zekerheid begrensd of geïdentificeerd kunnen worden. Wij stellen dat dit waarschijnlijk verzoorzaakt wordt door incompleet sorteren van genetische lijnen en genetische introgressie (het laatste mogelijk doormiddel van *Wolbachia* endosymbionten) tussen recent ontstane *Cymothoe* soorten. Om te bepalen welke DNA barcode methoden het beste werken in recent ontstane soorten hebben we zes methoden vergeleken aan de hand van hun bruikbaarheid voor het correct identificeren van DNA barcodes uit geselecteerde publicaties alsook uit gesimuleerde gegevens (Hoofdstuk 3). Onze resultaten lieten zien dat recent ontstane soorten een inderdaad problematisch zijn voor DNA barcoding, maar dat gevoelige methoden gebaseerd op similariteit of diagnostiek desalniettemin de toepasbaarheid substantieel kunnen verbeteren ten opzichte van de veelgebruikte boom-gebaseerde methoden.

Om de classificatie van *Rinorea* te verbeteren en haar biogeografische geschiedenis op te helderen hebben we een herziene fylogenetische boom gepresenteerd; met bemonstering van meer taxa en gebaseerd op zowel plastide als nucleaire DNA

sequenties (Hoofdstuk 5). Verwantschappen die werden afgeleid op basis van nucleair DNA waren grotendeels in overeenstemming met resultaten van eerdere studies gebaseerd op plastide haplotypen. Bovendien hielpen ze aanvullende claden op te lossen, waarbij voor sommigen verdere taxonomische studie wenselijk is. Een inschatting van divergentietijden gaf aan dat *Rinorea* in de Neotropen ontstaan is en Afrika bereikte middels trans-Atlantische verspreiding in het Eoceen. Vanuit Afrika verspreidde *Rinorea* zich vervolgens naar Azië in het Oligoceen of vroege Mioceen, en koloniseerde Madagascar meerdere malen onafhankelijk binnen een relatief korte tijdspanne (Plioceen), wat de indruk wekt dat deze onafhankelijke kolonisaties door gelijkaardige factoren mogelijk werden gemaakt.

In Hoofdstuk 4 schatten wij in of het verschil in netto soortsdiversificatie tussen de Afrikaanse zustergeslachten van tropische regenwoudvlinders *Harma* (1 soort) en *Cymothoe* (ongeveer 82 soorten) het best verklaard kan worden door een verschuiving naar nieuwe waardplanten (van Achariaceae naar *Rinorea*) of door omgevingsfactoren zoals een veranderend klimaat. We produceerden de eerste fylogenetische boom van *Harma* en *Cymothoe*, op soortsniveau en gecalibreerd in tijd. Op basis daravan vonden we dat, na hun splitsing in het Mioceen rond 15 Mega-annum (Ma), netto diversificatie gedurende de eerste 7 Ma laag was. Samenvallend met het begin van diversificatie in *Cymothoe* in het late Mioceen (rond 7.5 Ma) zagen we een scherpe en significante versnelling van netto soortsdiversificatie die de suggestie wekt van een snelle radiatie. Deze verhoogde snelheid correleerde niet met een verschuiving van Achariaceae naar *Rinorea* waardplanten, maar juist met een periode van wereldwijde verkoeling en verdroging van het klimaat, wat aangeeft dat de versnelling van soortsdiversificatie in *Cymothoe* wellicht is veroorzaakt door fragmentatie van tropisch regenwoud.

Als laatste integreerden we in Hoofdstuk 6 de in tijd gecalibreerde fylogenetische resultaten voor *Cymothoe* en *Rinorea* uit hoofdstukken 4 en 5 met herziene waardplant associatie gegevens, met het doel verschillende scenario's van de evolutie van insect-waardplant associates te testen. Onze resultaten lieten zien dat: (i) divergentietijden tussen huidige *Cymothoe* recenter zijn dan die tussen hun *Rinorea* waardplanten, wat een asynchrone diversificatie suggereert van *Cymothoe* herbivoren op reeds gediversificeerde claden binnen Afrikaanse *Rinorea*; (ii) fylogenetische bomen van *Cymothoe* en *Rinorea* niet in overeenstemming zijn en de huidige associaties tussen *Cymothoe* herbivoren en hun *Rinorea* waardplanten het resultaat zijn van voornamelijk waardplant verwisselingen in plaats van gezamenlijke soortsvorming; en (iii) verwante *Cymothoe* meestal verwante *Rinorea* als waardplant hebben. Op basis van de beschikbare gegevens stellen wij dat de associatie met *Rinorea* waardplanten recent is ontstaan binnen *Cymothoe*, met een enkele kolonisatie van reeds bestaande lijnen in het late Mioceen. De huidige associaties worden het best verklaard door een overmaat van waardplant verwisselingen tussen verwante planten, waarschijnlijk vanwege beperkingen in larvale fysiologie en ovipisitiegedrag. In het algemeen stemmen onze bevindingen overeen met het groeiende bewijs dat divergenties van herbivore

insecten en hun waardplanten asynchroon zijn en dat de evolutionaire krachten bij parasieten en hun waarden geen gezamenlijke soortsvorming bevorderen.

Er is groeiende aandacht voor insect-plant interacties omdat ze van belang zijn voor gewasproductie en -bescherming. Tegelijkertijd komen momenteel steeds meer insecten en planten die gescheiden zijn geëvolueerd voor het eerst in contact vanwege menselijke activiteit en klimaatverandering. Het is daarom verleidelijk om vanuit onze bevindingen gevolgtrekkingen te maken voor agronomische systemen (Hoofdstuk 7). Op basis van onze resultaten zou verwacht kunen worden dat insecten enkel een plaag worden voor gewassen die nauw verwant zijn aan hun natuurlijke waardplant. Het is echter moeilijk om onze bevindingen te extrapoleren naar een agronomische situatie omdat de selectieve krachten erg verschillend zijn in agronomische en natuurlijke ecosystemen.

Acknowledgements

This is probably the most often read section of this book. You may be wondering whether you are acknowledged, and indeed you are: Thank you, dear reader! For your interest in my thesis and for taking part in this important episode of my life.

But obviously, according to tradition, I should first thank my promoter Marc Sosef for supporting my PhD project. I am thankful for the freedom you gave me in performing my research and for your contributions to improving my chapters.

I am deeply grateful to my co-promotor Freek Bakker: you have been a great mentor and I appreciate your support and advice - on many levels. I really enjoyed our scientific discussions and brainstorming sessions, as well as our joint crafting of manuscripts. It has been great to have such a supportive supervisor!

I also thank the other members of the PhD committee for their time and attention paid to reading and critically assessing my thesis: Thank you Bas Zwaan, Niklas Wahlberg, Steph Menken and Marcel Dicke; I look forward to our discussions at the defence of my thesis. Together with Steph, Joop van Loon has given valuable advice on my project.

Marleen, I cannot thank you enough for your tremendous support, especially in the last year. You have also given great advice for improvement of early versions of nearly al my chapters. Nathan, thanks for being playful!

To all my colleagues at the Biosystematics Group and at the herbarium, thanks for the nice talks and interesting discussions during the coffee or lunch breaks! I still miss those quite regularly. Lars Chatrou deserves a special mention for having been a great colleague and friend. All the students who have been working at *the Island* created a great atmosphere!

I had the pleasure to supervise two talented students who contributed to my project: Renske Onstein, thank you for collecting essential *Cymothoe-Rinorea* host plant associations in Ghana and for generating *Rinorea* sequence data. Gerco Niezing, many thanks for your study on the chemical compounds of *Rinorea*.

I very much enjoyed the collaboration with people from many different countries and diverse backgrounds. This has been a main source of joy and inspiration throughout my project and without their help I would not have been able to present such a wealth of data in this thesis! First and foremost I would like to thank Jean-Louis Amiet: My project would not have been possible without your many years of excellent work on the associations *Cymothoe* and *Rinorea* in Cameroon, this

work would not even have been conceived. It would have been great to learn more from you and I regret not having had the time to accept your kind invitation to visit you in person. Torben Larsen, you have been a great mentor and it was an honour to for me to describe my first new species, *Cymothoe druryi* together with you. More importantly you have been instrumental in introducing me to the community of African butterfly experts and enthusiasts. Many people generously donated specimens or samples used in this study, thank you Tinde van Andel, Julian Bayliss, Lars Chatrou, Hein Boersma, Dries Bonte, Thierry Bouyer, Oskar Brattström, Steve Collins, Frans Desmet, Robert Ducarme, Carel Jongkind, Erik Koenen, David Lees, Quentin Luke, Dino Martins, Carolyn McBride, Freerk Molleman, Philippe Oremans, Tomasz Pyrcz, Sáfián Szabolcz, Gael vandeWeghe, Eric Vingerhoedt, Robert Warren, Haydon Warren-Gash, Jan Wieringa and Steve Woodhall. It has been a great pleasure working with all of you and I consider especially Sáfián Szabolcz and Oskar Brattström as good friends.

Greg Wahlert, thank you for your cooperation and for sharing your *Rinorea* sequence and morphological data. It has been a pleasure working with you. I really hope you can continue working on the dearly-needed taxonomic treatment of African *Rinorea*!

I thank Steve Collins (African Butterfly Research Institute), Blanca Huartas (Natural History Museum) and Ugo Dall'Asta (Royal Museum for Central Africa) for facilitating access to collections. Wouter Dekonick and Stefan Kerkhof (Royal Belgian Institute of Natural Sciences), Teresa di Micco de Santo (ABRI) and Suzanne Rab Green (American Museum of Natural History) kindly provided specimen photographs for my study.

My collection trips would not have been so successful (and so much fun!) without the wonderful help in organizing and performing of fieldwork by Gaston Achoundong and Olivier Sene Belinga in Cameroon, Steve Collins in Kenya and Oskar Brattström and Robert Warren in Nigeria. Many thanks also to Sáfián Szabolcz, Stephen Davey, Andrew Walker and Andrews Kankam Amankwah, for helping Renske and Gerco collect samples in Ghana.

I owe a lot to several people for their help in generating an important part of the sequencing data: Ria Vrielink, you were very helpful in training me in lab techniques. I am also very thankful for your help and support in the lab generating most of the sequence data for *Rinorea*. Paul Hebert and staff of the Canadian Centre for DNA Barcoding; Niklas Wahlberg and Pavel Matos from University of Turku: thank you for granting access to your sequencing facilities. I also thank Camiel Doorenweerd for generating sequences at Naturalis Biodiversity Centre. Thanks also to Bram Lestrade and Sandrina Pardoel for their help in the lab.

Paola Bertolazzi and Emanuel Weitschek are acknowledged for implementing significant improvements in their BLOG software. It was great working with you on

testing and improving of BLOG. I would like to thank Guillome Achaz and Sophie Brouillet for help with running ABGD, and Leandro Jones and Wayne Matten for their advice on configuration of TNT and BLAST, respectively. Marc Suchard kindly helped me with editing the BEAST ancestral host plant reconstruction .xml files. Arne Mooers provided excellent suggestions about how to correct for missing data in my diversification analyses. The following people generously shared their empirical data sets for DNA barcoding analysis: Brian Golding and Melanie Lou provided data of *Drosophila*, Kyle Dexter provided data of *Inga*, Chris Meyer provided data of Cypraeidae.

I thank my fellow members of the *Wageningen Evolution and Ecology Seminars* organizing committee. It has been great to be part of its inception and I believe we have collectively made an important contribution to academic life and training at our university.

I am really lucky to have Setareh Mohammadin and Pulcherie Bissiengou by my side when I defend my thesis. I wish you both all the best with your own PhD projects at the Biosystematics Group. Projectleader Rene Geurts has been very generous in giving me time off from my duties at the Laboratory of Molecular Biology to finalize the last parts of my thesis. Ansa Fiaz has greatly helped me with mastering LaTeX in order to typeset this book.

Finally I'd like to thank my family and friends for all those years of support, interest and most importantly, the good times together!

Publications

van Velzen, R., Wahlberg, N., Sosef, M.S.M. and Bakker, F.T. (2013). Effects of changing climate on species diversification in tropical forest butterflies of the genus *Cymothoe* (Lepidoptera: Nymphalidae). *Biological Journal of the Linnean Society* **108**(3): 546–564.

Yang, M.Q., van Velzen, R., Bakker, F.T., Sattarian, A., Li, D.Z. and Yi, T.S. (2013) Molecular phylogenetics and character evolution of Cannabaceae. *Taxon* **62**(3): 473–485.

Weitschek, E., van Velzen, R., Felici, G. and Bertolazzi, P. (2013) BLOG 2.0: a software system for character-based species classification with DNA Barcode sequences. What it does, how to use it. *Molecular Ecology Resources* **13**: 1043–1046.

Lin, K., Limpens, E., Zhang, Z., Ivanov, S., Saunders, D., Mu, D., Pang, E., Cao, H., Cha, H., Lin, T., Zhou, Q., Shang, Y., Li, Y., Sharma, T., van Velzen, R.,de Ruijter, N., Aanen, D., Win, J., Kamoun, S., Bisseling, T., Geurts, R. and Huang, S. (submitted) Single nucleus genome sequencing reveals high similarity among nuclei of an endomycorrhizal fungus.

van Velzen, R. and Wieringa, J.J. (submitted) *Rinorea calcicola* (Violaceae), an endangered new species from south-eastern Gabon.

Chatrou, L.W., Pirie, M.D., van Velzen, R. and Bakker, F.T. (in press) Annonaceae substitution rates – a codon model perspective. *Revista Brasileira de Fruticultura.*

van Velzen, R., Weitschek, E., Felici, G. and Bakker, F.T. (2012). DNA barcoding of recently diverged species: relative performance of matching methods. *PLoS ONE* **7**(1): e30490.

Weitschek, E., van Velzen, R. and Felici, G. (2011) Species classification using DNA Barcode sequences: A comparative analysis. *IASI-CNR Report* **11**(7): 14pp.

McBride, C.S., van Velzen, R. and Larsen, T.B. (2009). Allopatric origin of cryptic butterfly species that were discovered feeding on distinct host plants in sympatry. *Molecular Ecology* **18**(17): 3639–3651.

van Velzen, R., Larsen, T.B. and Bakker, F.T. (2009). A new hidden species of the *Cymothoe caenis*-complex (Lepidoptera: Nymphalidae) from western Africa. *Zootaxa* **2197**: 53–63.

van Velzen, R., Bakker, F.T. and van Loon, J.J.A. (2007). DNA barcoding reveals hidden species diversity in *Cymothoe* (Nymphalidae) *Proceedings of the Netherlands Entomological Society Meeting* **18**: 95–103.

van Velzen, R., Bakker, F.T., Sattarian, A. and van der Maesen, L.J.G. (2006) Evolutionary relationships of Celtidaceae. in: *Contribution to the Biosystematics of Celtis L. (Celtidaceae) with Special Emphasis on the African Species.* A. Sattarian. PhD thesis, Wageningen University, Wageningen. 7–29.

Curriculum Vitae

Robin van Velzen was born in Mussel (Stadskanaal), The Netherlands, on the 5th of November 1979. In 1998 he completed his secondary school education at Vrije School VO in Groningen, and in 1999 he qualified for his pre-university diploma at Drenthe College in Emmen.

In September 1999 he started his university studies in Biology at Wageningen University; specializing in population biology. As part of an internship at the National Herbarium Nederland in Utrecht he joined a botanical expedition to Peru together with Michael Pirie and Marleen Botermans, in collaboration with the San Marcos Herbarium in Lima. He then completed his first MSc thesis project *Exploring Current Methods for Phylogenetic Inference: Celtidoid Evolutionary Relationships within the Urticalean Rosids* at the Herbarium Vadense, Wageningen University, resulting in a publication (van Velzen *et al.* 2006). His second MSc thesis entitled *Evolution of Host-Plant use in* Cymothoe *(Nymphalidae) feeding on* Rinorea *(Violaceae)* was a joint project of the Biosystematics Group and the Laboratory of Entomology. The project included the organization of 2 months fieldwork in Cameroon, in collaboration with the Herbier National du Cameroun in Yaoundé and resulted in another publication (van Velzen *et al.* 2007). He gained the degree of Master of Science in December 2006 (cum laude).

He re-joined the Biosystematics Group in 2007 as Junior Researcher to write a PhD proposal based on his MSc work and started his PhD project on the *Evolution of associations between* Cymothoe *butterflies and their* Rinorea *host plants in tropical Africa* in June 2008. As part of his research he organized two additional field expeditions: to Kenya, in collaboration with the African Butterfly Research Institute in 2008 and to Nigeria together with Oskar Brattström from Cambridge University in 2010. He participated in the international workshops *High performance computing for phylogenetics* (2010, Knoxville, United States) and *Molecular Evolution* (2011, Český Krumlov, Czech Republic). In From 2009 to 2012 he was member of the organizing committee of the monthly *Wageningen Evolution and Ecology Seminars.* He presented his work at the *International Conference on the Biology of Butterflies* in Rome, Italy; the joint *European Systematics Congress* in Leiden, The Netherlands; and the *ESEB Congress* in Tübingen, Germany. He has been awarded bursaries to visit natural history collections in Tervuren, London, and Nairobi and to present his work at the *EDIT Young Taxonomists Symposium* in Carvoeirho, Portugal. His lecture at the *Biodiversity PhD Symposium* in Leiden, The Netherlands was selected as best oral presentation.

In December 2012 he began a four-year postdoctoral project at the laboratory of Molecular Biology in Wageningen, where he studies the evolution of rhizobium symbiosis in the non-Legume *Parasponia.*

PE&RC PhD Training Certificate

With the educational activities listed below the PhD candidate has complied with the educational requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 32 ECTS (= 22 weeks of activities).

Writing of project proposal (4.5 ECTS)

Host tracking by *Cymothoe* (Nymphalidae) feeding on *Rinorea* (Violaceae): evolution & mechanisms of association (2007)

Post-graduate courses (7.4 ECTS)

High performance computing for phylogenetics; NIMBioS; USA (2010)

R for statistical analysis; PE&RC (2010)

Molecular evolution workshop; Cummings, M. et al.; Czech republic (2011)

Laboratory training and working visits (4.4 ECTS)

DNA Barcoding of *Cymothoe* (Nymphalidae); CCDB, Canada (2008)

Invited review of (unpublished) journal manuscript (1 ECTS)

Mol. Ecol. Res.: VIP DNA barcoding (2013)

Competence strengthening / skills courses (1.5 ECTS)

Techniques for writing and presenting a scientific paper; SENSE (2010)

Talentendag (workshops for PhD candidates); NWO (2010)

PE&RC Annual meetings, seminars and the PE&RC weekend (3 ECTS)

Wageningen Evolution and Ecology Seminars; organizing committee (2009-2012)

PhD Day Research School Biodiversity; organizing committee (2009)

PhD Day Research School Biodiversity (2011)

Discussion groups / local seminars / other scientific meetings (7.5 ECTS)

Biosystematics Group journal club (2008-2012)

NHN Wageningen / Biosystematics colloquia (2008-2012)

PhD Symposium; University of Amsterdam, the Netherlands (2011)

International symposia, workshops and conferences (6.7 ECTS)

Systematics; Leiden, the Netherlands (2009)

EDIT Young Taxonomist Symposium; Carvoeirho, Portugal (2009)

ESEB; Tuebingen, Germany (2011)

Lecturing / supervision of practicals/ tutorials; (3 ECTS)

Biosystematics & Biodiversity (2009)

Capita Selecta (2009)

Flora of the Netherlands (2009-2010)

Evolution & Systematics (2009-2011)

Fieldcourse European Flora and Fauna (2009, 2011)

Advanced Biosystematics (2011)

Supervision of 2 MSc students

Evolution of *Rinorea* (Violaceae) as host plant for *Cymothoe* (Nymphalidae)

The chemistry of specificity in the *Rinorea Cymothoe* host-herbivore association

The research described in this thesis was financed by The National Herbarium of the Netherlands and carried out at the Biosystematics Group, Wageningen University, The Netherlands.

Fieldwork was financially supported by the Systematics Research Fund, Alberta Mennega Foundation, and the Hugo de Vries Foundation. The European Commission's Research Infrastructure Action financed visits to natural history collections in Tervuren and London (SYNTHESYS grant numbers BE-TAF-3810 and GB-TAF-4003). Financial support for the printing of this thesis was kindly provided by the Biosystematics Group at Wageningen University.

Cover design and layout: Robin van Velzen