THE EVOLUTIONARY HISTORY OF *CRYPTOBLEPHARUS* LIZARDS: RECENT DIVERSIFICATION ACROSS CONTINENTS AND OCEANS

Mozes Pil Kyu Blom

The Australian National University

Supervised by

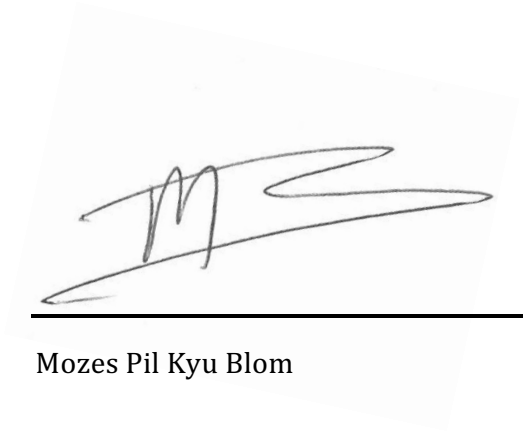
Professor Craig Moritz

A thesis submitted for the degree of Doctor of Philosophy

Canberra 2016

DECLARATION

The research presented in this thesis is my own original work except where due reference is given in the text. All of the chapters are co-authored and appendices are single-author. Unless otherwise indicated, the authorship order indicates the intellectual input and workload. No part of this thesis has been submitted for any previous degree.



Mozes Pil Kyu Blom

September 2016

ACKNOWLEDGEMENTS

First I would like to thank my supervisor and mentor, Craig Moritz, for his tremendous help and enthusiasm during the past four years. Throughout my PhD, I have greatly valued the academic freedom to explore my own ideas, while he has simultaneously provided all the support needed. His love for outback Australia and all its intriguing critters has been infectious and I hope that we will be able to team up for another trip in the Troopie!

I would like to thank Jason Bragg, Sally Potter and Matt Fujita in particular for their intellectual support and ongoing friendship, and all my other close collaborators; Paul Horner, Nick Matzke, Dan Rosauer, Robert Fisher, Jim Mcguire, Eric Rittmeyer, Sara Rocha, Sean Reilly and Alexander Stubbs. Furthermore, I would like to thank all the other past and current members of the Moritz' lab; Gaye Bourke (both inside and outside the lab!!!), Renae Pratt, Paul Oliver, Maxine Piggot, Christiana McDonald-Spicer, Jessica Antunes, Leonardo Tedeschi, Jeana Wong, Maria Strangas, Luisa Teasdale, Catherine Noble, Lauren Ashmann, Rosa Agudo, Rebecca Laver, Martha Muñoz, Felipe Martins, Stewart MacDonald, Richard Harrison, Sarah Catalano, Maggie Leigh, Gaynor Dolman, Arthur Georges, Leo Joseph, Kerensa McElroy, Angela McGaughran and Pascal Title.

I would also like to thank my official supervisory panel; Scott Keogh, Sylvain Foret, Lindell Bromham, and my unofficial panel; Huw Ogilvie, Marta Vidal-Garcia, Damien Esquerre, Mitzy Pepper, Ian Brennan, Lisa Schwanz, Sonal Singhal, Liam Bailey, Glib Mazepa and Kevin Mulder. They have provided me with help and advice throughout.

I thank the various museums in Australia (MAGNT, WAM, SAM, MV, QM, AM, ANWC), National Parks Australia, and properties across the monsoonal tropics for

their help and in particular: Mark Adams, Mark Hutchinson, Steve Donellan, Paul Doughty, the Australian Wildlife Conservancy and everyone at Theda station.

I would like to thank the Mohamed Bin Zayed Fund for Species Conservation and The National Geographic Society for funding my fieldwork in the Pacific. Special thanks to Jean-Yves Meyer, Neil Davies and Hinano Murphy at the Gump Field Station.

Finally, I would like to thank all the wonderful people in EEG for making this such a special and social environment. I'd like to thank in particular: Cat, Meredith, Anna H, Sonya, Regina, Dani, Bo, Iliana, Coni, Pip, Jessie, Carlos, Amit, Dave K, Dave H, Dani, Hannah, Dan H, Danswell, Robyn, Virginia, Sarah, David D, Megan, Emily H, Emma S, Nina, Alyssa, Tom S, Jared, Kristal, Thomas M, Frances, Zoe, Will F, Carsten, Gaby, Thom W, Hee Jin, Jono, Elly, Belinda, Wes and Jan.

My time in Canberra would not have been the same without my two academic siblings, labmates and above all dear friends; Ana Silva and Josh Penalba. They have made Canberra both fun, tasty and four years seem incredibly short. Last, but not least I want to thank Berry van den Pol for the past 24 years of friendship, Foteini Spagopoulou for all her help and patience during the past four years and my family for supporting me from the moment they picked me up from Schiphol in 1988.

ABSTRACT

Understanding the evolutionary processes that generate and maintain biodiversity is a fundamental objective in ecology and evolution. In this dissertation, I characterize phylogenetic patterns in a recent radiation of Australian skinks, discuss the ecological context of diversification and how this has translated into macroevolutionary change across the continent. By also reconstructing the evolutionary history of all *Cryptoblepharus* species globally, I shed further light on the evolutionary and biogeographic processes that have shaped the diversity of the genus. This dissertation project has generated an empirical framework for future studies into the continuous nature between micro- and macroevolutionary change.

To infer the phylogeny of Australian *Cryptoblepharus*, I generated an exon-capture dataset and designed a bioinformatic pipeline to generate quality filtered sequence alignments (Appendix A). Multi-locus datasets are required to confidently infer species trees for rapidly speciating clades due to a high prevalence of gene tree incongruence among loci. In Chapter I, I use the *Cryptoblepharus* radiation as an empirical example and describe how to account for differences in gene tree resolution when employing summary-coalescent methods for species tree inference. Our study highlights the importance of phylogenetically informative loci but simultaneously demonstrates that the addition of non-informative loci does not introduce phylogenetic noise.

In Chapter II, I then use comparative methods and morphological measurements for over 800 individuals, to examine the ecological context of diversification in Australian *Cryptoblepharus*. Specifically, I focus on whether habitat specialisation can explain current patterns of variation in ecologically relevant traits. I observed significant differences in morphology between species that occur in distinct environments (rock, arboreal and littoral) and species that occur within the same

habitat are often cryptic. These findings suggest that isolated analogous habitats have provided ecological opportunity and repeatedly promoted adaptive diversification, while speciation within habitat has accrued without ecomorphological change. In contrast to well known adaptive radiations in insular environments, continental radiations are likely driven by alternative diversification processes that jointly stimulate species proliferation.

In Chapter III, I explore patterns of introgression between phylogenetically divergent species. I combine population and phylogenetic tools, to quantify the extent of introgression between ecomorphologically distinct and similar taxa. I describe the frequent occurrence of mitochondrial haplotype sharing across species boundaries and the complete replacement of the mitochondrial genome in one species. Furthermore, non-sister species often share more nuclear variants than as expected under a model of incomplete lineage sorting only, suggesting substantial historical introgression.

Finally, *Cryptoblepharus* skinks are renowned for their widespread distribution, across continents and many island archipelagoes, while they have only emerged and diversified recently (i.e. since late Miocene/early Pliocene). In Chapter IV, I reconstruct the global phylogeny and discuss the importance of trait-based dispersal. Large scale range expansions across the Indian and Pacific Ocean have only occurred relatively recently, after an ancestor adapted to a more littoral habitat, and many extralimital taxa still only occur in close vicinity to coastal areas (Appendix B). These lizards therefore exemplify how ecological traits can increase the propensity of dispersal and that disjunct geographic distributions are not solely explained by a vicariance model.

THESIS OUTLINE

The following chapters compose this thesis:

1. Accounting for uncertainty in gene tree estimation: Summary-coalescent species tree inference in a challenging radiation of Australian lizards
   Blom MPK, Bragg JG, Potter S. and Moritz C.
   Systematic Biology (2017) 66: 352-366

2. Convergence across a continent: adaptive diversification in a recent radiation of Australian lizards
   Blom MPK, Horner, P. and Moritz C.
   Proceedings B (2016) 283: 20160181

3. Gene flow across species boundaries despite extensive ecological and temporal divergence in a recent radiation of Australian lizards.
   Blom MPK and Moritz C.
   In Prep

4. When ecology shapes biogeography: Habitat preference modulates dispersal probability in *Cryptoblepharus* lizards
   Blom MPK, Matzke NJ, Bragg JG, Arida E, Austin C, Backlin A et al.
   In Prep

Two additional manuscripts have been published during my PhD and are included as Annexes.

5. EAPhy: A flexible tool for high-throughput quality filtering of exon-alignments and data processing for phylogenetic methods.
   Blom MPK
   PLoS Currents Tree of Life (2015):1

6. Habitat use and new locality records for *Cryptoblepharus poecilopleurus* (Squamata: Scincidae) from French Polynesia
   Blom MPK
   Herpetology Notes (2015) 8: 579-582

TABLE OF CONTENTS

INTRODUCTION

In order to sustain biodiversity, studying processes of diversification in an evolutionary framework is of critical importance. Microevolutionary forces, such as selection and drift, induce population divergence and can ultimately promote the formation of novel species. As such, microevolutionary processes generate the building blocks for large-scale macroevolutionary change (Charlesworth et al. 1982). While evolution is therefore in essence a continuous process, the continuous nature between both units of evolutionary change is often disregarded. Research programs either focus on micro- or macroevolutionary patterns of diversification but seldom address both ends of the spectrum. Microevolutionary studies tend to focus on genotype-phenotype interactions and how processes that induce divergence might ultimately drive speciation, while macroevolutionary studies mostly dissect the tempo and mode of lineage diversification. However, incipient species are frequently merely ephemeral (Rosenblum et al. 2012; Singhal and Moritz 2013; Dynesius and Jansson 2014) and it generally remains challenging to accurately infer diversification rates due to the unknown effects of past extinction in most (ancient) clades (Stadler 2013). The number of biological systems where we have a thorough comprehension of the evolutionary path between divergent populations and emerging genera is therefore limited and most of our understanding stems from the study of adaptive radiations.

Adaptive radiation is the evolution of ecological diversity within a rapidly multiplying lineage. It is the differentiation of a single ancestor into an array of species that inhabit a variety of environments and that differ in traits used to exploit those environments (Glor 2010). Speciation during adaptive radiation is guided by differential resource use (Schluter 2000) and/or sexual selection (Wagner et al. 2012). Subsequent persistence is facilitated via niche sorting and specialization. However, adaptive radiations often occur in insular systems and it remains unclear to

what extent the processes that govern adaptive radiations on isolated islands or in lakes, are responsible for shaping biodiversity in non-insular settings. Indeed, in contrast to adaptive radiations that are the product of divergent selection alone, most species assemblages on continental landmasses seem to have been formed due to a combination of both adaptive and stochastic processes. Thus our understanding of the processes that promote and sustain diversification is incomplete and a comprehensive consideration of such mechanisms in evolutionary radiations other than the classic adaptive radiation, can provide important insight (Simões et al. 2016).

However, inferring species relationships in evolutionary radiations is not a trivial exercise and has traditionally been notoriously difficult due to the genealogical incongruence between genetic markers (Giarla and Esselstyn 2015). While these complications were relatively obscure in the outset of the phylogenetic era, topological incongruence among loci has become a major concern in many phylogenomic studies. Topological discordance due to biological sources such as incomplete lineage sorting and introgression can be expected under a wide variety of realistic evolutionary scenarios and therefore needs to be accounted for when inferring species trees (Maddison 1997; Degnan and Rosenberg 2009). Most species tree methods incorporate the Multi-Species Coalescent (MSC) to model the observed heterogeneity in coalescent histories. Yet full-coalescent based approaches, that simultaneously estimate gene- and species tree, are computationally demanding and their use often remains unfeasible with large phylogenomic datasets. Summary-coalescent based approaches provide an alternative that is less computationally demanding but requires a two-step approach, where gene trees are initially inferred and the species tree is estimated independently based on the resulting collection of gene trees (Mirarab et al. 2014). The accuracy of summary-coalescent approaches will therefore depend strongly on the quality of the underlying gene trees (Liu et al. 2015). The inference of phylogenetic relationships thus remains a dynamic field and there is much

scope for future research that addresses ongoing concerns such as the impact of gene tree estimation error on species tree inference or how to simultaneously model incomplete lineage sorting and introgression. Nonetheless, quantifying topological heterogeneity within large nuclear datasets can provide further insight in the microevolutionary processes that have shaped the evolutionary history of empirical species groups of interest and can be used in combination with contemporary coalescent based methods to improve the accuracy of both topology and branch length estimation. Here I have used such a holistic approach to study the challenging radiation of *Cryptoblepharus* skinks.

Skinks within the genus *Cryptoblepharus* are small diurnal lizards that occur in a variety of different habitats and are geographically widespread. They are the most widely distributed genus within the *Scincidae* family and occur across the Indo-Australian continent, the Malagasy region and throughout many island archipelagoes of the Pacific Ocean (Rocha et al. 2005; Horner 2007; Hayashi et al. 2009; Blom 2015). A recent taxonomic revision using both allozymes and morphological characteristics, elevated the number of Australian species from 6 to 25 but interestingly only 55% of the species were diagnosed based on both morphological and genetic differences, whereas the other lineages were morphologically or genetically indistinguishable from congeners (Horner and Adams 2007). This mosaic of both cryptic and phenotypically diverse lineages suggests that the underlying diversification processes are not straightforward but potentially influenced by multiple evolutionary forces of both adaptive and stochastic nature. In this dissertation, I have examined the evolutionary history of the *Cryptoblepharus* genus with a combined assessment of phylogenetic (Ch. I, III), ecological (Ch. II, III) and biogeographic (Ch. IV) patterns. I discuss both population level processes of divergence (ecological diversification; Ch. II) and persistence (introgression following secondary contact; Ch. III) and how ecology can play an important role in promoting macroevolutionary change (habitat

specialization; Ch. II, trait dependent dispersal; Ch. IV). As such, this dissertation project addresses both timely (Ch. I) and longstanding questions (Ch. II – IV) in evolutionary biology and simultaneously generates an empirical framework for future studies into the continuous nature between micro- and macroevolutionary change.

References

Blom M.P.K. 2015. Habitat use and new locality records for Cryptoblepharus poecilopleurus (Squamata: Scincidae) from French Polynesia. Herp. Notes 8:579-582.

Charlesworth B., Lande R., Slatkin M. 1982. A neo-Darwinian commentary on macroevolution. Evolution. 36:474–498.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Dynesius M., Jansson R. 2014. Persistence of within-species lineages: a neglected control of speciation rates. Evolution. 68:923–934.

Giarla T.C., Esselstyn J.A. 2015. The challenges of resolving a rapid, recent radiation: Empirical and simulated phylogenomics of Philippine shrews. Syst. Biol. 64:727–740.

Glor R.E. 2010. Phylogenetic insights on adaptive radiation. Annu. Rev. Ecol. Evol. Syst. 41:251–270.

Hayashi F., Shima A., Horikoshi K., Kawakami K., Segawa R.D., Aotsuka T., Suzuki T. 2009. Limited overwater dispersal and genetic differentiation of the snake-eyed skink (Cryptoblepharus nigropunctatus) in the oceanic Ogasawara Islands, Japan. Zool. Sci. 26:543–549.

Horner P. 2007. Systematics of the snake-eyed skinks, Cryptoblepharus Wiegmann (Reptilia: Squamata: Scincidae)–an Australian based review. The Beagle Supp. 3:21–198.

Horner P., Adams M. 2007. A Molecular-systematic assessment of species boundaries in Australian Cryptoblepharus (Reptilia: Squamata: Scincidae): A case study for the combined use of allozymes and morphology to explore cryptic biodiversity. The Beagle Supp. 3:1–20.

Liu L., Xi Z., Wu S., Davis C.C., Edwards S.V. 2015. Estimating phylogenetic trees from

genome-scale data. Ann. N. Y. Acad. Sci. 1360:36–53.

Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

Mirarab S., Bayzid M.S., Warnow T. 2014. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. Syst. Biol. 65.:366–380.

Rocha S., Carretero M., Vences M., Glaw F. 2005. Deciphering patterns of transoceanic dispersal: the evolutionary origin and biogeography of coastal lizards (Cryptoblepharus) in the Western Indian Ocean region. J. of Biogeogr. 33:13-22

Rosenblum E.B., Sarver B.A.J., Brown J.W., Roches Des S., Hardwick K.M., Hether T.D. 2012. Goldilocks meets Santa Rosalia: An ephemeral speciation model explains patterns of diversification across time scales. Evol Biol. 39:255–261.

Simões M., Breitkreuz L., Alvarado M., Baca S., Cooper J.C., Heins L. 2016. The evolving theory of evolutionary radiations. Trends Ecol. Evol. 31:27–34.

Singhal S., Moritz C. 2013. Reproductive isolation between phylogeographic lineages scales with divergence. Proc. R. Soc. B. 280:20132246–20132246.

Stadler T. 2013. Recovering speciation and extinction dynamics based on phylogenies. J. Evol. Biol. 26:1203–1219.

Wagner C.E., Harmon L.J., Seehausen O. 2012. Ecological opportunity and sexual selection together predict adaptive radiation. Nature. 487:366–369.

CHAPTER 1

Accounting for Uncertainty in Gene Tree Estimation: Summary-Coalescent Species Tree Inference in a Challenging Radiation of Australian Lizards

Accounting for Uncertainty in Gene Tree Estimation: Summary-Coalescent Species Tree Inference in a Challenging Radiation of Australian Lizards

Mozes P.K. Blom[1], Jason G. Bragg[1], Sally Potter[1] & Craig Moritz[1]

[1]*Research School of Biology, The Australian National University, Canberra ACT 0200, Australia*

ABSTRACT

Accurate gene tree inference is an important aspect of species tree estimation in a summary-coalescent framework. Yet, in empirical studies, inferred gene trees differ in accuracy due to stochastic variation in phylogenetic signal between targeted loci. Empiricists should therefore examine the consistency of species tree inference, while accounting for the observed heterogeneity in gene tree resolution of phylogenomic datasets. Here, we assess the impact of gene tree estimation error on summary-coalescent species tree inference by screening ~2000 exonic loci based on gene tree resolution prior to phylogenetic inference. We focus on a phylogenetically challenging radiation of Australian lizards (genus *Cryptoblepharus*, Scincidae) and explore effects on topology and support. We identify a well-supported topology based on all loci and find that a relatively small number of high-resolution gene trees can be sufficient to converge on the same topology. Adding gene trees with decreasing resolution produced a generally consistent topology, and increased confidence for specific bipartitions that were poorly supported when using a small number of informative loci. This corroborates coalescent-based simulation studies that have highlighted the need for a large number of loci to confidently resolve challenging relationships and refutes the notion that low-resolution gene trees introduce phylogenetic noise. Further, our study also highlights the value of quantifying changes in nodal support

across locus subsets of increasing size (but decreasing gene tree resolution). Such detailed analyses can reveal anomalous fluctuations in support at some nodes, suggesting the possibility of model violation. By characterizing the heterogeneity in phylogenetic signal among loci, we can account for uncertainty in gene tree estimation and assess its effect on the consistency of the species tree estimate. We suggest that the evaluation of gene tree resolution should be incorporated in the analysis of empirical phylogenomic datasets. This will ultimately increase our confidence in species tree estimation using summary-coalescent methods and enable us to exploit genomic data for phylogenetic inference.

INTRODUCTION

With the development of novel sequencing methods, phylogenomic datasets are generated that can contain hundreds to thousands of orthologous loci (Blair and Murphy 2011; Yang and Rannala 2012; McCormack *et al.* 2013). This steep change in the availability of genetic markers has led to a focus on phylogenetic methods that accommodate the frequently observed incongruence in evolutionary histories among loci (Jennings and Edwards 2005; Edwards 2009; Lemmon and Lemmon 2013; Nater *et al.* 2015). Topological discordance due to biological sources such as incomplete lineage sorting, can be expected under a wide variety of realistic evolutionary scenarios and therefore needs to be evaluated when inferring species trees (Pamilo and Nei 1988; Maddison 1997; Degnan and Rosenberg 2009). Most species tree methods aim to account for this heterogeneity in coalescent histories by incorporating the multi-species coalescent (MSC) model (Knowles 2009; Liu *et al.* 2009; Degnan and Rosenberg 2009; Nakhleh 2013; Liu *et al.* 2015a). Using the MSC and independent genetic markers, these methods weigh distinct species tree hypotheses by comparing the observed distribution in gene trees with an expected distribution given a species tree hypothesis. Full-coalescent sequence based methods that jointly infer gene trees and species tree, such as *BEAST (Heled and Drummond 2010) or BEST (Liu 2008), are preferable but also remain computationally intractable with a large number of loci (Leaché and Rannala 2011). Thus, the use of such methods remains unfeasible for most phylogenomic studies and alternative approaches have been developed that alleviate computational burden (Kubatko *et al.* 2009; Liu *et al.* 2010; Mirarab *et al.* 2014b).

Coalescent methods based on summary statistics ('summary-coalescent'), use a two-step approach where individual gene trees are initially inferred and species tree inference is conducted by summarizing across the resulting collection of gene trees.

Summary-coalescent methods are computationally cheap and have become increasingly popular as the species tree method of choice for empirical phylogenomic studies using full-sequence data (Lemmon *et al.* 2012; Ilves and López-Fernández 2014; Bond *et al.* 2014; Leaché *et al.* 2014; Giarla and Esselstyn 2015). However, summary-coalescent methods have been criticized for disregarding uncertainty in gene tree estimation and their unrestricted adoption has been questioned (Gatesy and Springer 2013; 2014; Springer and Gatesy 2015). Although statistically inconsistent (Roch and Steel 2015), concatenating loci can under some circumstances result in a better estimate of the underlying species tree than a summary-coalescent tree that is based on gene trees with poor phylogenetic signal (Mirarab *et al.* 2014a). The potential problems associated with gene tree accuracy has motivated some phylogeneticists to favor concatenation over summary-coalescent approaches (Gatesy and Springer 2013; 2014; Springer and Gatesy 2015; but see Edwards *et al.* 2016). However, the relative importance of gene tree inaccuracy remains undetermined (Huang *et al.* 2010), since other studies have suggested that accurate species trees can still be inferred even in the presence of gene tree estimation error (Roch and Warnow 2015). Thus, empirical studies incorporating summary-coalescent methods should validate the consistency of species tree inference, while accounting for the potential uncertainty in gene tree estimation.

There are two main sources underlying gene tree estimation error. First, model misspecification can result in systematic error, where alternative gene trees are consistently recovered due to the fit of erroneous models of sequence evolution (Jeffroy *et al.* 2006; Kumar *et al.* 2012; Doyle *et al.* 2015). Second, if the phylogenetic information content (PIC) of an individual locus is relatively low, the inferred gene tree will be inaccurate and conflicting topologies are equally likely ('low resolution'). The distribution of PIC in current phylogenomic studies is often uneven, since genetic markers frequently vary in length and/or mutation rate (Faircloth *et al.* 2012;

Lemmon *et al.* 2012; Lanier and Knowles 2012). The inferred gene trees in empirical studies will therefore differ in precision and gene tree estimation error can be regarded as a stochastic artifact inherent to the method used for generating phylogenomic sequence data, the loci targeted and the success of reassembling contigs.

Whereas the performance of species tree methods has been extensively tested (McCormack *et al.* 2009; Leaché and Rannala 2011; Mirarab *et al.* 2014a), these simulations often do not address the observed heterogeneity in PIC among loci and the corresponding variation in resolution of inferred gene trees. Simulation studies that have explicitly accounted for variation in PIC show that the use of genes with higher mutation rates (Huang *et al.* 2010; Lanier *et al.* 2014; Giarla and Esselstyn 2015) or longer length (Liu *et al.* 2015b) can significantly increase the accuracy of species tree estimation. Yet, the effect of adding low-resolution loci to an informative dataset remains unclear; some simulations suggest that such gene trees do not affect the species tree estimate (Lanier *et al.* 2014), whereas others report a decrease in performance rather than an improvement in accuracy (Liu *et al.* 2015b).

The concerns regarding gene tree estimation error and the observed variation of gene tree resolution in empirical datasets have motivated us to explore the performance of summary-coalescent species tree inference while explicitly considering the resolution of the included gene trees. Recently developed target-capture methods provide opportunity to generate large-scale DNA sequence datasets (Faircloth *et al.* 2012; Lemmon *et al.* 2012; Bragg *et al.* 2016; Jones and Good 2015) and to use empirical data for characterizing the effect of gene tree resolution on species tree inference**.** Here we thoroughly explore an exon-capture dataset (~2000 loci) to address this issue and simultaneously aim to resolve a challenging radiation of Australian lizards.

Lizards of the genus *Cryptoblepharus* Wiegmann (Reptilia: Squamata: Scincidae) are small scansorial skinks that range through three broad, geographic regions: The Ethiopian-Malagasy (southwest Indian Ocean), Indo-Pacific and Australian (Ineich and Blanc 1988; Rocha *et al.* 2006; Hayashi *et al.* 2009; Blom 2015a). A thorough revision of 396 Australian *Cryptoblepharus* individuals (Horner and Adams 2007), using 45 genetic (allozyme) and 33 morphological markers, increased the number of described Australian species from seven to 25. Based on genetic differences between taxa, the diversification of Australian lineages has likely occurred in two discrete radiations (clades A and B). The crown age for each clade is around 5 Ma., but clade B contains more lineages (11 and 17 taxa respectively) and so has diversified more rapidly (Blom *et al.* 2016). Phylogenetic relationships in recent, rapid radiations can be notoriously difficult to resolve due to the widespread presence of incomplete lineage sorting and the potential reliance on genetic markers that evolve slowly relative to the rate of speciation (Giarla and Esselstyn 2015). Hence, this Australian group of skinks provides an excellent opportunity to study potential sensitivities in the inference process, while resolving radiations with distinct rates of diversification (i.e. clades A and B).

Here we quantify the impact of stochastic gene tree estimation error on summary-coalescent species tree inference by screening loci based on gene tree resolution prior to species tree estimation. We use a recently developed measure (Salichos and Rokas 2013; Salichos *et al.* 2014) to quantify the degree of conflict among all bipartitions present in a set of bootstrapped trees as a proxy for gene tree resolution. It is not the aim of this study to explicitly compare the performance between species tree methods (Leaché and Rannala 2011), investigate the effects of missing data (Streicher *et al.* 2015) or explore the effect of systematic gene tree estimation error (Doyle *et al.* 2015). Rather we aim to characterize the heterogeneity in gene tree resolution that is frequently observed in empirical datasets and to quantify its effect on topology and support of species trees. Applying such thorough

evaluations will enable empiricists to benefit from the additional value that novel molecular approaches offer to the field of phylogenetics.

MATERIALS AND METHODS

Taxon Sampling

Based on results from Horner and Adams (2007), we selected a single representative for each of the 28 identified allozyme lineages of Australian *Cryptoblepharus* (Supplementary Table 1). Three of these lineages have not been diagnosed as species, due to morphological and geographic overlap with currently described species (Horner 2007). Here they are treated as separate taxa because the available samples exhibit a considerable degree of genetic differentiation from other species (Horner and Adams 2007).

Most specimens included in this study are held in the collections of the Museum and Art Gallery of the Northern Territory (MAGNT), Western Australian Museum (WAM) or Queensland Museum (QM) and were used in the initial taxonomic revision. We used tissues available from the Australian Biological Tissue Collection at the South Australian Museum, unless tissues previously analyzed were depleted. For these species, we used recently collected field samples. Museum specimens from Horner and Adams (2007) have associated allozyme profiles, but we screened recently collected samples for morphological characteristics and sequenced a mitochondrial marker (*ND2*) to verify correct species assignment. We did not verify lineage assignment for one species, *C. gurrmul,* since the amount of available tissue was very limited. *C. gurrmul* is endemic to a small number of islands and is the only *Cryptoblepharus* species present in the sampled location (Horner 2007). We chose another taxon within

the Eugongylus group (*Bassiana duppereyi*) as an outgroup, based on existing phylogenetic hypotheses (Brandley *et al.* 2015).

Exon Capture Design, Library Preparation and Sequencing

The design of the exon-capture kit used in this study is outlined in detail in Bragg *et al.* (2016). Briefly, we identified a set of single-copy exon targets with a balanced base composition, in the *Anolis* genome and identified their orthologs in transcriptomes of three species from genera related to *Cryptoblepharus* (*Carlia rubrigularis*, *Lampropholis coggeri* and *Saproscincus basiliscus*; Singhal 2013). A total of 3320 loci were targeted (>200 base pairs), with a total target length of 4.31 Mb (including a representative of each exon from each of the three species). Based on these exon targets, Roche NimbleGen designed and synthesized a SeqCap EZ Developer Library as our probe set. These probes capture homologous targets with high efficiency across the entire Eugongylus group, to which *Cryptoblepharus* belongs (Bragg *et al.* 2016).

We extracted genomic DNA from liver tissue stored either frozen or in RNALater, following the salting-out method of Sunnucks and Hales (1996). We prepared genomic libraries with ~1400 ng. input DNA per sample and according to the protocol of Meyer and Kircher (2010), using modifications of Bi *et al.* (2012). In brief, library preparation consisted of blunt-end repair, adapter ligation, adapter fill-in and was followed by two independent index-PCRs to reduce PCR bias. Each sample had a unique barcode for pooling DNA for the hybridization. We assessed DNA concentrations using a Nanodrop (Thermo scientific) and the distribution of fragment lengths on 1.5% agarose gels. Barcoded libraries were pooled in equimolar ratios prior to hybridization. The exon-capture hybridization was performed following SeqCap EZ Developer Library user's guide (Roche NimbleGen). We assessed the quality of the hybridizations using qPCR following methods of Bi *et al.* (2012). The qPCR assays used specific primers to assess

enrichment of targeted regions, and de-enrichment of non-targeted regions, of the genome. In addition, the quantity and quality of the hybridizations were measured using a Bioanalyzer (Agilent Technologies), to quantify the concentration of the pre- and post-capture libraries. Once the libraries passed all aforementioned quality checks (i.e. successful enrichment), they were submitted for sequencing (Functional Genomics Lab QB3 core facility, UC Berkeley). We sequenced the enriched libraries (100bp paired-end) on a single Illumina HiSeq 2000 lane.

Read Processing and Assembly

Illumina sequencing reads were processed using a workflow that was described previously by Singhal (2013), and is available at https://github.com/MVZSEQ. The workflow removes duplicate, low complexity and contaminant reads, and trims adaptors and low quality bases (TRIMMOMATIC v0.22, Bolger *et al.* 2014). Overlapping reads were merged using FLASH (v1.2.2, Magoc and Salzberg 2011).

Cleaned sequencing reads were assembled using an approach described by Bragg *et al.* (2016). Briefly, for each sample, each locus was assembled separately, after identifying reads with homology to the encoded protein (using blastx, v2.2.25, expectation value = 1E-9, Altshul *et al.* 1990). These reads were then assembled with VELVET (K values 31, 41, 51, 61, 71 and 81; v1.2.08, Zerbino and Birney 2008). Contigs for a locus were combined using CAP3 (parameter values –o 20 –p 99, version date 08/06/13, Huang and Madan 1999), and trimmed to the exon boundaries using EXONERATE (v2.2.0, Slater and Birney 2005). Where multiple contigs were assembled, we used a reciprocal best blastx hit criterion to select a contig orthologous to the targeted *Anolis* exon (see Bragg *et al.* 2016).

Alignment and Data Pre-processing for Species Tree Inference

High-quality sequence alignments are essential for correct phylogenetic inference (Zwickl *et al.* 2014) but visual inspection of alignment quality, as traditionally conducted, is challenging with thousands of loci (Lemmon and Lemmon 2013; Blom 2015a). We developed a flexible bioinformatic workflow for alignment and alignment filtering of exonic sequences, EAPhy (v0.9, Blom 2015b), and generated high-quality alignments for each specific analysis. In brief, EAPhy aligned sequences using MUSCLE (v3.8.31, Edgar 2004), performed checks to ensure coding of amino acids and removed missing data from the ends of the alignments. We visually inspected alignments after filtering if individual sequences still deviated significantly (more than three amino-acids in a seven base window) from the alignment consensus or contained more than one stop codon. EAPhy generates the desired input format (i.e. Phylip, Fasta) for most species tree methods and the complete pipeline can be found at (https://github.com/MozesBlom/EAPhy).

Assessing Stochastic Gene Tree Estimation Error

Stochastic gene tree estimation error is driven by a lack of phylogenetic signal, and the phylogenetic informativeness of loci is often characterized as the total number or relative ratio of parsimony informative sites (e.g. Leaché *et al.* 2014). However, the relationship between gene tree accuracy and the number of informative sites is not necessarily proportional. For example, only a small proportion of sites might prove informative for challenging clades with high rates of diversification because most informative sites will only differentiate between major clades or in- and outgroup taxa (Townsend 2007). Thus, whereas the number of informative sites might be high for a given locus, the gene tree accuracy in some parts of the tree might still be relatively poor. An alternative way to evaluate gene tree estimation is to compare the

consistency of bootstrap replicates. This could be calculated as the average bootstrap support (BS) value across all internal nodes. This would provide an indication of the overall consistency, but does not necessarily estimate the accuracy of the inferred gene tree across bootstrap replicates. For example, if a certain bipartition is observed in 51/100 bootstrapped trees, it is unknown whether the second most common, but conflicting, bipartition is supported by the remaining 49 bootstrapped trees or only a few. Here we assume that the estimate of each bipartition is more accurate when the second most common, conflicting, bipartition is observed at low frequency. To estimate the accuracy of each gene tree, we used the 'Tree Certainty' (TC) score introduced by Salichos *et al.* (2013; 2014). The TC score is the sum of 'Internode Certainty' estimates, which represent the support for each bipartition of a given topology by considering its frequency in a set of bootstrapped trees, jointly with that of the most prevalent conflicting bipartition in the same set of trees. If gene tree estimation is precise, most bipartitions across the tree are consistently recovered in much higher frequencies than conflicting bipartitions and TC, the sum of internode certainty estimates, is large. Alternatively, TC will be small if many internodes have a low resolution, suggesting that a high frequency of conflicting bipartitions have been inferred across bootstrap replicates. For each locus, we used RAxML (v8.1, Stamatakis 2014) to infer the gene tree with the highest likelihood out of 10 replicates (GTR + Γ) and estimated the TC score based on 100 bootstrap replicates. We calculated the TC score by inferring the majority rule consensus tree across bootstrap replicates and summing the internode certainty scores for each of the inferred bipartitions of the consensus tree. Finally, to assess a potential correlation between locus length and TC score, we used a linear model in R ('lm' function, v3.1.2, R Core Team 2014).

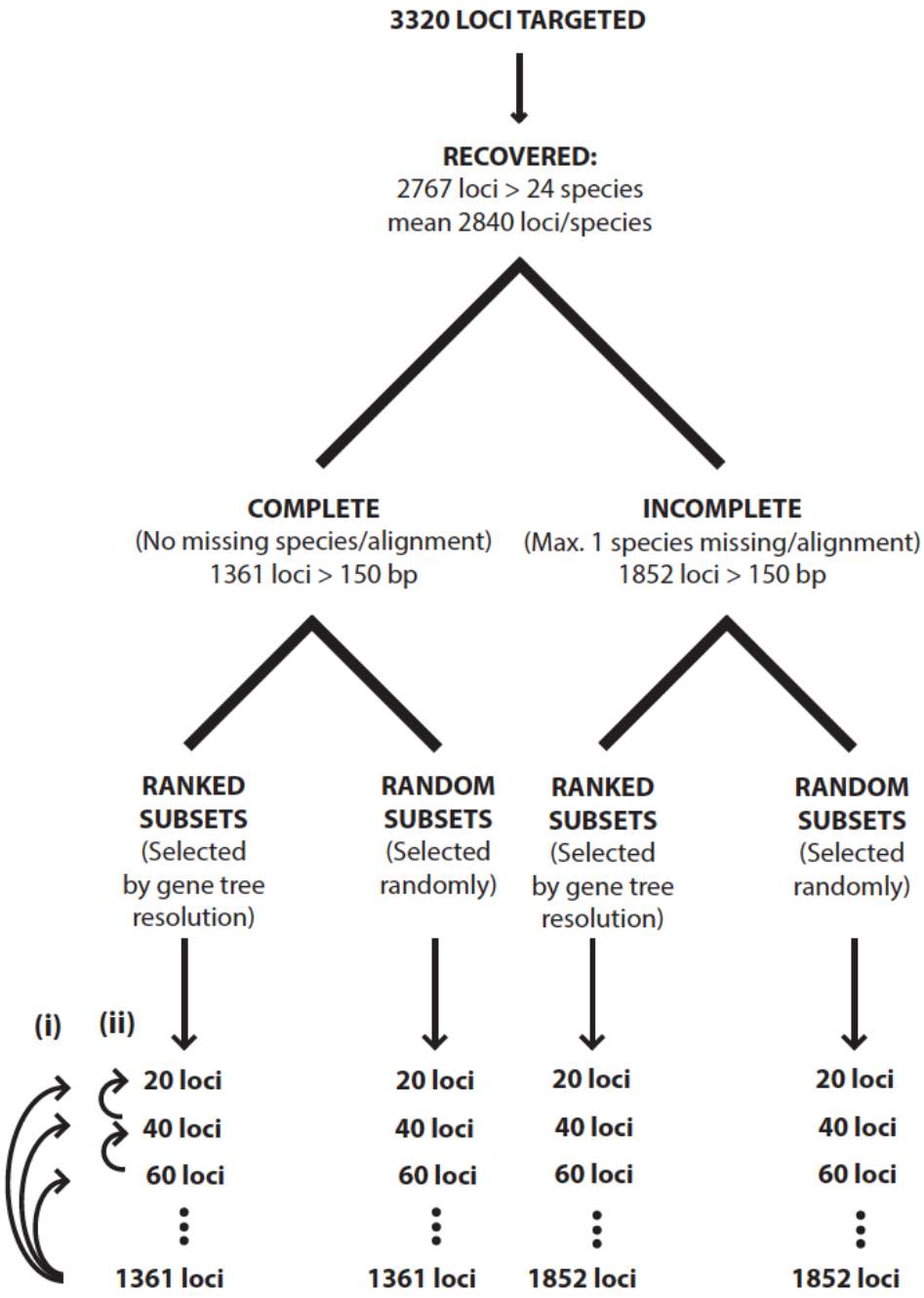


Figure 1. Schematic of the data structure used for analyses. Analyses were conducted on a dataset that contained alignments without missing species and a dataset that also included alignments that missed up to one species. Each dataset was divided in subsets, containing gene trees that were either picked randomly or ranked by Tree Certainty (TC) scores. After inferring an ASTRAL species tree for each subset of loci, Robinson-Foulds distances were calculated by comparing the species tree for each locus subset with the species tree based on all loci (i) or the species tree based on the previous locus subset (ii).

Species Tree Inference

Prior to estimating species trees with coalescent based methods, we inferred a maximum-likelihood (ML) phylogeny based on a concatenated alignment of all loci. When loci are concatenated, variation in genealogical histories is not considered explicitly, and this can result in inflated support metrics (Kubatko and Degnan 2007; Edwards *et al.* 2007; Knowles 2009; Roch and Steel 2015). However, in addition to providing an initial phylogenetic hypothesis, we generated a concatenated ML phylogeny to test whether the inferred branch lengths were predictive for the degree of discordance among loci. That is, we expected that a shorter interval between species splits would result in a higher degree of incomplete lineage sorting and thus require more loci to confidently infer the underlying species history in a coalescent-based framework (Degnan and Rosenberg 2009). Lastly, we assessed whether species tree methods that incorporate a full-coalescent model benefit from the identification of loci with a high TC-score. Full-coalescent based methods are computationally intractable with large numbers of loci and the ranking of loci by TC score can prioritize the inclusion of more informative loci over others, potentially resulting in more accurate inference than when using a random subset.

Concatenated ML species tree.— We generated a dataset where all *Cryptoblepharus* species were represented and where each alignment had a minimum length of a 150 bp. We then used PartitionFinder (rcluster search, v1.1.1, Lanfear *et al.* 2012) to identify the optimal partitioning scheme and substitution model for the concatenated alignment by considering both gene and codon position. Using the optimal partitioning strategy and appropriate substitution model (GTR + Γ), we inferred a ML phylogeny from the concatenated dataset using RAxML. The tree with the highest likelihood score was selected out of 100 replicate searches and BS values calculated for each bipartition based on 1000 bootstrap replicates.

Summary-coalescent species trees.— To evaluate the impact of adding loci with decreasing phylogenetic resolution, we used the RAxML gene tree and TC score for each locus, and inferred a summary-coalescent species tree for subsets of loci using ASTRAL II v.4.7.6 (Fig. 1; Mirarab and Warnow 2015). We used ASTRAL II since it performs better or equally well in comparison to other summary-coalescent methods (Chou *et al.* 2015; Ogilvie *et al.* 2016), is computationally efficient and uses unrooted gene trees.

Locus subsets of equal size were either chosen randomly from all available loci ('random') or by highest remaining TC value ('ranked'). For the first 200 loci, we iteratively increased the size of each subset with 20 additional loci. During subsequent iterations, the size of each subset was increased with 100 loci until all available loci were included (Fig. 1). We inferred a summary-coalescent species tree during each iteration for both the random and ranked locus subsets and used multi-locus bootstrapping, as incorporated in ASTRAL II, to calculate BS values for each bipartition (100 replicates).

We inferred species trees for complete gene trees where all species were present (Fig. 1; 'complete' dataset) and for incomplete gene trees where up to one of the 28 species could be absent (Fig. 1**;** 'incomplete' dataset). It is possible to include incomplete gene trees since ASTRAL II uses all quartet trees supported in each gene tree for scoring the species tree, whether the taxa are all present or not (S. Mirarab, pers. comm.). We set the criterion of one missing species for two reasons. First, we calculated TC scores for gene trees where data for up to three species were missing (i.e. ~ 10%), and there was not a substantial improvement in the number of alignments with high TC scores when allowing for two or three missing sequences. Second, we have been conservative regarding the degree of missing data permitted since the heuristics of ASTRAL II, which account for the uncertainty that is potentially

introduced by tolerating missing taxa, seem to generate consistent results with various degrees of missing data (Xi *et al.* 2016) but have not been tested exhaustively.

For the bipartitions that were supported in both the concatenated ML phylogeny and the summary-coalescent species trees, we extracted the internode branch length (in mean number of nucleotide substitutions per site) defining each bipartition in the ML phylogeny and calculated the corresponding average BS value across summary-coalescent trees. To calculate the average BS values for each supported bipartition, we took the summary-coalescent trees based on the alignments from the incomplete dataset and used species trees based on ranked subsets of loci. The size of these locus subsets increased iteratively with 100 loci in the range of 100 to 1800 loci. We assessed a potential correlation between branch length and average BS for each bipartition using a linear model in R ('lm' function), but excluded the bipartitions that are strongly supported (BS = 100) across all subsets since these are uninformative to the model.

Full-coalescent species trees.— We ranked loci by TC score and generated subsets of the 20, 30, 40 and 50 highest ranked loci (complete dataset – i.e. no missing taxa). Furthermore, we generated eight random subsets of 20 loci and five random subsets of 30 loci, from the 200 loci with the highest TC score. We then used *BEAST v2.1.3. (Bouckaert *et al.* 2014) to infer both gene and species trees. We ran each *BEAST run in duplicate, with separate starting seeds, using a GTR + Γ substitution model with four Γ rate categories, a strict clock model and applied a birth-death species tree prior. Each analysis was run for 200 million generations and we sampled the MCMC chain every 200,000 generations. We discarded the first 10% as burn-in, used Tracer v1.5 to check for convergence (estimated sample size; ESS > 200) and LogCombiner v2.1.3. to combine the posterior sample of trees across runs. *BEAST analyses that failed to converge were rerun with 400 million generations (required for two random subsets

of 30 loci). Since subsets of loci with the 40 and 50 highest ranked TC scores still failed to converge, we only report the inferred species trees based on 20 and 30 loci. The species tree for each subset was individually summarized using TreeAnnotator v2.1.3.

Evaluating the Effect of Gene Tree Resolution

With ASTRAL II, we inferred a summary-coalescent species tree for each locus subset. This resulted in a sequence of species trees, where each tree is inferred based on a larger (ranked or random) subset of loci (see Fig. 1). To evaluate potential changes in topology across locus subsets of increasing size, we compared species tree topologies using a Robinson-Foulds (RF) tree distance calculation (Robinson and Foulds 1981). The RF distance represents the number of bipartition changes between two trees. By comparing the species tree for each locus subset with the species tree based on all loci for the complete or incomplete dataset, we quantified a) the number of changes in topology when adding gene trees randomly or b) the number of changes in topology when adding ranked gene trees with decreasing resolution. Since the 'true' species tree was unknown (in comparison to typical simulation studies), we compared the species tree inferred for each locus subset to the tree based on all loci, assuming that the species tree based on 'complete evidence' was the best species tree given the dataset. In addition, we also compared the species tree for each subset of loci to the species tree based on the previous subset of loci (e.g. subset with 300 loci vs. subset with 200 loci), to circumvent the inevitability of increasing topological concordance as the number of loci approaches the full dataset (see Fig. 1). Lastly, as for the ASTRAL comparisons, we evaluated topological changes across *BEAST runs by comparing the species tree for each subset to the ASTRAL species tree based on all loci (incomplete dataset – i.e. 1852 loci, one individual missing).

To calculate RF distances between trees, we rooted each species tree (both ASTRAL and *BEAST) on the *B. duppereyi* outgroup and used the *'symmetric_difference'* function in the Python module Dendropy v.4.0.2 (Sukumaran and Holder 2010). All analyses were conducted for both the complete and incomplete dataset.

RESULTS AND DISCUSSION

Characteristics of Loci

We used a custom exon-capture system that was designed for related Eugongylus skinks, and observed considerable capture success across most species of *Cryptoblepharus*. On average, 2840 out of 3320 targeted loci were successfully assembled for each individual, with a mean coverage of 140X (Supplementary Table 1). These results are in line with Bragg *et al.* (2016), who achieved similar capture success across lizard genera with up to ~40 million years of divergence.

The majority of loci were sequenced across many species (i.e. 2767 loci – 24 species or more; Supplementary Fig. S1), but a significant proportion of the loci (~60%) was not recovered across all species in the study. In total, 1484 loci were successfully recovered across all target species and 1361 of these were longer than 150 bp (complete dataset; mean length 384 bp/alignment). If we relaxed the criterion for missing data and allowed up to one species missing per alignment, an additional 491 loci (each > 150 bp) were included, yielding a total dataset of 1852 genes (incomplete dataset; mean length 410 bp/alignment).

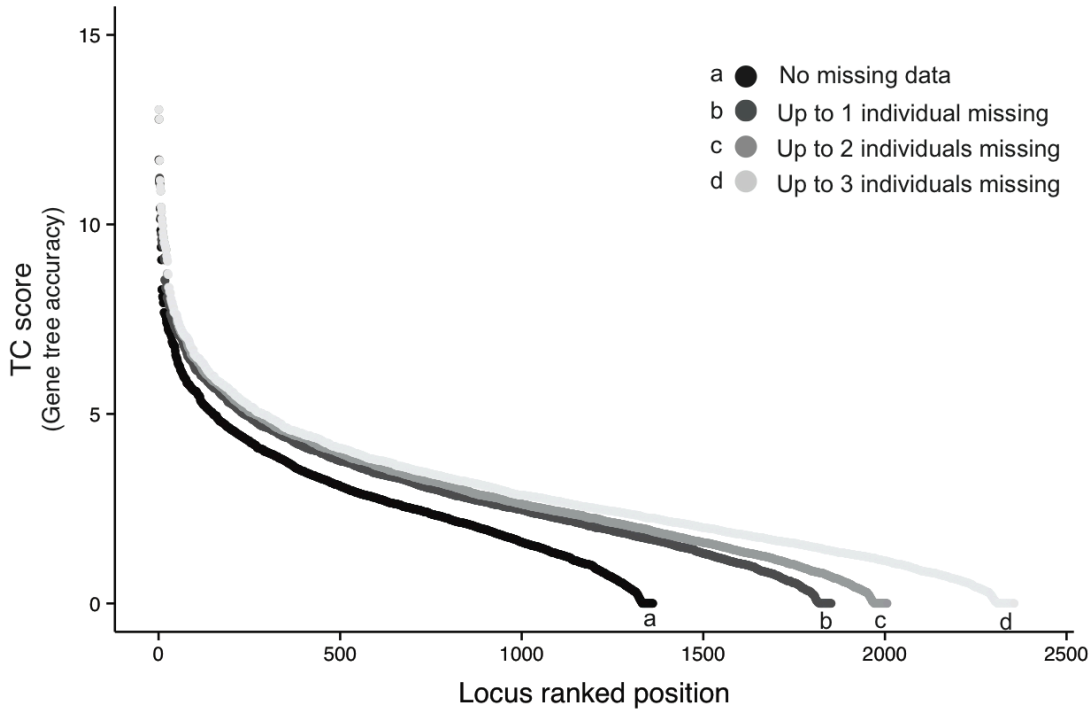


Figure 2. The inferred TC score for each alignment and their ranked position. Different shades (a-d), correspond to the maximum number of missing species allowed per alignment. By relaxing the criterion for missing species, the total number of recovered loci and the average TC score, gene tree accuracy, for a given number of loci increased.

We found substantial variation in TC scores for both the complete and incomplete dataset (Fig. 2), illustrating the heterogeneity of resolution among gene trees. This motivated us to further examine how such variation might affect species tree inference. In comparison to the complete dataset, the incomplete dataset yielded more loci and the average gene tree resolution increased for each similar sized locus subset ranked by TC score (Fig. 2). To verify whether allowing for additional missing species (n = 2 and 3 missing) would result in an even further increase in average TC score per locus subset, we also calculated the TC scores for genes that had up to three species missing per alignment. This did not result in a sizeable gain in gene tree resolution across similar sized locus subsets (Fig. 2) and we therefore limited our analyses to the complete and incomplete (up to one species missing) dataset.

TC score was positively correlated with locus length (linear model, N = 1361, r^2 = 0.369, $p < 0.01$; Supplementary Fig. S2), suggesting that longer loci on average result in more accurate gene trees (Liu *et al.* 2015b). Previous studies have recommended several alternative features that increase the PIC of loci, such as high mutation rate (Lanier *et al.* 2014; Giarla and Esselstyn 2015), but we expect that locus length is likely the most unbiased surrogate to select for and, as demonstrated here, can significantly increase gene tree accuracy. While aiming for loci with a high mutation rate would induce an increase in the number of parsimony informative characters, it might also bias branch length estimation or result in erroneous inference due to saturation. These results (Supplementary Fig. S2) suggest it might be advantageous to target long loci where possible in sequence-capture studies and we encourage further technical development on this front (e.g. Faircloth *et al.* 2012; Lemmon *et al.* 2012). The use of long loci might have some adverse effects if it increases the risk of intra-locus recombination; but simulation experiments suggest that a limited amount of recombination within loci has a relatively modest effect on the inferred species tree (Lanier and Knowles 2012).

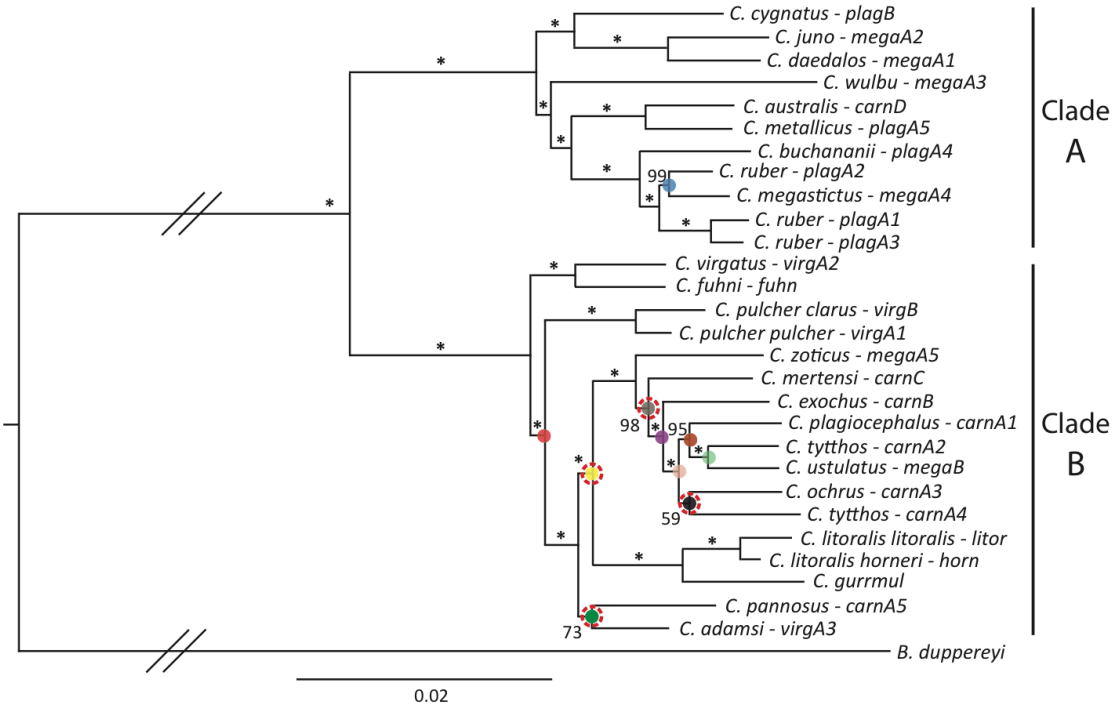


Figure 3. Phylogeny of Australian *Cryptoblepharus* based on a concatenated maximum-likelihood analysis of 1361 exons (complete dataset). Values at internode branches reflect bootstrap support (BS) and an asterisk (*) denotes BS = 100. The nodes annotated with colored orbs, were poorly supported (average BS < 95) using ASTRAL with 80 ranked loci and color scheme used matches across figures (Fig. 4 and 5). Colored nodes were encircled with a dotted red line if the topology differed in the ASTRAL analyses. The current species name and allozyme group (sensu Horner and Adams, 2007) are provided at each tip.

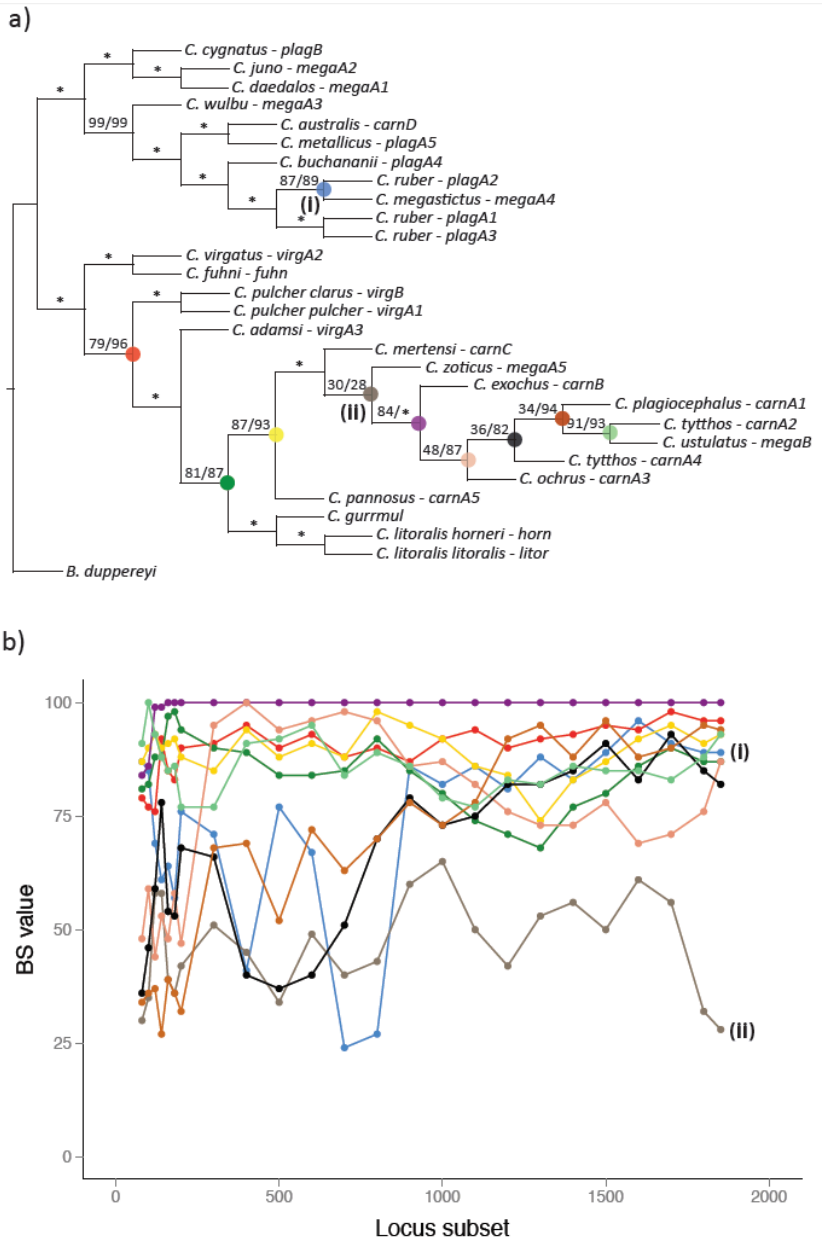


Figure 4. a) Species topology of Australian *Cryptoblepharus*, based on summary-coalescent (ASTRAL) analyses of 80 or 1852 ranked exons (incomplete dataset). The species tree topology changed between initial subsets (20, 40 and 60 loci), but is generally stable from subsets with 80 ranked exons and above. Values at internode branches reflect BS and an asterisk (*) denotes BS = 100. Where two BS values are denoted, the first one is for the species tree based on 80 ranked exons and the second value is for the species tree based on all 1852 exons. The nodes annotated with colored orbs, were poorly supported (average BS < 95) using 80 ranked loci. The current species name and allozyme group (sensu Horner and Adams, 2007) is provided at each tip. b) The change in BS across ranked locus subsets, for each node that was poorly supported (average BS < 95) using 80 ranked loci. The support for two nodes changed erratically, due to potential model violation of the structured coalescent (i.e. introgression) as described in the main text and are labeled (i) and (ii). The color scheme used matches across figures (Fig. 3, 4a and 5).

Concatenated and Summary-Coalescent Species Trees

The ML phylogeny (Fig. 3), based on concatenation of the complete dataset (1361 exons, no missing taxa: 522,592 bp), provides a significant gain in resolution of the evolutionary relationships of Australian *Cryptoblepharus* in comparison to the Neighbor-Joining (NJ) tree based on allozyme distances (Horner and Adams 2007). As in the allozyme analysis, the Australian representatives of the *Cryptoblepharus* genus are clearly separated in two distinct clades. The few well-supported bipartitions in the allozyme NJ tree are also present in the ML phylogeny, and there are only two nodes in the concatenated exon-capture analysis where the interspecific relationships are poorly resolved (BS < 90). However, the short branch lengths between many nodes in clade B warrant the application of species tree estimation methods that account for deep coalescence.

ASTRAL summary-coalescent species trees were inferred based on all loci for both the complete (Supplementary Fig. S3) and incomplete dataset (Fig. 4a). The species tree inferred for each dataset generally agrees with the concatenated ML phylogeny and there are no well-supported differences, except for the placement of *C. pannosus*. In the ASTRAL trees, *C. pannosus* is within the *C. gurrmul et al.* clade rather than a sister to this group, as suggested by the concatenated analysis (Fig. 3). However, the ASTRAL species tree based on the complete dataset is not fully resolved, with multiple unresolved nodes (BS < 95) that were highly supported in the concatenated ML phylogeny. For this species tree, the relationships remain unclear between *C. pulcher pulcher/C. pulcher clarus* and *C. virgatus*/*C. fuhni*, *C. zoticus* and *C. mertensi*, and within the clade involving *C. exochus/C. tytthos(carnA4)/C. ochrus/C. plagiocephalus/C. ustulatus/C. tytthos(carnA2)* (Supplementary Fig. S3). In comparison, the ASTRAL species tree based on the incomplete dataset, with 491 additional loci, has better overall support. The polytomy between *C. pulcher pulcher/C. pulcher clarus* and

C. virgatus/C. fuhni is resolved and is consistent with the concatenated ML tree (Fig. 4a). In addition, the resolution within the clade *C. exochus/C. tytthos(carnA4)/C. ochrus/C. plagiocephalus/C. ustulatus/C. tytthos(CarnA2)* is improved with increased BS for each inferred bipartition (Fig. 4b).

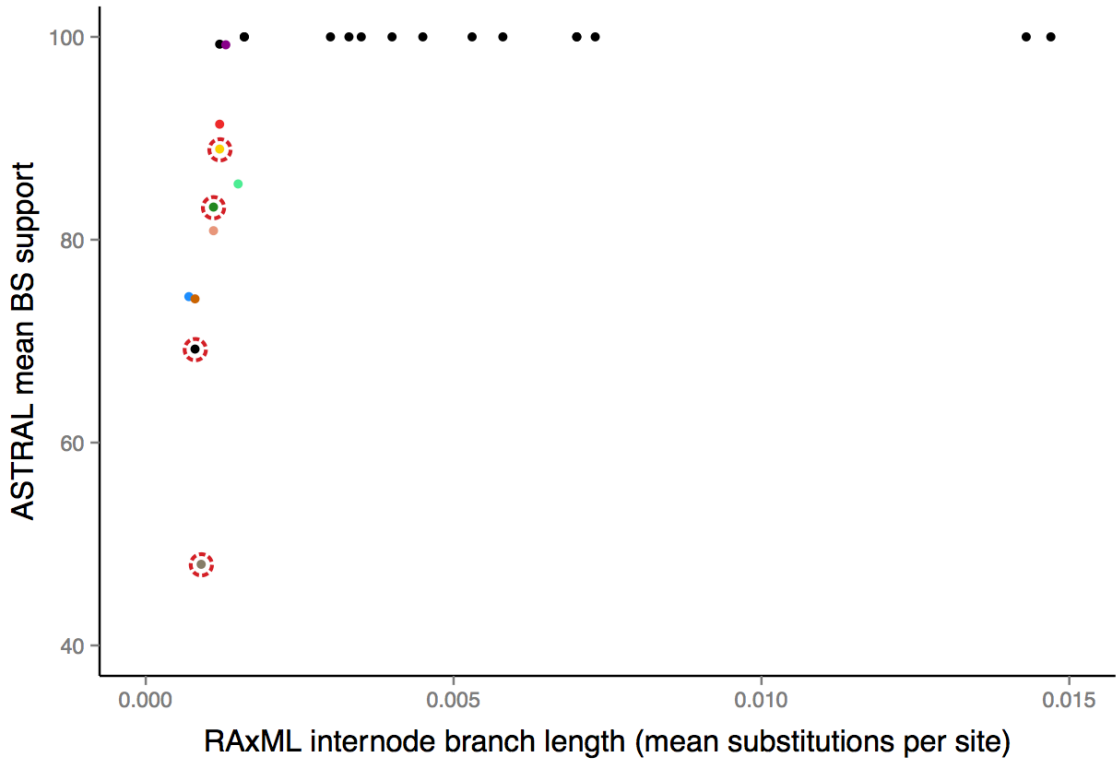


Figure 5. The RAxML internode branch length and mean BS across ranked locus subsets (range 100 – 1800 loci, incomplete dataset) for each node. Colors match scheme used in other figures (Fig. 3 and 4). Nodes not recovered in the ASTRAL analyses are encircled with a dotted red line.

Internode branch lengths in the RAxML phylogeny were positively correlated (N = 6, $r^2 = 0.56$, $p = 0.09$, Fig. 5) with average BS values across ASTRAL species trees, suggesting that longer periods between species splits are associated with greater concordance. Each bipartition in the RAxML phylogeny with an internode branch length shorter than 0.0015 (average expected substitutions per site) had an average

BS value below 100 in the summary-coalescent analysis. However, the majority of these challenging bipartitions were still strongly supported in the concatenated ML phylogeny (Fig. 3). Even though our ASTRAL species tree based on the incomplete dataset largely agrees with the RAxML phylogeny, caution is warranted regarding the strong support in the ML phylogeny for those bipartitions that are not consistently recovered across species tree inference methods (i.e. placement of *C. pannosus*). Numerous previous studies (e.g. Kubatko and Degnan 2007; Edwards *et al.* 2007; Knowles 2009; Roch and Steel 2015) have highlighted the potential for inflated BS in phylogenomic size datasets and our results support a similar conclusion. Nonetheless, a concatenated ML phylogeny is computationally cheap in comparison to many other species tree methods and our results demonstrate that, as expected, short branch lengths are indicative for diversification histories where a coalescent-based method is preferred.

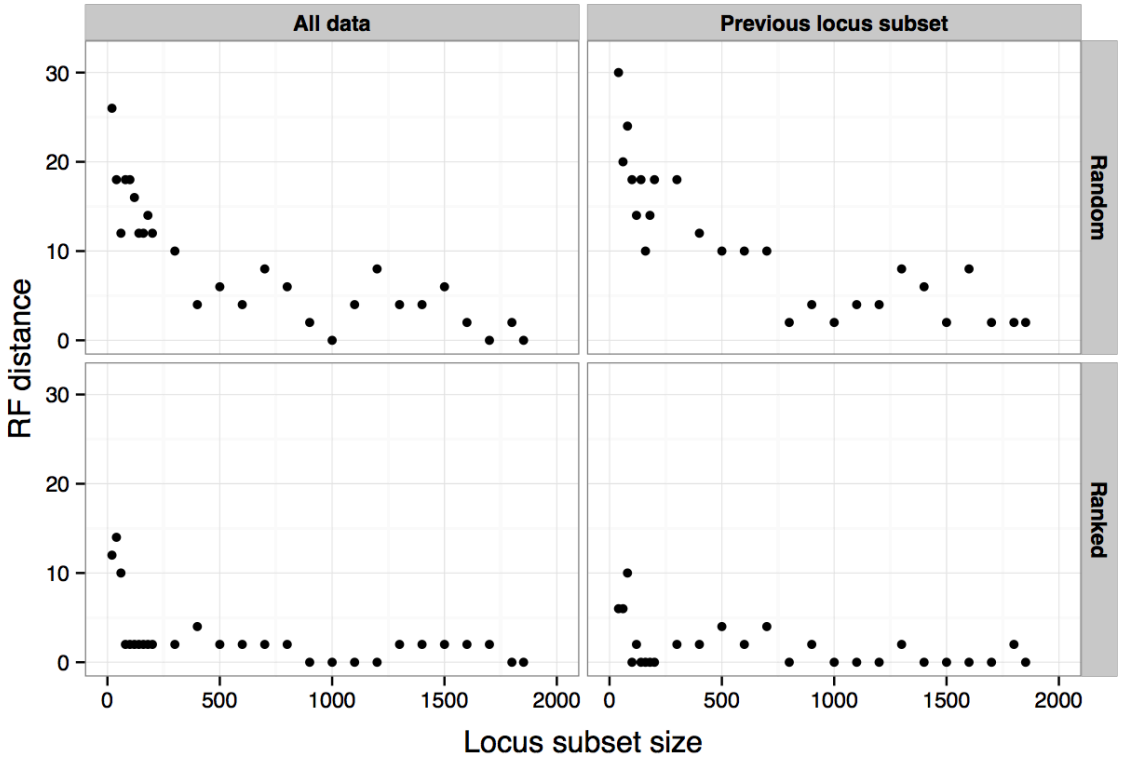


Figure 6. The Robinson-Foulds distance between the inferred ASTRAL species tree for each locus subset (incomplete dataset) and the ASTRAL species trees based on 1852 exons (left) or the previous locus subset (right). Subsets of loci were either chosen randomly (top) or ranked (bottom) by Tree Certainty score. The addition of loci with poor gene tree resolution does not increase topological discordance and limited numbers of ranked loci already converge on the most optimal topology given the dataset.

Effect of Locus Subsampling by TC score

To evaluate the effect of including gene trees with an increased probability of inaccurate inference on summary-coalescent species tree estimation, we characterized changes in both topology and support. First, we discuss changes in topology with an increase in number of loci while disregarding gene tree resolution. Second, we specifically examine the contrasting patterns of topological change between ranked and random locus subsets of increasing size. Third, in addition to topological variation, we also consider the changes in BS with increasing numbers of ranked loci. Fourth, we discuss whether the addition of low-resolution gene trees introduces phylogenetic noise or, in fact, can increase node support. Finally, we evaluate whether full-

coalescent species tree methods such as *BEAST benefit from the inclusion of loci with high TC scores, and whether joint estimation of gene and species trees lead to improved species tree estimation with similar numbers of loci.

Topology: number of loci.— The emergent patterns of topological change with increasing numbers of loci are similar for both the complete (all taxa present; Supplementary Fig. S4) and incomplete dataset (up to one taxa missing; Fig. 6). The RF distance between the species tree for each locus subset and the optimal species tree (i.e. based on all loci) decreases when species trees were inferred based on larger subsets of loci. There was no clear difference if we used the species tree inferred for each previous locus subset, instead of the optimal species tree, for comparison. The decrease in RF distance with additional loci, for both tree comparisons, suggests that topological similarity is not solely induced by an increasing similarity in number of loci between each subset and all loci. If variation in topology was driven by similarity in locus number alone, we would not expect the same decrease in RF distance with additional loci, when comparing the species tree for each subset with the species tree based on the previous locus subset. Even with the observed decrease in RF distance with a larger number of loci, some polytomies remain unresolved. The RF distance between trees therefore does not reduce to zero; even when large numbers of loci are included (Fig. 6). This remaining variation in RF distance depends on how polytomies (i.e. poor support) are 'arbitrarily' resolved in each tree, to represent a bifurcating tree. The fluctuation in RF scores is larger for the complete dataset (Supplementary Fig. S4), than for the incomplete dataset (Fig. 6), since the species tree based on the incomplete dataset is more confidently resolved (Fig. 4).

Topology: ranked vs. random gene trees.— When inferring species trees based on ranked subsets of loci, the most optimal topology given the dataset is inferred with fewer loci than if equal sized subsets were picked randomly (Supplementary Fig. S4**,**

Fig. 6). For the incomplete dataset, the inferred species tree based on the 80 loci with the highest TC scores is almost identical in topology to the species tree based on all 1852 loci (Fig. 4a). These trees only differ in how the polytomy involving *C. zoticus* and *C. mertensi* is resolved. The difference between ranked and randomly selected loci dissipates when the size of the locus subset increases, potentially due to the increased likelihood that the randomly selected subsets also contain high-resolution gene trees. In line with previous simulation studies that have evaluated summary-coalescent methods (Lanier *et al.* 2014; Liu *et al.* 2015b), high-resolution gene trees seem to have a stronger effect on the inferred topology than low-resolution gene trees. This is further illustrated when comparing the species trees based on small subsets of ranked loci for the incomplete dataset, with the species trees inferred for the complete dataset. The RF distances between trees in the complete dataset are higher, even for reasonably large subsets of loci, than when comparing between trees based on small subsets of ranked loci for the incomplete dataset (Fig.6, Supplementary Fig. S4). This difference is likely due to the larger number of high-resolution gene trees in the incomplete dataset (Fig. 2).

Bootstrap support.— Due to the minimal differences in topology between subsets of ranked loci (incomplete dataset), we evaluated the change in BS values for each bipartition (Fig. 4b) across species trees based on subsets of increasing size. Whereas topological changes are limited, the BS values for some nodes change substantially with increasing numbers of loci. Nodes that are well supported (BS = 100) remain high, but different patterns emerge for the more challenging nodes. For the nodes that had a BS value < 90 in the species tree based on 80 ranked loci (Fig. 4a), the accuracy increases on average with additional loci (Fig. 4b). But whereas BS increases coherently for some difficult nodes (e.g. the placement of the *C. pulcher* clade), BS estimates for other nodes fluctuate considerably with increasing size of locus subsets. Most notably, the BS for the polytomy involving *C. zoticus* and *C. mertensi* on average

fluctuates from 40 - 60 and ultimately, remains unresolved. Secondly, the BS values for the node involving *C. ruber (plagA2)* and *C. megastictus* change markedly between subsets, up to 1000 ranked loci, and then steadily increases for larger consecutive subsets. Interestingly the taxa for which BS values vary erratically with increasing subset size also match observations of mitochondrial introgression between distinct ecomorphs; specifically in these two species pairs (*C. ruber/C. megastictus, C. zoticus/C. mertensi*; Blom *et al.*, *in prep*.). These results suggest possible nuclear introgression and violation of the structured coalescent model. Most coalescent-based species tree methods assume that incongruence is solely due to incomplete lineage sorting, but introgression could result in similar patterns of topological discordance (Kutschera *et al.* 2014; Nater *et al.* 2015). By quantifying patterns of change in topology or support across locus subsets, such underlying signals can be distilled, inspected for biological relevance and potentially studied in a coalescent framework that also aims to model alternative sources of gene tree – species tree incongruence (i.e. a phylogenetic network approach (Yu *et al.* 2014)). Alternatively, SNP based approaches such as ABBA-BABA tests could potentially be employed when focusing on specific taxa that have putatively exchanged genetic variants (Blom *et al.*, *in prep*).

Phylogenetic noise.— The addition of gene trees with low TC scores increases the probability of including gene trees with stochastic estimation error. However, this did not result in any well-supported topological changes (Fig. 4a). For both the complete and incomplete dataset, a number of nodes remain unresolved and some topological variance between trees was generated depending on how such uncertain relationships were (arbitrarily) resolved. The remaining variation in RF distances between trees does not imply that the addition of low-resolution loci reduces the accuracy of the species tree estimate since we would expect that this would have a similar effect on well-supported nodes. The topological uncertainty driving the persistent variability in

RF distances between trees based on large locus subsets only occurs at nodes that were poorly supported throughout the analyses.

Rather than introducing phylogenetic noise (Liu *et al.* 2015b), our results suggest that the inclusion of low-resolution gene trees increases the consistency of phylogenetic inference with average BS values increasing across several of the most challenging bipartitions (Fig. 4b). Previous simulation studies (Leaché and Rannala 2011; Mirarab *et al.* 2014a) have highlighted that for most relationships, barring the 'anomaly zone' (Liu and Edwards 2009; Huang and Knowles 2009), coalescent-based approaches will infer the correct phylogenetic history with sufficient numbers of independent loci. Even though most of these analyses assume a homogeneous distribution of PIC across genes, our results suggest that even relatively uninformative loci still improved BS across some of the most challenging bipartitions (Fig. 4b). This was surprising to us and we encourage future simulation studies to further examine this empirical observation. We do not expect that this is an artifact of the method used since other nodes, such as the placement of *C. zoticus/C. mertensi,* remain unresolved; regardless of the number of gene trees included.

Full-coalescent analyses with ranked loci.— Lastly, we evaluated whether a small subset of the most informative loci could potentially be sufficient to infer the optimal species tree using a full-coalescent species tree analysis. Although the joint estimation of gene and species tree is generally preferred over summary-coalescent methods (Leaché and Rannala 2011; Ogilvie *et al.* 2016), we did not infer a more accurate species tree using *BEAST (Supplementary Fig. S5). The *BEAST species trees differed from the optimal species tree to a similar extent as ASTRAL trees based on ranked loci subsets of similar size (Supplementary Fig. S5). Furthermore, *BEAST trees based on similar sized (20 or 30 loci) random subsets of loci with the highest TC scores vary substantially (Supplementary Fig. S5). Currently, we cannot determine with certainty whether the differences between the *BEAST trees and the ASTRAL tree based on all

loci, occur due to methodological differences between full and summary-coalescent analyses, or if it is because *BEAST can only be run with a much smaller number of loci. Though the discordance among the *BEAST trees suggests the latter. However we expect this to be a fruitful area of research in the future, with improvements in the implementation of full-coalescent species tree analyses.

CONCLUSION

Species trees are commonly estimated based on gene trees. It is well known that those gene trees can be different from the species tree due to a variety of biological processes. For example, incomplete lineage sorting can be explicitly incorporated into species tree analyses, using coalescent-based methods. However, gene trees can also differ from the underlying species tree due to estimation error and unlike most simulation analyses, empirical phylogenomic studies often generate gene trees that widely vary in resolution. Here we quantified the impact of stochastic gene tree estimation error on summary-coalescent species tree inference by screening loci based on gene tree resolution prior to species tree estimation. Our results indicate that longer loci yield higher resolution gene trees, that a relatively small number (~80, incomplete dataset) of high resolution gene trees can already converge on the optimal topology given the dataset and that convergence on this topology occurs with fewer loci if gene trees are first ordered by resolution. Moreover, the addition of low-resolution gene trees did not introduce phylogenetic noise (Liu *et al.* 2015b) and in fact increased support for several challenging nodes. These empirical findings highlight the importance of gene tree resolution for species tree inference and are in line with previous simulation studies (Lanier *et al.* 2014; Mirarab *et al.* 2014a; Liu *et al.* 2015b).

Unlike most simulations, empirical phylogenomic datasets are highly heterogeneous in terms of phylogenetic information content. Rather than treating each sequenced locus as equal, our study suggests that by characterizing the heterogeneity in gene tree resolution and ranking them accordingly, we can account for stochastic gene tree estimation error when using summary-coalescent methods. Although it remains unclear to what extent our empirical findings can be generalized across taxa with different diversification histories, we have presented a conceptual approach that can be applied without difficulty in any other phylogenomic study where sequenced loci vary in PIC. By using the TC score of gene trees as a proxy for gene tree resolution, we have demonstrated how the ranking of loci by gene tree resolution prior to summary-coalescent species tree estimation can yield valuable insights. In addition to providing the means to evaluate the effect of including gene trees with an increased probability of inaccurate inference, it also enabled us to assess the importance of gene tree resolution in general and to identify characteristics of loci that can predict the accuracy of gene tree estimates (i.e. length). Thus our results strongly suggest that the evaluation of gene tree resolution is an informative and relevant practice that can greatly benefit empiricists.

Whereas summary-coalescent methods might prove effective in estimating a tree topology, full-coalescent sequence based methods can simultaneously infer a topology and branch lengths, and are therefore ultimately more appropriate for many purposes. A question that remains unexplored is to what extent a hybrid approach that incorporates both summary- and full-coalescent analyses can improve the computational efficiency and accuracy of species tree inference? Our results show that the majority of the *Cryptoblepharus* topology was unambiguously supported across most subsets of loci and that a small subset of informative loci already inferred the most optimal topology. If computationally efficient approaches such as summary-coalescent methods can resolve 'easy' nodes confidently, the search space of possible

trees could be reduced significantly for a full-coalescent method. It remains unknown whether these implementations would reduce the computational challenge of full-coalescent species tree estimation and allow for the inclusion of more loci, but it merits further investigation. In conclusion, our findings suggest there is much scope for future theoretical and empirical research into the best approaches for estimating species trees with large phylogenetic datasets. Such studies should explicitly consider the heterogeneity in phylogenetic signal among loci and how this translates into an accurate inference of the underlying species tree.

FUNDING

This work was supported by a grant from the Australian Research Council to C.M. (FL110100104).

ACKNOWLEDGEMENTS

We thank Paul Horner and Mark Adams who generously shared the morphological and allozyme data, which were beneficial to target appropriate representatives of each species. Ke Bi and Sonal Singhal provided guidance with both laboratory and bioinformatic procedures. M.P.K.B greatly benefited from the Computational Molecular Evolution EMBO workshop. We thank Huw Ogilvie and Robert Lanfear for insightful discussions regarding species tree inference analysis approaches. We thank the Australian Biological Tissue Collection at the South Australian Museum (Stephen Donnellan), the Western Australian Museum (Paul Doughty), the Queensland Museum

(Jessica Worthington Wilmer) and the Museum and Art Gallery of the Northern Territory (Stephen Richards) for access to tissues and specimens.

REFERENCES

Altschul S.F., Gish W., Miller W., Meyers E.W., Lipman D.J. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Bi K., Vanderpool D., Singhal S., Linderoth T., Moritz C., Good J.M. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC genomics. 13:403.

Blair C., Murphy R.W. 2011. Recent trends in molecular phylogenetic analysis: where to next? J. Hered. 102:130–138.

Blom M.P.K. 2015a. Habitat use and new locality records for *Cryptoblepharus poecilopleurus* (Squamata: Scincidae) from French Polynesia. Herp. Notes 8: 579-582.

Blom M.P.K. 2015b. EAPhy: A flexible tool for high-throughput quality filtering of exon-alignments and data processing for phylogenetic methods. PLoS Curr. 1 doi: 10.1371/currents.tol.75134257bd389c04bc1d26d42aa9089f.

Blom M.P.K., Horner, P., Moritz, C. 2016. Convergence across a continent: Adaptive diversification in a recent radiation of Australian lizards. Proc. R. Soc. B. 283:20160181.

Bolger A.M., Lohse M., Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30:2114–2120.

Bond J.E., Garrison N.L., Hamilton C.A., Godwin R.L., Hedin M., Agnarsson I. 2014. Phylogenomics resolves a spider backbone phylogeny and rejects a prevailing paradigm for orb web evolution. Curr. Biology. 24:1765–1771.

Bragg J.G., Potter S., Bi K., Moritz C. 2016. Exon capture phylogenomics: efficacy across scales of divergence. Mol. Ecol. Res. 16: 1059–1068.

Brandley M.C., Bragg J.G., Singhal S., Chapple D.G., Jennings C.K., Lemmon A.R., *et al.* 2015. Evaluating the performance of anchored hybrid enrichment at the tips of the tree of life: a phylogenetic analysis of Australian Eugongylus group scincid lizards. BMC Evol. Biol. 15:62.

Chou J., Gupta A., Yaduvanshi S., Davidson R., Nute M., Mirarab S. and Warnow T. 2015. A comparative study of SVDquartets and other coalescent-based species tree estimation methods. BMC Genomics 16:S2

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Doyle V.P., Young R.E., Naylor G.J.P., Brown J.M. 2015. Can we identify genes with increased phylogenetic reliability? Syst. Biol. 64:824–837.

Edgar R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 32:1792–1797.

Edwards S.V. 2009. Is a new and general theory of molecular systematics emerging? Evolution. 63:1–19.

Edwards S.V., Liu L., Pearl D.K. 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. U.S.A. 104:5936–5941.

Faircloth B.C., McCormack J.E., Crawford N.G., Harvey M.G., Brumfield R.T., Glenn T.C. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61:717–726.

Gatesy J., Springer M.S. 2013. Concatenation versus coalescence versus "concatalescence." Proc. Natl. Acad. Sci. U.S.A. 110:E1179.

Gatesy J., Springer M.S. 2014. Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. Mol. Phylogenet. Evol. 80:231–266.

Giarla T.C., Esselstyn J.A. 2015. The challenges of resolving a rapid, recent radiation: Empirical and simulated phylogenomics of Philippine shrews. Syst. Biol. 64:727–740.

Hayashi F., Shima A., Horikoshi K., Kawakami K., Segawa R.D., Aotsuka T., Suzuki T. 2009. Limited overwater dispersal and genetic differentiation of the snake-eyed skink (*Cryptoblepharus nigropunctatus*) in the oceanic Ogasawara islands, Japan. Zool. Sci. 26:543–549.

Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. Mol Biol Evol. 27:570–580.

Horner P. 2007. Systematics of the snake-eyed skinks, *Cryptoblepharus* Wiegmann (Reptilia: Squamata: Scincidae)–an Australian based review. The Beagle Supplement 3:21–198.

Horner P., Adams M. 2007. A molecular-systematic assessment of species boundaries in Australian *Cryptoblepharus* (Reptilia: Squamata: Scincidae): A case study for the combined use of allozymes and morphology to explore cryptic biodiversity. The Beagle Supplement 3:1–19.

Huang X., Madan A. 1999. CAP3: A DNA sequence assembly program. Genome Res. 9:868–877.

Huang H., He Q., Kubatko L.S., Knowles L.L. 2010. Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. Syst. Biol. 59:573–583.

Huang H., Knowles L.L. 2009. What is the danger of the anomaly zone for empirical phylogenetics? Syst. Biol. 58:527–536.

Ilves K.L., López-Fernández H. 2014. A targeted next-generation sequencing toolkit for exon-based cichlid phylogenomics. Mol. Ecol. Res. 14:802–811.

Ineich I., Blanc C.P. 1988. Distribution des reptiles terrestres en Polynesie Orientale. Atoll Res. Bull. 318:1–75

Jeffroy O., Brinkmann H., Delsuc F., Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225–231.

Jennings W.B., Edwards S.V. 2005. Speciational history of Australian grass finches (*Poephila*) inferred from thirty gene trees. Evolution. 59:2033–2047.

Jones M.R., Good J.M. 2016. Targeted capture in evolutionary and ecological genomics. Mol. Ecol. 25:185–202.

Knowles L.L. 2009. Estimating species trees: Methods of phylogenetic analysis when there is incongruence across genes. Syst. Biol. 58:463–467.

Kubatko L.S., Carstens B.C., Knowles L.L. 2009. STEM: Species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25:971–973.

Kubatko L.S., Degnan J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56:17–24.

Kumar S., Filipski A.J., Battistuzzi F.U., Kosakovsky Pond S.L., Tamura K. 2012. Statistics and truth in phylogenomics. Mol. Biol. Evol. 29:457–472.

Kutschera V.E., Bidon T., Hailer F., Rodi J.L., Fain S.R., Janke A. 2014. Bears in a forest of gene trees: Phylogenetic inference is complicated by incomplete lineage sorting and gene flow. Mol. Biol. Evol. 31:2004–2017.

Lanfear R., Calcott B., Ho S.Y.W., Guindon S. 2012. Partitionfinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29:1695–1701.

Lanier H.C., Knowles L.L. 2012. Is recombination a problem for species-tree analyses? Syst. Biol. 61:691–701.

Lanier H.C., Huang H., Knowles L.L. 2014. How low can you go? The effects of mutation rate on the accuracy of species-tree estimation. Mol. Phylogenet. Evol. 70:112–119.

Leache A.D., Rannala B. 2011. The accuracy of species tree estimation under simulation: A comparison of methods. Syst. Biol. 60:126–137.

Leaché A.D., Wagner P., Linkem C.W., Böhme W., Papenfuss T.J., Chong R.A., *et al.* 2014. A hybrid phylogenetic-phylogenomic approach for species tree estimation in African Agama lizards with applications to biogeography, character evolution, and diversification. Mol. Phylogenet. Evol. 79:215–230.

Lemmon A.R., Emme S.A., Lemmon E.M. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61:727–744.

Lemmon E.M., Lemmon A.R. 2013. High-throughput genomic data in systematics and phylogenetics. Annu. Rev. Ecol. Evol. Syst. 44:99–121.

Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics. 24:2542–2543.

Liu L., Edwards S. 2009. Phylogenetic analysis in the anomaly zone. Syst. Biol. 58:452–460.

Liu L., Wu S., Yu L. 2015a. Coalescent methods for estimating species trees from phylogenomic data. J. Syst. Evol. 53:380–390.

Liu L., Xi Z., Wu S., Davis C.C., Edwards S.V. 2015b. Estimating phylogenetic trees from genome-scale data. Ann. N. Y. Acad. Sci. 1360:36–53.

Liu L., Yu L., Edwards S.V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10:302.

Liu L., Yu L., Kubatko L., Pearl D., Edwards S.V. 2009. Coalescent methods for estimating phylogenetic trees. Mol. Phylogenet. Evol. 53:320–328

Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536

Magoc T., Salzberg S. 2011. FLASH: Fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 27:2957–2963

McCormack J.E., Huang H., Knowles, L.L. 2009. Maximum likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. Syst. Biol. 58:501–508.

McCormack J.E., Hird S.M., Zellmer A.J., Carstens B.C., Brumfield R.T. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol. Phylogenet. Evol. 66:526–538.

Meyer M., Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc.

Mirarab S., Bayzid M.S., Warnow T. 2014a. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. Syst. Biol. 65: 366–380.

Mirarab S., Reaz R., Bayzid M.S., Zimmermann T., Swenson M.S., Warnow T. 2014b. ASTRAL: Genome-scale coalescent-based species tree estimation. Bioinformatics. 30:541–548.

Mirarab S., Warnow T. 2015. ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics. 31:44–52.

Nakhleh L. 2013. Computational approaches to species phylogeny inference and gene tree reconciliation. Trends Ecol. Evol. 28:719–728.

Nater A., Burri R., Kawakami T., Smeds L., Ellegren H. 2015. Resolving evolutionary relationships in closely related species with whole-genome sequencing data. Syst. Biol. 64: 1000-1017.

Ogilvie H.A., Heled J., Xie D., Drummond A.J. 2016. Computational performance and statistical accuracy of *BEAST and comparisons with other methods. Syst. Biol. 65: 381–396.

Pamilo P., Nei M. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.

R Core Team. 2014. R: a language and environment for statistical computing. Available from: URL http://www.R-project.org/ (last accessed August 25, 2015)

Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Roch S., Steel M. 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theor. Pop. Biol. 100:56–62.

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

Rocha S., Carretero M., Vences M., Glaw F. 2006. Deciphering patterns of transoceanic dispersal: the evolutionary origin and biogeography of coastal lizards (*Cryptoblepharus*) in the Western Indian Ocean region. J. Biogeog. 33:13–22.

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

Salichos L., Stamatakis A., Rokas A. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol. Biol. Evol. 31:1261–1271.

Singhal S. 2013. De *novo* transcriptomic analyses for non-model organisms: an evaluation of methods across a multi-species data set. Mol. Ecol. Res. 13:403–416.

Slater G.S., Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31.

Springer M.S., Gatesy J. 2016. The gene tree delusion. Mol. Phylogenet. Evol. 94: 1–33.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312–1313.

Streicher J.W., Schulte II J.A., Wiens J.J. 2016. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in Iguanian lizards. Syst. Biol. 65: 128–145.

Sukumaran J., Holder M.T. 2010. DendroPy: A Python library for phylogenetic computing. Bioinformatics. 26:1569–1571.

Sunnucks P., Hales D.F. 1996. Numerous transformed sequences of mitochondrial cytochrome oxidase I-II in aphids of the genus *Sitobion* (Hemiptera: Aphididae). Mol. Biol. Evol. 13:510–524.

Townsend J.P. 2007. Profiling phylogenetic informativeness. Syst. Biol. 56:222–231.

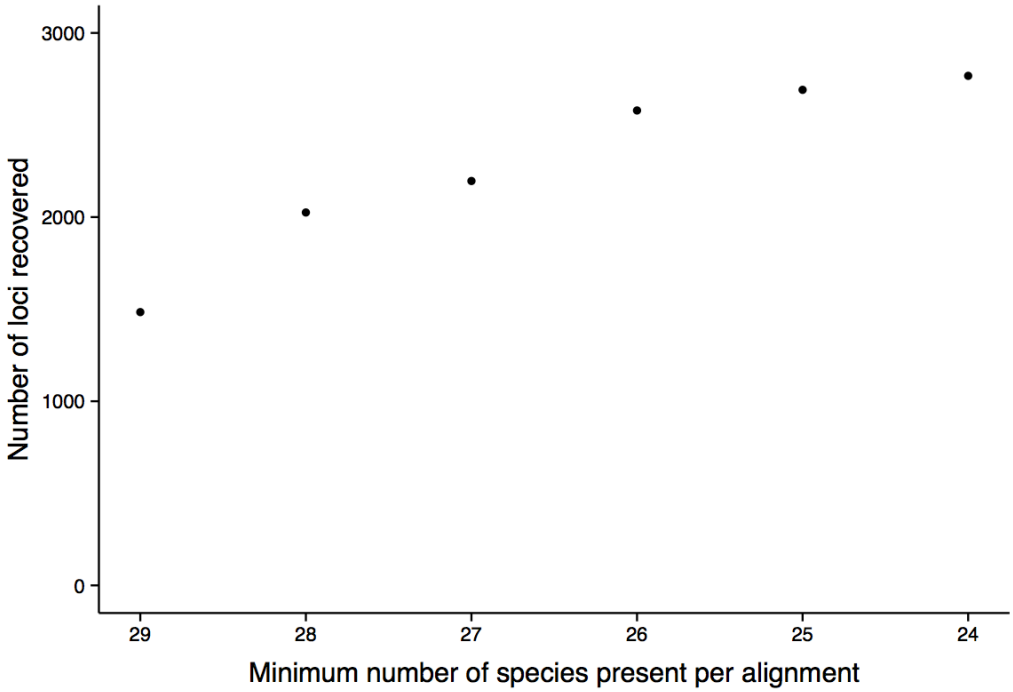
Xi Z., Lian L., Davis C.C. 2016. The impact of missing data on species tree estimation. Mol. Biol. Evol. 33:838–860.

Yang Z., Rannala B. 2012. Molecular phylogenetics: Principles and practice. Nat. Rev. Gen. 13:303–314.

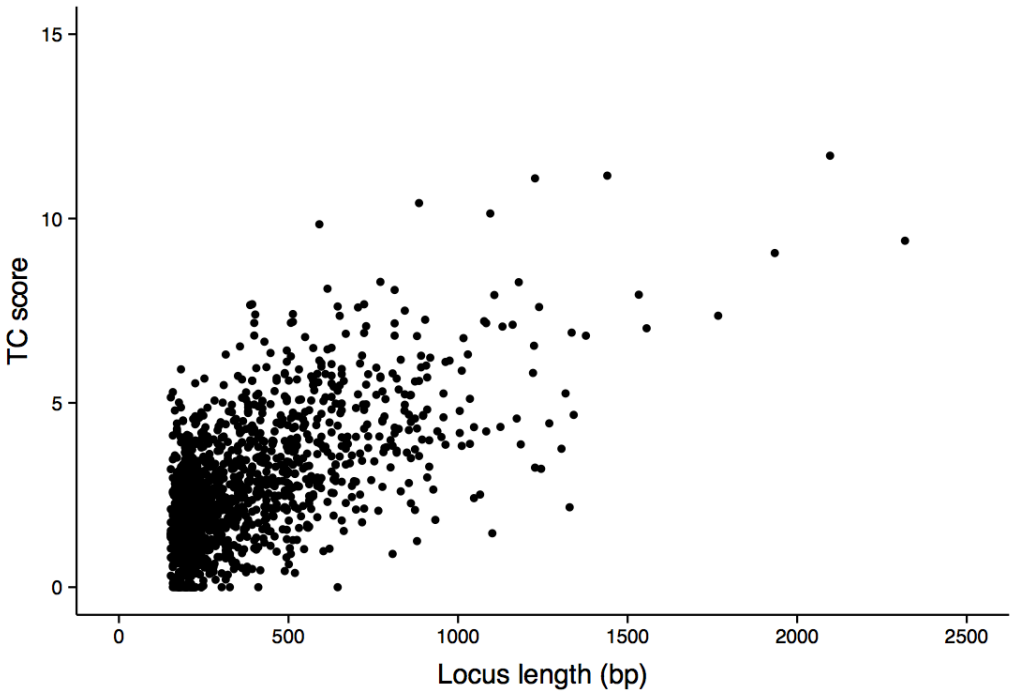
Yu Y., Dong J., Liu K.J., Nakhleh L. 2014. Maximum likelihood inference of reticulate evolutionary histories. Proc. Natl. Acad. Sci. 46:16448–16453.

Zerbino D.R., Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome. Res. 18:821–829.

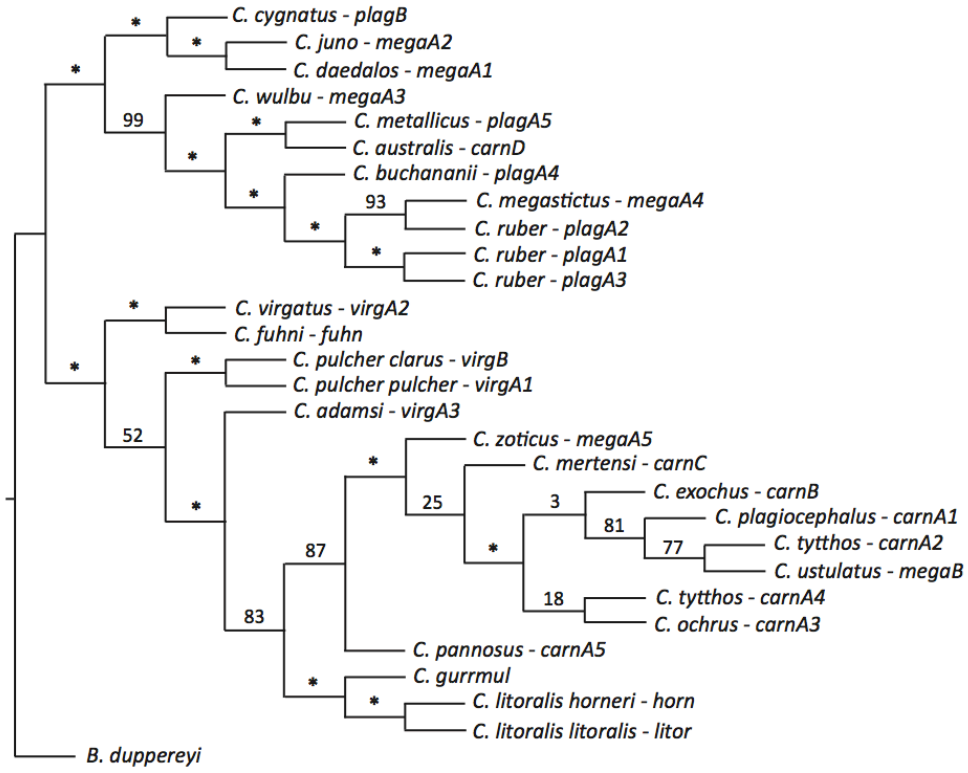
Zwickl D.J., Stein J.C., Wing R.A., Ware D., Sanderson M.J. 2014. Disentangling methodological and biological sources of gene tree discordance on *oryza* (poaceae) chromosome 3. Syst. Biol. 63:645–659.



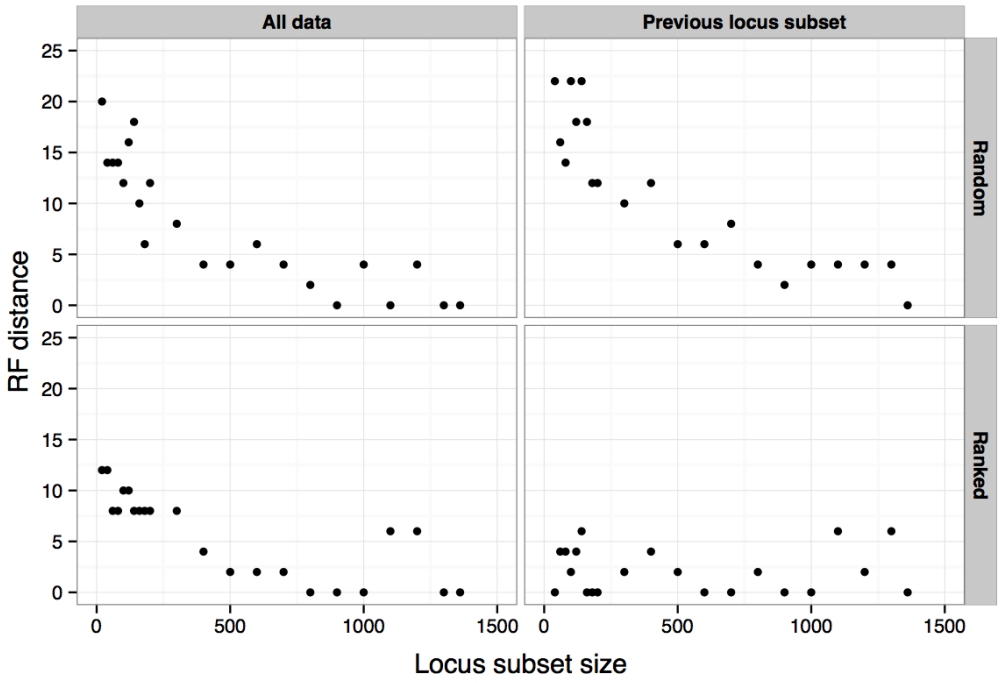
Supp. Fig. 1. The number of alignments recovered (regardless of length) when allowing the minimum number of individuals present to vary.



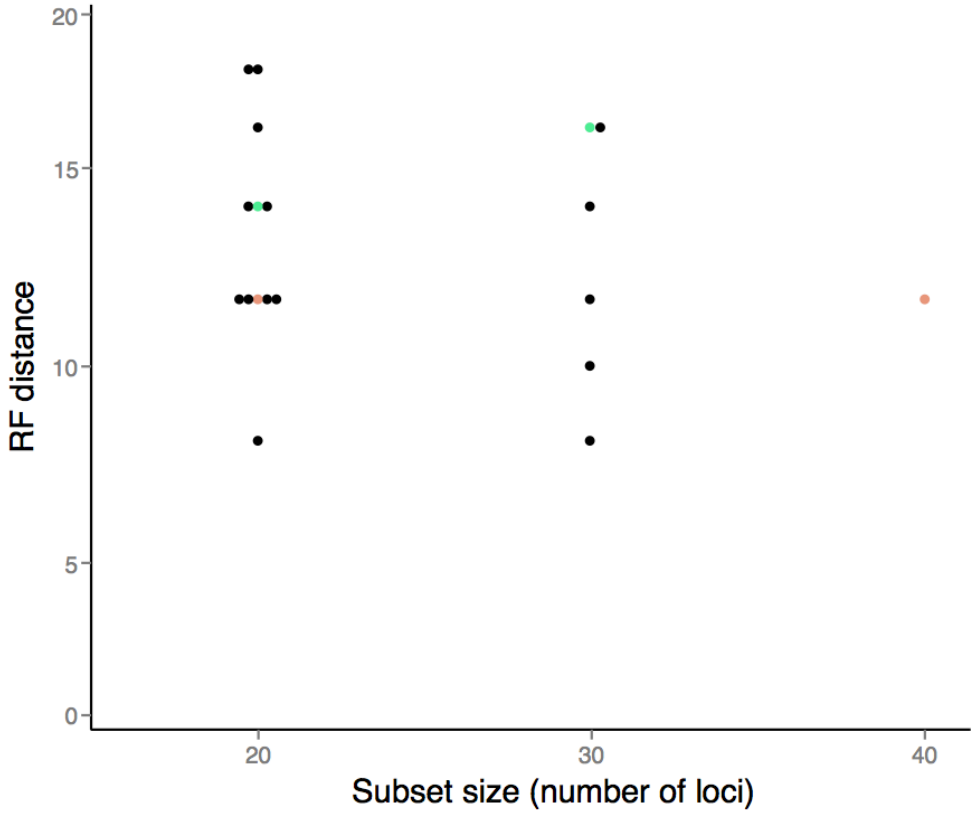
Supp. Fig. 2. The locus length and Tree Certainty (TC) score foreach alignment over 150 bp long. Locus length is correlated with TC score.



Supp. Fig. 3. Species topology of Australian Cryptoblepharus, based on summary- coalescent analyses of 1361 exons (complete dataset). Values at internode branches reflect bootstrap support (BS) and an asterisk (*) denotes BS = 100. The current species name and allozyme group (sensu Horner and Adams, 2007) are provided at each tip.



Supp. Fig. 4. .The Robinson-Foulds distance between the inferred ASTRAL species tree for each locus subset (complete dataset) and the ASTRAL species trees based on 1361 exons or the previous locus subset. Subsets of loci were either chosen randomly or ranked by TC score. The addition of loci with poor gene tree resolution does not result in an increase in topological changes and limited numbers of ranked loci already converge on the most optimal topology given the dataset. Considerable variation in RF distances persists, since the species trees contain a number of unresolved nodes (See supplementary Fig. S3).



Supp. Fig. 5. The Robinson-Foulds distance between the inferred *BEAST species tree and the ASTRAL species tree based on 1852 exons (incomplete dataset – up to one taxon missing per alignment). *BEAST trees were either inferred based on subsets of loci with the highest ranked TC score (green orbs), or on random subsets of the 200 loci with the highest ranked TC score (black orbs). For reference, the RF distance for ASTRAL trees based on similar sized ranked loci subsets are presented as well (light-pink orbs). A full-coalescent species tree approach did not infer a more accurate topology, in comparison to a summary-coalescent tree based on the same loci. Furthermore, *BEAST trees based on random subsets of loci vary considerable, suggesting that more loci are required to infer the most optimal topology.

CHAPTER 2

Convergence across a Continent: Adaptive Diversification in a recent Radiation of Australian Lizards

PROCEEDINGS B

rspb.royalsocietypublishing.org

Research

Cite this article: Blom MPK, Horner P, Moritz C. 2016 Convergence across a continent: adaptive diversification in a recent radiation of Australian lizards. *Proc. R. Soc. B* **283**: 20160181.
http://dx.doi.org/10.1098/rspb.2016.0181

Received: 27 January 2016
Accepted: 19 May 2016

Subject Areas:
evolution, ecology, genomics

Keywords:
ecomorphology, adaptive radiation, continental radiation, convergence, *Cryptoblepharus*, speciation

Author for correspondence:
Mozes P. K. Blom
e-mail: mozes.blom@gmail.com

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2016.0181 or via http://rspb.royalsocietypublishing.org.

THE ROYAL SOCIETY
PUBLISHING

Convergence across a continent: adaptive diversification in a recent radiation of Australian lizards

Mozes P. K. Blom[1], Paul Horner[2] and Craig Moritz[1]

[1]Research School of Biology, The Australian National University, Canberra ACT 0200, Australia
[2]Museum and Art Gallery of the Northern Territory, GPO Box 4646, Darwin NT 0801, Australia

MPKB, 0000-0002-6304-9827

Recent radiations are important to evolutionary biologists, because they provide an opportunity to study the mechanisms that link micro- and macroevolution. The role of ecological speciation during adaptive radiation has been intensively studied, but radiations can arise from a diversity of evolutionary processes; in particular, on large continental landmasses where allopatric speciation might frequently precede ecological differentiation. It is therefore important to establish a phylogenetic and ecological framework for recent continental-scale radiations that are species-rich and ecologically diverse. Here, we use a genomic (approx. 1 200 loci, exon capture) approach to fit branch lengths on a summary-coalescent species tree and generate a time-calibrated phylogeny for a recent and ecologically diverse radiation of Australian scincid lizards; the genus *Cryptoblepharus*. We then combine the phylogeny with a comprehensive phenotypic dataset for over 800 individuals across the 26 species, and use comparative methods to test whether habitat specialization can explain current patterns of phenotypic variation in ecologically relevant traits. We find significant differences in morphology between species that occur in distinct environments and convergence in ecomorphology with repeated habitat shifts across the continent. These results suggest that isolated analogous habitats have provided parallel ecological opportunity and have repeatedly promoted adaptive diversification. By contrast, speciation processes within the same habitat have resulted in distinct lineages with relatively limited morphological variation. Overall, our study illustrates how alternative diversification processes might have jointly stimulated species proliferation across the continent and generated a remarkably diverse group of Australian lizards.

1. Introduction

Understanding the processes that promote biological diversity is a major challenge in evolutionary biology. In this context, much has been gleaned from the study of adaptive radiations; the rise of diverse ecological roles and phenotypic disparity due to role-specific adaptations within a lineage [1,2]. Adaptive radiations have drawn the attention of evolutionary biologists, because they exemplify the mechanisms that link micro- and macroevolution. During an adaptive radiation, ecological opportunity facilitates speciation and ecological diversification [3,4]. Such ecological opportunity can arise when an isolated area with a depauperate biota is colonized (i.e. islands or lakes) or when the evolution of a key trait opens a new adaptive zone (i.e. key innovation) [5,6]. Young adaptive radiations in isolated geographical entities such as islands or lakes are particularly well studied [7] and have highlighted the role of ecological speciation [8] and sexual selection [9].

However, evolutionary radiations can be triggered by a wide range of biotic and abiotic factors [10], and not all evolutionary radiations can be characterized as adaptive radiations [1,2,11]. Allopatric speciation for instance, though unlikely to occur within a small island or lake, can be an important driver of

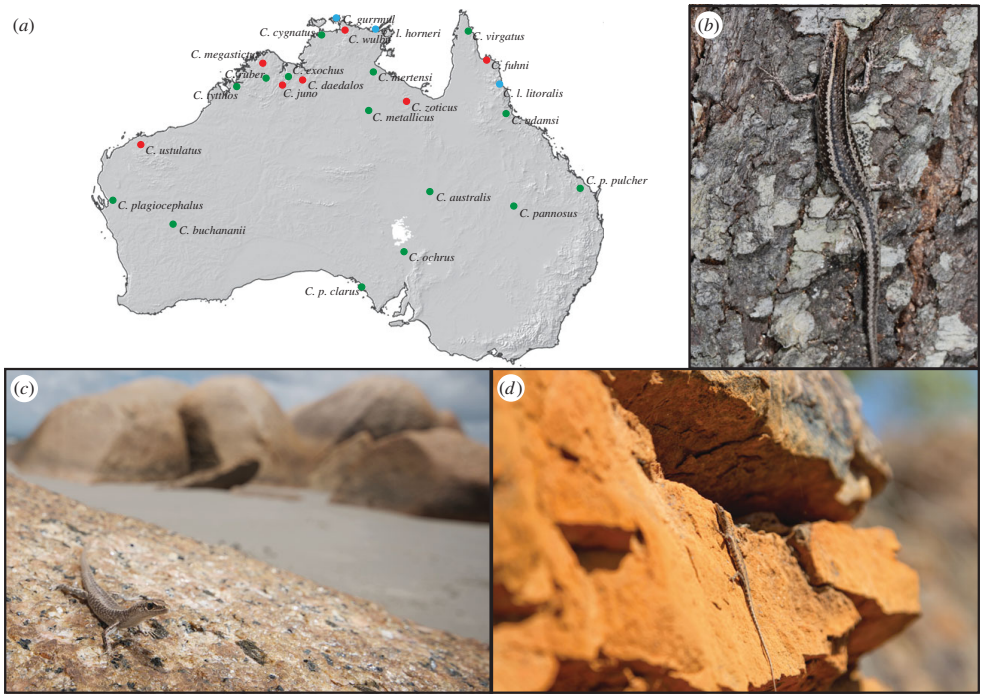


Figure 1. Distribution of Australian *Cryptoblepharus* and the three habitat specialists. (*a*) Topographic map of Australia with the mean point of each species' distribution plotted and coloured according to habitat type (for complete distribution maps, see [25]). *In situ* photographs of (*b*) arboreal, (*c*) littoral, and (*d*) rock specialists (green, blue, and red dots on the topographic map, respectively).

evolutionary radiation on large continental landmasses. Lineages within a new region might diversify via allopatric speciation when an ancestral species crosses a geographical barrier or when habitats become fragmented following a climatic shift [10]. Speciation in such geographical isolates can be driven via means other than ecologically mediated divergent selection and precede significant ecological differentiation. Ecological differentiation could subsequently still arise in isolation, or via character displacement between reproductively isolated lineages that come into secondary contact. Thus, processes other than ecological speciation can promote evolutionary radiation at a continental scale and even generate patterns that resemble adaptive radiation [11]. Indeed, the study of continental-scale radiations has provided wonderful examples of species proliferation and adaptive phenotypic change across major taxonomic groups [12–16]. Yet it can be challenging to identify the key evolutionary processes that have promoted speciation in such older radiations. It is therefore important to establish a phylogenetic and ecological framework for recently emerged continental clades that are widespread, species-rich, and ecologically diverse. The study of such recent radiations can ultimately provide further insight into the factors that have shaped macroevolutionary patterns across a continent [10,11,17–20].

The study of species diversification and phenotypic radiation has greatly benefited from the development of phylogeny-aware comparative methods and molecular approaches to generate large sequence datasets [4,21]. The phylogenetic structure underlying rapid radiations has been notoriously difficult to resolve due to a lack of phylogenetic signal or incongruence in phylogenetic history between loci [20–22]. This observed discordance emphasizes the need to use multi-locus datasets for species tree inference and to incorporate coalescent-based methods that account for heterogeneity in coalescent histories [23]. In addition, analysing large numbers of genetic markers will also optimize branch length estimation and this increase in accuracy is particularly relevant for comparative analyses [24]. With a complete phylogeny in place and information on contemporary phenotypic variation, patterns of adaptive diversification can be identified by comparing the fit of distinct models of phenotypic change [13]. In this study, we use a phylogenomic approach to generate a time-calibrated ultrametric tree for a recent radiation of Australian lizards, and subsequently employ comparative methods to test how habitat specialization may have influenced morphological diversification in the course of this radiation.

Skinks of the genus *Cryptoblepharus* have radiated across the entire Australian continent (figure 1*a*) while simultaneously colonizing different scansorial habitats (rocks and trees; figure 1*b*,*d*) which are largely unoccupied by other species of the rich diurnal lizard fauna. Furthermore, although terrestrial habitats are dominated by other genera of ground-dwelling skinks, there are three littoral species of *Cryptoblepharus* that are found in close association with rocks on beaches (figure 1*c*)—another unique habitat. A recent taxonomic revision using 45 (allozyme) loci and 33 morphological markers increased the number of recognized Australian species from seven to 25 [26,27]. Although this analysis was

unable to resolve the phylogeny of the genus, a recent phylogenomic analysis revealed that rock and arboreal specialists are dispersed across the phylogeny [28]. Interestingly, the geographical distributions of rock and arboreal specialists are often overlapping and many instances of sympatry have been reported [27]. This genus therefore provides an excellent opportunity to study the role of habitat specialization in promoting adaptive diversification at a continental scale.

The overarching goal of this study is to examine the ecological context of diversification of Australian *Cryptoblepharus* with a combined assessment of morphological, ecological, and phylogenetic patterns. Specifically, we focus on whether habitat specialization can explain current patterns of phenotypic variation in ecologically relevant traits. A statistical correlation between phenotype and environment is a first indication of adaptive diversification, but here we build further upon a rich literature of ecomorphological research in lizards [25,29–31]. By explicitly focusing on traits that are known to improve performance within specific habitats [3], we examine the adaptive consequences of habitat specialization and quantify convergent change across the Australian continent.

2. Material and methods

(a) Taxon sampling for phylogenetic inference

A previous allozyme and morphological analysis identified 28 lineages of Australian *Cryptoblepharus* [26], of which 25 were recognized as separate species. Three genetically divergent lineages were morphologically, ecologically, and geographically indistinguishable—two in *C. ruber* and one in *C. tytthos*—and were therefore not elevated to species status. We selected a single representative for each of the 28 lineages using mostly the same individuals as examined allozymically by Horner & Adams [26], except where tissues were depleted. For those species, we used recently collected field samples where species identification was verified based on morphological characteristics and a mitochondrial marker (*ND2*). Further details can be found in Blom *et al*. [28].

(b) Exon capture

We used a custom-designed exon capture approach [32], to generate a large multi-locus dataset of orthologous genetic markers suitable for phylogenetic inference. The designs of the exon capture kit, sequencing strategy, and sequencing success, are outlined in detail in Blom *et al*. [28] and references therein. Briefly, our capture design included exon-targets based on orthologues from seven transcriptomes of three genera closely related to *Cryptoblepharus* (*Carlia rubrigularis*, *Lampropholis coggeri*, and *Saproscincus basiliscus*; [33]). We used an in-solution hybridization capture (Roche NimbleGen) and sequenced (100 bp paired-end) the enriched libraries on a single Illumina HiSeq 2000 lane. We filtered and assembled the sequence data using a workflow that was described previously by Singhal *et al*. [33] and is available at https://github.com/MVZSEQ.

We have developed and applied a flexible bioinformatic workflow for alignment and alignment filtering of exonic sequences, EAPhy (v. 1.1, [34]). EAPhy automatically aligns sequences using MUSCLE (v. 3.8.31, [35]), performs checks to ensure coding of amino acids and removes missing data from the end of the alignments. EAPhy assesses each alignment individually and automatically generates either locus-specific or concatenated alignments. We only concatenated loci where each lineage with morphological data was present and the alignment of each individual locus was longer than 150 bp.

(c) Phylogenetic inference

Estimating a species tree is particularly challenging for rapid radiations [22,36]. We have previously employed summary-coalescent methods and a thorough gene tree estimation sensitivity analysis to infer the *Cryptoblepharus* species tree topology [28]. In brief, we first screened loci based on gene tree resolution and subsequently quantified the impact of stochastic gene tree estimation error on summary-coalescent species tree inference. Here, we use the inferred species tree topology that was well-supported across analyses. However, we excluded two lineages from the dataset for which no morphological measurements were available (sub-species: *C. pulcher clarus* and a divergent lineage of *C. tytthos*—'carnA4' in Horner & Adams [26]).

We generated an ultrametric tree from the concatenated alignment with BEAST v. 2.1.3 [37], while constraining the topology to that of the species tree (*sensu* [28]) and therefore only fitted branch lengths. We used a GTR + Γ substitution model with four Γ rate categories, a strict clock, and estimated each substitution rate from the data. In the absence of suitable fossil calibrations or a previous estimate of crown age for the genus, we scaled branch lengths from a number of expected substitutions per site to years, using an empirically obtained molecular clock estimation (0.001 substitutions/site/Myr) for another genus of lizards within the same family (*Scincidae*, [38]). We ran the BEAST analyses in duplicate with separate starting seeds. Each analysis was run for 10 million generations and we sampled chains every 10 000 generations. We discarded the first 10% of trees as burnin, used Tracer v. 1.5 to check for convergence, and LogCombiner v. 2.1.3 to combine the posterior sample of trees across runs. The ultrametric species tree was summarized using TreeAnnotator v. 2.1.2. Lastly, we tested whether the rate of lineage accumulation changed over time using the *ltt* function in the R-package 'Phytools' [39].

(d) Morphospace construction

To examine phenotypic changes in Australian *Cryptoblepharus*, we combined our species tree estimate based on genomic data with the morphological characters recorded during the last major taxonomic revision [27]. Horner [27] recorded complete metric and meristic measurements for 863 Australian *Cryptoblepharus* specimens. Morphometric measurements were taken under an illuminated magnifying lens, with electronic digital calipers to the nearest 0.01 mm. Across the 26 taxa for which morphological data were recorded, the 863 individuals represented an average of 33 individuals per species and all species were represented by four measurements or more (electronic supplementary material, table S1). From the suite of characters recorded for the taxonomic revision, we selected metric characters known or suspected to be relevant to the habitats used by *Cryptoblepharus* [30]. These include snout–vent length (SVL), forelimb length (FL), rear-limb length (RL), snout length (SE), eye to ear length (CHEEK), ear to limb length (NECK), head height (HH), and head width (HW). We used SVL as a measure for overall body size and divided the head length estimate into three separate metrics (SE, CHEEK, NECK) to account for differences that might not be appropriately captured by head length alone. To identify size-independent axes of trait variation, we calculated residual values from phylogenetic regressions of each log-transformed trait against log-SVL. For each trait, we first calculated the mean species values and then used the function *phyl.resid* (Phytools; [40]), to infer the size-independent trait values. We used the λ correction to avoid bias due to non-Brownian evolution, during the estimation of the phylogenetically corrected regression.

To identify the major axes of variation and reduce the multidimensionality of the data, we used a phylogenetic principal component analysis (pPCA) on the size-corrected species trait data (all traits excluding SVL), while simultaneously optimizing λ (*phyl.pca*, Phytools). We used a scree plot to visualize and

examine the amount of variation explained by each individual component. The first two components jointly explained over 85% of the variation in the data and were retained for further in-depth analysis.

We examined the degree of morphospace occupied by each habitat type, using a three-dimensional phylomorphospace plot (*phylmorphospace3d*, Phytools) where each axis represents a trait that loaded strongly on PC1. By plotting the size-corrected residual scores for each species, superimposing phylogenetic relationships, and highlighting species by habitat type, the phylomorphospace plot illustrates morphological variation between and within habitat types for the traits that jointly explained most phenotypic variation. In addition to visualizing habitat specific differences in phylomorphospace, we also estimated the degree of phenotypic disparity between all species combined and within each habitat type, by calculating the average squared Euclidean distance among all pairs of PC1 scores using the *disparity* function from the R-package 'Geiger' [41].

(e) Associations between morphology and habitat

We examined whether trait values along the two major axes of trait variation differed between species occurring in different habitat types (rock, arboreal, and littoral). We used a multivariate analysis of variance (MANOVA) with habitat as a predictor variable and the species' PC scores for the first two components as the dependent variables. To conduct a MANOVA in a phylogenetic context, we used the *aov.phylo* function (1 000 simulations, Wilks' λ) in Geiger. We subsequently examined differences for each individual PC separately. Based on 1 000 simulations, we calculated the probability of the observed differences in PC scores between each group (*phylANOVA*, Phytools).

The ANOVA for each PC provides an overall view of differences along the two major axes of trait variation between species occurring in different habitats. However, to examine changes in individual morphological traits, we repeated the phylo-ANOVA's using the size-independent residual scores from phylogenetic regression for each individual trait. Lastly, we also tested differences in overall size (SVL) by comparing the log-transformed species means.

(f) Morphological evolution

To assess whether habitat shifts can explain current patterns of phenotypic diversification, we used two different approaches to estimate the evolutionary trajectory of phenotypic change. Firstly, we quantified whether morphology varies following a Brownian Motion (BM) process, where phenotypic differences accumulate at random with time, or whether morphological diversification is constrained around one or more optima (OU, Ornstein–Ühlenbeck process). Recent expansions in the class of OU models allow variable rates and strengths of selection around the trait optima [42]. However, parameter estimation in such complex models requires large numbers of taxa [42]. Because the number of *Cryptoblepharus* species and the frequency of habitat shifts are relatively limited, we only evaluated the presence or the absence of multiple phenotypic optima rather than estimating selection strength as well. We first estimated ancestral states for each internal node (*rerootingMethod*, Phytools) and then used the R-package OUwie [42] to fit three distinct models of character evolution on PC1 scores. We fitted a single rate BM model (BM1) and OU models with either a single optimum for all species (OU1) or with multiple optima, but single rates of selection (α) and stochastic motion around all optima (σ^2).

Secondly, we evaluated whether independent lineages converged on similar phenotypic optima by using a comparative approach implemented in R. SURFACE [43] uses a stepwise corrected AICc (Akaike's information criterion corrected for sample size) approach to fit Hansen models and evaluates the most optimal set of evolutionary regimes and regime shifts. SURFACE analyses consist of two distinct phases; a 'forward' phase during which regime shifts are added to the tree and a 'backward' phase during which shifts towards the same peaks are identified and collapsed. The addition and collapsing of shifts is reiterated until AICc scores cease to improve. SURFACE can identify cases of convergence across a clade without the subjective *a priori* designation of candidate convergent taxa, and only takes the phylogeny and multidimensional trait data as input. We ran SURFACE on the size-corrected residuals for each log-transformed trait and log-transformed species means for SVL. Finally, to visualize how species that belong to distinct regimes differ in phenotype, we plotted the size-corrected residual scores for each species in two-dimensional trait space across all traits that were inferred as significantly different between regimes.

3. Results

(a) Phylogenetic analyses

We used 1 195 loci (618 860 bp) and an empirically estimated molecular clock to generate a time-calibrated ultrametric tree (figure 3*a*). We fitted branch lengths on a summary-coalescent species tree that is well supported, except for the node that involves *C. zoticus* and *C. mertensi*. This analysis infers that Australian *Cryptoblepharus* have diversified recently, within the last 10 million years overall, across the Australian continent. Two distinct clades, with 11 and 15 taxa, respectively, have proliferated since the Pliocene and rock and arboreal specialists have emerged within each radiation. The species accumulation rate was elevated during the beginning of the continental radiation and then decreased over time ($\gamma = -2.070$, $p = 0.04$; electronic supplementary material, figure S1).

(b) Morphospace and association with ecology

The main variation in size-corrected morphology was between species from different habitats and in particular between rock and other specialists. This is shown by the separation of the rock species from all others along the first axis of the PCA (figure 2*a*). PC1 (66.45%) and PC2 (18.93%) jointly explain over 85% of all morphological variation, but only PC1 is significantly correlated with habitat type (phylogenetic MANOVA on both PC1 and PC2, d.f. = 2, Wilks' $\lambda = 0.13$, $p < 0.01$; phylogenetic ANOVAs on PC1, d.f. = 2, $F = 59.19$, $p < 0.01$, and on PC2, d.f. = 2, $F = 1.43$, $p = 0.40$). PC1 loads strongly on fore- and hindlimb length and HH (electronic supplementary material, table S2), suggesting that species that occur on rock substrates tend to have longer limbs and a more compressed head shape than species that occur on trees or in a littoral habitat. PC2 loads most strongly on features pertaining to head length but does not correlate with habitat type (electronic supplementary material, table S2).

In addition to assessing the correlation between morphology and habitat type along the two major axes of trait variation (i.e. PC1 and PC2), we also employed phylogenetic ANOVAs and post hoc *t*-tests to examine differences in individual traits. The overall pattern observed is similar for HH, with species that occur on rock substrates having dorsally compressed skulls (d.f. = 2, $F = 28.12$, $p < 0.01$) and no difference between arboreal and littoral specialists (d.f. = 2, $T = -0.06$, $p = 0.97$). Rock and arboreal species consistently differ in limb length (forelimb, d.f. = 2, $T = -8.78$,

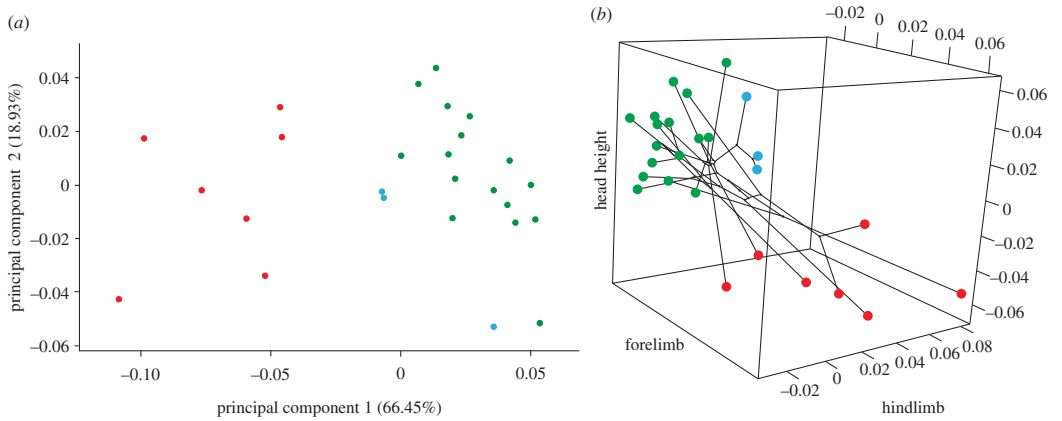


Figure 2. pPCA and three-dimensional phylomorphospace plot. (*a*) Morphological variation among species along pPC1 and pPC2, where each dot represents a species and is coloured by that species' habitat type (figure 1). PC1 separates the rock specialists from the other habitat specialists. (*b*) Size-corrected residual scores from phylogenetic regression are plotted for each of the traits that were identified as different between habitats. Colours correspond to the habitat type for each respective species (figure 1) and the phylogeny is mapped onto morphospace. Regardless of phylogenetic association, species are more closely clustered in morphospace by habitat.

$p < 0.01$ and hindlimb, d.f = 2, T = −7.35, $p < 0.01$). Whereas (ground-dwelling) littoral specialists have similar scores for PC1 as arboreal species, phylogenetic ANOVA of individual traits show that the former have significantly longer hindlimbs than their arboreal counterparts (d.f. = 2, $T = -3.24$, $p = 0.04$), but do not differ from rock specialists (d.f. = 2, $T = -1.88$, $p = 0.23$). Arboreal and littoral species did not differ in FL (d.f. = 2, $T = -2.20$, $p = 0.10$). There was no significant correlation between habitat and any other trait (SVL, HW, SE, CHEEK, or NECK).

Significant differences between habitat types in the three divergent traits (fore-, hindlimb, and HH) are apparent when mean species scores are visualized in morphospace (figure 2*b*). Most notably, arboreal species tend to cluster closely and rock species are clearly distinct in terms of HH, regardless of phylogenetic association between species. These results were confirmed when comparing morphological disparity metrics, which were more than twofold lower for the arboreal species (electronic supplementary material, table S3) than for the other habitat categories.

(c) Morphological evolution

An OU model with multiple phenotypic optima (OUM: AICc score −121.54) was substantially better supported than a BM (AICc score −73.38) or OU1 model (AICc score −78.43), suggesting that there is more than one phenotypic optimum for traits that strongly load on PC1. The estimated optima were found within the values realized for the extant species (electronic supplementary material, table S4), indicating that the model is a realistic description of current morphological patterns and is not negatively biased by factors of uncertainty such as potentially spurious ancestral state reconstruction. In addition, the standard errors around the optima reflect the observed variation in morphospace and disparity metrics for each habitat category, further confirming that the inferred optima of the model represent realistic differences between habitat specialists (electronic supplementary material, table S4).

Having inferred multiple phenotypic optima, we tested whether independent lineages converged on the same optima and quantified the frequency of such shifts. The Hansen model returned by the SURFACE analysis highlights the presence of two regimes and four convergent regime shifts. The AICc score improved by 29.34 units during the forward phase and by another 25.6 units during the backward phase. The AICc score for the final model suggests that the phenotypic data match an OU model with two phenotypic optima much better than a similar model under BM or an OU model with a single optimum (electronic supplementary material, figure S2), and is therefore consistent with the result from the OUwie analysis.

The two optima identified coincide with phenotypes that match species found on either an arboreal or a rock substrate, with the four regime shifts corresponding to transitions from trees to rocks (figure 3*a*). However, and interestingly, two of the seven rock species did not converge in overall phenotype with the other saxicolous species, nor did littoral specialists occupy a distinct phenotypic optimum (figure 3*a*). Instead, the two rock (*C. zoticus* and *C. ustulatus*) species in question and all three littoral specialists (*C. l. litoralis*, *C. l. horneri*, and *C. gurrmul*) belong to the same regime as the arboreal specialists. When focusing on the three functional traits that are distinct between rock and other specialists, it is apparent why these species are matched to the arboreal regime by the SURFACE analysis. Whereas the two rock species are convergent in HH with other rock species, the length of their limbs is not (figure 3*b*). Furthermore, as previously identified with a phylogenetic ANOVA, littoral species differ in the length of their hindlimbs but not in any other trait (figure 3*b*).

4. Discussion

(a) Habitat specialization

Cryptoblepharus skinks are the most prominent scansorial specialists within the most species-rich family of lizards in Australia and globally (*Scincidae*). Species that occur on distinct substrates differ significantly in functionally relevant phenotypic traits and our analyses suggest that the evolutionary

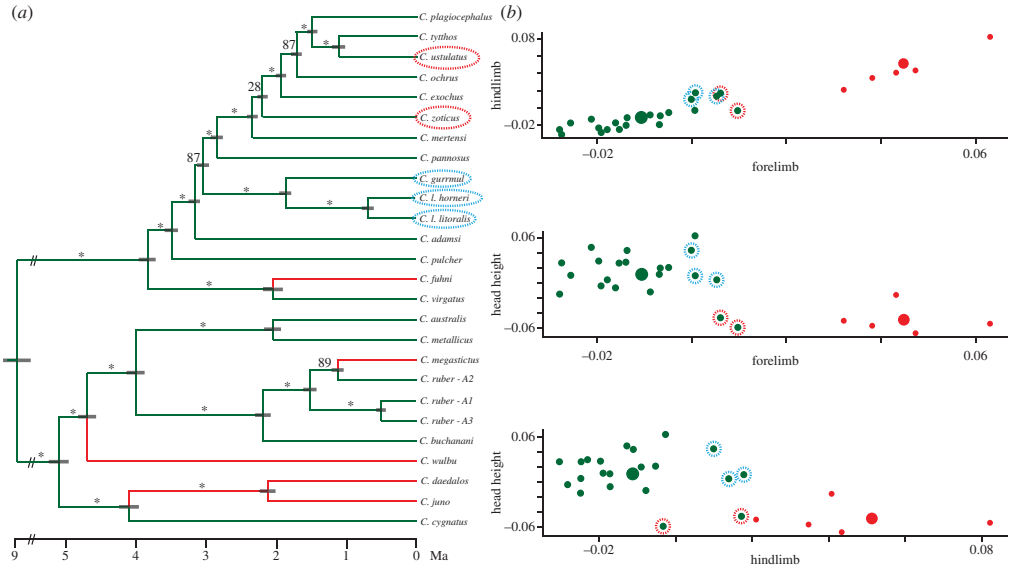


Figure 3. Time-calibrated phylogeny of Australian *Cryptoblepharus* and phenotypic convergent regimes. (*a*) The branches of the phylogeny are coloured according to the convergent regime inferred with SURFACE. There are two adaptive peaks and at least four independent regime shifts. Bootstrap support is given for each bipartition [32], with asterisks indicating bootstrap support of more than 90, and grey boxes around each node representing the 95% confidence interval for the node age estimate. Non-arboreal species that belong to the arboreal regime are highlighted with a coloured dotted circle. (*b*) Size-corrected residual scores from phylogenetic regression are plotted for each of the traits that were identified as distinct and coloured by their inferred convergent regime. Non-arboreal species that belong to the arboreal regime are highlighted with a coloured dotted circle. Two rock specialists have not converged on the same adaptive peak as the other rock dwelling species. These species are convergent in terms of HH, but have relatively short limbs. The three littoral species only differ from the arboreal specialists in hindlimb length and are likely therefore not identified as a separate adaptive peak.

trajectories of such traits have changed in a predictable direction based on biomechanical performance tests in other lizards [30,31]. Rock species of *Cryptoblepharus* occupy steep sandstone escarpments, where they move rapidly on perpendicular cliffs and hide in shallow crevices between the rocks. The reduction in HH enables species to shelter in narrow cracks, sometimes just a few millimetres wide. Arboreal species however, tend to move up into trees and hide between the foliage, and likely do not experience a similar degree of selection for reduced HH. The increase in limb length of rock species aids locomotion on flat vertical surfaces by maintaining the centre of mass (and balance) close to the substrate. By contrast, long legs might present challenges for arboreal lizards by increasing the distance between the centre of mass and perch [44]. The littoral species, that tend to climb less than arboreal and rock specialists, mostly resemble arboreal species except for the length of their hindlimbs, in which they are more similar to rock species (figures 2*b* and 3*b*). Whereas scansorial species tend to rely equally on fore- and hindlimbs for locomotion, ground-dwelling lizards are expected to have relatively long hindlimbs since they mostly use their back legs to thrust forward for movement in a horizontal direction [45]. Overall, the morphology of *Cryptoblepharus* skinks tends to match habitat closely, such that these variants can be considered as ecomorphs.

(b) Convergence

Morphological differences are not only correlated with environment, but they also have evolved independently and resulted in at least four convergent shifts towards the same phenotypic optimum (figure 3*a*). Given the strong correlation between habitat and functionally relevant traits, the trait value optima of the SURFACE regimes can be interpreted as adaptive peaks for an arboreal and a rock ecomorph. Interestingly, on visual inspection of the three-dimensional phylomorphospace plot (figure 2*b*) and the phenotypic disparity metrics, the phenotypic variation surrounding the adaptive peaks is more limited for the arboreal type than the rock type. Even though there are less than half as many rock as arboreal lineages (7 versus 16), and both ecomorphs span the phylogeny, phenotypic disparity between the rock species is twofold greater than between the arboreal lineages (electronic supplementary material, table S3). Although we are unable to accurately estimate the strength of selection surrounding phenotypic optima with a limited number of species [42], this prominent difference in phenotypic disparity within ecomorph groups might indicate that the strength of selection is more variable towards the rock optimum and more stringent for arboreal species. This is exemplified by *C. zoticus* and *C. ustulatus*; the two rock species that were identified as belonging to the arboreal regime by the SURFACE analysis. These species are clearly convergent with other rock species for HH (figures 2*b* and 3*b*), but their fore- and hindlimbs are relatively short compared to other rock species (figures 2*b* and 3*b*) and in particular *C. fuhni*, the rock specialist with the most pronounced degree of limb elongation [27]. Further ecological studies and performance tests should investigate whether this difference has any consequence in terms of

locomotion performance or whether this is mitigated by alternative use of a similar habitat [46]. Similarly, the three littoral species are joined with the arboreal regime because the difference in hindlimb length alone is not significant enough to identify a separate adaptive peak. Previous simulation analyses have confirmed that SURFACE performs well with datasets that include multiple convergent traits, but instances of convergence were not always identified with single (convergent) traits [43]. Thus, even though these two rock and three littoral species significantly differ along specific trait axes from arboreal types, they have not been diagnosed as distinct.

Examples of lizard species that have repeatedly adapted to rock environments and converged in limb length and head depth, have been reported previously (i.e. [47]). However, the frequency of convergence within this recent radiation of *Cryptoblepharus* is striking and likely correlated with the rapid spread across the Australian continent. Independent adaptive peak shifts between arboreal and rock habitat have occurred in isolated sandstone ranges that are surrounded by vast stretches of savannah woodland or desert. It is unlikely that low-dispersal rock specialists can easily migrate across such distinct habitats, such that these sandstone ranges resemble islands in a sea of unsuitable habitat where parallel evolutionary change has resulted in convergent ecomorphological lineages. These repeated convergent outcomes in functional traits strongly suggest that adaptation to different habitats has promoted an increase in ecological specialization and associated phenotypic disparity in ecologically relevant traits between sister taxa [48]. Furthermore, adaptation to distinct substrates has often facilitated the sympatric coexistence of closely related (sister) species [27]. In these respects, the recent radiation of Australian *Cryptoblepharus* bears strong similarity to the patterns that characterize adaptive radiations [2,3,49].

(c) Continental radiation

Our findings suggest that each of the two clades of *Cryptoblepharus* skinks have proliferated rapidly during the last approximately 5 Myr (figure 3*a*) and have repeatedly developed phenotypic traits that are known to be of functional importance within specific environments [30]. This strong correlation between habitat and morphology underlines the importance of ecologically mediated selection within this system. Repeated habitat shifts have resulted in divergent selection that has increased morphological disparity, while uniform selection across geographical isolates on similar substrates has likely limited morphological differentiation, especially within the arboreal taxa. Indeed, the taxonomy of *Cryptoblepharus* has traditionally been viewed as exceptionally challenging due to the limited morphological differences between species that occur in similar habitat. Hence, many of these cryptic lineages were only identified with the aid of genetic screening [26]. The presence of strong selection, either uniform or divergent, can accelerate the speciation process and is more likely to have promoted species proliferation in this genus than neutral processes [50] alone, especially given the recency of the radiation [51].

Interestingly, examination of the temporal pattern of diversification suggests that the rate of species accumulation was elevated during the beginning of the continental radiation (electronic supplementary material, figure S1). Although this is often interpreted as evidence of adaptive radiation (i.e. 'early-burst' signal), processes other than initial niche filling could also result in a slowdown of diversification over time [52]. Whereas a novel lineage on an isolated island or lake might rapidly have access to all available niche space, it might take a significant amount of time before a continental clade has spread across all available habitats. Indeed, our results indicate that habitat shifts have not predominantly occurred in the beginning of the radiation, but both in the deep and more recent past (figure 3*a*). The rapid accumulation of lineages during the beginning of the radiation could therefore simply reflect geographical isolation following an early range expansion, rather than initial diversification of niche use.

Our analyses highlight two seemingly contrasting patterns of diversification: speciation within versus between distinct habitat substrates. This suggests that the evolutionary radiation of Australian *Cryptoblepharus* is not solely driven by ecologically mediated divergent selection, as observed in some sympatric systems that reside on isolated islands or lakes [8,9], but rather parallels another enigmatic radiation, the Anoles of the West Indies [31]. *Anolis* lizards have radiated spectacularly and many distinct ecotypes occur in sympatry on islands across the Caribbean basin. However, there is no direct evidence that ecological specialists have emerged in sympatry and within-island cladogenesis appears to be limited to larger islands in the Greater Antilles, even though some of the smaller islands exhibit the same degree of environmental heterogeneity [31]. Furthermore, deep intraspecific divergence within widespread species such as *A. cybotes* [53] indicates that significant genetic differentiation has accrued without extensive ecomorphological divergence and highlights the potentially important role of macrohabitat in the speciation history of *Anolis* lizards [54]. As such, with contrasting patterns of differentiation between and within habitat types, the continental radiation of Australian *Cryptoblepharus* resembles the radiation of Caribbean Anoles and perhaps many other continental (or large island, e.g. [55]) systems of different ages (i.e. [17,49]).

Whereas the radiation of Caribbean Anoles might span 40–60 Myr [31], the relatively young age of the Australian *Cryptoblepharus* radiation invites further investigation into the mechanisms that have promoted species diversification. Of particular interest, is to understand whether the contrasting patterns of phenotypic diversification within and between habitats, also represent alternative speciation dynamics or whether reproductive isolation has developed in a similar manner and ecological differentiation has only occurred via character displacement in secondary contact [11]. By modelling demographic and divergence history, future studies can quantify the evolution of reproductive isolation between lineages in ecologically similar refugia, such as has been inferred for rainforest skinks from related genera [56]. Or alternatively, such studies can explicitly examine the geographical context of diversification and gene flow [57], between ecomorphologically distinct young sister species with a parapatric distribution (i.e. *C. ruber* and *C. megastictus*; figure 3*a*). Hence, the outcomes of our study will function as a phylogenetic and ecological framework, and invite further investigation into the proximate mechanisms that have driven speciation within and between habitats.

The study of adaptive radiations within isolated insular systems has provided important insights on the role of ecology in driving species proliferation. However, it is

important to ask to what extent similar mechanisms promote the radiation of continental biota. Because it is challenging to address this question by focusing on older radiations, the evaluation of recent continental radiations can shed further light on how commonly speciation precedes significant ecological differentiation. Our analysis of the *Cryptoblepharus* radiation for instance, highlights the importance of ecological selection both within and between habitat types, but simultaneously suggests that species proliferation is not driven by divergent selection alone. The importance of examining other recent widespread radiations (e.g. the *Sigmodontinae* of South America [58,59]) is therefore evident and will inform our general understanding on the process of continental diversification. As such, the study of recent radiations can provide a window into the origin of biodiversity and how microevolutionary processes ultimately induce macroevolutionary change at a continental scale.

Ethics. The study was performed under ethics approval from the Australian National University—A2012/14.

Data accessibility. Raw data are deposited in Dryad under: http://dx.doi.org/10.5061/dryad.j4r56.

Authors' contributions. M.P.K.B. and C.M. conceived and planned the project and wrote the manuscript. M.P.K.B. and P.H. collected the data; M.P.K.B. analysed the data and prepared the figures. All authors gave final approval for the publication.

Competing interests. We declare we have no competing interests.

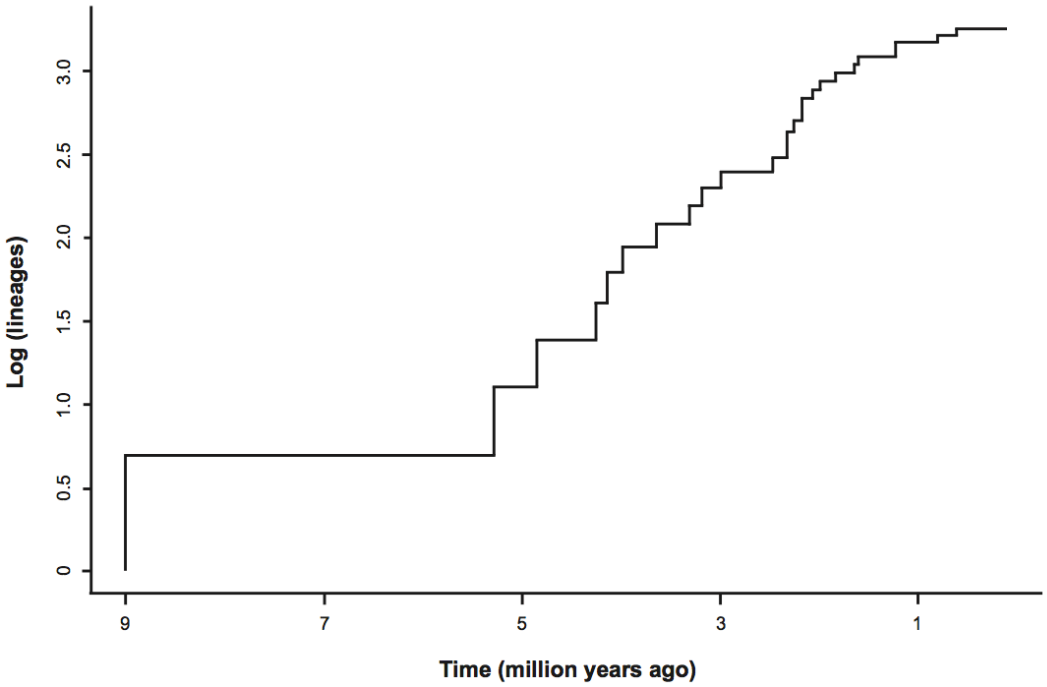
Funding. This work was supported by a grant from the Australian Research Council to C.M. (FL110100104).

Acknowledgements. We thank Mark Adams, Jason Bragg, and Sally Potter for their ongoing support to the *Cryptoblepharus* project. We thank the museum curators for access to tissues and specimens, Damien Esquerré, Martha Muñoz, and Marta Vidal-García for providing helpful advice during the analyses. We thank Stewart Macdonald and Pascal Title for sharing images used in figure 1. Finally, we thank the associate editor and two anonymous reviewers for their helpful suggestions to improve the manuscript.

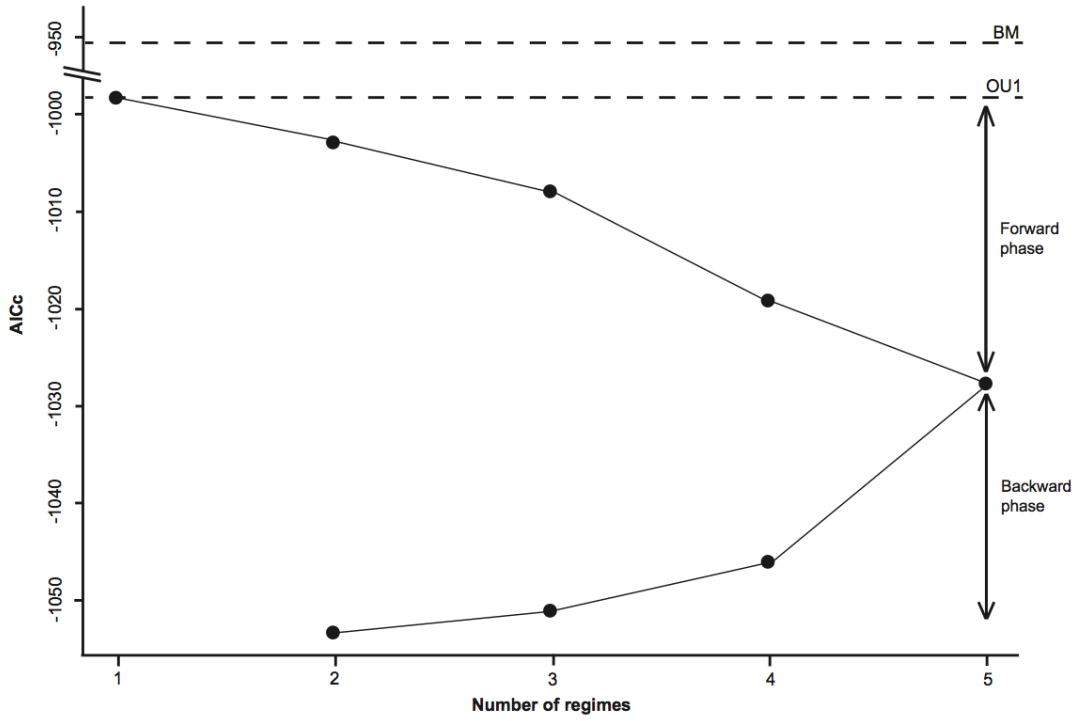
References

1. Losos JB, Mahler DL. 2010 Adaptive radiation: the interaction of ecological opportunity, adaptation and speciation. In *Evolution Since Darwin: The first 150 years* (eds MA Bell, DJ Futuyma, WF Eanes, JS Levinton), pp. 381–420. Sunderland, MA: Sinauer Associates.
2. Givnish TJ. 2015 Adaptive radiation versus 'radiation' and 'explosive diversification': why conceptual distinctions are fundamental to understanding evolution. *New Phytol.* **207**, 297–303. (doi:10.1111/nph.13482)
3. Schluter D. 2000 *The ecology of adaptive radiation*. Oxford, UK: Oxford University Press.
4. Losos JB. 2010 Adaptive radiation, ecological opportunity, and evolutionary determinism. *Am. Nat.* **175**, 623–639. (doi:10.1086/652433)
5. Simpson GG. 1953 *The major features of evolution*. New York, NY: Columbia University Press.
6. Wright S. 1982 Character change, speciation, and the higher taxa. *Evolution* **36**, 427–443. (doi:10.2307/2408092)
7. Losos JB, Ricklefs RE. 2009 Adaptation and diversification on islands. *Nature* **457**, 830–836. (doi:10.1038/nature07893)
8. Rundle HD, Nagel L, Boughman JW, Schluter D. 2000 Natural selection and parallel speciation in sympatric sticklebacks. *Science* **287**, 306–308. (doi:10.1126/science.287.5451.306)
9. Wagner CE, Harmon LJ, Seehausen O. 2012 Ecological opportunity and sexual selection together predict adaptive radiation. *Nature* **487**, 366–369. (doi:10.1038/nature11144)
10. Simões M, Breitkreuz L, Alvarado M, Baca S, Cooper JC, Heins L, Herzog K, Lieberman BS. 2016 The evolving theory of evolutionary radiations. *Trends Ecol. Evol.* **31**, 27–34. (doi:10.1016/j.tree.2015.10.007)
11. Rundell RJ, Price TD. 2009 Adaptive radiation, nonadaptive radiation, ecological speciation and nonecological speciation. *Trends Ecol. Evol.* **24**, 394–399. (doi:10.1016/j.tree.2009.02.007)
12. Gould SJ. 1989 *Wonderful life: the Burgess shale and the nature of history*. New York, NY: WW Norton & Company.
13. Lapiedra O, Sol D, Carranza S, Beaulieu JM. 2013 Behavioural changes and the adaptive diversification of pigeons and doves. *Proc. R. Soc. B* **280**, 20122893. (doi:10.1098/rspb.2012.2893)
14. Givnish TJ, *et al.* 2014 Adaptive radiation, correlated and contingent evolution, and net species diversification in Bromeliaceae. *Mol. Phylogenet. Evol.* **71**, 55–78. (doi:10.1016/j.ympev.2013.10.010)
15. Moen DS, Morlon H, Wiens JJ. 2016 Testing convergence versus history: convergence dominates phenotypic evolution for over 150 million years in frogs. *Syst. Biol.* **65**, 146–160. (doi:10.1093/sysbio/syv073)
16. Esquerré D, Keogh JS. 2016 Parallel selective pressures drive convergent diversification of phenotypes in pythons and boas. *Ecol. Lett.* **19**, 800–809. (doi:10.1111/ele.12620)
17. Hughes C, Eastwood R. 2006 Island radiation on a continental scale: exceptional rates of plant diversification after uplift of the Andes. *Proc. Natl Acad. Sci. USA* **103**, 10 334–10 339. (doi:10.1073/pnas.0601928103)
18. Rowe KC, Aplin KP, Baverstock PR, Moritz C. 2011 Recent and rapid speciation with limited morphological disparity in the genus *Rattus*. *Syst. Biol.* **60**, 188–203. (doi:10.1093/sysbio/syq092)
19. Rabosky DL, Donnellan SC, Grundler M, Lovette IJ. 2014 Analysis and visualization of complex macroevolutionary dynamics: an example from Australian scincid lizards. *Syst. Biol.* **63**, 610–627. (doi:10.1093/sysbio/syu025)
20. Ebel ER, DaCosta JM, Sorenson MD, Hill RI, Briscoe AD, Willmott KR, Mullen SP. 2015 Rapid diversification associated with ecological specialization in neotropical *Adelpha* butterflies. *Mol. Ecol.* **24**, 2392–2405. (doi:10.1111/mec.13168)
21. Glor RE. 2010 Phylogenetic insights on adaptive radiation. *Annu. Rev. Ecol. Evol. Syst.* **41**, 251–270. (doi:10.1146/annurev.ecolsys.39.110707.173447)
22. Giarla TC, Esselstyn JA. 2015 The challenges of resolving a rapid, recent radiation: empirical and simulated phylogenomics of Philippine shrews. *Syst. Biol.* **64**, 727–740. (doi:10.1093/sysbio/syv029)
23. Edwards SV. 2009 Is a new and general theory of molecular systematics emerging? *Evolution* **63**, 1–19. (doi:10.1111/j.1558-5646.2008.00549.x)
24. Garamszegi LZ. 2014 *Modern phylogenetic comparative methods and their application in evolutionary biology*. Berlin, Germany: Springer.
25. Kaliontzopoulou A, Carretero MA, Llorente GA. 2010 Intraspecific ecomorphological variation: linear and geometric morphometrics reveal habitat-related patterns within *Podarcis bocagei* wall lizards. *J. Evol. Biol.* **23**, 1234–1244. (doi:10.1111/j.1420-9101.2010.01984.x)
26. Horner P, Adams MA. 2007 Molecular-systematic assessment of species boundaries in Australian *Cryptoblepharus* (Reptilia: Squamata: Scincidae): A case study for the combined use of allozymes and morphology to explore cryptic biodiversity. *Beagle Suppl.* **3**, 1–21.
27. Horner P, Adams MA. 2007 A molecular systematic assessment of species boundaries in Australian *Cryptoblepharus* (Reptilia: Squamata: Scincidae) – a case study for the combined use of allozymes and morphology to explore cryptic biodiversity. *Beagle Suppl.* **3**, 1–19.
28. Blom MPK, Bragg JG, Potter S, Moritz C. 2016 Accounting for uncertainty in gene tree estimation: summary-coalescent species tree inference in a challenging radiation of Australian lizards. In review. (bioRxiv http://dx.doi.org/10.1101/056085)
29. Goodman B, Isaac J. 2008 Convergent body flattening in a clade of tropical rock-using lizards (Scincidae: Lygosominae). *Biol. J. Linn. Soc.* **94**, 399–411. (doi:10.1111/j.1095-8312.2008.00988.x)

30. Goodman BA, Miles DB, Schwarzkopf L. 2008 Life on the rocks: habitat use drives morphological and performance evolution in lizards. *Ecology* **89**, 3462–3471. (doi:10.1890/07-2093.1)
31. Losos JL. 2011 *Lizards in an evolutionary tree: ecology and adaptive radiation of Anoles*. Berkeley, CA: University of California Press.
32. Bragg JG, Potter S, Bi K, Moritz C. 2015 Exon capture phylogenomics: efficacy across scales of divergence. *Mol. Ecol. Resour*. (doi:10.1111/1755-0998.12449)
33. Singhal S. 2013 *De novo* transcriptomic analyses for non-model organisms: an evaluation of methods across a multi-species data set. *Mol. Ecol. Resour*. **13**, 403–416. (doi:10.1111/1755-0998.12077)
34. Blom MPK. 2015 EAPhy: A flexible tool for high-throughput quality filtering of exon-alignments and data processing for phylogenetic methods. *PLoS Curr. Tree of Life* **1**. (doi:10.1371/ currents.tol.75134257bd389c04bc1d26d42aa9089f)
35. Edgar RC. 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. **32**, 1792–1797. (doi:10.1093/nar/gkh340)
36. Maddison WP. 1997 Gene trees in species trees. *Syst. Biol*. **46**, 523–536. (doi:10.1093/sysbio/46.3.523)
37. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014 BEAST 2: A software platform for bayesian evolutionary analysis. *PLoS Comput. Biol*. **10**, e1003537. (doi:10.1371/journal.pcbi.1003537)
38. Brandley MC, Wang Y, Guo X, Nieto Montes de Oca A, Fería-Ortíz M, Hikida T, Ota H. 2011 Accommodating heterogenous rates of evolution in molecular divergence dating methods: an example using intercontinental dispersal of *Plestiodon* (Eumeces) lizards. *Syst. Biol*. **60**, 3–15. (doi:10.1093/sysbio/syq045)
39. Revell LJ. 2011 Phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol*. **3**, 217–223. (doi:10.1111/j.2041-210X.2011.00169.x)
40. Revell LJ. 2009 Size-correction and principal components for interspecific comparative studies. *Evolution* **63**, 3258–3268. (doi:10.1111/j.1558-5646.2009.00804.x)
41. Pennell MW, Eastman JM, Slater GJ, Brown JW, Uyeda JC, FitzJohn RG, Alfaro ME, Harmon LJ. 2014 Geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics* **30**, 2216–2218. (doi:10.1093/bioinformatics/btu181)
42. Beaulieu JM, Jhwueng DC, Boettiger C, O'Meara BC. 2012 Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution* **66**, 2369–2383. (doi:10.1111/j.1558-5646.2012.01619.x)
43. Ingram T, Mahler DL. 2013 SURFACE: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike Information Criterion. *Methods Ecol. Evol*. **4**, 416–425. (doi:10.1111/2041-210X.12034)
44. Sinervo B, Losos JB. 1991 Walking the tight rope: arboreal sprint performance among *Sceloporus occidentalis* lizard populations. *Ecology* **72**, 1225–1233. (doi:10.2307/1941096)
45. Snyder RC. 1954 The anatomy and function of the pelvic girdle and hindlimb in lizard locomotion. *Am. J. Anat*. **95**, 1–45. (doi:10.1002/aja.1000950102)
46. Schulte II JA, Losos JB, Cruz FB, Núnez H. 2004 The relationship between morphology, escape behaviour and microhabitat occupation in the lizard clade *Liolaemus* (Iguanidae: Tropidurinae: Liolaemini). *J. Evol. Biol*. **17**, 408–420. (doi:10.1046/j.1420-9101.2003.00659.x)
47. Revell LJ, Johnson MA, Schulte II JA, Kolbe JJ, Losos JB. 2007 A phylogenetic test for adaptive convergence in rock-dwelling lizards. *Evolution* **61**, 2898–2912. (doi:10.1111/j.1558-5646.2007.00225.x)
48. Harmon LJ, Kolbe JJ, Cheverud JM, Losos JB. 2005 Convergence and the multidimensional niche. *Evolution* **59**, 409–421. (doi:10.1111/j.0014-3820.2005.tb00999.x)
49. Pincheira-Donoso D, Harvey LP, Ruta M. 2015 What defines an adaptive radiation? Macroevolutionary diversification dynamics of an exceptionally species-rich continental lizard radiation. *BMC Evol. Biol*. **15**, 153. (doi:10.1186/s12862-015-0435-9)
50. Mayr E. 1970 *Populations, species and evolution*. Cambridge, MA: Harvard University Press.
51. Coyne JA, Orr HA. 2004 *Speciation*. Sunderland, MA: Sinauer Associates.
52. Moen D, Morlon H. 2014 Why does diversification slow down? *Trends Ecol. Evol*. **29**, 190–197. (doi:10.1016/j.tree.2014.01.010)
53. Glor RE, Kolbe JJ, Powell R, Larson A, Losos JB. 2003 Phylogenetic analysis of ecological and morphological diversification in Hispaniolan trunk-ground anoles (*Anolis cybotes* group). *Evolution* **57**, 2383–2397. (doi:10.1111/j.0014-3820.2003.tb00250.x)
54. Wollenberg KC, Wang IJ, Glor RE, Losos JB. 2013 Determinism in the diversification of Hispaniolan trunk-ground anoles (*Anolis cybotes* species complex). *Evolution* **67**, 3175–3190. (doi:10.1111/evo.12184)
55. Paun O, Turner B, Trucchi E, Munzinger J, Chase MW, Samuel R. 2016 Processes driving the adaptive radiation of a tropical tree (*Diospyros, Ebenaceae*) in New Caledonia, a biodiversity hotspot. *Syst. Biol*. **65**, 212–227. (doi:10.1093/sysbio/syv076)
56. Singhal S, Moritz C. 2013 Reproductive isolation between phylogeographic lineages scales with divergence. *Proc. R. Soc. B* **280**, 20132246. (doi:10.1098/rspb.2013.2246)
57. Pinho C, Hey J. 2010 Divergence with gene flow: models and data. *Annu. Rev. Ecol. Evol. Syst*. **41**, 215–230. (doi:10.1146/annurev-ecolsys-102209-144644)
58. Leite RN, Kolokotronis SO, Almeida FC, Werneck FP, Rogers DS, Weksler M. 2014 In the wake of invasion: tracing the historical biogeography of the South American Cricetid Radiation (Rodentia, Sigmodontinae). *PLoS ONE* **9**, e100687. (doi:10.1371/journal.pone.0100687)
59. Parada A, D'Elia G, Palma RE. 2015 The influence of ecological and geographical context in the radiation of Neotropical sigmodontine rodents. *BMC Evol. Biol*. **15**, 172. (doi:10.1186/s12862-015-0440-z)



Supp. Fig. 1. Lineage through time plot.



Supp. Fig. 2. SURFACE results. Change in the corrected Akaike's Information Criterion (AICc) during the forward and backward phases of the SURFACE analysis. The dashed lines indicate the AICc scores for the OU model with a single adaptive peak and a Brownian Motion model.

Supplementary Table 1: Sample sizes for the morphological analysis, the habitat type each species belongs to and the museum registration number for the individual used for sequence capture and species tree inference (ABTC: Australian Biological Tissue Collection, QM: Queensland Museum and CCM: Moritz' et al. field collection).

Species	Sample size	Ecotype	Phylo rep
C. plagiocephalus	27	arboreal	ABTC63597
C. tytthos	26	arboreal	ABTC70865
C. ochrus	21	arboreal	ABTC29743
C. pannosus	61	arboreal	ABTC29754
C. exochus	29	arboreal	ABTC67985
C. mertensi	23	arboreal	ABTC30122
C. australis	98	arboreal	ABTC29749
C. fuhni	12	saxicolous	QMJ58845
C. gurrmul	13	littoral	ABTC28475
C. l. horneri	13	littoral	ABTC29264
C. l. litoralis	33	littoral	ABTC30378
C. daedalos	15	saxicolous	ABTC70694
C. juno	36	saxicolous	CCM1765
C. wulbu	11	saxicolous	ABTC72544
C. megastictus	9	saxicolous	CCM1178
C. zoticus	16	saxicolous	ABTC11972
C. ustulatus	28	saxicolous	ABTC29787
C. ruber - A1	22	arboreal	ABTC30319
C. ruber - A2	5	arboreal	ABTC30156
C. ruber - A3	4	arboreal	ABTC28730
C. buchananii	44	arboreal	ABTC63616
C. metallicus	117	arboreal	ABTC30152
C. cygnatus	70	arboreal	ABTC29638
C. p. pulcher	76	arboreal	ABTC30450
C. virgatus	30	arboreal	ABTC30392
C. adamsi	24	arboreal	ABTC30376

Supplementary Table 2: Loadings for the first two components of the phylogenetic principal components analysis (pPCA).

Character	pPC1	pPC2
Fore limb length	**-.89678899**	-.4126609
Hind limb length	**-.85257573**	-.4668857
Head depth	**.92359736**	-.3344562
Head width	.06457198	**-.4855113**
Snout to eye length	.18055190	**-.5765652**
Eye to ear length	-.07400464	**-.4652479**
Ear to arm length	-.51115946	**.6340225**
Variance explained (%)	66.5	18.9

Note: Boldface indicates traits used to interpret the axes

Supplementary Table 3: Disparity indices for each habitat category and between all points combined, based on PC1 species scores.

Data	Average squared Euclidean distance
All species	0.00455420
Arboreals	0.00055221
Rock	0.00130304
Littoral	0.00121110

Supplementary Table 4: Estimates of θ, phenotypic trait optima, and their standard error, as inferred under the OUM model with OUwie.

Habitat category	Theta	S.E.
Arboreals	0.02909755	0.004557275
Rock	-0.07700898	0.007361843
Littoral	0.00755553	0.01057598

CHAPTER 3

Gene Flow across Species Boundaries Despite Extensive Ecological and Temporal Divergence in a Continental Radiation of Australian Lizards.

Gene Flow across Species Boundaries Despite Extensive Ecological and Temporal Divergence in a Continental Radiation of Australian Lizards.

Mozes P.K. Blom[1] & Craig Moritz[1]

[1]*Research School of Biology, The Australian National University, Canberra ACT 0200, Australia*

ABSTRACT

We have become increasingly aware that introgression, interspecific gene flow via hybridization, is a widespread phenomenon across the tree of life. Introgression occurs most frequently between closely related sister-species and eventually ceases or results in the fusion of lineages. Yet, in contrast with this view, instances of hybridization are occasionally recorded among distantly related taxa and can even occur between non-sister species. It remains unclear whether introgression in such situations is evolutionary meaningful and results in the actual exchange of functional variants or whether the realized degree of nuclear introgression is negligible. To address this question, further empirical examples are required that span the evolutionary timescales along which introgression continues to proceed. Here we focus on a group of Australian lizards and explore patterns of interspecific introgression in both a temporal and ecological framework. We focus on three species pairs within the genus *Cryptoblepharus,* where cytonuclear discord varies from complete fixation of an alternative mitochondrial haplotype to no cytonuclear discord at all. We employ a genomic approach (exon-capture) to generate a large number of orthologous loci and utilize both sequence and SNP based approaches to infer the genealogical history of individual loci. We find marked differences in shared genetic variation between the three species pairs and outcomes are congruent across

inference approaches. Our results highlight that in this system: I) Cytonuclear discord per se was predictive of shared nuclear variation, II) the degree of introgression is independent from divergence time and III) that gene flow persists between species shaped by ecologically mediated divergent selection. Overall, this study highlights that (ancient) introgression can result in pervasive patterns of shared genetic variation, regardless of phylogenetic distance or ecological differentiation, and its functional relevance needs to be evaluated.

INTRODUCTION

We have become increasingly aware that well-established species boundaries can remain semipermeable for prolonged periods of time. Occasional hybridization can result in interspecific gene transfer and the sharing of genetic variants without the complete break down of species barriers. While it is long known that introgression occurs frequently in plants (Anderson 1949), the evolutionary significance of introgression among animals has historically been largely ignored (Mayr 1963; Dowling and Secor 1997; Arnold 1997). Yet, with an increasing ability to characterize the genetic background of species, the evolutionary importance of interspecific hybridization between distinct forms is now recognized as a potentially effective mechanism for introducing novel genetic variants (Mallet 2007; Schwenk et al. 2008; Twyford and Ennos 2011; Hedrick 2013). Modern humans for example, might have benefited from introgressed Neanderthal genes while colonizing non-African environments (Green et al. 2010) and Tibetans have a higher frequency of specific Denisovan alleles that increase respiratory efficiency in high-altitude regions (Huerta-Sánchez et al. 2014). Occasional introgression can thus benefit the adaptive potential of species by increasing genetic variation, while rampant gene flow should be selected against due to the break up of species-specific beneficial allele combinations (Racimo et al. 2015). Interspecific gene flow therefore introduces a conundrum of which the exact circumstances remain relatively unclear but that potentially holds important implications for speciation, adaptation and conservation.

Introgression frequently occurs between sister-species that have only diverged relatively recently (Rheindt et al. 2014; Nater et al. 2015). It is therefore not unexpected that interspecific gene flow has been frequently recorded in well-known adaptive radiations of recent origin (Seehausen 2004) and these examples provide rich case studies (Grant and Grant 2008; The Heliconius Genome Consortium 2013;

Smith and Kronforst 2013; Kozak et al. 2015; Paun et al. 2016). In fact, a propensity for hybridization can result in hybrid swarms and has even been suggested as the origin of several adaptive radiations (Seehausen 2004; Glor 2010). Hybridization does not have to be continuous in time but can rather be characterized in a fission and fusion framework, where fission (divergence via selection against hybrids) and fusion (the merging of species through hybridization) of lineages occurs periodically and is environmentally dependent (Grant and Grant 2008). Hybridization is therefore an important component in a dynamic process, where species might arise frequently but are often ephemeral in nature due to subsequent hybridization (Rosenblum et al. 2012; McKay and Zink 2015) or eventually develop pre- and postzygotic isolating mechanisms that increasingly limit the opportunity for introgression. Grant and Grant (2008) suggested that fission and fusion dynamics are therefore ultimately only a transient stage during the early stages of radiations and are eventually reduced to faint echoes in the genomes of older species.

Yet, in contrast with this view, instances of hybridization are occasionally also recorded among distantly related species. In Birds-of-Paradise for example, hybrids have been observed between parents from different genera that have been phylogenetically distinct for millions of years (Frith & Beehler 2008; Irestedt et al. 2009). Similarly, mimicry genes have been exchanged across species boundaries in Heliconius butterflies, in groups that separated as long as 30 million generations ago (Kronforst 2008; Zhang et al. 2016). Such observations defy the notion that full reproductive isolation (i.e. complete separation of gene pools) should uniformly emerge with time across the animal kingdom, but instead suggests that in some taxa there is a lack of strong selection against occasional introgression. This is not to say that incipient species barriers do not develop uniformly with a similar reduction in gene flow (i.e. see Roux et al. 2016), but rather that incidental introgression across well-established barriers might persist at different time scales of divergence between

taxa. Therefore, it is important to ask how often interspecific introgression across distant taxa occurs (Hedrick 2013), how long interspecific hybridization continues (Stuglik and Babik 2016) and at what stage it is reduced to represent mere echoes without much evolutionary significance. To address these questions, further empirical examples are required that span the evolutionary timescales along which introgression continues to proceed.

When studying the temporal dimension of interspecific gene flow, it is important to take the ecological context of introgression into account (Pereira et al. 2016). Ecologically distinct species could be less likely to hybridize or to exchange genes than are ecologically similar species that come into secondary contact after an extensive period of allopatric divergence. Variation in hybridization propensity can be expected due to differing underlying mechanisms that both drive reproductive isolation. Speciation due to ecologically mediated divergence should rapidly result in extrinsic reproductive isolation if the viability or the fertility of hybrids with intermediate genotypes is reduced compared to the parental species (Nosil 2012; Seehausen et al. 2014). In contrast, uniform selection across ecologically similar geographical isolates can lead to intrinsic reproductive isolation if Bateson-Dobzhansky-Muller or other genetic incompatibilities fix between lineages. However, it likely takes much longer for such genomic incompatibilities to emerge (Gavrilets 2003) and alternative speciation dynamics can therefore result in different rates of introgression between species pairs that have been separated across similar time scales. The focus of the current study is to explore empirical patterns of interspecific introgression in both a temporal and ecological framework.

Lizards of the genus *Cryptoblepharus* are small, scansorial and globally widely distributed skinks (Rocha et al. 2005; Horner 2007; Hayashi et al. 2009; Blom 2015a), that have radiated across the Australian continent in the last ~5 million years (Blom et

al. 2016). In a recent study (Blom et al. 2016), we examined the ecological context of diversification with a combined assessment of morphological, ecological and phylogenetic patterns. We highlighted the important role of ecologically mediated divergent selection between rock and tree habitats and the repeated evolution of convergent phenotypes ('ecomorphs') between species that independently switched between habitat. Simultaneously, species that emerged within the same habitat are often highly cryptic and these findings suggest that ecologically imposed selection constrains morphological divergence away from the optimal tree or rock phenotype. Diversification within this genus is therefore likely the joint product of divergent selection between distinct habitats and uniform selection in isolated populations across similar habitats.

Although alternative diversification processes might have simultaneously stimulated species proliferation, the contemporary distribution of ecologically similar and distinct species frequently overlap. *C. ruber* and *C. metallicus* for example, are two tree species that are widely distributed across the monsoonal tropics of Northern Australia (Fig. 1). Even though they are morphologically highly cryptic (Horner 2007; Horner and Adams 2007; Blom et al. 2016), they are genetically well differentiated and form a paraphyletic group with two other species which are more closely related to *C. ruber* than to *C. metallicus* (*C. buchanani* and *C. megastictus*; Blom et al. 2017). *C. juno* is a rock-adapted species that is locally sympatric with *C. ruber* and *C. metallicus*, but is markedly different in morphology and genetically highly divergent. These three species have been phylogenetically distinct from one another for about three to five million years; yet, while characterizing phylogeographic variation within each species, we observed extensive mitochondrial haplotype sharing across species boundaries (see below). Observations of cytonuclear discordance are often used to support claims of purported introgression (Toews and Brelsford 2012), but such incongruence does not always equate to a detectable degree of nuclear introgression and therefore has

proven to be an inconsistent proxy for interspecific gene flow (i.e. Alves et al. 2008; Good et al. 2015).

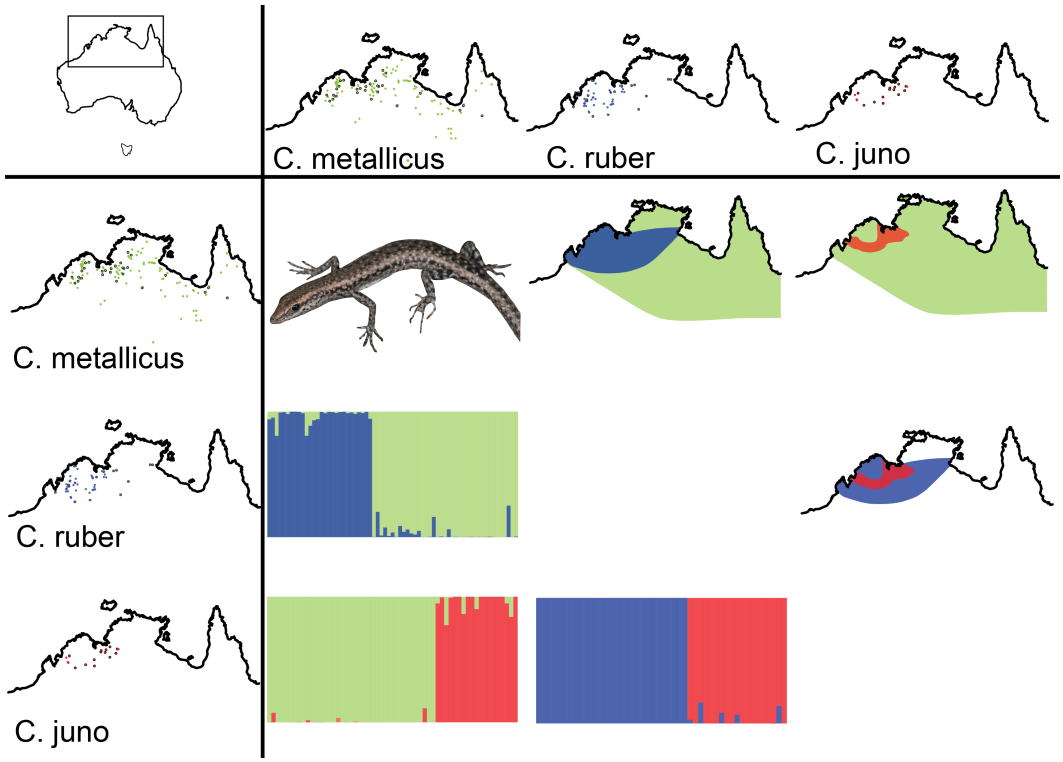


Figure 1. Distribution of *Cryptoblepharus* species examined for introgression and genetic clustering results for each pairwise species comparison. The species outline maps of Northern-Australia include the localities of the samples that were sequenced for an ND2 marker and overlapping species distributions (based on Horner (2007) and our findings) are illustrated in the cross sections above the diagonal of the diagram. Enlarged versions of the sampling locality maps are included in the supplementary material (Supp. Fig. 7 (*C. metallicus*), 8 (*C. ruber*), 9 (*C. juno*)). Bayesian clustering for each individual based on allele frequencies of nDNA as inferred by the program STRUCTURE are illustrated below the diagonal.

To examine whether the observed cytonuclear discordance reflects recent interspecific gene flow, we used an exon-capture approach and assayed genetic variation within each species for over 100 individuals. Large numbers of genetic markers are required to confidently discern between incomplete sorting of ancestral variation (incomplete lineage sorting; ILS) and post speciation gene flow. By

employing an exon-capture approach, we reconstruct whole mitochondrial genomes using an iterative baiting and mapping strategy, then generate a highly complete nuclear data matrix across divergent species and ultimately utilize both sequence and SNP based approaches to infer the genealogical history of individual genes. We then evaluate the distribution of genealogical histories and weigh alternative hypotheses that can explain patterns of shared genetic variation.

MATERIAL AND METHODS

Taxon Sampling

Based on the currently known range of each species (Horner 2007), we targeted individuals from unique localities across their distribution in northern Australia. Tissues were either sampled from museums (frozen or in alcohol; Museum and Art Gallery of the Northern Territory, South Australian Museum, Western Australian Museum, Queensland Museum or Museum Victoria) or from individuals collected during recent field expeditions (RNA later; 2013 – 2015). We first characterized mitochondrial diversity for a total of 390 individuals by sequencing the ND2 gene and then used a subset of 122 individuals for further analyses of nuclear sequence variation (exon-capture; see Supp. Table 1 for list of all samples). We selected individuals for inclusion in our exon capture experiment to represent i) the geographic distribution of each species, ii) mtDNA lineage diversity discerned by the ND2 screen and iii) areas of sympatry among species.

We also included individuals from the species *C. buchanani* (2x), *C. cygnatus* (2x) and *C. wulbu* (1x) which were sequenced in a previous exon capture experiment (Blom et al. 2017) and are lineages within the same radiation as the focal species.

Furthermore, we used *C. leschenault* (2x) to polarize SNPs in the ABBA-BABA tests and as outgroup in the phylogenetic analyses, since it belongs to a strongly supported alternative clade of *Cryptoblepharus,* is equidistant related to all the focal species in the present study (Blom et al. 2017) and only occurs outside of Australia (the lesser Sunda islands; Indonesia).

DNA Extractions, mtDNA Sanger Sequencing & Phylogenetic Inference

We extracted genomic DNA using the salting-out method of Sunnucks and Hales (1996). We amplified part of the mitochondrial gene encoding for NADH dehydrogenase 2 (ND2) using the PCR primers (5' – 3'): L4437b - AAGCAGTTGGGCCCATRCC, and ND2R_102 – CAGCCTAGGTGGGCGATTG (Smith et al. 2007) or ND2_f_jA – CACTCATACTAACTAACCTTGC, and ND2_r_jB - GTCTATCTAGGAGGCTTTAGC. PCRs were performed in 25 uL reactions containing ~100 ng DNA, 2.5 uL 10x PCR buffer, 0.2 mM dNTP, 2.5 mM MgCl2, 10 pmol each forward and reverse primer and 0.5 U *Taq* DNA Polymerase (Invitrogen). Each PCR reaction was run on a Corbett PC-960C cooled thermal cycler: initial denaturation step of 94 °C (1 min.), followed by 38 cycles of 94 °C (45 sec.), 52 °C (45 sec.) and 72 °C (1 min.). Each PCR run concluded with a final extension period of 1 min. at 72 °C. PCR products were subsequently purified with an exo-nuclease sequence clean-up (5 ul of PCR product, 0.4 ul Exonuclease 1, 1.6 ul Shrimp Alkaline Phosphatase and 3 ul of distilled water) at 37 °C for 45 min. followed by 80 °C for 15 min. Purified PCR products were sequenced in 20 ul reactions containing 0.8 ul BigDye Terminator v3.1 (Applied Biosystems), 4.5 ul 5x sequencing buffer, 3.2 pmol primer, 1 ul purified PCR product and 13.5 ul double-distilled water (see Potter et al. 2016 for Big Dye cycle sequencing conditions). Products were washed with sodium acetate and eluted in 20 ul of HiDi formamide, prior to sequencing on an ABI 3100 DNA analyzer.

We visually inspected each sample and manually edited sequences in GENEIOUS (v. 6.1). Our main goal for sequencing ND2, was to characterize the phylogeographic diversity within and across species. Accurately characterizing relationships between such distinct lineages is challenging with a single mitochondrial locus of limited length and we therefore only inferred a maximum-likelihood tree using RAxML (v.8.0**;** Stamatakis 2014). We inferred the tree with the highest likelihood out of 100 replicates, assuming a GTR + Γ model of sequence evolution, and subsequently generated 1000 bootstrapped trees to estimate bipartition support across bootstrap replicates.

Exon Capture Design, Library Preparation & Illumina Sequencing

We used two different sequence capture designs to enrich target loci of interest. The first capture design is the same as outlined in detail in Bragg et al. (2016) and the second capture is a modified version of the original design but excluded target regions that were not consistently recovered. Whereas the initial design was solely based on transcriptomes of three species from genera related to *Cryptoblepharus* (*Carlia rubrigularis, Lampropholis coggeri* and *Saproscincus basiliscus*; Singhal 2013), the modified version also included exonic information of a recently generated *Cryptoblepharus ruber* transcriptome (individual CMWA61, El Questro WA).

In brief, we identified the capture targets based on exonic regions with a balanced base composition in the *Anolis* genome and only selected orthologs that were also present in the transcriptomes of the aforementioned skink species. A total of 3320 exon targets were selected in the first design (each > 200 base pairs) and 2920 in the second design. Based on these exon targets, Roche NimbleGen designed and synthesized a SeqCap EZ Developer Library as our probe set. These probes capture

homologous targets with high efficiency across the entire Eugongylus group, to which *Cryptoblepharus* belongs (Bragg et al. 2016).

We prepared genomic libraries with ~1400 ng. input DNA per sample and according to the protocol of Meyer and Kircher (2010), using modifications of Bi et al. (2012). In brief, library preparations consisted of blunt-end repair, adapter ligation, adapter fill-in and was followed by two separate index-PCRs to reduce PCR bias. All individuals that were enriched with the original capture design had a single unique bar code, whereas 70 individuals (exon-capture ID MBCAP05_*; see Supp. Table 1) had a double bar code combination and were targeted with the updated capture design. The updated capture design targets less loci and the applicability of double-indexing was explored to examine whether a higher number of individuals could be pooled and enriched within a single capture. We assessed DNA concentrations using a Nanodrop (Thermo Scientific) and the distribution of fragment lengths on 1.5% agarose gels. Barcoded libraries were pooled in equimolar ratios prior to hybridization and the exon-capture hybridization was performed following the SeqCap EZ Developer Library user guide (Roche Nimblegen). We assessed the quality of the hybridizations using qPCR following methods of Bi et al. (2012). The qPCR assays used specific primers to assess enrichment of targeted regions, and de-enrichment of non-targeted regions of the genome. In addition, the quantity and quality of the hybridizations were measured using a Bioanalyzer (Agilent technologies), to quantify the concentration of the pre- and post-capture libraries. Once the libraries passed all aforementioned quality checks (i.e. successful enrichment), they were submitted for sequencing (ACRF Biomolecular Resource Facility, ANU). We sequenced the enriched libraries (100 bp. paired-end) on a single Illumina HiSeq 2500 lane.

Alignment and Data Pre-processing for Phylogenetic Inference

A detailed description of the read processing and subsequent assemblies can be found in Bragg et al. (2016). In brief, we cleaned sequencing reads following a workflow developed by Singhal (2013) which removes low quality, low complexity and duplicate reads. Cleaned reads were initially mapped to the original exons used for target design and libraries were subsequently individually assembled using Velvet (v.3; (Zerbino and Birney 2008), K = 31, 41, 51, 61, 71, 81). The assembled contigs from different K values were merged with CAP3 (Huang and Madan 1999) and trimmed to their respective exon boundaries using EXONERATE (v.2.2; Slater and Birney 2005). We then used the assembled contigs for each individual as its own reference and mapped cleaned reads back to these best contigs. Mapping was performed using BOWTIE2 (v.2.2.2; Langmead & Salzberg 2012) and resulting SAM files processed with SAMTOOLS (v.0.1.19; Li et al. 2009). We employed GATK (McKenna et al. 2010) to identify heterozygous sites, mask sites with a low quality genotype call (GQ<20) and generated phased haplotypes using the individual sequencing reads ('read backed phasing').

To assess sequencing success and assembly quality, we calculated the number of exons recovered, the proportion of missing data (number of missing sites relative to total assembly length) and proportion of heterozygous sites for each individual. Due to the efficacy of our exon-capture approach across divergent species (Bragg et al. 2016), we were able to enforce strict limits on library quality and still retain a relatively high proportion of individuals. For downstream analyses we only used individuals with more than 1000 exons sequenced and less than 20% missing data. Furthermore, we removed the top 5% of individuals with the highest proportion of heterozygous sites to avoid inclusion of libraries that potentially contain cross sample contamination. We are aware that a high degree of heterozygosity might reflect recent hybridization, but

we elected to take a conservative approach and minimize the possibility of false positives.

We then used the best contigs for each individual to generate high-quality sequence alignments using a recently developed workflow for alignment and alignment filtering of exonic sequence data. While high-quality sequence alignments are essential for phylogenetic inference, the visual inspection of individual alignments is impractical with sub-genomic datasets. EAPhy (v. 1.0; Blom 2015b) uses MUSCLE (v.3.8.31; Edgar 2004), or any other preferred aligner, to generate basic alignments and then reviews alignment quality based on a user-defined set of criteria. We removed missing data from the ends of alignments and removed alignments with more than one stop codon or where more than three amino acids in a seven-codon window differed from the alignment consensus. We included each alignment over 150 bp. in length for further downstream analyses. EAPhy generates both individual gene alignments and a concatenated alignment of all filtered loci. Lastly, we also used EAPhy to generate a concatenated alignment of the polymorphic sites across all loci and to randomly sample a single SNP for each locus. We only included polymorphic sites that are biallelic, parsimony informative (i.e. minimal two individuals with a deviating genotype) and minimized the degree of missing data (i.e. random selection only considered the polymorphic sites with the fewest missing genotype calls).

Phylogenetic History of Mitochondrial Genome

Although we succeeded in outlining the major phylogeographic lineages with the ND2 sequence data alone; the evolutionary history between such lineages remained unresolved. We therefore strived to reconstruct whole or partial mitochondrial genomes ('WMG') from the exon-capture 'by-catch' (i.e. non-target sequenced reads) and used an iterative baiting and mapping approach; MITObim (Hahn et al. 2013).

MITObim finds initial regions of similarity between a target library and a distant reference, and then uses an iterative mapping strategy to find reads that overlap with these initial segments. This strategy is repeated until the complete mitochondrial genome is assembled or if no further overlapping reads are identified.

We used an annotated mitochondrial genome of another skink, *Lygosoma sundevalli* (Fujita et al., in prep.), as an initial reference and generated a *Cryptoblepharus* specific reference genome based on reads from three individuals outside our clade of interest (*C. zoticus;* Blom et al. 2017). We then repeated the MITObim mapping strategy for each individual library, using the *Cryptoblepharus* reference, and generated a reference mitogenome for all target individuals. We subsequently mapped the reads of each individual back to its own reference using Bowtie2 and generated a consensus sequence with BCF tools (v.1.3.1; Li et al. 2009). We masked sequence positions with a read depth below 10x coverage and calculated the overall proportion of missing sites for each consensus sequence. We aligned mitogenome sequences with less than 35% missing data using MUSCLE and visually inspected the resulting alignment in GENEIOUS. We removed unique insertions (i.e. present in less than two individuals) since such sites are not phylogenetically informative and identified a start and endpoint of the alignment based on data completeness across all individuals. We reconstructed the mitochondrial phylogeny using RAxML and inferred the tree with the highest likelihood out of ten tree searches (GTR + Γ). We estimated bipartition support across the tree with 100 bootstrap replicates.

Phylogenetic History of Nuclear Loci

Concatenation.— We first generated concatenated alignments where we either included one nuclear haplotype for each individual or where both haplotypes were

collapsed in a single sequence and ambiguous sites coded according to the IUPAC scheme. We took both approaches since heterozygous positions are not inferred as such in the likelihood calculations but are rather treated as uncertainty in the genotype call. We observed no major differences (in topology) between the two trees and therefore focus on the concatenated nuclear haplotype tree in subsequent analyses. We used RAxML to infer the maximum likelihood tree for each concatenated alignment. We inferred the maximum likelihood tree using a (GTR + Γ) model of sequence evolution, ten tree searches and estimated bipartition support with 100 bootstrap replicates.

Genealogical History of Individual Genes.— For analyses that required individual gene trees, we removed alignment columns with missing data and first used JModeltest v.2.1.0 (Darriba et al. 2012) to identify the substitution model with the best fit for each specific locus. We quantified the best fitting model for individual genes because many exonic loci do not harbor a sufficient number of variable sites to empirically estimate substitution rates as required for a GTR model. We used RAxML to infer individual gene trees using the best tree out of ten tree search replicates. We also estimated 100 bootstrap replicates, so we could evaluate the accuracy of the gene tree estimate using a Tree Certainty (TC) assessment (Salichos et al. 2014).

Introgression Analyses

By comparing the phylogenetic tree based on nuclear loci with the tree based on the mitogenome data, we characterized the frequency of cytonuclear discordance and identified two distinct species pairs (i.e. ecologically distinct *C. juno* – *C. metallicus* and ecologically similar *C. metallicus* – *C. ruber*) that contain signatures of putative introgression. In addition to the species pairs with signals of cytonuclear discord, we also compared a third species pair (*C. juno* and *C. ruber*) since they have overlapping

ranges and are both part of the aforementioned species pairs. The species pair involving *C. megastictus* and *C. ruber* also exhibited instances of cytonuclear discordance but these are very closely related sister taxa (Horner and Adams 2007; Blom et al. 2017) and the focus of the current study is specifically on interspecific gene flow between distinct species. Nonetheless, the evolutionary dynamics that have promoted diversification between these distinct ecomorphological lineages will be explored in future studies (Blom et al., in prep.).

We employed a number of alternative approaches to asses whether the observed cytonuclear discord reflects genealogical patterns of interspecific gene flow or ILS. We first used STRUCTURE (v.2.3.4; Pritchard et al. 2000) to evaluate signals of recent admixture and then combined these results with the observed patterns of cytonuclear discordance, to identify individuals within species pairs that have potentially inherited a higher proportion of introgressed genes. Simultaneously, we also identified putatively 'pure' individuals; individuals without evidence for recent admixture or cytonuclear discord. By selecting four individuals across this range (putative introgression vs pure) for each species, we could conduct pairwise comparisons (n = 16) and quantify potential changes in introgression patterns for each species pair (Supp. Fig. 1).

Structure analyses.— Since a phylogeny is strictly bifurcating, it can be inadequate for evaluating whether occasional introgression occurred and we used STRUCTURE to generate more fine scale estimates of genetic clustering. We first employed STRUCTURE independently for each species pair to reduce the complexity of the model and then ran STRUCTURE on a dataset that included all three species. We generated alignments that included all individuals from each species pair, randomly sampled a single SNP from each locus and then repeated the random sampling to generate a second independent dataset to account for differences due to stochastic

sampling of SNP's. Furthermore, we created an additional SNP alignment that sampled all available SNP's and therefore greatly increased the number of sites but also included multiple SNP's from the same exon. The STRUCTURE analyses were conducted using the admixture ancestry and independent allele frequency models for 10 iterations at K = 2 for each species pair comparison and K = 3 for the STRUCTURE run that included all species. We used a burnin period of 10000 generations and subsequently sampled an additional 100000 generations. We used a custom R-script to assess convergence of each run and discarded incidental runs that were stuck in a local optimum with a considerably lower likelihood value. We used STRUCTURE HARVESTER (Earl and von Holdt 2011) to generate CLUMMP input files and CLUMMP (Jakobsson and Rosenberg 2007) to align replicate runs.

Phylonetwork analyses.— Species tree methods infer bifurcating trees and assume that all interspecific haplotype sharing is due to ILS. Recent advances in the likelihood calculation of the multispecies network coalescent (Yu et al. 2014) have been incorporated in programs such as PhyloNet (Than et al. 2008) and can now efficiently model both ILS and reticulation for networks with limited numbers of taxa. We used PhyloNet (v.3.6.0) for each species pair and examined whether a species tree with no reticulation or a network with one or more reticulation events is better supported by the data. We generated four taxon alignments that included the individual combinations for each pairwise comparison of species (as detailed above), *C. buchanani* and *C. leschenault*. *C. buchanani* is included because in each comparison it is more closely related to one of the focal taxa and therefore should be its sister species under neutral expectations. *C. leschenault* was included as an outgroup since PhyloNet requires rooted gene trees. We inferred individual gene trees and ranked them based on TC score. Since it remains unknown to what extent PhyloNet analyses are sensitive to gene tree estimation error, we opted a conservative approach and only included gene trees that were well resolved. An unrooted four taxon tree only has one

internal branch and a TC score of 0.4 equates to a bootstrap support above 80 (Salichos et al. 2014). We included all gene trees with a TC score above 0.4 and inferred the species network using maximum likelihood. We estimated the likelihood score for each network, with zero to three reticulation events, and the best network at each reticulation value was chosen out of ten iterations.

Frequency of topologies.— Given our expectations of potential introgression within specific species pairs, our four-taxon trees can have three possible topologies: i) The correct species tree, ii) an introgressed topology (either due to introgression or ILS) or iii) an alternative topology (due to ILS only; see Supp. Fig. 2 for schematic). If reticulation has occurred between our candidate taxa, we would expect that the frequency of the introgressed topology should be much higher than the frequency of the alternative topology. Similarly, if reticulation has not occurred the frequencies of the introgressed and alternative topology should be approximately equal, under a null-model of ILS only. In essence, this is the same principle as the ABBA-BABA test (see Sousa and Hey 2013) but the frequency of ABBA's and BABA's specifically refers to ancestral and derived SNP's. Here we employ a similar approach but use estimated gene trees rather than SNP's alone as the divergence time between our species is relatively old and gene trees should therefore be equally informative. We estimated Patterson's-*D* (Green et al. 2010) by calculating the frequency of introgressed and alternative topologies and will subsequently use D_g to refer to the *D*-statistic based on gene trees and *D* to refer to the *D*-statistic as estimated based on single nucleotide variants (SNV's).

Frequency of ABBA-BABA SNV's.— We also estimated *D* based on SNV's, because i) SNV's are not subject to potential gene tree estimation error, ii) additional loci can be included regardless of gene tree estimation error and iii) a difference between D_g and *D* can highlight potential within locus recombination. We used the alignments for

each combination of individuals across species pair comparisons and polarized SNP's using the *C. leschenault* outgroup. Having specified the *C. leschenault* genotype as ancestral, we then calculated Patterson's *D* based on the distribution of ancestral and derived alleles at each polymorphic site. Based on the well-resolved species tree, we assigned species for each species pair comparison as either P2 or P3, *C. buchanani* as P1 and *C. leschenault* as O (see Supp. Fig. 2). We used the R-package 'evobiR' (Blackmon and Adams 2015) to calculate the frequency of 'BBAA', 'ABBA' and 'BABA', infer the D-statistic, conduct 1000 bootstraps with replacements to calculate Z-values and estimated the probability of the observed distribution of site patterns.

RESULTS

Exon Capture Results

In the present study, we specifically targeted a large number of individuals from our focal species and simultaneously used individuals that have been sequenced for prior phylogenetic studies (Blom et al. 2017). We have previously presented a detailed evaluation of our sequence capture approach, capture success and how this changes while targeting taxa at different evolutionary timescales (Bragg et al. 2016). Overall, our results are congruent with these findings (see Supp. Table 1) and we therefore only focus on the key results pertaining to the current study.

We retained 105 high-quality libraries for which on average 2149 loci (110x) were recovered. This dataset with high-quality libraries included 19 *C. juno*, 11 *C. megastictus*, 39 *C. metallicus* and *29 C. ruber* individuals. The other libraries were non-focal species and were sequenced prior. We successfully recovered a high proportion of loci (see Supp. Table 1), regardless of which capture design was used. The mean

coverage per locus dropped significantly (157x to 37x) when we pooled a larger number of individuals using the double-barcoding approach, even though fewer loci were targeted in our updated capture design. Nonetheless, 70% of these libraries could still be included.

Cyto-Nuclear Discordance

Phylogeographic Diversity based on ND2 and WMG phylogeny.— Our phylogeographic analysis based on ND2 revealed several well-supported mitochondrial clades but relationships between clades remained obscure (Electronic Supp. Material 1). Inspection of the ND2 phylogeny already indicated extensive mitochondrial haplotype sharing across species boundaries, in particular between the ecomorphologically distinct species *C. juno* and *C. metallicus*. Phylogenetic analyses based on WMG's confirmed preliminary observations based on ND2 alone and furthermore resolved relationships between clades with high support (Fig. 2b; Supp. Fig. 3). A few individuals from the Keep and Kimb clade aside, most *C. juno* individuals appear to have mtDNA from *C. metallicus* and *C. juno* individuals frequently cluster within clades of *C. metallicus* from the same geographic region rather than by original species assignment based on morphology (Fig. 2b; Supp. Fig. 3 and 4). The small number of *C. juno* representatives with a distinct mitochondrial haplotype, still have a haplotype that is embedded within the *C. metallicus*/*C. ruber*/*C. megastictus*/*C. buchanani* group rather than being a sister-taxa to *C. cygnatus,* suggesting that the *C. juno* haplotype is completely replaced (Fig. 2b; Supp. Fig. 3). No mitochondrial haplotype sharing was observed between the distinct ecomorphological species *C. ruber* and *C. juno*, and mitochondrial introgression between *C. ruber* and *C. metallicus* is not directly evident without the nuclear data due to a high degree of morphological similarity (and therefore a high probability of species misidentification).

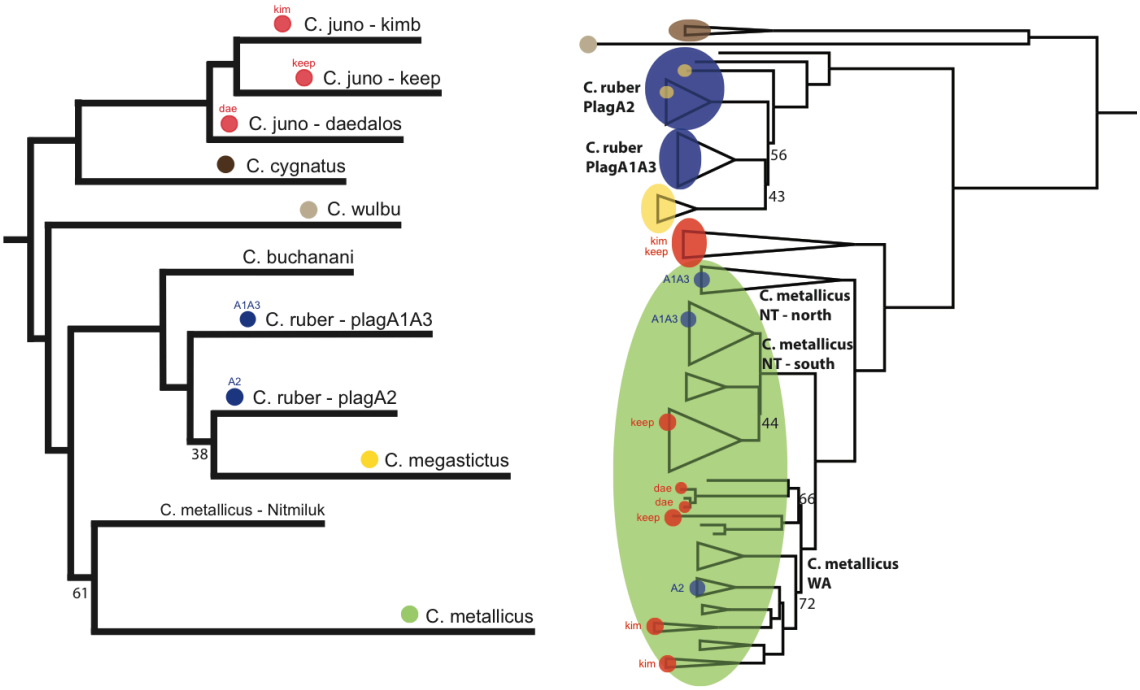


Figure 2. Maximum-likelihood (ML) species phylogeny based on concatenation of 421 nuclear loci and ML phylogeny based on complete mitochondrial genomes. a) Graphical representation of species relationships based on nuclear loci for 104 individuals, where all individuals belonging to the same species are collapsed in a single branch. b) Graphical representation of species relationships based on complete mitochondrial genomes for 75 individuals and well-supported phylogeographic clades are collapsed. Clades are colored by nuclear species classification and the color palette matches across trees. Bipartitions with bootstrap support below 95 are annotated at the corresponding node in each tree. Both phylogenies were rooted with two *C. leschenault* individuals and these were subsequently trimmed from each tree.

We did not manage to reconstruct WMG's for all individuals since the mitochondrial genome was not targeted in the first place. Yet, due to the high copy number of mitochondrial templates, a proportion of sequence reads that originate from mitochondrial fragments were still present in the final sequence libraries and were utilized for reconstruction of the mitochondrial genome. Unsurprisingly, the assembly of WMG's was particularly successful for libraries with poor capture efficiency and vice versa (Supp. Table 1). We managed to recover 75 WMG's that were on average 15122 bp's in length and contained less than 35% missing genotype calls under a 10x coverage cut-off.

Phylogeny based on Concatenation of Nuclear Loci.— The species phylogeny based on concatenation of 421 nuclear loci (169,671 bp's; Fig. 2a) is highly congruent with the topology as inferred with a previous extensive summary-coalescent based analyses that only included a single representative for each species (Blom et al. 2017). The relationships among species are highly supported, except for the placement of a *C. metallicus* individual (CCM5451) from Nitmiluk National Park and the relationships between the two distinct lineages of *C. ruber* and *C. megastictus*. Nonetheless, *C. megastictus* and *C. ruber* form a highly supported monophyletic group and is the sister-clade to *C. buchanani*; an allopatric arboreal species from south west Australia. *C. ruber* consists of two main lineages with distinct allozyme profiles (Horner and Adams 2007) and these distinct lineages are recovered with both our nuclear (Fig. 2a) and mitochondrial data (Fig. 2b). However, they have not been delineated as separate species since they are closely related, morphologically cryptic and their exact distributions remain unknown (Horner 2007). Since there does not seem to be a difference in introgression patterns (see Fig. 2b; mitochondrial introgression from *C. metallicus* into both clades of *C. ruber*) they are treated as a single species but representatives of each lineage are included in the introgression analyses (Supp. Fig. 1).

C. juno forms a well-supported distinct clade together with *C. cygnatus*, from all the other target species in the current study. In this study, we have clustered *C. juno* and *C. daedalos,* even though *C. daedalos* has been described as a distinct species based on a limited degree of morphological variation and four fixed allozyme differences (Horner and Adams 2007). *C. daedalos* and *C. juno* both occur on a saxicolous substrate and are closely distributed in space. By delineating *C. daedalos* as a distinct species, nuclear divergence suggests that *C. juno* likely contains two distinct forms (*C. juno – kimb* and *C. juno – keep*) that are almost equally divergent from one another as from *C. daedalos* (Fig. 2a). Yet, mitochondrial introgression patterns seem similar across all

three forms (Fig. 2b). Given the focus of the current study and uncertainty regarding exact species boundaries, we cluster the three lineages in a *C. juno* species complex but account for differences between lineages by including representatives of each in our introgression analyses (Supp. Fig. 1).

Introgression Analyses

STRUCTURE.— We used a concatenated SNP alignment containing an average of 7267 polymorphic sites for our analyses of fine-scale genetic structure across our three species comparisons. Individual species were consistently recovered and most individuals exhibit marginal amounts of shared ancestry between species (e.g. only 7% of individuals in the *C. juno*/*C. metallicus* comparison have over 10% shared ancestry) regardless of geographic origin or when sampling only a single SNP per exon (Supp. Fig. 5). Yet, some genetic admixture between species was observed in each species comparison (Fig. 1). Whereas this is bi-directional for the *C. ruber* and *C. metallicus* comparison, *C. juno* individuals tend to contain a higher degree of admixed individuals than *C. ruber* and *C. metallicus* in each respective species comparison. Nonetheless, the overall degree of admixture is marginal and suggests that no recent hybrids (i.e. F1's) were included in the analyses, but that some signal of (ancient) admixture is present. Similar patterns are observed when employing a genetic clustering approach with all three species simultaneously included (Supp. Fig. 4). However, in the three species analysis the degree of admixture between *C. juno* and *C. ruber* is reduced, with most admixture within *C. juno* individuals being shared with *C. metallicus*.

PhyloNetwork.— For each species comparison, we used 16 combinations of individuals with varying degrees of cyto-nuclear discordance and/or nuclear admixture (Supp. Fig. 1). PhyloNetwork analyses were on average based on 580 well-

resolved gene trees (TC score > 0.4) and results were highly congruent within each species comparison. The likelihood score of a network improved markedly (mean likelihood change = 33.6) for each *C. juno* and *C. metallicus* comparison with a single reticulation event added, in comparison to a species tree without any reticulation events (Fig. 3a). The inference of phylogenetic networks with additional reticulation events modeled did not improve the likelihood score any further, suggesting that a single reticulation event accounts for most incongruence among genealogies. Accounting for a single reticulation event yielded limited improvement of the likelihood score for the *C. ruber*/*C.metallicus* comparisons (mean likelihood change = 9.3; Fig. 3b) and was even more limited for most *C. ruber*/*C. juno* comparisons (mean likelihood change = 7; Fig. 3c).

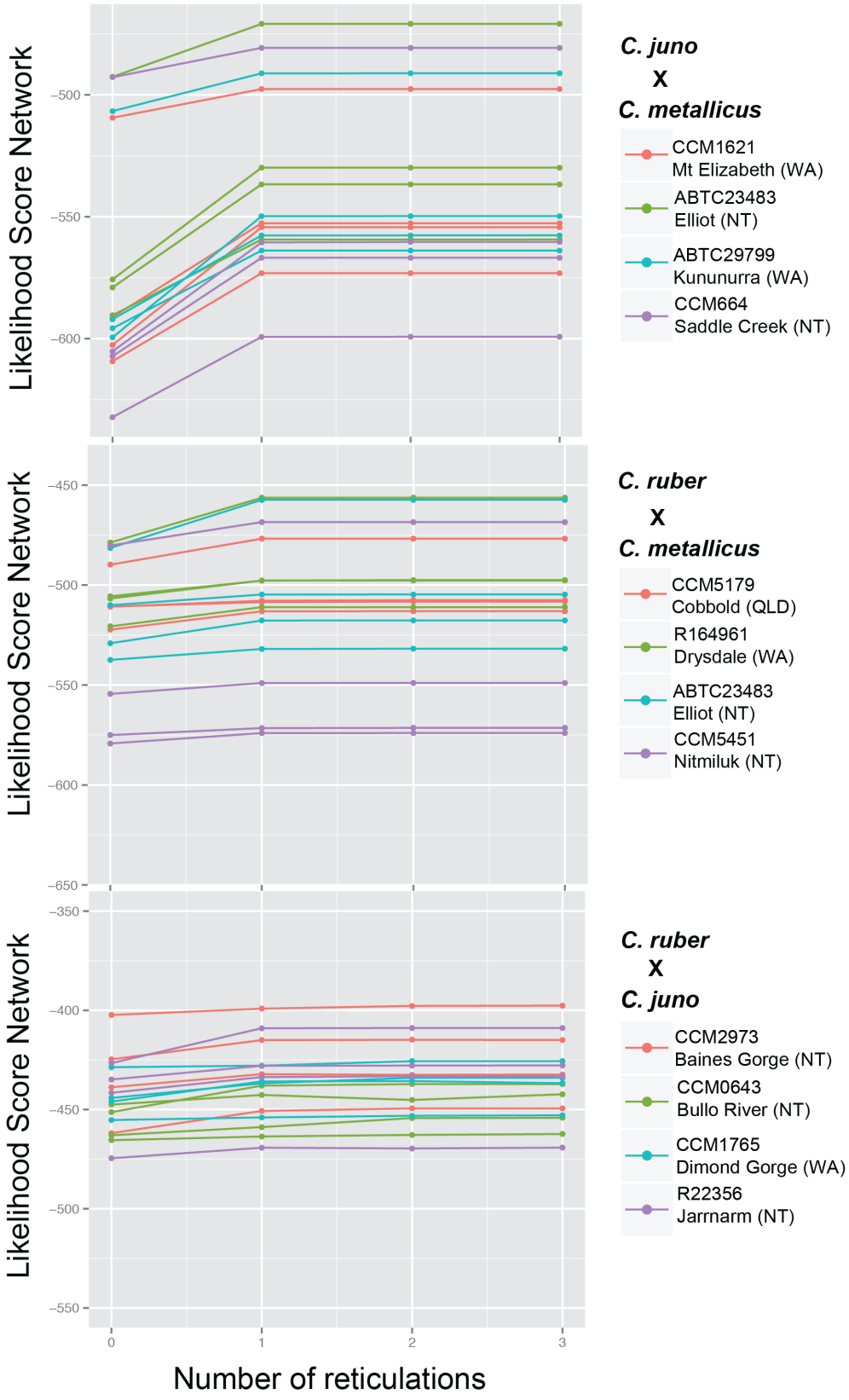


Figure 3. Likelihood scores for phylonetworks, as inferred using PhyloNet, for each individual combination across species comparisons. The likelihood score for each of the 16 combinations of individuals, across the three species comparisons with putative introgression. Likelihood scores were calculated under a model of zero, one, two or three reticulation events.

Frequency of Topologies.— We scored the frequency of topologies that supported the species tree, a putatively introgressed topology or an alternative topology, using the same collection of four taxon gene trees as included in the PhyloNet analyses. The high support for a single reticulation event in the *C. juno* – *C. metallicus* comparison is paralleled by a high frequency of topologies that reflect past introgression or ILS (Fig. 4a). However, the ratio of introgressed and alternative topologies is markedly skewed (mean D_g = 0.50) towards a higher prevalence of putative introgressed topologies where *C. juno* and *C. metallicus* are more closely related to one another than to *C. buchanani*. Furthermore, D_g remains remarkably stable regardless of the combination of *C. juno*/*C. metallicus* individuals examined. Across individual combinations, the main difference is the number of well-resolved gene trees recovered; each comparison involving *C. juno* from Jarrnarm (ABTC22356) results in a decrease in well-resolved gene trees but interestingly, this decrease only seems to manifest itself in a decrease of topologies that support the species tree, while the ratio of ABBA and BABA topologies does not deviate.

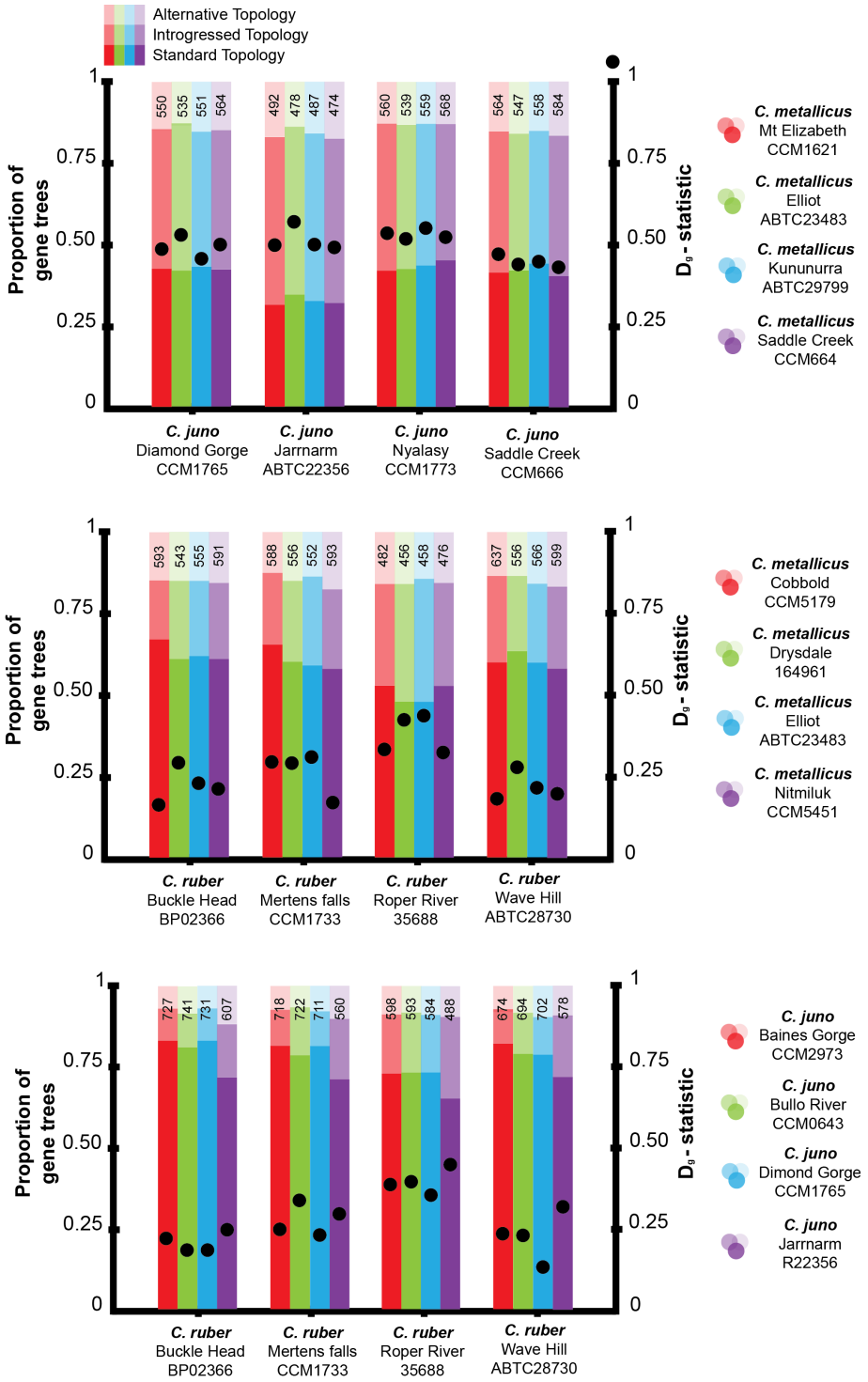


Figure 4. Frequencies of true species tree, introgressed and alternative topology for each of the individual combinations across species comparisons. Unrooted four taxon alignments can yield three possible topologies and contain one internal branch. Gene trees were inferred for each individual combination across species comparisons and included *C. buchanani* and *C. leschenault*. Vertical bars represent the frequency of gene trees supporting each possible topology and total number of gene trees are listed at the top of each bar. The inferred estimate of D_g is included as a black dot for each individual combination across species comparisons.

Estimates of D_g for the *C. metallicus*/*C. ruber* and *C. ruber*/*C. juno* comparisons are on average much lower than for the *C. metallicus*/*C. juno* comparison (Fig. 4). Furthermore, the frequency of topologies that support the species tree is much higher for both of these comparisons, but in particular for *C. juno*/*C. ruber* where on average 77% show a genealogical pattern that is concordant with the known species tree. This is a remarkable shift, when considering the proportion of gene trees that support the species tree in the *C. juno*/*C. metallicus* comparison; 41%. Different combinations of individuals for *C. metallicus*/*C. ruber* and *C. ruber*/*C. juno* comparisons also results in a greater variation of D_g, suggesting that the relative frequency of ABBA and BABA topologies differ between combinations of individuals within each species comparison. Lastly, similar as for *C. metallicus*/*C. juno*, for both comparisons individual combinations that include *C. ruber* from Roper River (ABTC35688) have a marked decrease in the number of well resolved topologies but this decrease seems mainly manifested in a reduced number of genealogies that follow the species tree. For the *C. juno* and *C. ruber* comparison, this pattern can also be observed for every species comparison that include *C. juno* from Jarrnarm (ABTC22356).

D-statistic based on SNV's.— Patterns of introgression between topology based analyses and SNP based approaches are largely congruent across species comparisons. The number of SNP sites that are informative with respect to the ABBA-BABA test varies between 1824 and 3670 (Supp. Table 2). Regardless of the combination of individuals used for the species comparison involving *C. juno* and *C. metallicus*, there is an excess of shared variants between *C. juno* and *C. metallicus* (mean D = 0.22; Fig. 5a; Supp. Table 2). For all individual combinations this excess is significantly higher than would be expected under a model of ILS alone (mean Z-score = 8.08, $p < 0.05$). Furthermore, the ratio of sites with an ABBA or BABA pattern remains consistent as highlighted by the stable estimate of the D-statistic across individual combinations (Supp. Table 2). Estimates of Patterson's D are on average two-fold lower for the other

two comparisons; *C. ruber*/*C. metallicus* (mean D = 0.11; Fig. 5b; Supp. Table 2) and *C. ruber*/*C.juno* (mean D = 0.07; Fig. 5c; Supp. Table 2). While the estimates for Patterson's D are markedly lower for *C. ruber*/*C. metallicus*, bootstrapping with replacement suggests that the frequency of observed sites with an ABBA pattern is still significantly higher than the number of BABA sites for most individual combinations (94%). By comparison, the observed ABBA-BABA ratio for 56% of the *C. ruber/C.juno* combinations do not significantly deviate from the null-expectation of no introgression and an equal numbers of ABBA-BABA sites. Finally, the estimates of D (based on SNV's) and D_g vary considerably and are on average two-fold greater when inferred based on genetrees.

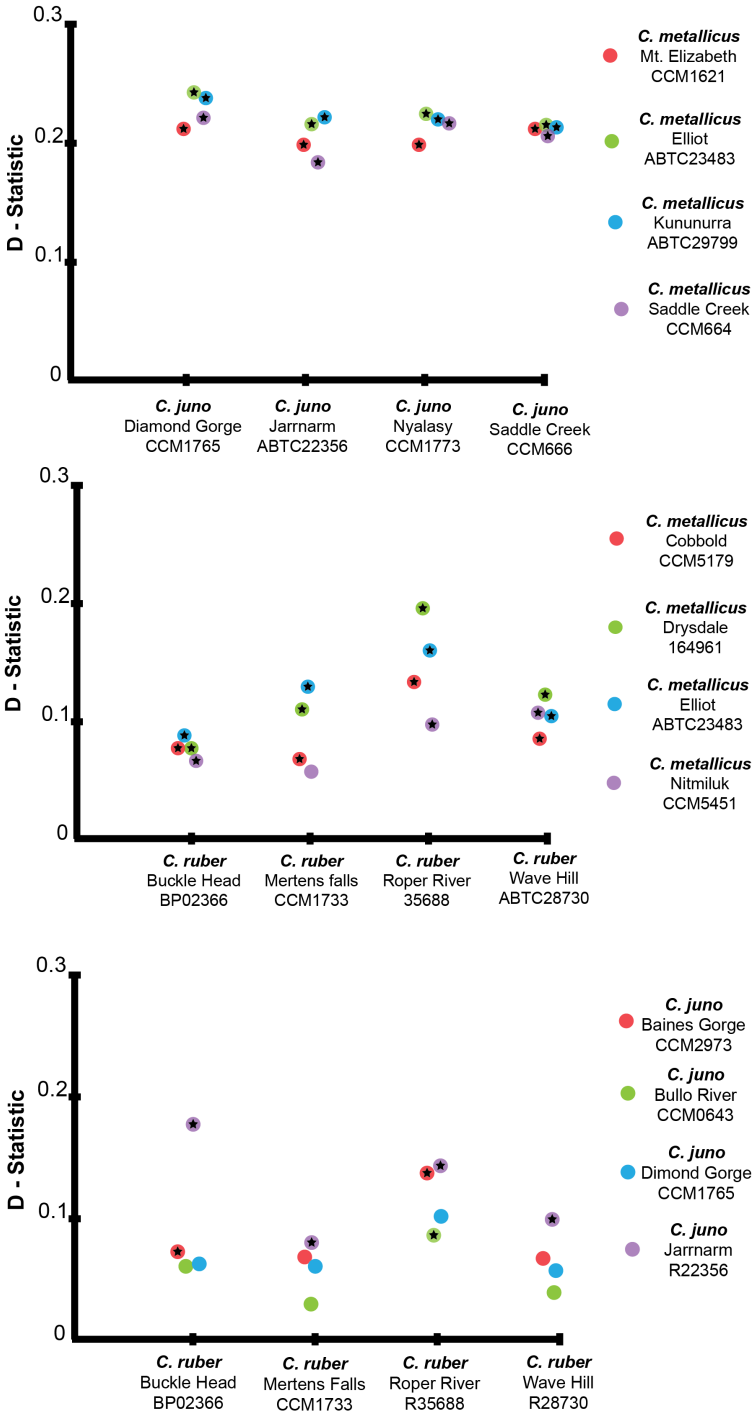


Figure 5. Inferred *D*-values following ABBA-BABA tests for each individual combination across species comparisons.

D-statistic values highlighting the presence of an excess of shared derived variants, as inferred for each individual combination across species comparisons. Single nucleotide variants (SNV) were polarized by the genotype of *C. leschenault* and the frequency of ABBA and BABA sites used to calculate Patterson's *D*. Each alignment of SNV included the two focal individuals for each combination, *C. buchanani* and *C. leschenault*. Significant deviations from zero, an equal ratio of ABBA and BABA site patterns, are highlighted with a black star.

DISCUSSION

It is now well established that hybridization among animal species can result in shared genetic variation across species boundaries. Yet, most introgression tends to occur between closely related sister taxa and gene flow either ceases with increasing time of divergence or results in the fusion of lineages. Here we have shown that this does not always have to be the case and that in some taxa interspecific gene flow can persist for a remarkably long time after initial divergence.

Quantifying Introgression

Observations of cytonuclear discordance during a phylogeographic screen of mitochondrial diversity primarily motivated the current study. For mtDNA, individuals of the distinct ecomorphological species *C. juno* and *C. metallicus* clustered more frequently by geography than species (Fig. 2, Supp. Fig. 3 and 4). Similarly, when comparing our nuclear and mitochondrial phylogeny, a second instance of mitochondrial haplotype sharing was observed between *C. ruber* and *C. metallicus,* which was initially undetected due to the high degree of morphological similarity between species. The mitochondrial haplotype of most *C. ruber* individuals from the Northern-Territory was completely replaced by a *C. metallicus* equivalent from the same region. The overwhelming majority of reported cases on introgression stem from similar observations of cytonuclear discordance as presented here, but such discord does not always reflect a meaningful degree of nuclear introgression (Alves et al. 2008; Good et al. 2015). A recent study (Good et al. 2015) on two hybridizing chipmunk species for example, reported complete fixation of introgressed mitochondrial DNA in some populations but very minor levels of interspecific nuclear gene flow (~1%). They therefore concluded that introgressive hybridization had little

impact on the overall genetic composition of each species. Such studies highlight the importance of using genome wide approaches to quantify the relative contribution of introgression to the genetic make-up of species, rather than inferring introgression from mitochondrial loci alone (Toews and Brelsford 2012). However, our results suggest that although phylogeographically structured mitochondrial capture is a poor proxy for the degree of introgression (i.e. syntopic species that shared similar mitochondrial haplotypes were not necessarily more admixed for nDNA), cytonuclear discordance per se was predictive of shared nuclear variation across species boundaries and therefore remains a valuable tool for detecting potential instances of introgression.

Although initial clustering analyses of nuclear sequence variation identified a relatively modest degree of admixture, a detailed assay of introgression using both well-resolved gene trees and SNP data suggests that the genealogical history of a considerable number of loci conflicts with the species tree. Instead, a large proportion of gene trees (across all individual combinations of *C. juno* and *C. metallicus,* and to a lesser extent for *C. metallicus* and *C. ruber*), match a topology that is congruent with an introgression scenario.

One of the main challenges in the study of introgression, is to disentangle signals of introgression from shared ancestry due to incomplete sorting of ancestral variation (Twyford and Ennos 2011). Network based analyses aim to simultaneously account for both sources of incongruence by incorporating the multi-species network coalescent (Yu et al. 2014). We found considerable support for a network that included one reticulation event for each individual combination of the *C. metallicus* and *C. juno* comparison, suggesting that the observed distribution of topologies is much better explained by a reticulation model than an ILS model alone (Fig. 3). The increase in support for a network is much less prevalent for the other two species

comparisons and a bifurcating model that only accounts for ILS already explains a considerable degree of the observed incongruence.

The relative frequency of each topology further supports these findings (Fig. 4). The frequency of introgressed topologies is approximately equal to the species tree across individual combinations of the *C. juno*/*C. metallicus* comparison, whereas the ratio of introgressed to species tree topologies is much lower for the other species groups. In all these comparisons, we used the placement of *C. buchanani* as a calibration to each topology; following the species tree *C. buchanani* should be more closely related to *C. ruber* and *C. metallicus,* than to *C. juno* (Blom et al. 2017). We specifically chose *C. buchanani* because it is the only species with a completely allopatric distribution from each of the focal species and therefore minimizes the probability of undetected introgression between our calibration and focal species (Horner 2007). However, *C. buchanani* and *C. ruber* are more closely related to one another and are separated from *C. metallicus* by a relatively long branch (Fig. 2a). It is therefore not unexpected that the absolute number of topologies which support the species trees is higher for each species comparison that involves C. ruber, since genes have had more time to sort accordingly (i.e. this also explains the overall higher number of well-resolved gene trees for the *C. juno*/*C. ruber* comparison). Nonetheless, the recurrent pattern that prevails across all comparisons is that the relative frequency of alternative topologies changes minimally; it is the ratio of topologies that either support the species tree or an introgression scenario which fluctuates.

The ratio of introgressed to alternative topologies is stable within the *C. juno*/*C. metallicus* comparison but fluctuates within the other two comparisons. However more remarkably, there are two specific individuals that systematically tend to have a lower number of well-resolved topologies, but this decrease is only manifested in a reduction of topologies that support the species tree. *C. juno* from Jarrnarm

(ABTC22356) is an individual that was selected because it showed an elevated proportion of shared ancestry in the STRUCTURE analyses and *C. ruber* from Roper River (35688) was included because it has a *C. metallicus* mitochondrial haplotype. These observations were surprising to us and future analyses are required to examine why these individuals systematically tend to have a reduction in well-resolved topologies and why this reduction is only affecting topologies that support the species tree (Fig. 4).

In addition to a gene tree focused approach, we also examined instances of interspecific haplotype sharing using SNV's. We were specifically motivated to explore both types of data, because introgression signals from gene tree topologies could be masked if intra-locus recombination is prevalent in many loci. Furthermore, loci that yielded poorly resolved gene trees where excluded and thus resulted in an overall dataset with less genetic markers. We employed ABBA-BABA tests and compared the frequency of introgressed (ABBA) and alternative (BABA) topologies for each SNV. We found that across each species combination, the frequency of ABBA topologies is greater than the frequency of BABA topologies (Fig. 4 and 5). This translated into positive estimates of Patterson's *D* across all individual combinations and species comparisons, and is most evident for *C. metallicus* and *C. juno*. Estimated values of Patterson's *D* are significantly different from zero, equal ratio of ABBA and BABA sites, for the majority of individual species combinations except for the species comparison *C. juno* and *C. ruber*. Although the *D* statistic values are relatively similar between the *C. metallicus*/*C. ruber* and *C. ruber*/*C. juno* comparisons, 9 out of the 16 individual combinations for *C. juno* and *C. ruber* are not significantly different and therefore indicates that shared haplotype variation across species is not higher than expected under a model of ILS only. Interestingly, although using a similar rationale, the estimates of *D* and D_g differ considerably and D_g estimates are on average almost two-fold higher (Fig. 4 and 5). Individual introgressed regions can be relatively short,

particularly under a model of ancient hybridization, and given the evolutionary timescale of our species comparisons the discrepancy between D and D_g is likely the result of intra-locus recombination. The overall introgression patterns across species comparisons remain similar nonetheless and do not differentiate between D and D_g. This suggests that disregarding intra-locus recombination likely resulted in an overestimation of the degree of introgression, but that this overestimation is proportionally equal across all species comparisons.

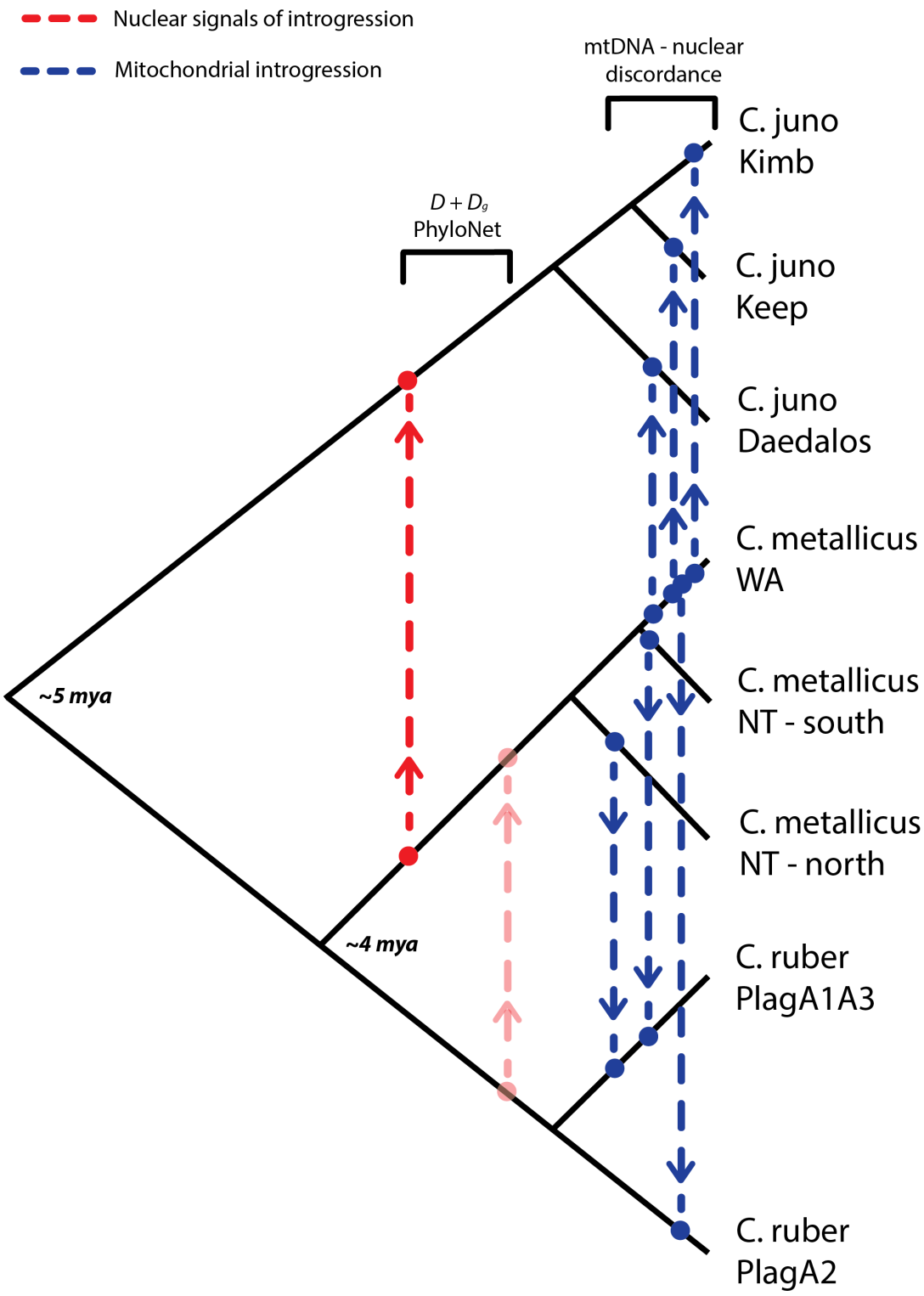


Figure 6. Schematic overview of the observed introgression patterns across analyses. Reduced representation of the phylogeny with observed introgression patterns plotted, for both the mitochondrial and the nuclear data.

Gene Flow across Species Boundaries

Our genomic characterization of genealogical variation within and between species comparisons, provides an intriguing perspective on the evolutionary history of *Cryptoblepharus* (Fig. 6) and simultaneously stimulates further evaluation of the importance of interspecific gene flow. All of our analyses highlight that *C. metallicus* and *C. juno* have hybridized after the split of *C. metallicus* and *C. buchanani*. Furthermore, our phylogenetic tree based on complete mitochondrial genomes suggests that hybridization has occurred after the split of *C. juno* and *C. cygnatus*, since the *C. juno* mitochondrial haplotype is completely replaced while a *C. cygnatus* haplotype is still present (Fig. 2b; Supp. Fig. 3). Based on these observations and a recently published time-calibrated phylogeny (Blom et al. 2016), hybridization has occurred at least one million years after the original split between the ancestral lineages of *C. metallicus* and *C. juno*. This is most likely the lower limit however, because divergence time is likely underestimated if occasional introgression has occurred since the original species split and *C. juno* individuals tend to have mitochondrial *C. metallicus* haplotypes that are phylogeographically structured (Fig. 2b; Supp. Fig. 3 and 4), hence postdating the splits among *C. metallicus* mtDNA lineages. The phylogeographic structure of mitochondrial haplotypes suggests that hybridization has likely continued after the initial hybridization process that led to the complete fixation of the *C. metallicus* mitochondrial haplotype in the *C. juno* ancestor. The relative similarity in the degree of shared haplotype variation (Patterson's *D* and D_g) across all individual combinations within the *C. juno* and *C. metallicus* comparison highlights that if this is the case, recent hybridization has not resulted in an increase of introgressed regions across the genome (but potentially a replacement of already introgressed regions). It currently remains challenging to disentangle the timing and the amount of introgression based on datasets such as ours. Historical migration rates

could potentially be estimated, using the full-length sequence data, in an 'isolation-with-migration' framework (Sousa and Hey 2013). However, we are uncertain how violation of the underlying assumptions would affect demographic inference in empirical systems such as *Cryptoblepharus*, where gene flow has likely occurred between non-sister taxa and thus migration needs to be modeled across multiple unique species. With the continuous ease to generate sequence data, it will soon be feasible to sequence whole genomes for datasets with phylogeographic sampling and analysis of linkage decay in introgressed regions across the genome might provide a more detailed insight on the timing of hybridization.

Introgression Between and Within Ecomorphs

We have previously shown that species that occur in different environments can be regarded as distinct ecomorphs since they differ in a variety of ecologically functional traits. Yet, our characterization of introgression patterns suggests that introgressive hybridization has been more prevalent between two species that occupy alternative peaks in an adaptive landscape, than species that are ecologically similar and more closely related. Complete reproductive isolation between *Cryptoblepharus* species therefore does not emerge uniformly with time and the ecological context of diversification needs to be considered. Introgressive hybridization between ecologically distinct forms have seemingly not resulted in the break-up of beneficial allelic combinations and neither resulted in the fusion of lineages. Furthermore, our concatenated and summary-coalescent phylogenetic analyses (Blom et al. 2017) consistently recover the underlying species tree and therefore highlight that species remain diverged across the majority of the genome. These results are consistent with other studies that have highlighted the semi-permeability of species boundaries and have shown that introgressive hybridization can result in a heterogeneous distribution

of gene flow across the genome (Larson et al. 2013; Poelstra et al. 2014). Ecological distinctiveness is retained by localized selection on genomic regions that harbor genetic variants of functional importance while other parts of the genome interchange relatively freely (Lindtke and Buerkle 2015). But where most of these studies focus on closely related sister taxa, our study of *Cryptoblepharus* indicates that these patterns might persist for prolonged periods of time.

Whereas the genetic signature of introgression between *C. metallicus* and *C. juno* seems markedly high, the support for past reticulation in the other ecologically distinct species pair, *C. ruber* and *C. juno,* is minimal. It remains unclear why there is less evidence of introgression between *C. ruber* and *C. juno*, even though divergence times between *C. juno* and *C. metallicus*/*C.ruber* are similar and both arboreal taxa are locally sympatric with *C. juno* across its range. *C. ruber* and *C. metallicus* occupy the same environmental niche and are morphologically highly cryptic, yet their (past) propensity for hybridization with *C. juno* is clearly different. The contrast between the two species pairs is remarkable and provides an exciting opportunity to further investigate the evolution of pre- and postzygotic isolating mechanisms, which remains largely unknown for the majority of taxa within the most species rich family of lizards, *Scincidae*.

Whereas ecologically distinct species such as *C. juno* and *C. metallicus* retained strong signals of past hybridization, the introgression signal was much less pronounced for *C. ruber* and *C. metallicus*; ecologically similar species with overlapping distributions. Previous work on Eugongylus skinks, the group to which the genus *Cryptoblepharus* belongs, has shown that reproductive isolation between cryptic lineages scales with divergence time and how lineages that have been separated for different amounts of time represent various stages along the speciation continuum (Singhal and Moritz 2013). Lineages that come into secondary contact after

relatively short periods of isolation seem to fuse in hybrid zones, whereas others that have been separated for prolonged periods only admix marginally. With a rough estimate of a species split around 4 Mya between *C. metallicus* and *C. ruber* (Blom et al. 2016), our observations of limited introgression seem to concur with the empirical estimates of Singhal and Moritz (2013). Although much further research is required and any inferences remain speculative, it is salient that patterns of introgression and the evolution of complete reproductive isolation can differ depending on ecological context. Whereas reproductive isolation might scale with divergence time between lineages that have emerged in isolated uniform environments (*C. metallicus* – *C. ruber*), this can be different for species that have emerged due to ecologically mediated divergent selection (*C. juno* – *C. metallicus*) and can even vary between species that have likely diverged in similar ways (*C. juno* – *C. metallicus* vs. *C. juno* – *C. ruber*).

CONCLUSION

In summary, by characterizing patterns of introgression in an empirical group of Australian vertebrates, our study provides a striking example of interspecific gene flow between taxa that have been ecologically and phylogenetically distinct for millions of years. Our results highlight that the emergence of complete reproductive isolation does not necessarily always scales with divergence time, that gene flow can persist between species shaped by ecologically mediated divergent selection and that more studies are required to address the importance of occasional gene flow between phylogenetically distinct taxa. Finally, given our observations of introgressive hybridization, future studies should examine whether these introgressed regions introduce novel genetic variation and have been essential in promoting adaptive

evolution, or whether they really just represent mere echoes of a former transient stage in the genomes of older species.

ACKNOWLEDGEMENTS

We thank Jason Bragg and Sally Potter for their ongoing support to the *Cryptoblepharus* project. We thank the museum curators and Paul Oliver for access to tissues and specimens, and Matthew Fujita for providing the annotated *Lygosoma sundevalli* mitochondrial genome. We also like to thank all the volunteers who helped out during the various expeditions across Northern-Australia. Finally, we'd like to thank Ana-Catarina Silva, Joshua Penalba, Sonal Singhal and Huw Ogilvie for helpful advice and discussions during analyses.

REFERENCES

Alves P.C., Melo-Ferreira J., Freitas H., Boursot P. 2008. The ubiquitous mountain hare mitochondria: multiple introgressive hybridization in hares, genus Lepus. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 363:2831–2839.

Anderson E. 1949. Introgressive hybridization. John Wiley and Sons, New York.

Arnold M.L. 1997. Natural Hybridization and Evolution. Oxford University Press, Oxford.

Blom M. 2015a. Habitat use and new locality records for Cryptoblepharus poecilopleurus (Squamata: Scincidae) from French Polynesia. Herpetol. Notes 8:579-582.

Blom M.P.K. 2015b. EAPhy: A Flexible Tool for High-throughput Quality Filtering of Exon-alignments and Data Processing for Phylogenetic Methods. PLoS Curr.:1–12; doi: 10.1371/currents.tol.75134257bd389c04bc1d26d42aa9089f.

Blom M.P.K., Horner P., Moritz C. 2016. Convergence across a continent: adaptive diversification in a recent radiation of Australian lizards. Proc. R. Soc. B. 283:20160181.

Blom M.P.K., Bragg J., Potter S., Moritz C. 2017. Accounting for uncertainty in gene tree estimation: Summary-coalescent species tree inference in a challenging radiation of Australian lizards. Syst. Biol. 66:352-366.

Bragg J.G., Potter S., Bi K., Moritz C. 2016. Exon capture phylogenomics: efficacy across scales of divergence. Mol. Ecol. Res. 16: 1059–1068.

Bi K., Vanderpool D., Singhal S., Linderoth T., Moritz C., Good J.M. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics 13:403.

Blackmon, H., Adams. R.A. 2015. EvobiR: Tools for comparative analyses and teaching evolutionary biology. doi:10.5281/zenodo.30938.

The Heliconius Genome Consortium G. 2013. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature. 487:94–98.

Darriba D., Taboada G.L., Doalla R. Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nature Methods. 9:772.

Dowling T.E., Secor C.L. 1997. The role of hybridization and introgression in the diversification of animals. Annu. Rev. Ecol. Evol. Syst. 28:593–619.

Earl D.A. & VonHoldt B.M. 2011. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv. Genet. Resources 2:359–361.

Edgar R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Frith, C.B. and Beehler, B.M. (1998) The Birds of Paradise. Oxford University Press, Oxford

Gavrilets S. 2003. Perspective: Models of speciation: What have we learned in 40 years? Evolution 57:2197-2215.

Glor R.E. 2010. Phylogenetic Insights on Adaptive Radiation. Annu. Rev. Ecol. Evol. Syst. 41:251–270.

Good J.M., Vanderpool D., Keeble S., Bi K. 2015. Negligible nuclear introgression despite complete mitochondrial capture between two species of chipmunks. Evolution. 69:1961–1972.

Grant B.R., Grant P.R. 2008. Fission and fusion of Darwin's finches populations. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 363:2821–2829.

Green R.E., Krause J., Briggs A.W., Maricic T., Stenzel U., Kircher M., et al. 2010. A Draft Sequence of the Neandertal Genome. Science. 328:710–722.

Hahn C., Bachmann L., Chevreux B. 2013. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads--a baiting and iterative mapping approach. Nucleic Acids Res. 41:e129–e129.

Hayashi F., Shima A., Horikoshi K., Kawakami K., Segawa R.D., Aotsuka T., Suzuki T. 2009. Limited overwater dispersal and genetic differentiation of the snake-eyed skink (Cryptoblepharus nigropunctatus) in the Oceanic Ogasawara Islands, Japan. Zool. Sci. 26:543–549.

Hedrick P.W. 2013. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. Mol Ecol. 22: 4606-18.

Horner P. 2007. Systematics of the snake-eyed skinks, Cryptoblepharus Wiegmann (Reptilia: Squamata: Scincidae)–an Australian based review. The Beagle Suppl. 3:21–198.

Horner P., Adams M. 2007. A Molecular-systematic Assessment of Species Boundaries in Australian Cryptoblepharus (Reptilia: Squamata: Scincidae): A Case Study for the Combined Use of Allozymes and Morphology to Explore Cryptic Biodiversity. The Beagle Suppl. 3:1–20.

Huang X., Madan A. 1999. CAP3: A DNA sequence assembly program. Genome Res. 9:868–877.

Huerta-Sánchez E., Jin X., Asan, Bianba Z., Peter B.M., Vinckenbosch N., Liang Y., *et al.* 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature. 512:194–7.

Irestedt M., Jønsson K.A., Fjeldså J., Christidis L., Ericson P.G. 2009. An unexpectedly long history of sexual selection in birds-of-paradise. BMC Evol Biol. 9:235–11.

Jakobsson M., Rosenberg N.A. 2007. CLUMMP: A clustering matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23: 1801-1806.

Kozak K.M., Wahlberg N., Neild A.F.E., Dasmahapatra K.K., Mallet J., Jiggins C.D. 2015. Multilocus Species Trees Show the Recent Adaptive Radiation of the Mimetic Heliconius Butterflies. Syst. Biol. 64:505–524.

Kronforst M.R. 2008. Gene flow persists millions of years after speciation in Heliconius butterflies. BMC Evol Biol. 8:98–8.

Langmead B., Salzberg S. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357-359.

Larson E.L., White T.A., Ross C.L., Harrison R.G. 2013. Gene flow and the maintenance of species boundaries. Mol. Ecol. 23:1668-1678.

Li H., Handsaker B., Wysoker A., Fennel T., Ruan J., Homer N. et al. 2009. The sequence alignment/Map format and SAMtools. Bioinformatics 25:2078-2079.

Lindtke D., Buerkle C.A. 2015. The genetic architecture of hybrid incompatibilities and their effect on barriers to introgression in secondary contact. Evolution. 69:1987-2004.

Mallet J. 2007. Hybrid speciation. Nature. 446:279–283.

Mayr E. 1963. Animal species and evolution. Belknap Press, Cambridge

McKay B.D., Zink R.M. 2015. Sisyphean evolution in Darwin's finches. Biol. Rev. 90:689–698.

McKenna A.H., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A. et al. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20:1297–1303.

Meyer M., Kicher M. (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protocol. 6, t5448. doi:10.1101/pdb.prot5448.

Nosil P. 2012. Ecological Speciation. Oxford University Press, Oxford.

Nater A., Burri R., Kawakami T., Smeds L., Ellegren H. 2015. Resolving Evolutionary Relationships in Closely Related Species with Whole-Genome Sequencing Data. Syst. Biol. 64:1000-1017.

Paun O., Turner B., Trucchi E., Munzinger J., Chase M.W., Samuel R. 2016. Processes Driving the Adaptive Radiation of a Tropical Tree (Diospyros, Ebenaceae) in New Caledonia, a Biodiversity Hotspot. Syst. Biol. 65:212–227.

Pereira R.J., Martínez-Solano I., Buckley D. 2016. Hybridization during altitudinal range shifts: nuclear introgression leads to extensive cyto-nuclear discordance in the fire salamander. Mol. Ecol. 25:1551–1565.

Poelstra J.W., Vijay N., Bossu C.M., Lantz H., Ryll B., Muller I., Baglione V., Unneberg P., Wikelski M., Grabherr M.G., Wolf J.B.W. 2014. The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. Science. 344:1410–1414.

Potter S., Bragg J.G., Peter B.M., Bi K., Moritz, C. 2016. Phylogenomics at the tips: Inferring lineages and their demographic history in a tropical lizard, Carlia amax. Mol. Ecol. 25: 1367-1380

Pritchard J.K., Stephens M., Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics. 155:945–959.

Racimo F., Sankararaman S., Nielsen R., Huerta-Sánchez E. 2015. Evidence for archaic adaptive introgression in humans. Nat. Rev. Genet. 16:359–371.

Rheindt F.E., Fujita M.K., Wilton P.R., Edwards S.V. 2014. Introgression and phenotypic assimilation in Zimmerius flycatchers (Tyrannidae): population genetic and phylogenetic inferences from genome-wide SNPs. Syst. Biol. 63:134–152.

Rocha S., Carretero M., Vences M., Glaw F. 2005. Deciphering patterns of transoceanic dispersal: the evolutionary origin and biogeography of coastal lizards (Cryptoblepharus) in the Western Indian Ocean region. J. Biogeogr. 33:13-22.

Rosenblum E.B., Sarver B.A.J., Brown J.W., Roches Des S., Hardwick K.M., Hether T.D., et al. 2012. Goldilocks Meets Santa Rosalia: An Ephemeral Speciation Model Explains Patterns of Diversification Across Time Scales. Evol Biol. 39:255–261.

Roux C., Fraisse C., Romiguier J., Anciaux Y., Galtier N., Bierne N. 2016. Shedding light on the grey zone of speciation along a continuum of genomic divergence. PLoS Biol. 14:e2000234.

Salichos L., Stamatakis A., Rokas A. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol. Biol. Evol. 31:1261–1271.

Schwenk K., Brede N., Streit B. 2008. Introduction. Extent, processes and evolutionary impact of interspecific hybridization in animals. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 363:2805–2811.

Seehausen O. 2004. Hybridization and adaptive radiation. Trends Ecol. Evol. 19:198–207.

Seehausen O., Butlin R.K., Keller I., Wagner C.E., Boughman J.W., Hohenlohe P.A. et al. 2014. Genomics and the origin of species. Nat. Rev. Genet. 15:176–192.

Singhal S. 2013. De novotranscriptomic analyses for non-model organisms: an evaluation of methods across a multi-species data set. Mol. Ecol. Res. 13:403–416.

Singhal S., Moritz C. 2013. Reproductive isolation between phylogeographic lineages scales with divergence. Proc. Roy. Soc. B. 280:20132246.

Sunnucks P., Hales DF. (1996) Numerous transposed sequences of mitochondrial cytochrome oxidase I-II in aphids of the genus Sitobion (Hemiptera: Aphididae). Mol. Biol. Evol. 13:510-524.

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

Smith J., Kronforst M.R. 2013. Do Heliconius butterfly species exchange mimicry alleles? Biol. Lett. 9:20130503.

Smith S.A., Sadlier R.A., Bauer A.M., Austin C.C., Jackman T. 2007. Molecular phylogeny of the scincid lizards of New Caledonia and adjacent areas: evidence for a single origin of the endemic skinks of Tasmantis. Mol. Phylogenet. Evol. 43:1151–1166.

Sousa V., Hey J. 2013. Understanding the origin of species with genome-scale data: modelling gene flow. Nat. Rev. Genet. 14:404–414.

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

Stuglik M.T., Babik W. 2016. Genomic heterogeneity of historical gene flow between two species of newts inferred from transcriptome data. Ecol Evol. 6:4513–4525.

Toews D.P.L., Brelsford A. 2012. The biogeography of mitochondrial and nuclear discordance in animals. Mol Ecol. 21:3907–3930.

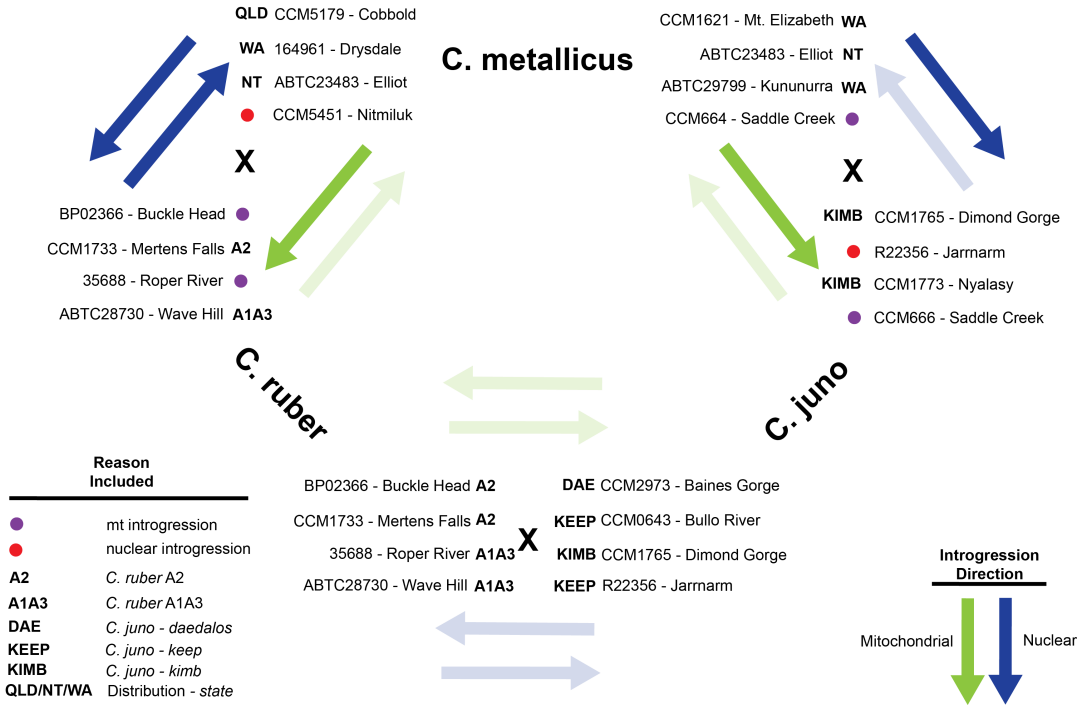
Than C., Ruths D., Nakhleh L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC informatics 9:322.

Twyford A.D., Ennos R.A. 2011. Next-generation hybridization and introgression. Heredity. 108:179–189.

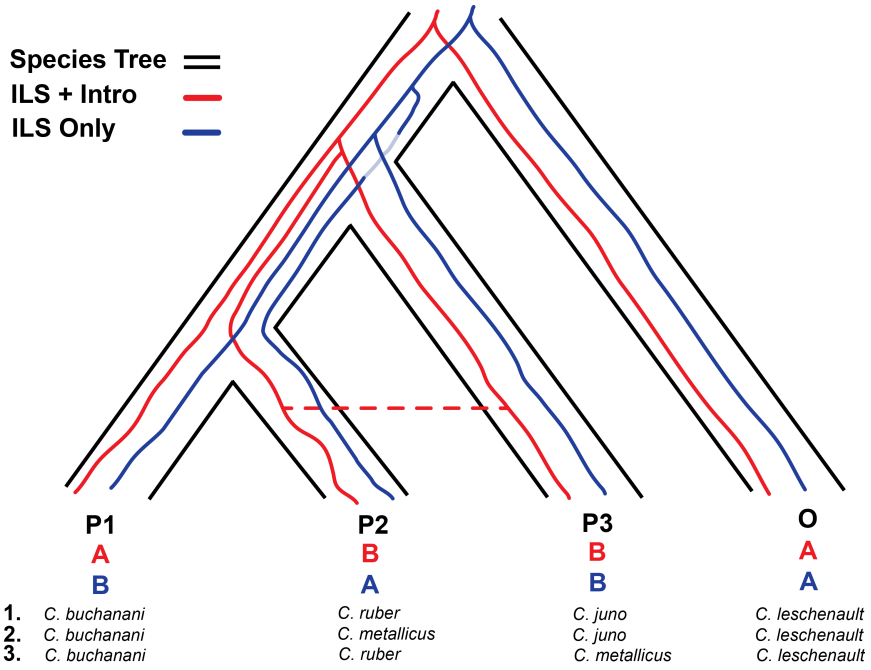
Yu Y., Dong J., Liu K.J., Nakhleh L. 2014. Maximum likelihood inference of reticulate evolutionary histories. Proc. Natl. Acad. Sci. U.S.A. 111:16448–16453.

Zerbino D.R., Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821–829.

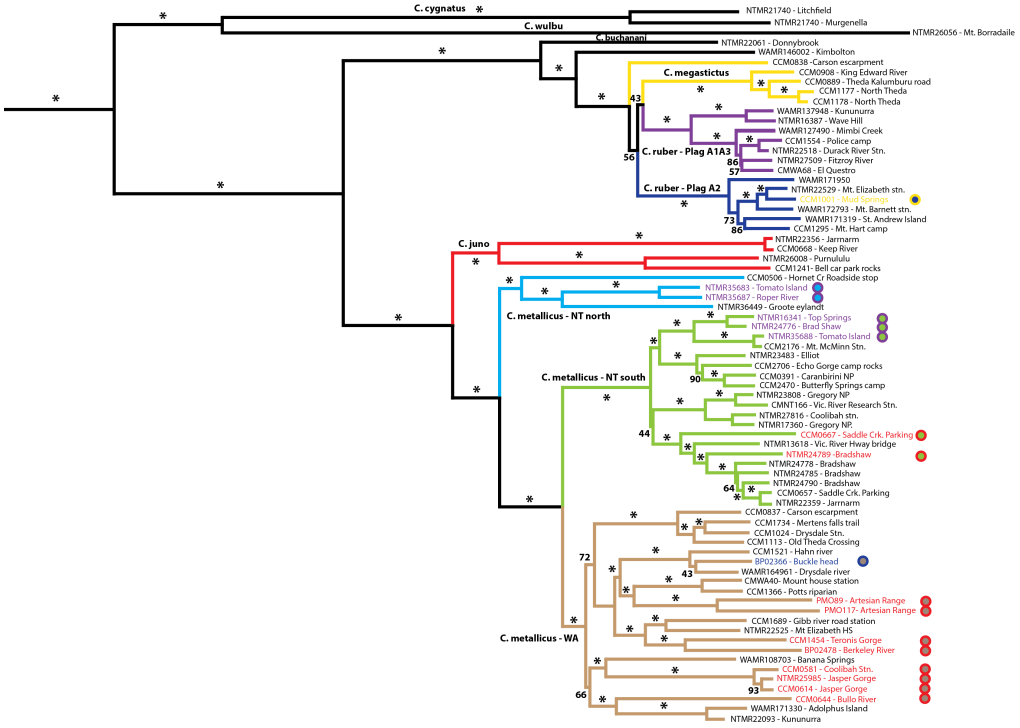
Zhang W., Dasmahapatra K.K., Mallet J., Moreira G.R.P., Kronforst M.R. 2016. Genome-wide introgression among distantly related Heliconius butterfly species. Genome Biol. 17:1–15.



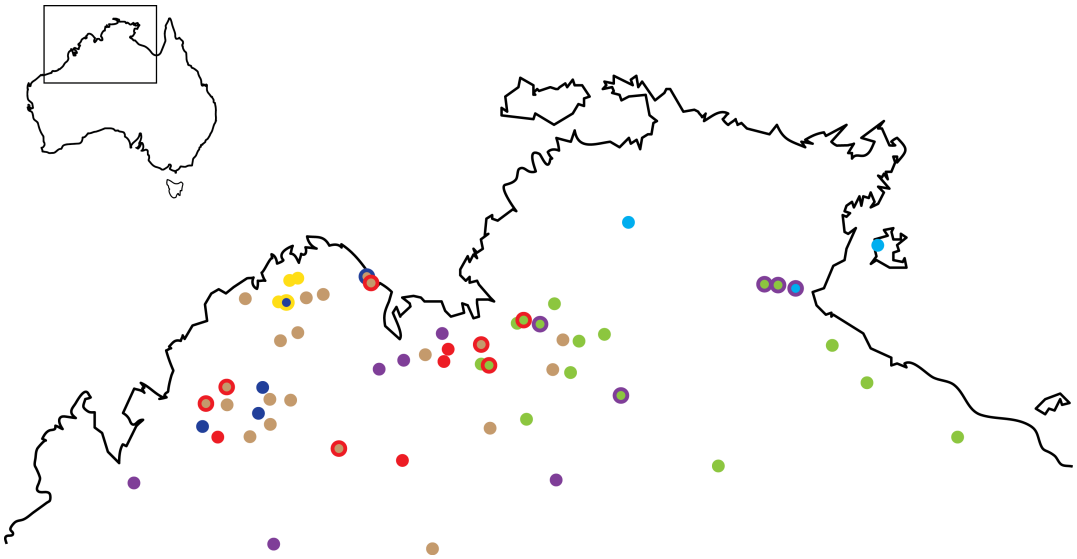
Supp. Fig. 1. Schematic overview of the species pair comparisons and the directionality of gene flow for both the mitochondrial and nuclear genome. Individuals were chosen based on the STRUCTURE plots, evidence of cytonuclear discord or biogeographic overlap between populations, to evaluate potential nuclear signatures of past introgression.



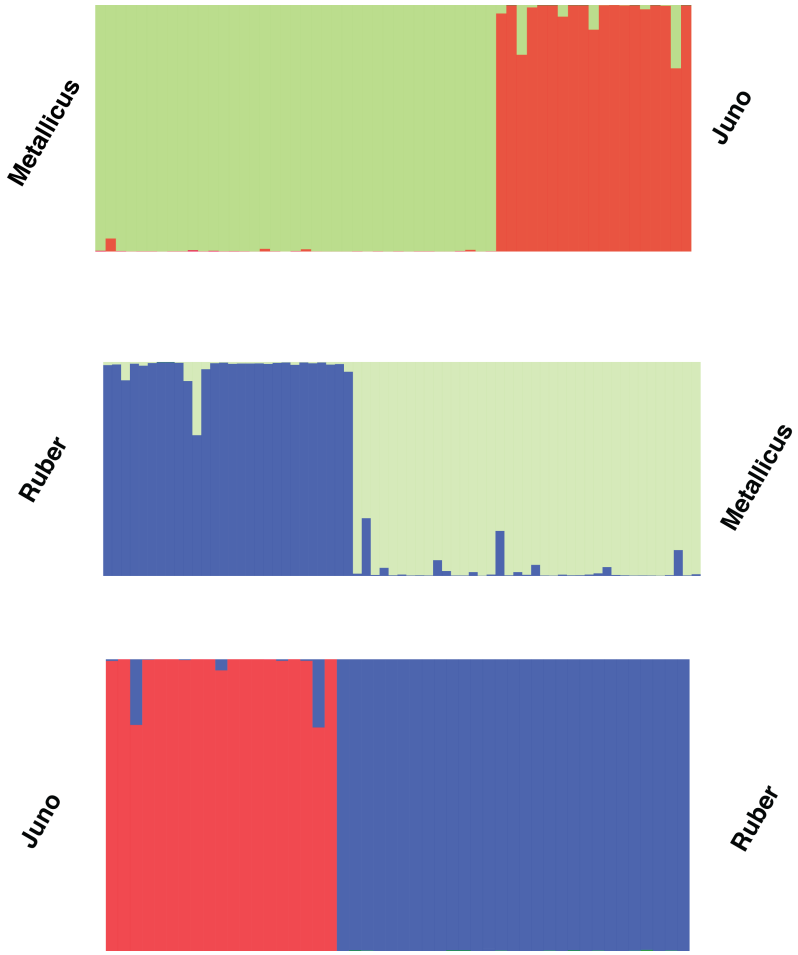
Supp. Fig. 2. The three possible topologies for the four taxon trees as inferred for each combination of individuals per species pair comparison. A high proportion of ABBA topologies would suggest that individual P2 and P3 have an excess of shared haplotype variation, potentially due to past introgression.



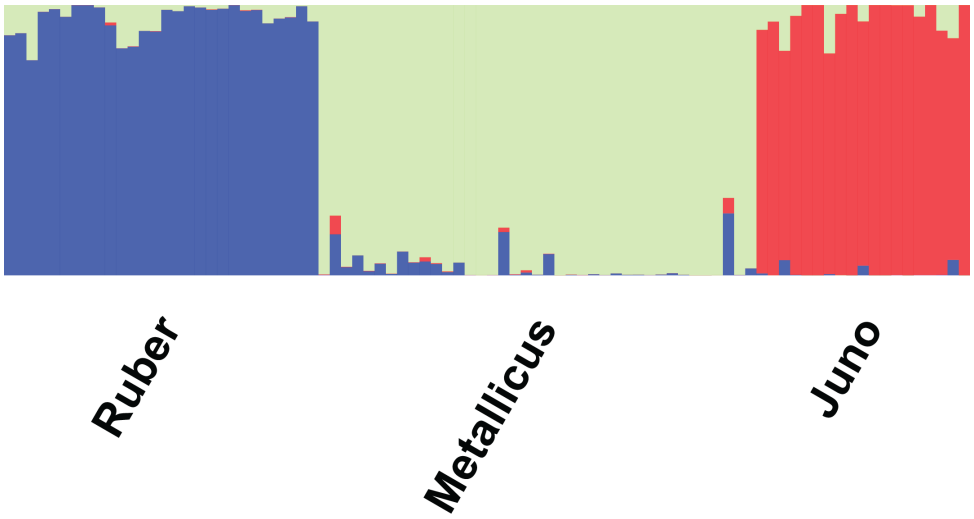
Supp. Fig. 3. Mitogenome phylogeny as inferred with a maximum likelihood approach. Branches are colored by respective lineage and match with the colour scheme used in Supp. figure 4. Individuals with a discordant combination of mitonuclar haplotypes are widespread across the tree and incongruence is geographically structured (see Supp. Fig. 4)



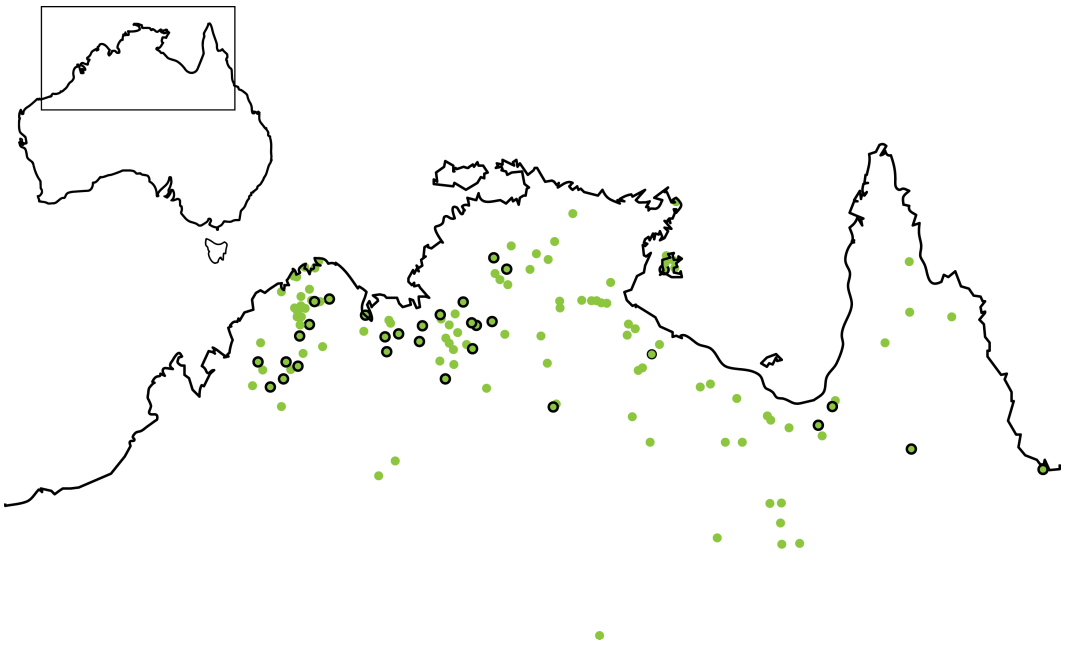
Supp. Fig. 4. A geographic map with the sampling localities and the genetic lineages based on both the nuclear and mitochondrial data. Cytonuclear discordance often tend to be locally structured.



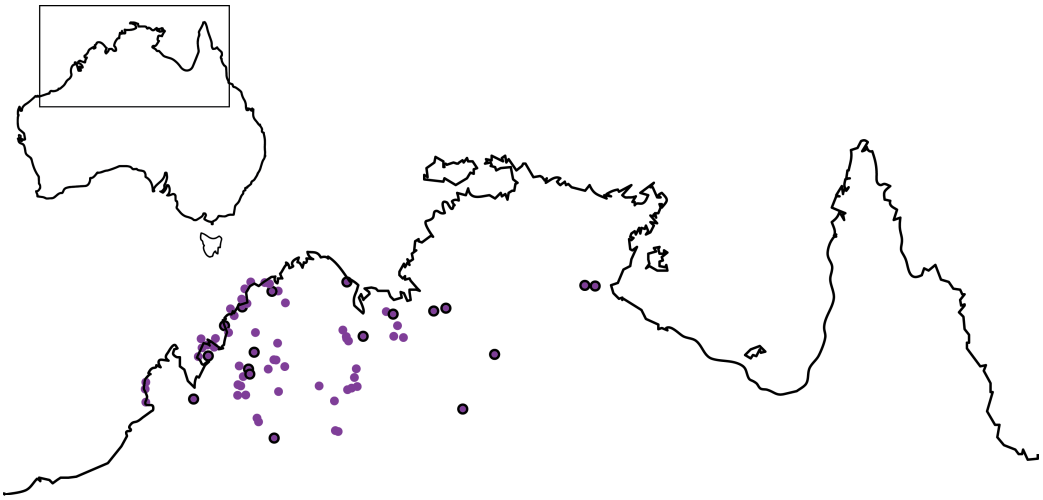
Supp. Fig. 5. STRUCTURE plots based on all individuals sequenced for each species comparison. In contrary to the STRUCTURE plots in the main text, here wo only sampled a single marker per exon.



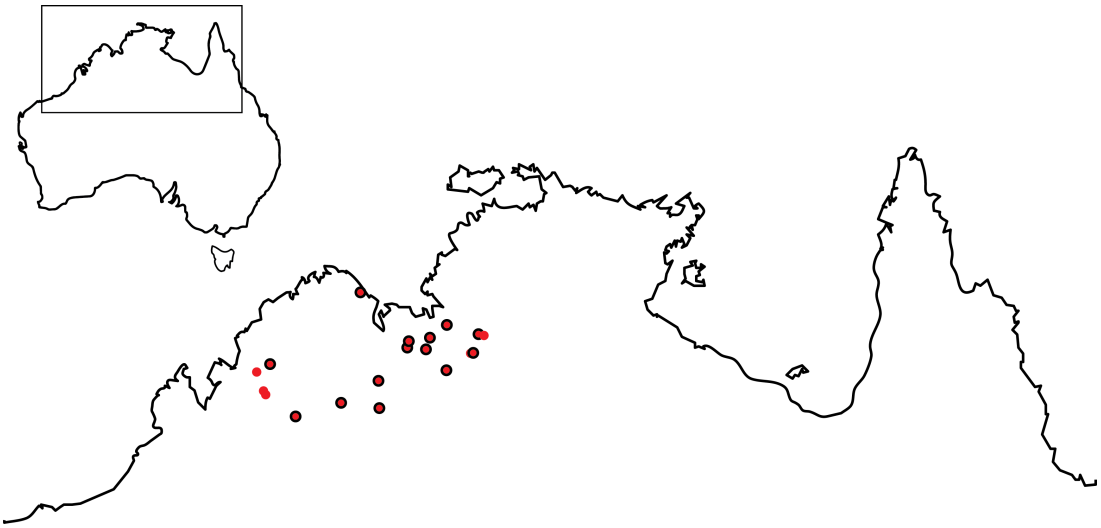
Supp. Fig. 6. Combined STRUCTURE analysis for all individuals across the three focal taxa.



Supp. Fig. 7. Distribution map with all *C. metallicus* individuals that have been sequenced (either ND2, sequence capture or both)



Supp. Fig. 8. Distribution map with all *C. ruber* individuals that have been sequenced (either ND2, sequence capture or both.



Supp. Fig. 9. Distribution map with all *C. juno* individuals that have been sequenced (either ND2, sequence capture or both.

Supplementary Table 2. Summary statistics for multiple introgression tests, across species comparisons.

Comparison	Individual combination focal species comparison	Δ lk score 0 - 1 reticulation events	Total number of gene trees with TC score above 0.4	Number of SNV ABBA-BABA informative - ALL	Patterson's D - ALL	Z-score - ALL	Number of SNV ABBA-BABA informative - ONE snp per gene	Patterson's D - ONE snp per gene	Z-score - ONE SNP per gene
C. juno - C. metallicus	Diamond Gorge (CCM175) - Mt elizabeth (CCM1621)	-38,5188404	550	2804	0,2098592	8,186672	1283	0,163233	4,206412
C. juno - C. metallicus	Diamond Gorge (CCM175) - Elliot (ABTC23483)	-45,8143301	535	2561	0,2458234	8,690964	1185	0,25	6,433665
C. juno - C. metallicus	Diamond Gorge (CCM175) - Kununurra (ABTC29399)	-34,3670784	551	2821	0,2415254	9,389068	1260	0,1887676	4,875325
C. juno - C. metallicus	Diamond Gorge (CCM175) - Saddle Crk (CCM664)	-40,3555552	564	2832	0,2215321	8,259158	1273	0,1993958	5,155277
C. juno - C. metallicus	Jarrnarm (ABTC22356) - Mt elizabeth (CCM1621)	-11,7598642	492	2127	0,2003439	7,073664	1070	0,3191489	8,128098
C. juno - C. metallicus	Jarrnarm (ABTC22356) - Elliot (ABTC23483)	-21,7110613	478	2014	0,2171482	7,332313	1019	0,269962	6,28724
C. juno - C. metallicus	Jarrnarm (ABTC22356) - Kununurra (ABTC29399)	-15,4943327	487	2151	0,2242114	7,728154	1064	0,215971	5,123606
C. juno - C. metallicus	Jarrnarm (ABTC22356) - Saddle Crk (CCM664)	-12,0719601	474	2122	0,1918506	6,547718	1050	0,1819757	4,48025
C. juno - C. metallicus	Nyalasy (CCM1773) - Mt elizabeth (CCM1621)	-48,2227286	560	2840	0,2077562	8,240437	1309	0,1568047	4,237721
C. juno - C. metallicus	Nyalasy (CCM1773) - Elliot (ABTC23483)	-42,3027018	539	2589	0,2317361	8,591493	1194	0,1914894	4,562526
C. juno - C. metallicus	Nyalasy (CCM1773) - Kununurra (ABTC29399)	-49,6772975	559	2873	0,2284334	8,865468	1283	0,2480499	6,415581
C. juno - C. metallicus	Nyalasy (CCM1773) - Saddle Crk (CCM664)	-44,6837129	568	2877	0,2111337	7,907907	1288	0,1679274	4,428653
C. juno - C. metallicus	Saddle Creek (CCM666) - Mt elizabeth (CCM1621)	-36,0737704	564	2834	0,2127512	8,065652	1296	0,1993958	5,340118
C. juno - C. metallicus	Saddle Creek (CCM666) - Elliot (ABTC23483)	-31,0110542	547	2575	0,2133228	7,654567	1199	0,1336717	3,320131
C. juno - C. metallicus	Saddle Creek (CCM666) - Kununurra (ABTC29399)	-31,8879043	558	2840	0,2237762	8,661314	1270	0,1671827	4,143502
C. juno - C. metallicus	Saddle Creek (CCM666) - Saddle Crk (CCM664)	-32,9217996	584	2879	0,2069892	8,150722	1286	0,2411765	6,425019
C. ruber - C. metallicus	Buckle head (BP02366) - Cobbold (CCM5179)	-2,206927391	593	2977	0,08182784	2,520971	1260	0,04449649	0,925882
C. ruber - C. metallicus	Buckle head (BP02366) - Drysdale (164961)	-9,00618416	543	2582	0,08170311	2,376128	1168	0,1187335	2,256004
C. ruber - C. metallicus	Buckle head (BP02366) - Elliot (ABTC23483)	-5,33916521	555	2571	0,09237875	2,80989	1161	0,01	0,1950576
C. ruber - C. metallicus	Buckle head (BP02366) - Nitmiluk (CCM5451)	-5,43759295	591	2849	0,06818182	2,133265	1266	0,135255	2,835311
C. ruber - C. metallicus	Mertens Falls (CCM1733) - Cobbold (CCM5179)	-9,182918684	588	2886	0,06930693	2,127924	1246	0,09047619	1,83242
C. ruber - C. metallicus	Mertens Falls (CCM1733) - Drysdale (164961)	-9,621745428	556	2522	0,1176471	3,503175	1144	0,1767554	3,533812
C. ruber - C. metallicus	Mertens Falls (CCM1733) - Elliot (ABTC23483)	-11,35249876	552	2487	0,1398104	4,148362	1131	0,1498771	2,94622
C. ruber - C. metallicus	Mertens Falls (CCM1733) - Nitmiluk (CCM5451)	-3,492735257	593	2810	0,05857741	1,765443	1246	0,1121718	2,321357
C. ruber - C. metallicus	Roper River (35688) - Cobbold (CCM5179)	-13,01703352	482	2070	0,1307902	3,541989	999	0,09433962	1,871053
C. ruber - C. metallicus	Roper River (35688) - Drysdale (164961)	-22,41339806	456	1862	0,2039106	5,350353	932	0,1506024	2,739825
C. ruber - C. metallicus	Roper River (35688) - Elliot (ABTC23483)	-24,0803855	458	1824	0,1601732	4,24406	938	0,09392265	1,842809
C. ruber - C. metallicus	Roper River (35688) - Nitmiluk (CCM5451)	-11,70376018	476	2044	0,1001335	2,716338	1001	0,05291005	0,9954603
C. ruber - C. metallicus	Wave Hill (ABTC28730) - Cobbold (CCM5179)	-2,944797452	637	2902	0,09529025	2,826045	1258	0,0821256	1,68226

Comparison	Individual combination focal species comparison	Δ lk score 0 - 1 reticulation events	Total number of gene trees with TC score above 0.4	Number of SNV ABBA-BABA informative - ALL	Patterson's D - ALL	Z-score - ALL	Number of SNV ABBA-BABA informative - ONE snp per gene	Patterson's D - ONE snp per gene	Z-score - ONE SNP per gene
C. ruber - C. metallicus	Wave Hill (ABTC28730) - Drysdale (164961)	-7,84923472	556	2500	0,1284514	3,782186	1139	0,07936508	1,559877
C. ruber - C. metallicus	Wave Hill (ABTC28730) - Elliot (ABTC23483)	-5,446325277	566	2520	0,1097852	3,192407	1149	0,1859838	3,743097
C. ruber - C. metallicus	Wave Hill (ABTC28730) - Nitmiluk (CCM5451)	-5,189574459	599	2837	0,1131313	3,552557	1256	0,2	4,263563
C. ruber - C. juno	Buckle head (BP02366) - Baines Gorge (CCM2973)	-9,689115581	727	3670	0,07358263	2,055756	1437	0,06666667	1,27416
C. ruber - C. juno	Buckle head (BP02366) - Bullo River (CCM0643)	-1,877913513	741	3603	0,05972046	1,754245	1421	0,03921569	0,6734475
C. ruber - C. juno	Buckle head (BP02366) - Diamond Gorge (CCM1765)	-0,801820473	731	3668	0,06166868	1,792116	1428	0,130719	2,291581
C. ruber - C. juno	Buckle head (BP02366) - Jarrnarm (R22356)	-5,348314336	607	2761	0,08802309	2,312187	1188	0,1827957	3,099798
C. ruber - C. juno	Mertens Falls (CCM1733) - Baines Gorge (CCM2973)	-6,569791996	718	3575	0,06516291	1,884205	1410	0,06626506	1,188473
C. ruber - C. juno	Mertens Falls (CCM1733) - Bullo River (CCM0643)	-4,116272341	722	3531	0,02077922	0,5705425	1414	0,02970297	0,5108144
C. ruber - C. juno	Mertens Falls (CCM1733) - Diamond Gorge (CCM1765	-10,14504917	711	3566	0,06683168	1,905521	1401	0,02298851	0,4271236
C. ruber - C. juno	Mertens Falls (CCM1733) - Jarrnarm (R22356)	-6,879437766	560	2692	0,08052709	1,992495	1163	0,0915493	1,522641
C. ruber - C. juno	Roper River (35688) - Baines Gorge (CCM2973)	-11,18942625	598	2648	0,1355422	3,46142	1162	0,1619048	2,840497
C. ruber - C. juno	Roper River (35688) - Bullo River (CCM0643)	-13,3455068	593	2584	0,08737864	2,151285	1151	0,1304348	2,354748
C. ruber - C. juno	Roper River (35688) - Diamond Gorge (CCM1765)	-7,362855787	584	2616	0,1063174	2,759855	1148	0,1313131	2,220835
C. ruber - C. juno	Roper River (35688) - Jarrnarm (R22356)	-17,57160324	488	2057	0,1423423	3,411742	991	0,1740891	2,672408
C. ruber - C. juno	Wave Hill (ABTC28730) - Baines Gorge (CCM2973)	-3,169444676	674	3605	0,06765068	1,937262	1413	0,07886435	1,442817
C. ruber - C. juno	Wave Hill (ABTC28730) - Bullo River (CCM0643)	-4,858599576	694	3554	0,03084833	0,8747928	1421	0,02769231	0,4872299
C. ruber - C. juno	Wave Hill (ABTC28730) - Diamond Gorge (CCM1765)	-1,313132839	702	3589	0,05131414	1,485822	1408	-0,01607717	0,2807485
C. ruber - C. juno	Wave Hill (ABTC28730) - Jarrnarm (R22356)	-7,932412499	578	2681	0,1007519	2,596054	1162	0,06312292	1,061851

CHAPTER 4

When Ecology shapes Biogeography: Habitat Preference Modulates Dispersal Probability in *Cryptoblepharus* lizards

When Ecology shapes Biogeography: Habitat Preference Modulates Dispersal Probability in *Cryptoblepharus* lizards

Mozes P.K. Blom[1], Nicholas J. Matzke[1], Jason G. Bragg[1], Evy Arida[2], Christopher C. Austin[3], Adam Backlin[4], Robert N. Fisher[5], Stacie Hathaway[5], Djoko T. Iskandar[6], Jimmy A. McGuire[7], Sean Reilly[7], Eric N. Rittmeyer[1,3], Sara Rocha[8], Alexander Stubbs[7], & Craig Moritz[1]

[1]*Research School of Biology, The Australian National University, Canberra, Australia*

[2]*Research Center for Biology, The Indonesian Institute of Sciences, Cibinong, Indonesia*

[3]*Museum of Natural Science, Louisiana State University, Baton Rouge, USA*

[4]*U.S. Geological Survey, Western Ecological Research Center, Santa Ana, USA*

[5]*U.S. Geological Survey, Western Ecological Research Center, San Diego, USA*

[6]*School of Life Sciences and Technology, Institut Teknologi, Bandung, Indonesia*

[7]*Museum of Vertebrate Zoology, University of California Berkeley, Berkeley, USA*

[8]*Department of Biochemistry, Genetics and Immunology, University of Vigo, Vigo, Portugal*

ABSTRACT

The relative importance of vicariance and dispersal in shaping geographic range evolution has been the center of debate among biogeographers for decades. However, this debate has largely focused on dating results and as a consequence, possible predictors of long-distance dispersal have been understudied. An important example is the question whether dispersal probability is dependent on the ecology of species? Here we combine a genomic dataset suitable for phylogenomic inference with probabilistic modeling of biogeographic history and ask whether habitat preference explains the current widespread distribution of a recent squamate radiation. We find that a habitat-dependent dispersal model is significantly better supported than any

other biogeographic model. Lizards of the genus *Cryptoblepharus* that occur in a littoral habitat are 6 to 8 times more likely to disperse than non-littoral species, and this has ultimately resulted in an extraordinarily widespread distribution that includes multiple continents and isolated island archipelagoes. While species that occur in the Indo-Australian region are relatively distinct, the littoral species are more closely related and have spread across the Pacific and Indian-Ocean within the past 3 million years. As such, our empirical study demonstrates that natural dispersal is a potent mechanism for widespread range evolution, but moreover that ecology needs to be accounted for since it can significantly modulate the probability of dispersal.

INTRODUCTION

The relative importance of dispersal and vicariance in shaping the biogeographic distributions of modern day taxa has been the focus of debate among biogeographers for many decades. From the outset of evolutionary biology, Darwin and Wallace have both argued that long-distance dispersal can be an important component in shaping the distribution of many living species (Darwin 1859; Wallace 1869). Following the confirmation of continental drift in the mid 20th century, the study of biogeography was strongly influenced by proponents of vicariance, who suggested that most widely distributed taxa were already present at the time of continental break-up and that subsequent long-distance dispersal has only played a minor role (Wiley 1988). However, a growing number of recent molecular phylogenetic studies suggest that continental break up predates the divergence times of many widespread clades and that long distance dispersal is therefore a more parsimonious explanation (e.g. Vences et al. 2003; Longrich et al. 2015).

Debates about long-distance dispersal have largely focused on dating results and the difficulty of studying very rare "chance" dispersal events. As a result, possible predictors of long-distance dispersal have been understudied. An important example is the question whether dispersal probability is dependent on the ecology of species. Many biogeographers have proposed that range evolution can be conditional on specific traits or habitat affinities (Wilson 1961; Carstensen et al. 2013; Sukumaran et al. 2016), but empirical examples where the relationship between ecology and dispersal have been described in detail remain scarce. Quantifying such relationships would generate major insight on organismic prerequisites that have induced macroevolutionary change because geography and species diversification are often intimately linked (Coyne and Orr 2004). Furthermore, ecology-dependent dispersal could be a major extension to the theory of island biogeography since island

colonization would no longer be the sole product of island size and distance to a nearest source population (Whittaker 1998). Under an ecology-dependent dispersal model, the colonization probability of a given island would also strongly depend on the focal taxa in question and the dispersal probability of such taxa given its ecological state. Understanding whether range evolution can depend on the ecology of species is therefore a major objective with potentially important implications for our understanding of species diversification.

Although biogeographers have long been aware of its potential significance, the study of ecology-based dispersal has long been hampered by the lack of an appropriate statistical framework to quantify such processes (Sukumaran et al. 2016). However, the field of historical biogeography has advanced markedly during recent years with the introduction of probabilistic model-based biogeographic inference (Ree 2005; Ree and Smith 2008; Ree and Sanmartín 2009; Matzke 2013). Likelihood-based methods such as BioGeoBEARS (Matzke 2014) infer the likelihood of data under explicit biogeographic models and can simultaneously estimate parameter values of interest. Alternatively, a recently introduced likelihood-free method uses a machine learning algorithm to optimize the simulation of phylogenies and then uses summary statistics to compare the simulated with the observed data (Sukumaran et al. 2016). With both approaches, the relative fit of alternative biogeographic models to the data can be quantified, and models that vary in the number of free parameters compared, to ultimately identify the model that best fits the data. Probabilistic modeling of geographic range evolution enables empiricists to evaluate explicit biogeographic models using standard model choice procedures such as AIC comparison or likelihood ratio tests (Matzke 2013). Here we extend these biogeographic models, to make the probability of dispersal dependent on a trait; such as the ecological state of a species. Sukumaran's likelihood-free approach can also specifically model trait-dependent dispersal, but here we use a likelihood based approach and calculate the likelihood of

the data under each specific scenario. We use this trait dependent dispersal model to specifically examine the biogeographic history of a diverse and relatively young squamate genus with a widespread distribution that includes multiple continents and island archipelagoes across the Pacific and Indian-Ocean.

Skinks of the genus *Cryptoblepharus* are small diurnal lizards that are globally widespread and occur in a variety of different habitats (Horner 2007; Blom et al. 2016). A recent analysis of the evolutionary history of Australian *Cryptoblepharus* highlighted the important role of habitat specialization in promoting a continental radiation (Blom et al. 2016). The Australian species diversified in two distinct clades since the late Miocene/early Pliocene, resulting in over 25 species and the repeated emergence of distinct habitat specialists ('ecomorphs') across the continent. However, this is in stark contrast with most species of *Cryptoblepharus* that extend beyond the Indo-Australian region and typically occur in coastal habitat (Horner 2007; Blom 2015a). Some species are true littoral specialists and inhabit crevices within tidal rocks next to crashing waves, whereas others also occur in mesic habitat adjacent to the beach. Nonetheless, these littoral species mostly occur in close proximity to the coast and have not adapted to habitats further inland to a similar extent as their Australian counterparts.

The aim of the current study is to examine the biogeographic history of the complete *Cryptoblepharus* genus and to characterize the underlying mechanisms that have shaped the extraordinary distribution of this relatively young clade of small stationary lizards. We sampled representatives of all described and proposed taxa across their global distribution, used a sequence-capture approach to generate a large multi-locus dataset and employed a coalescent-based method to infer the relationships among taxa. We used 1196 loci to fit branch lengths and time-calibrated the phylogeny with an empirically obtained molecular clock. This holistic approach

yields a phylogeny that is the basis for study of geographic range evolution, the probability of natural trans-oceanic dispersal and the importance of ecology for dispersal probability. Specifically, we ask whether the evolution of the littoral habit explains the current widespread distribution that includes multiple continents and isolated island archipelagoes.

MATERIAL AND METHODS

Taxon Sampling

In an extensive revision of the genus, Horner (2007) identified 62 *Cryptoblepharus* species globally of which 19 were previously undescribed. Furthermore, due to the high degree of cryptic diversity in the Australian radiation, the true species diversity is likely still an underestimation (Horner 2007). We therefore used an opportunistic approach and initially sampled as many *Cryptoblepharus* populations as possible. Lizards were either sampled from museum collections (frozen or in alcohol: South Australian Museum, MVZ Berkeley, LSU Museum of Natural History) or during recent field expeditions by coauthors (Supp. Table 1). We first sequenced a single mitochondrial marker (ND2) or used sequence data that was already available (Rocha et al. 2005), to discern the major phylogeographic lineages. We then selected two individuals for a detailed screen of nuclear diversity using exon-capture from each i) described species (Horner 2007), ii) distinct mitochondrial lineage and iii) biogeographic region. In example, if lineages seemed relatively closely related but occurred on island archipelagoes that are more than 1000 km. apart, we still included multiple representatives for each lineage. In contrast to Horner (2007), we regarded the Malagasy lineages of *Cryptoblepharus* as a polytypic species complex for the

purpose of the current study. The species status of the Malagasy *Cryptoblepharus* lineages remains unclear, with Brygoo (1986) only recognizing 13 forms (rather than species) and Horner (2007) delineating 13 distinct species of which 12 were distinguished based on 2 or more morphological differences. Yet, Rocha et al. (2005) highlighted that mitochondrial diversity between forms was minimal (<2% variation between the most distinct haplotype clades) and concluded that subspecific status of the lineages is more appropriate. We therefore included multiple representatives for each of the most distinct haplotype clades, rather than for each taxon.

The present study builds on a recent phylogenomic analysis of Australian *Cryptoblepharus* (Blom et al 2017). We used a similar exon capture approach for the Australian taxa and therefore included one individual for each Australian species from these initial experiments. Furthermore, our detailed analysis of introgression in a specific species group (Chapter 3) further characterized lineage diversity within *C. ruber* and *C. juno*, and identified putative new species. These were also included in the present study (*C. juno – Kimb*, *C. juno – Keep* and *C. ruber – PlagA2*). Lastly, two individuals from a closely related genus of skinks, *Emoia*, were included as outgroup species (Brandley et al. 2011).

ND2 Sanger Sequencing & Illumina Sequencing of Exon-Capture Samples

We extracted genomic DNA using the salting-out method of Sunnucks and Hales (1996) and amplified part of the mitochondrial gene encoding for NADH dehydrogenase (ND2) as described in (Ch.III). We visually inspected each sample and manually edited sequences in GENEIOUS (v.6.1). Our main goal for sequencing ND2, was to characterize phylogenetic diversity and identify potential cryptic lineages. We therefore only inferred a maximum-likelihood tree using RAxML (v.8.0**;** (Stamatakis

2014)), assuming a GTR + Γ model of sequence evolution and 1000 bootstrapped trees to estimate bipartition support across bootstrap replicates.

We then selected 91 individuals and generated individually bar-coded genomic libraries, suitable for exon-capture and subsequent Illumina sequencing. We used a modified version of our original Eugongylus group skink capture (design detailed in Bragg et al. 2016) that excluded target regions which were not consistently recovered and was partially designed using a *Cryptoblepharus ruber* transcriptome (CMWA61; El Questro WA). In brief, capture targets were based on exons with a balanced base composition in the *Anolis* genome and exons were only selected if corresponding orthologs were identified in the transcriptomes of three species from genera related to *Cryptoblepharus* (*Carlia rubrigularis, Lampropholis coggeri* and *Saproscincus basiliscus*; (Singhal 2013)) and the aforementioned *C. ruber* transcriptome. We targeted a total of 2920 exons and our capture probe set was synthesized by Roche NimbleGen in a SeqCap EZ Developer Library.

Genomic libraries were prepared with ~1400 ng. input DNA per sample and according to the protocol of Meyer and Kircher (2010), using modifications of Bi et al. (2012). A detailed description of the library preparation protocol can be found in (Bragg et al. 2016; Potter et al. 2016; Blom et al. 2017) but in brief, our protocol entailed blunt-end repair, adapter ligation, adapter fill-in and was followed by two index-PCRs to reduce PCR bias. We assessed DNA concentrations using a Nanodrop (Thermo Scientific) and the distribution of fragment lengths on 1.5% agarose gels. Barcoded libraries were pooled in equimolar ratios prior to hybridization and the exon-capture hybridization was performed following the SeqCap EZ Developer Library user guide (Roche Nimblegen). We assessed the quality of the hybridization as specified in (Blom et al 2017) and genomic libraries were sequenced (100 bp. paired

end) at the ACRF Biomolecular Resource Facility (Australian National University) on a single Illumina HiSeq 2500 lane.

Alignment and Data Pre-processing for Phylogenetic Inference

Each library was processed and assembled as outlined in detail previously (Bragg et al. 2016; Potter et al. 2016; Blom et al 2017). Once we had assembled contigs, we used these contigs as a reference and mapped cleaned reads back for each individual. Mapping was performed using BOWTIE2 (v.2.2; Langmead and Salzberg 2012) and resulting SAM files processed with SAMTOOLS (v.0.1.19; Li et al. 2009). We employed GATK (McKenna et al. 2010) to identify heterozygous sites, mask sites with a low quality genotype call (GQ<20) and used read-backed phasing that takes the physical linkage between sites within individual sequencing reads into account to generate phased haplotypes.

We assessed sequencing and assembly success by calculating the total number of exons recovered, the proportion of missing data (number of missing sites relative to total assembly length) and proportion of heterozygous sites for each individual. Due to the high efficacy of our exon-capture approach across divergent species (Bragg et al. 2016), we were able to enforce strict limits on library quality and still retain a relatively high proportion of individuals. For downstream analyses we only used individuals with more than 1000 exons sequenced and less than 20% missing data. Furthermore, we removed the top 5% of individuals with the highest proportion of heterozygous sites to avoid inclusion of libraries that potentially contain cross-sample contamination.

We used a bioinformatics workflow, EAPhy (v1.0; Blom 2015b), that uses Muscle (v3.6.1, Edgar 2004) for alignment and subsequently filters exonic sequences. For each gene, each individual with a contig over 150 bp's in length was included and

alignments were filtered based on the number of stop codons (exons with more than one stop codon excluded) and the degree of missing data (alignment columns were removed with more than 10% missing data). Furthermore, alignments that contained individual sequences where more than three codons in a seven-codon window differed from the alignment consensus were also removed. EAPhy generated both a concatenated alignment of all genes and alignments for each individual gene.

Phylogenetic Inference and Branch Fitting

We characterized the major phylogenetic lineages across all *Cryptoblepharus* with a maximum-likelihood inference of our concatenated dataset (1196 loci; 537,674 bp). We used RAxML (v.8.1; Stamatakis 2014) to estimate the most likely tree out of ten tree search replicates and subsequently generated 100 bootstrap replicates to quantify bipartition support. A GTR + Γ substitution model was employed to account for heterogeneity in substitution rates across the concatenated alignment. This exploratory analysis confirmed our prior conjecture that the current taxonomy is likely incomplete and we therefore chose two representatives from each species and/or major well-supported monophyletic lineage (i.e. divergence time > 1 Mya between terminal branches in dated phylogeny) for subsequent analyses. Future studies should examine whether these taxa represent unique species or mere phylogeographic diversity (often a challenging question for island representatives), but here we mainly focus on the biogeographic patterns between these distinct units. We repeated the alignment and alignment filtering process as described before, for the reduced dataset that only included two representatives for each major lineage.

In order to generate a time-calibrated phylogeny, we first inferred the species tree topology using a summary-coalescent approach and then fitted branch lengths a-posteriori using our concatenated alignment (Blom et al. 2016). We specifically

employed a coalescent-based species tree estimation method since our previous analyses of the *Cryptoblepharus* genus outlined an evolutionary history that included periods of rapid diversification (Blom et al. 2017). Instead of a two-step approach as presented here, the joint estimation of topology and branch lengths would be preferred (e.g. Ogilvie et al. 2016), but is currently insurmountable in a coalescent framework with NGS datasets that include this many taxa and loci.

We used a recently updated version of ASTRAL II (v.4.8; Mirarab and Warnow 2015) to infer the summary-coalescent species tree, where we assigned the two representatives of each major lineage as members of the same species. Whereas full-coalescent approaches such as *BEAST (Ogilvie et al. 2016) infer gene- and species tree simultaneously, summary-coalescent methods require gene trees to be inferred beforehand with alternative approaches. We first used JModelTest (v.2.1.0; Darriba et al. 2012) to identify the substitution model with the best fit for each individual locus and then used RAxML to infer the most likely gene tree out of ten replicates. To minimize the probability of erroneous inference due to missing data, we limited the number (two) of possible missing haplotypes for each alignment. We subsequently characterized differences in gene tree resolution between loci, by calculating the TC value for each locus using the most likely gene tree and 100 bootstrap replicates. In a previous study (Blom et al 2017), we showed that the inclusion of loci with a poor resolution does not improve the topological estimate of the species tree and we therefore used ASTRAL II on a dataset that included the 600 loci with the highest TC scores. We employed multi-locus bootstrapping to quantify bootstrap support for each bipartition in the resulting species tree topology. Finally, we used BEAST (v.2.1.3; Bouckaert et al. 2014) to fit branch lengths using a concatenated alignment of 1196 loci. We constrained the species tree topology and, without the availability of *Cryptoblepharus* fossils (Horner 2007), scaled the phylogeny with a molecular clock (0.001 substitutions/site/Myr) that was empirically calibrated in another group of

lizards within the same family (*Scincidae*; Brandley et al. 2011). We ran BEAST in duplicate for ten million generations, each run with a separate starting seed, a strict clock, and sampled the MCMC every ten thousand generations. Convergence of both runs was verified using Tracer, the first 20% of the MCMC discarded, and we merged the independent runs with LogCombiner.

Modeling Trait Dependent Dispersal

To test whether *Cryptoblepharus* species that occur in a coastal habitat have a higher propensity for dispersal, we first classified each lineage as being littoral or non-littoral. Horner (2007) provided a detailed description of the ecology of each species and we used our own field observations (i.e. for lineages that have not yet been delineated as distinct species) where no description was present. When ecological information was not available a-priori, lineages were assigned 'blindly' by collaborators with direct field experience; without them having seen the eventual phylogeny. We did not distinguish between true littoral specialists, in example species that frequently occur in the tidal zone, and species that only occur in habitat adjacent to the beach, but this might potentially be an interesting avenue for future analyses. Here, we specifically focused on the differences between inland and exclusively coastal species.

For historical biogeography analysis, we first defined discrete biogeographic regions within the overall distribution of the *Cryptoblepharus* genus. Ideally one would assign each island or alternative habitat as a distinct region, but likelihood-based biogeography analyses are computationally limited by the size of the transition rate matrix; 11 discrete regions means there are 2^{11} possible geographic ranges (states in the transition matrix), and the size of the transition matrix is 2^{11} X 2^{11}. The number of states can be reduced to a certain extent by specifying an upper maximum to range size. We chose to compromise between computational feasibility and the large number

of islands by using geological history across the *Cryptoblepharus* distribution as a proxy for biogeographic distinction between regions. Most island archipelagoes within the East Pacific for example are of very recent origin (~5 mya) whereas islands such as New Caledonia and Japan are much older (~30 mya; Neall and Trewick 2008). We clustered the islands in the East Pacific that are less than ten million years old in one region, whereas islands of continental origin such as New Caledonia were identified as unique regions. Distances between the 16 regions were estimated with Google Earth and were taken as the Great Circle distance between the center of each identified region (see Supp. Table 2).

We used maximum likelihood model fitting to examine whether habitat preference is a good predictor of dispersal probability and thus whether geographical range evolution is dependent on the habitat preference of a given species. Historical biogeography has advanced in recent years with the introduction of a probabilistic framework to model dispersal and vicariance processes, starting with the Dispersal-Extinction-Cladogenesis (DEC) model (Ree and Smith 2008). Matzke (2014) expanded the DEC model with the addition of a free parameter (+*J*) to accommodate jump-dispersal events and founder event speciation, and showed that DEC+J is the preferred model in many island radiations. Here we extend the DEC+J model so that dispersal probability can be a function of an evolving discrete character. In the trait-based dispersal model, the ecological trait character (state 1=non-littoral; state 2=littoral) evolves on the phylogeny along with geographic range. Two parameters describe the rates of transition forwards and backwards between the states: *t12* and *t21*. These rates are estimated along with the standard parameters *d*, *e*, and *j*. Multipliers on dispersal rate (and on jump dispersal weight, for cladogenetic founder-event dispersal) are applied based on the evolving character state. When a lineage is in state 1, the dispersal rate multiplier is *m1*, which is fixed to 1. When a lineage is in state 2 (littoral), the dispersal rate multiplier *m2* is used. The parameter *m2* is left free to

vary and estimated along with the other free parameters via maximum likelihood. A final free parameter is *x*, which modifies dispersal probability by multiplying with distance^x (Van Dam and Matzke, 2016). Therefore, models range in complexity from 2 free parameters (*d* and *e*, as in standard DEC) to seven (*d*, *e*, *j*, *x*, *t12*, *t21*, *m2*)

We used standard tools of statistical model comparison (Burnham and Anderson, 2001) as incorporated in BioGeoBEARS (v.2.1 with patches from wwww.phylowiki.com; Matzke 2014). We first evaluated the six standard biogeographic models (See Table 1 for overview) by assessing the maximum likelihood they confer on the geographic range data. Since the DEC+J model outperformed all other models (see Results), we then modified the standard DEC+J model and calculated the likelihood of four different versions of DEC+J, which ranged from the standard model to DEC+J+x+m2. In the most basic version of this model, habitat is a binary state and m2 is the increase or decrease in dispersal rate in state two relative to state one, when dispersal rate for state one is at a fixed value (1).

RESULTS

Exon Capture Results

We have previously presented a detailed evaluation of our sequence capture approach, capture success and how this changes while targeting taxa at different evolutionary timescales (Bragg et al. 2015). Overall, our results are congruent with these findings (see Supp. Table 1) and we therefore only focus on the key results pertaining to the current study.

We sequenced a total of 91 non-Australian individuals and retained 81 high quality libraries. We supplemented this new dataset with sequences from 27

Australian species that were obtained in a previous study (Blom et al 2017), resulting in a total dataset of 108 *Cryptoblepharus* individuals and two outgroup representatives (210 haplotypes). We recovered an average of 2469 loci for each library, with a mean coverage of 187x, and 1196 loci were sequenced for all individuals. We identified 53 unique lineages in the complete dataset, selected 2 individual haplotypes per lineage (if possible from different individuals within the same lineage – therefore supplemented with additional Australian representatives; Supp. Table 1) and generated new alignments for this subset of taxa. This dataset included 1225 loci that were sequenced for all species and 1775 loci with at least 104 individual haplotypes or more.

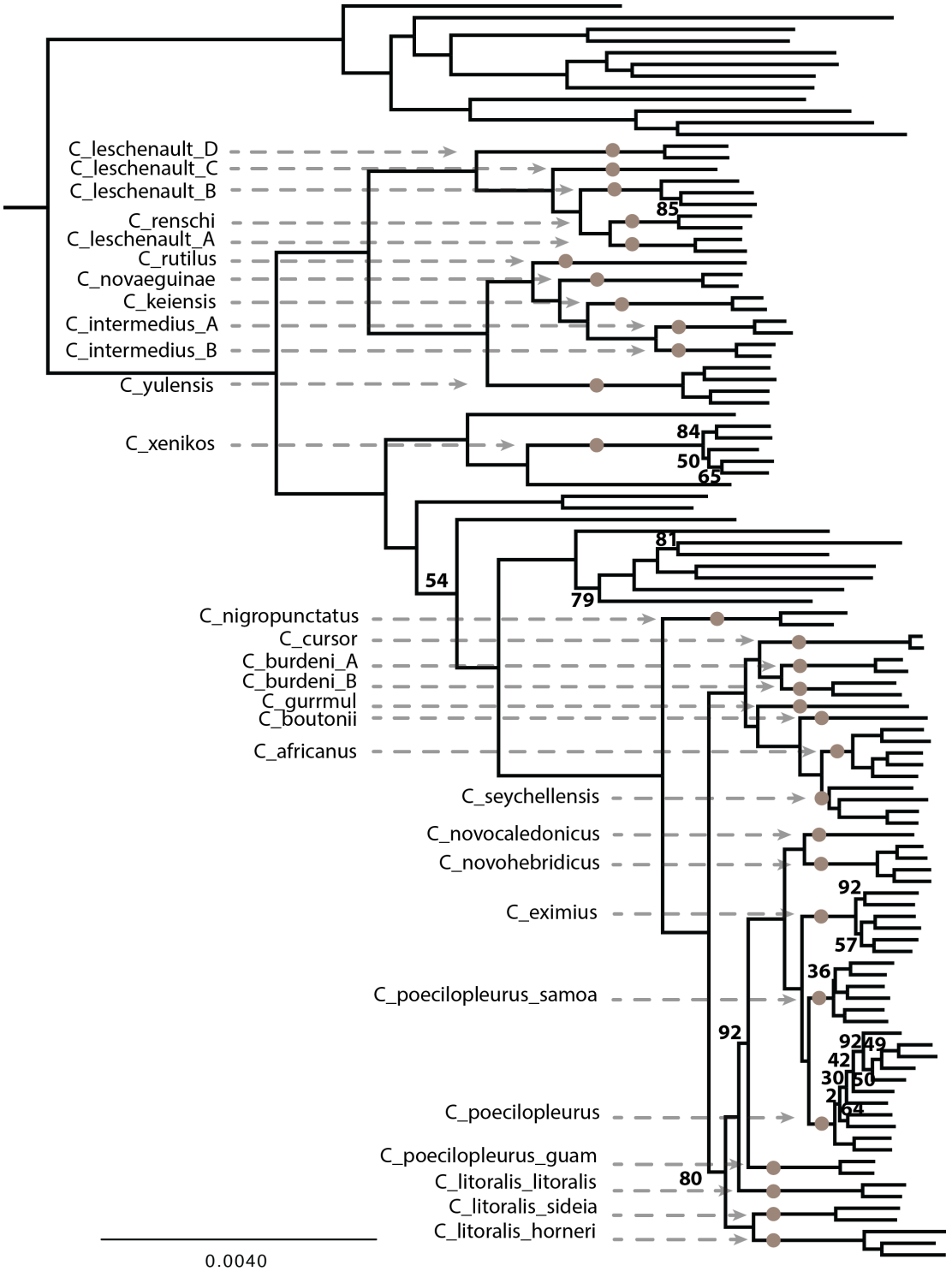


Figure 1. A genus level phylogeny that includes populations of *Cryptoblepharus* from across their distribution and is based on a concatenated maximum likelihood analysis of 1196 loci. We have highlighted only the non-Australian lineages because we have evaluated the diversity and relationships among the Australian species in previous studies (Blom et al. 2017). The non-Australian lineages that have been used for subsequent phylogenetic and biogeographic inference have been annotated with a small brown orb.

Phylogenetic Analyses

Both the species phylogeny based on concatenation of 1196 loci (537,674 bp; Fig. 1) and the summary-coalescent species tree based on the 600 most informative gene trees (with branch lengths fitted; Fig. 2), are well-supported and largely congruent with the backbone topology as inferred in previous studies that specifically focused on the Australian radiation (Blom et al. 2017). The inclusion of the non-Australian lineages highlights that the Indo-Australian lineages (Sahul shelf and Lesser Sunda islands) are relatively highly divergent, whereas the divergence between most Pacific and African lineages is more shallow. The Australian radiation is paraphyletic and suggests that *Cryptoblepharus* has colonized the Australian continent in multiple waves. Although the Australian radiation is paraphyletic overall, the two major clades remain monophyletic and the well-supported phylogeny suggests that the Australian *Cryptoblepharus* species have mainly diversified within the continent, rather than abroad with subsequent secondary introductions. There are two exceptions; *C. virgatus* from northern Queensland (Australia) is more closely related to *C. xenikos* from Papua New Guinea and the littoral Australian groups (*C. litoralis* and *C. gurrmul*) are more closely related to the other littoral species than to the Australian lineages (Fig. 3). Furthermore, both major clades of Australian *Cryptoblepharus* have a northern ancestor (Christmas Island and an Indonesian clade), confirming previous suggestions that *Cryptoblepharus* is likely of south-east Asian origin (Horner, 2007). Most lineages from the Indo-Papuan region form a distinct clade that is the sister group to the alternative Australian clade, that includes *C. egeriae*, *C. juno* etc. (Fig. 3), whereas the Pacific and African clades are a subgroup within this broader Indo-Papuan group.

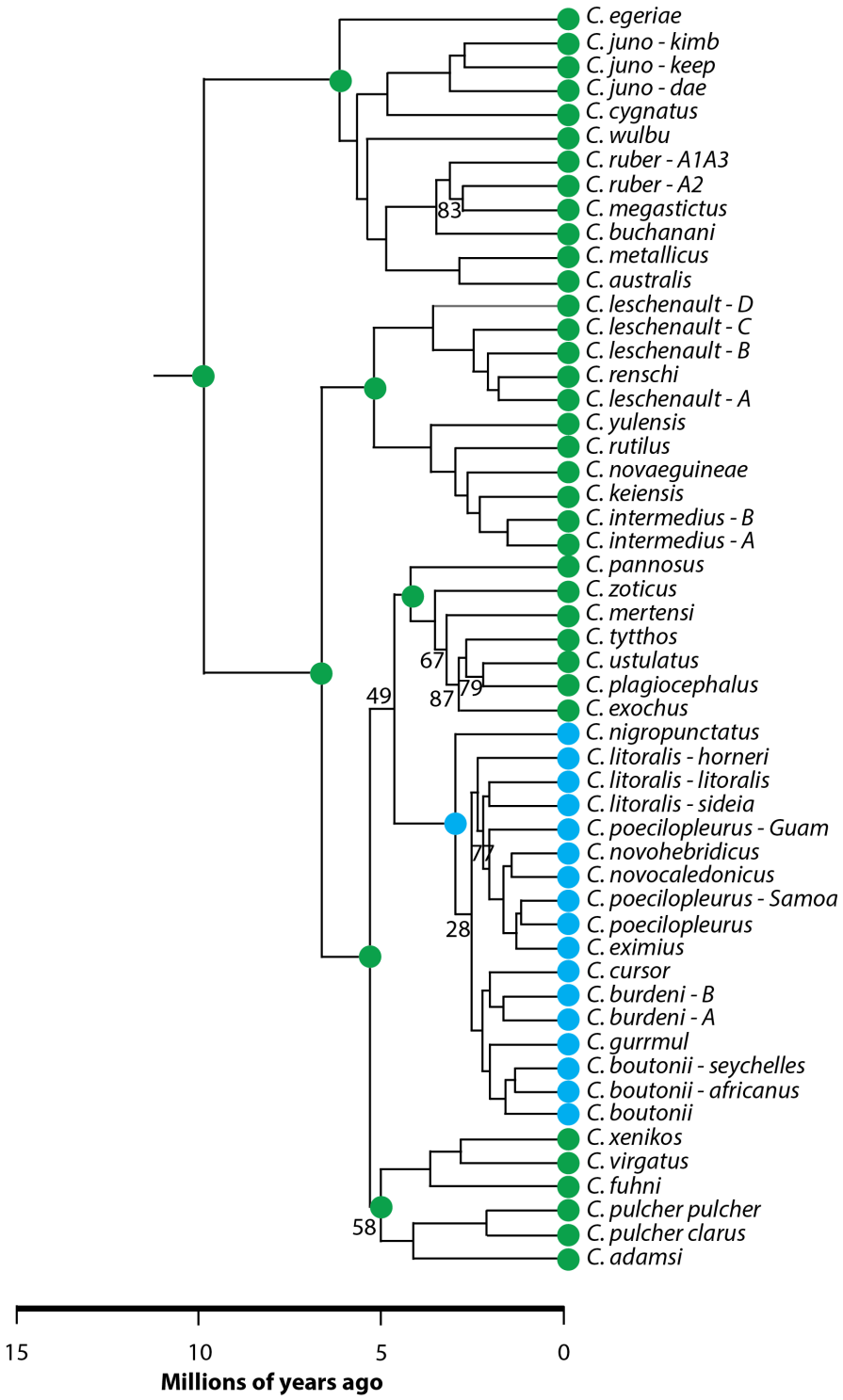


Figure 2. A time calibrated phylogeny for the major *Cryptoblepharus* lineages based on a summary coalescent analysis of the 600 loci with the highest TC scores. Branch lengths have been inferred independently with a concatenated alignment of 1225 loci and an empirically calibrated molecular-clock. Habitat occurrence has been plotted for each individual lineage and the major clades; blue orbs represent littoral specialists and green orbs represent non-littoral species

The divergence time estimates within the dated phylogeny are congruent with previous estimates (Blom et al. 2016) and this relatively young genus diversified during the Plio-Pleistocene after originating at the end of the Miocene (~10 Mya; Fig. 2). Interestingly, we did not recover any lineages that branched off during the first three million years and the evolutionary history during this initial period remains unclear.

The concatenated phylogeny strongly suggests that the current taxonomy is incomplete and that specific taxa require further revision. For example, *C. leschenault* consists of multiple deep phylogenetic lineages and is paraphyletic since *C. renschi* is clustered within the *C. leschenault* species group (Fig. 1, 2). These *C. renschi* individuals could have been misidentified, but even if so, the phylogenetic divergence between these major lineages equals other taxa that have been previously identified as distinct species and therefore merits further investigation. Our time-calibrated phylogeny further confirms conclusions from exploratory analyses that some of these lineages have been phylogenetically distinct for millions of years and that it is therefore likely more appropriate to treat such lineages as distinct entities in an analysis of biogeographic history.

Historical Biogeography

Ancestral range estimates reveal that *Cryptoblepharus* were mainly present on the Sahul shelf for the first 7 million years since emergence and then rapidly expanded across the Pacific and Indian Ocean, and diversified on small isolated islands (Fig. 3, 4). We therefore employed a range of biogeographic models that include both dispersal and vicariance.

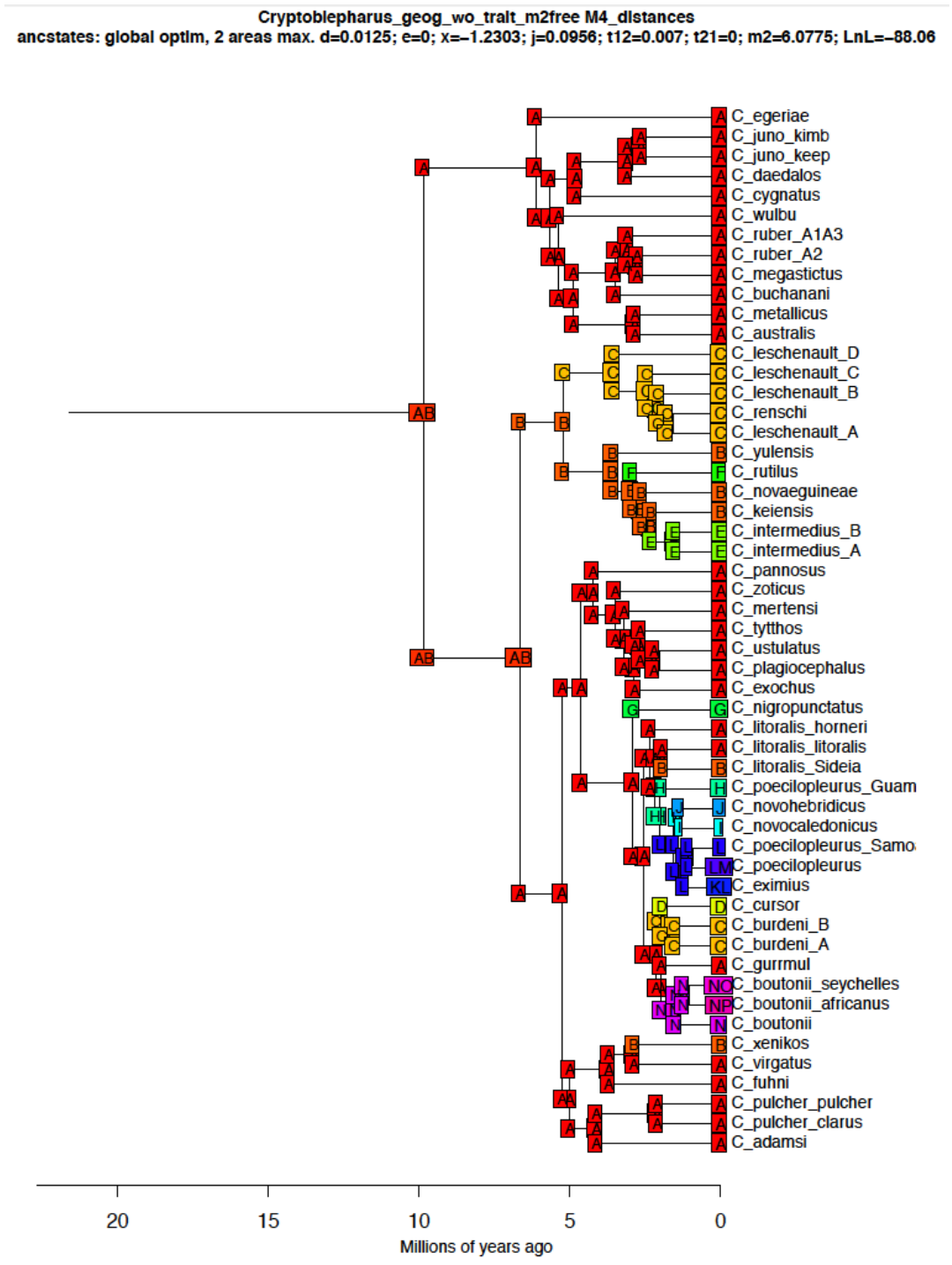


Figure 3. A time calibrated phylogeny for the major *Cryptoblepharus* lineages based on a summary coalescent analysis of the 600 loci with the highest TC scores and with the ancestral ranges for plotted for each contemporary and ancestral lineage. A) Australia, B) Papua New Guinea, C) Lesser Sunda Islands, D) Sulawesi, E) Molucca islands, F) Palau, G) Bonin islands, H) Guam, I) New Caledonia, J) Vanuatu, K) Fiji, L) Polynesia, M) Hawaii, N) Madagascar, O) Seychelles and P) Africa

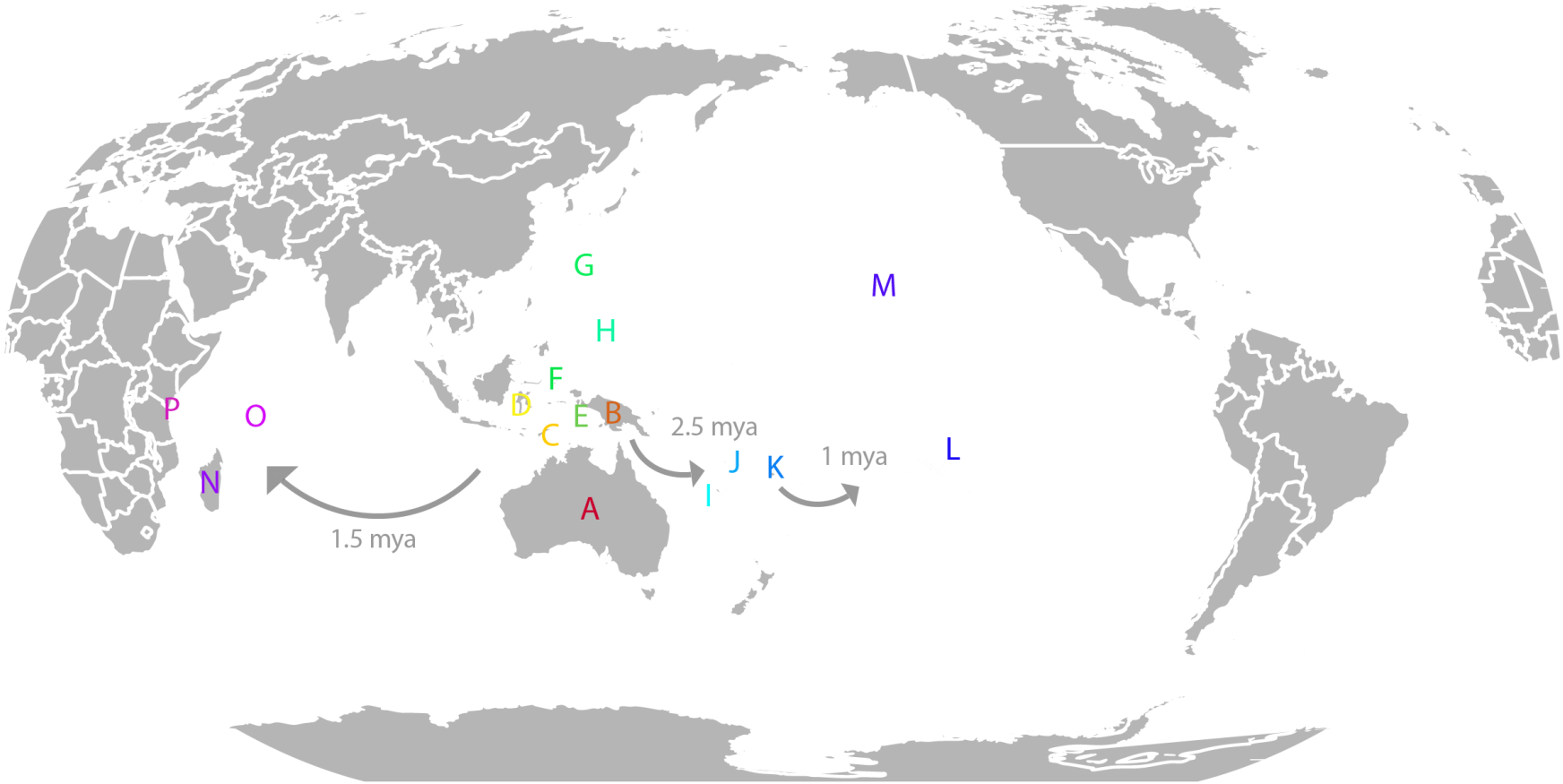


Figure 4. Overview of the main geographic regions within the genus wide distribution of *Cryptoblepharus*. Arrows indicate the inferred direction and timing of major range shifts based on the genus level phylogeny and, colours and characters correspond with the contemporary and ancestral ranges for species as specified in figure 3. A) Australia, B) Papua New Guinea, C) Lesser Sunda Islands, D) Sulawesi, E) Molucca islands, F) Palau, G) Bonin islands, H) Guam, I) New Caledonia, J) Vanuatu, K) Fiji, L) Polynesia, M) Hawaii, N) Madagascar, O) Seychelles and P) Africa

We first compared the DEC+J model with six alternative biogeographic scenarios. The DEC+J model resulted in a significant improvement in the likelihood score (Table A), with a considerable difference between DEC+J and the worst performing model (BAYAREALIKE; 47 log-likelihood units) and a marginal difference between DEC+J and the second best model (DIVALIKE+J; 1 log-likelihood unit). Likelihood ratio tests show that accounting for jump dispersal results in a significant improvement of the model across all biogeographic scenarios. AIC weights give a sense of the relative support for each model and the DEC+J model is strongly favored by the data over all other models (68%).

We then modified the DEC+J model and compared four different models where we specifically tested the improvement in the fit of our models with the addition of two extra parameters; distance dependent dispersal probability (*x*) and trait dependent dispersal rate (*m2*). A DEC+J model that included either of these two parameters was a significant improvement in comparison to the standard DEC+J model, with an improvement of 8 – 11 log-likelihood units (Table B). Moreover, even though our DEC+J+x+m2 model has two additional free parameters, the improvement in the fit of the model given our data is substantial (another 8 – 10 log-likelihood units). Likelihood ratio tests further confirm this with significant support for the most complex model (Table C) and the AICc model weights almost exclusively support the DEC+J+x+m2 model as the model that best explains the data (99.8%).

The parameter estimates inferred using maximum-likelihood provide further insight in the biogeographic history as estimated for each model. For the models where distance dependent dispersal probability is a free parameter, the probability of dispersal decreases with increasing distance (dispersal probability is proportional to distance^-1.23).

Table A. Statistical model comparison between six basic biogeographical models available in BioGeoBEARS (Matzke 2013), with Likelihood Ratio Test and sample-size corrected AIC (AICc). The data being fit are geographic ranges at the tips of the phylogeny (traits are not included).

Base model	LnL	# of free para-meters	*d*	*e*	*j*	Deviance	degrees of freedom	LRT *p*-value	AICc	AICc model weight	Percentage model weight
DEC	-131.7	2	0.0076	0.037	0	61.4	1	4.7E-15	267.6	1.0E-13	0.0%
DEC+J	-101	3	0.0017	1.00E-12	0.012				208.6	0.68	68.0%
DIVALIKE	-123.5	2	0.0072	1.00E-12	0	43.0	1	5.5E-11	251.3	3.5E-10	0.0%
DIVALIKE+J	-102	3	0.0019	1.00E-12	0.012				210.5	0.26	26.0%
BAYAREALIKE	-147.8	2	0.0092	0.12	0	88.8	1	4.4E-21	299.8	1.0E-20	0.0%
BAYAREALIKE+J	-103.4	3	0.0015	1.00E-07	0.012				213.3	0.064	6.4%

Table B. Maximum-likelihood estimates of parameters inferred with BioGeoBEARS for the DEC+J model, with and without distance-based dispersal (+*x*) and trait-based dispersal (+*m2*).

	Parameters (defined)							
	d	*e*	*j*	*x*	*t12*	*t21*	*m1*	*m2*
Model	dispersal/range gain (anagenetic)	range loss (anagenetic)	jump dispersal/found er-event weight (cladogenetic)	dispersal probability multiplied by distance^*x*	transition rate from nonlittoral to littoral	transition rate from littoral to nonlittoral	multiplier on dispersal when nonlittoral	multiplier on dispersal when littoral
1. DEC+J, with trait transitions	0.0017	0	0.012	0	0.0071	0	1	1
2. DEC+J+x (distance-based dispersal), with trait transitions	0.0244	0	0.174	-1.23	0.0071	0	1	1
3. DEC+J, with trait transitions, and dispersal multiplier for trait 'littoral'	0.0007	0	0.005	0	0.0070	0	1	8.51
4. DEC+J+x (distance-based dispersal), with trait transitions, and dispersal multiplier for trait 'littoral'	0.0125	0	0.095	-1.23	0.0070	0	1	6.12

(parameter is fixed to this value)

Table C. Likelihood Ratio Tests (LRT) for pairwise comparisons of nested models from Table B. The null hypothesis is that the fit of the more complex model to the data is no better than would be expected by addition of uninformative free parameter(s). Rejection of the null hypothesis suggests that addition of the free parameter(s) significantly improves model fit.

Simpler (nested) model	More complex model	Deviance	degrees of freedom	*p*-value	*p*-value	significance
Model 1	Model 2	15.3	1.0	0.00009320	9.3E-05	***
Model 1	Model 3	22.0	1.0	0.00000277	2.8E-06	***
Model 1	Model 4	36.8	2.0	0.00000001	1.0E-08	***
Model 2	Model 4	21.6	1.0	0.00000343	3.4E-06	***
Model 3	Model 4	14.9	1.0	0.00011595	1.2E-04	***

* = significant at p<0.05
** = significant at p<0.01
*** = significant at p<0.001

Table D. Statistical model comparison with sample-size corrected AIC (AICc).

Model	LnL	# of free parameters	AICc	AICc model weight	Percentage model weight
1. DEC+J, with trait transitions	-106.5	5	224.2	0.000	0.00%
2. DEC+J+x (distance-based dispersal), with trait transitions	-98.8	6	211.5	0.000	0.01%
3. DEC+J, with trait transitions, and dispersal multiplier for trait 'littoral'	-95.5	6	204.8	0.002	0.22%
4. DEC+J+x (distance-based dispersal), with trait transitions, and dispersal multiplier for trait 'littoral'	-88.0	7	192.6	0.998	99.77%

For the models where dispersal probability can be conditional on ecological state, being a littoral species results in a 6 to 8-fold increase in the probability of dispersal (Table D). Altogether, these results strongly suggest that a biogeographic model that accounts for jump dispersal and differences in dispersal probability between species that prefer different habitats is more supported than alternative models.

DISCUSSION

Range evolution is most likely an important determinant in promoting species diversification and has potentially shaped the macroevolutionary history of many taxa. Yet, the importance of long-distance dispersal and how ecology might promote or constrain such events, remains uncertain and has been debated among biogeographers for many decades. Quantifying such processes in empirical systems has proven challenging, since we are often limited in our ability to accurately retrace the biogeographic and temporal history of diversification in widespread ancient clades. Furthermore, the current distribution and prevalence of extant taxa might represent a biased sampling of past biogeographic history for a given clade. It is therefore more appropriate to address these questions in species groups that occur across vast geographic distances but have radiated relatively recently. We employed a genomic approach to infer a time-calibrated phylogeny based on 1196 genetic markers (>500,000 bp), to assess the ecological context of range evolution in a group of lizards that only started diversifying during the early Pliocene, but now occur across multiple continents and isolated oceanic islands.

Rather than inferring branch lengths based on a small number of loci, we specifically employed a genomic approach to improve the accuracy of our branch

length estimates. We used an empirically calibrated molecular clock since there are currently no fossils present for *Cryptoblepharus* or any other closely related genera. Nonetheless, *Cryptoblepharus* lizards inhabit a number of young oceanic islands that vary in geological age (Neall and Trewick 2008) and, without a prior constraint on node ages by island emergence times, the divergence dates in our tree do not surpass island age, suggesting that our divergence estimates are relatively accurate. Given that rate variation among lineages is not expected at a shallow phylogenetic scale (Yang 1996), we presume that our divergence estimates are therefore reasonable across the phylogeny. However, in comparison to our previous study that focused on the Australian lineages only, terminal branch lengths are longer, while the timing of branching events deeper in the tree remains similar. The extant lineages are older in our current tree since we fitted branch lengths with the use of haplotype rather than diplotype data where the two haplotypes of each individual are collapsed and heterozygous sites coded according to the IUPAC format. We changed from the use of diplotype data to phased haplotypes because a recent study has shown that terminal branch lengths might be underestimated with the use of diplotype data alone (Lischer et al. 2014).

Trans-Oceanic Dispersal

Where the relative importance of dispersal and vicariance in shaping the biogeographic ranges of ancient widely distributed taxa is often debated, our results confirm our prior expectation that the *Cryptoblepharus* genus is relatively young (~10 Myr), ruling out the possibility that the global distribution of *Cryptoblepharus* is shaped by past vicariance. *Cryptoblepharus* lizards were mainly confined to the Australasian region for the longest part of their evolutionary history and the major colonization of regions outside the Sahul shelf started approximately 3 Mya (Fig. 3, 4).

These divergence estimates are further corroborated by their presence on many island archipelagoes in the South Pacific that are of volcanic origin and only emerged within the last 5 Myr (i.e. the Austral islands; (Neall and Trewick 2008)), preventing them from colonizing these regions prior to this time. Given that the continental break-up significantly predates the crown age of the genus, the distribution of these lizards must therefore be the result of past dispersal; either natural or human mediated.

Given our current knowledge on the timing of human migration into the Pacific and Indian Ocean, *Cryptoblepharus* have likely colonized most regions throughout their current day distribution without human interference, as has been inferred for several other lizards across the region (see Tonione et al. 2011 and references therein). The minimum divergence time between the major lineages, as identified in our current study, exceeds 1 Myr and significantly predates the human migration into both the Pacific and Malagasy region. Furthermore, the phylogeographic structuring matches a stepping-stone model where migration has moved away from the Sahul shelf into both an eastern and western direction (Fig. 3, 4). For example, the oldest lineage from the Malagasy region is a population from Mauritius which is the sister lineage to the other two Malagasy species. Mauritius is the first major island between the Sahul shelf and the African continent and is placed on an ocean current that moves away from Australia in a western direction. Our results support a scenario in which members of an Indo-Australian lineage, (closely related to) *C. gurrmul*, dispersed around 1.5 Mya across the Indian Ocean and likely first settled on Mauritius. From there onwards they colonized the rest of the Malagasy region, within the following 250K years (Fig. 3, 4).

Whereas migration into the South Pacific is also predominantly via natural dispersal, it remains unclear to what extent human colonization of the furthest regions have potentially aided *Cryptoblepharus* migration. Migration into the South Pacific

follows a similar stepping-stone model from the Sahul shelf until Fiji and Samoa. From there onwards isolated island populations are closely related even though they span vast distances. For instance, *C. poecilopleurus* from French Polynesia and Hawaii are more closely related than *C. novocaledonicus* and *C. novohebridicus* from New Caledonia and Vanuatu respectively (Fig. 2). This is potentially due to the young age of many Polynesian islands and the recent arrival of *Cryptoblepharus* lizards or alternatively, because these populations have been translocated with the human migration into Polynesia (Austin 1999). Since phylogeographic structuring is marginal within *C. poecilopleurus*, we currently clustered distant populations into a single species for the purpose of further biogeographic analyses. However, with increased sampling across the East Pacific, future studies should dissect the phylogeographic patterns within this region to further characterize which insular areas might have been colonized via natural dispersal and/or in which islands they have been introduced via alternative means.

The conclusion that the widespread distribution of *Cryptoblepharus* is exclusively attained via natural dispersal, is not surprising given the young age of the genus. More importantly, it provides an empirical opportunity to explicitly analyze the factors involved in geographic range evolution via natural long-distance dispersal.

Ecology Dependent Dispersal

Cryptoblepharus inhabits a vast number of isolated islands and distant landmasses, and biogeographic models that include jump-dispersal/founder event speciation (+J models) are a significant better fit to the data than models that preclude such events (Table D). The importance of jump dispersal, in shaping the biogeographic history of the *Cryptoblepharus* genus, is congruent with many other taxa that are distributed across distant island archipelagoes (Matzke 2014). However, the inclusion of free

parameters that account for differences in dispersal probability based on an evolving trait is a significant improvement in model fit compared to models that only account for jump dispersal and founder event speciation.

We expanded the standard DEC+J model (Matzke 2013; 2014) with two additional free parameters and both parameters significantly improve the explanatory power of the model, suggesting that both dispersal distance and ecology are important factors that determine the likelihood of dispersal. It is long known that geographic distance is an important constituent for dispersal rate (Whittaker 1998) and it is not unexpected that dispersal is reduced with increasing distance. The geographic distance between the Sahul shelf and the African mainland for example is about 10,000 kilometers and, even though *Cryptoblepharus* seem to disperse frequently at a macroevolutionary scale, such dispersal events remain rare (Rocha et al. 2005). Nonetheless, these results also provide further indication that most geographic expansion within *Cryptoblepharus* has likely occurred via natural dispersal since dispersal via modern humans should be much less dependent on geographic distance.

We find strong support for a model where dispersal probability is conditional on ecological state; species that occur in littoral habitat are significantly more likely to disperse than non-littoral species. The inference of ancestral areas and ecological states highlights that the distribution of *Cryptoblepharus* was largely confined to the Sahul shelf until species adapted to a littoral habitat (Fig. 3). Once a littoral form emerged around 3 Mya, lineages dispersed in various directions and across vast geographic distances, ultimately resulting in the widespread contemporary distribution of the genus (Fig. 4). Littoral species are approximately six times more likely to disperse than non-littoral species (Table B). Although the accuracy of this estimate is probably dependent on many factors (e.g. the biogeographic regions that were assigned a-priori), statistical model choice unambiguously supports a model

where the dispersal probability is conditional on habitat and significantly deviates from a null-model with equal dispersal probabilities between habitat states (Table D). Thus, being littoral resulted in a significant increase in dispersal probability and has promoted widespread dispersal across both the Pacific and Indian Ocean.

Future studies should address which aspects of being littoral have ultimately increased dispersal probability. Some *Cryptoblepharus* species are truly littoral specialists and species have been observed that actually swim for short distances and feed on small crustaceans (Horner 2007). It remains unknown whether specific adaptations are required to exhibit such behaviours but these traits might potentially enable small lizards to sustain themselves during month-long voyages across oceans on flotsam or trees. Alternatively, the littoral forms might not have developed any traits that distinguish them from other habitat specialists, but inhabiting coastal habitat by itself might already have increased their propensity for dispersal. *Cryptoblepharus* lizards are small ectotherms that often lay their eggs underneath the bark of trees and perhaps some have ultimately just been transported before hatching. The exact mechanisms of dispersal remain unclear and require further study, but here we have characterized the biogeographic history of the genus and identified that habitat specialization has played an important role in shaping the widespread distribution of this relatively young clade of small lizards.

CONCLUSION

The study of historical biogeography is of fundamental importance for our understanding of large-scale macroevolutionary change. Dispersal and vicariance are two alternative hypotheses that can both explain the widespread distribution of major taxa and have been the focus of debate among biogeographers for many decades.

Although arguments for and against long-distance dispersal have mostly focused on the likelihood of such chance events to occur, factors that can modulate the probability of long-distance dispersal have not been thoroughly evaluated nor quantified. Here we have taken an alternative approach and used a probabilistic modeling framework to characterize the frequency of natural trans-oceanic dispersal and the importance of ecology for dispersal probability. Our results highlight that natural long-distance dispersal has been a recurring process in a recent radiation of lizards with a widespread distribution and that the evolution of geographic range has been strongly influenced by habitat occurrence. As such, our empirical study demonstrates that natural dispersal is a potent mechanism for widespread range evolution, but moreover that ecology needs to be accounted for since it can significantly modulate the probability of dispersal.

ACKNOWLEDGEMENTS

We thank Barnabus Wilmot, Georgia Kaipu, Jim Animiato, Bulisa Iova, Cathy Newman, Donna Dittman, Miguel Carretero, Miguel Vences, Frank Glaw, Mickael Sanchez, Jean-Yves Meyer, Matthew Fujita, Foteini Spagopoulou, Sally Potter, SIF, Gump Field Station, National Geographic Society and the Mohamed Bin Zayed Fund for Species Conservation, for helping in the field and/or supporting this work.

REFERENCES

Austin C.C. 1999. Lizards took express train to Polynesia. Nature 397:113-114

Bi K., Vanderpool D., Singhal S., Linderoth T., Moritz C., Good J.M. 2012. Transcriptome-

based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics 13:403

Blom M. 2015a. Habitat use and new locality records for Cryptoblepharus poecilopleurus (Squamata: Scincidae) from French Polynesia. Herpetol. Notes 8:579-582.

Blom M.P.K. 2015b. EAPhy: A Flexible Tool for High-throughput Quality Filtering of Exon-alignments and Data Processing for Phylogenetic Methods. PLoS Curr.:1–12. doi: 10.1371/currents.tol.75134257bd389c04bc1d26d42aa9089f

Blom M.P.K., Horner P., Moritz C. 2016. Convergence across a continent: adaptive diversification in a recent radiation of Australian lizards. Proc. R. Soc. B. 283:20160181–9.

Blom M.P.K., Bragg J., Potter S., Moritz C. 2017. Accounting for uncertainty in gene tree estimation: Summary-coalescent species tree inference in a challenging radiation of Australian lizards. Syst. Biol. 66: 352-366

Bouckaert R., Heled J., Kühnert D., Vaughan T., Wu C.-H., Xie D., Suchard M.A., Rambaut A., Drummond A.J. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Comput. Biol. 10:e1003537.

Bragg J.G., Potter S., Bi K., Moritz C. 2016. Exon capture phylogenomics: efficacy across scales of divergence. Mol. Ecol. Resour. 16:1059–1068.

Brandley M.C., Wang Y., Guo X., de Oca A.N., Fería-Ortíz M., Feria-Ortiz M., Hikida T., Ota H. 2011. Accommodating heterogenous rates of evolution in molecular divergence dating methods: an example using intercontinental dispersal of Plestiodon (Eumeces) lizards. Syst. Biol. 60:3–15.

Bromham L., Penn D. 2003. The modern molecular clock. Nature Rev. Gen. 4:216–224.

Brown R.P., Yang Z. 2011. Rate variation and estimation of divergence times using strict and relaxed clocks. BMC Evol. Biol. 11:271.

Brygoo E. 1986. Systematiques des lezards Scincides de la region malgache. XVIII. Les Cryptoblepharus. Bull. Mus. Natl. Hist. Nat. 8:643-690.

Burnham, K.P. and Anderson, D.R. (2002) Model selection and multimodel inference: A practical information-theoretic approach. Springer Verlag, New York.

Carstensen D.W., Dalsgaard B., Svenning J.-C., Rahbek C., Fjeldså J., Sutherland W.J., Olesen J.M. 2013. The functional biogeography of species: biogeographical species roles of birds in Wallacea and the West Indies. Ecography. 36:1097–1105.

Coyne, J.A. and Orr, H.A. (2005) Speciation. Sinauer Associates, Sunderland.

Darriba D., Taboada G.L., Doallo R., Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods. 9:772–772.

Darwin 1859. On the origin of species. John Murray, London.

Edgar R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Horner P. 2007. Systematics of the snake-eyed skinks, Cryptoblepharus Wiegmann (Reptilia: Squamata: Scincidae)–an Australian based review. The Beagle Supp. 3:21–198.

Langmead B., Salzberg S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods. 9:357–359.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078–2079.

Lischer H.E.L., Excoffier L., Heckel G. 2014. Ignoring Heterozygous Sites Biases Phylogenomic Estimates of Divergence Times: Implications for the Evolutionary History of Microtus Voles. Mol. Biol. Evol. 31:817–831.

Longrich N.R., Vinther J., Pyron R.A., Pisani D., Gauthier J.A. 2015. Biogeography of worm lizards (Amphisbaenia) driven by end-Cretaceous mass extinction. Proc. Roy. Soc. B. 282:20143034–20143034.

Matzke N.J. 2013. Probabilistic historical biogeography: new models for founder-event speciation, imperfect detection, and fossils allow improved accuracy and model-testing. Front. Biogeogr. 5:242-248

Matzke N.J. 2014. Model Selection in Historical Biogeography Reveals that Founder-Event Speciation Is a Crucial Process in Island Clades. Syst. Biol. 63:951–970.

Matzke N.J., Wright A. 2016. Inferring node dates from tip dates in fossil Canidae: the importance of tree priors. Biol. Lett. 12:20160328–4.

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M., DePristo M.A. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.

Meyer M., Kicher M. (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protocol. 6, t5448. (doi:10.1101/pdb.prot5448)

Mirarab S., Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics. 31:i44–52.

Neall V.E., Trewick S.A. 2008. The age and origin of the Pacific islands: a geological overview. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 363:3293–3308.

Ogilvie H.A., Heled J., Xie D., Drummond A.J. 2016. Computational Performance and Statistical Accuracy of *BEAST and Comparisons with Other Methods. Syst. Biol. 65:381–396.

Potter S., Bragg J.G., Peter B.M., Bi K., Moritz, C. 2016. Phylogenomics at the tips: Inferring lineages and their demographic history in a tropical lizard, Carlia amax. Mol. Ecol. 25: 1367-1380

Ree R.H. 2005. Detecting the historical signature of key innovations using stochastic models of character evolution and cladogenesis. Evolution 59:257-265

Ree R.H., Sanmartín I. 2009. Prospects and challenges for parametric models in historical biogeographical inference. J. Biogeogr. 36:1211–1220.

Ree R.H., Smith S.A. 2008. Maximum Likelihood Inference of Geographic Range Evolution by Dispersal, Local Extinction, and Cladogenesis. Syst. Biol. 57:4–14.

Rocha S., Carretero M., Vences M., Glaw F. 2005. Deciphering patterns of transoceanic dispersal the evolutionary origin and biogeography of coastal lizards (Cryptoblepharus) in the Western Indian Ocean region. J. Biogeogr. 33:13-22

Singhal S. 2013. De novotranscriptomic analyses for non-model organisms: an evaluation of methods across a multi-species data set. Mol. Ecol. Res. 13:403–416.

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

Sukumaran J., Economo E.P., Lacey Knowles L. 2016. Machine Learning Biogeographic Processes from Biotic Patterns: A New Trait-Dependent Dispersal and Diversification Model with Model Choice By Simulation-Trained Discriminant Analysis. Syst. Biol. 65:525–545.

Sunnucks P., Hales DF. 1996. Numerous transposed sequences of mitochondrial cytochrome oxidase I-II in aphids of the genus Sitobion (Hemiptera: Aphididae). Mol. Biol. Evol. 13:510-524

Tonione M.A., Reeder N., Moritz C.C. 2011. High Genetic Diversity Despite the Potential for Stepping-Stone Colonizations in an Invasive Species of Gecko on Moorea, French Polynesia. PLoS ONE. 6:e26874–6.

Vences M., Vieites D.R., Glaw F., Brinkmann H., Kosuch J., Veith M., Meyer A. 2003. Multiple overseas dispersal in amphibians. Proc. Roy. Soc. B. 270:2435–2442.

Wallace A.R. 1869. The Malay Archipelago. MacMillan and co., London.

Weir J.T., Schluter D. 2008. Calibrating the avian molecular clock. Mol. Ecol. 17:2321–2328.

Whittaker R.J. 1998. Island Biogeography: Ecology, Evolution and Conservation. Oxford University Press, Oxford

Wiley E.O. 1988. Vicariance biogeography. Annu. Rev. Ecol. Evol. Syst. 19:513-542

Wilson E.O. 1961. The nature of the Taxon Cycle in the Melanesian ant fauna. Am. Nat. 95:169-193

Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. 11:367–372.

SUMMARY & SYNTHESIS

SUMMARY & SYNTHESIS

Over 150 years ago, Darwin and Wallace outlined how population level processes can induce evolutionary diversification and we now recognize that species form the building blocks for large-scale macroevolutionary change. Yet, we also have become increasingly aware that speciation is a protracted process (Dynesius and Jansson 2014), that incipient species might be often mere ephemeral lineages (Rosenblum et al. 2012) and that species boundaries can remain semi-permeable for prolonged periods of time (Mallet 2005). The evolutionary processes that modulate the transition from population level divergence to eventual species variation, are therefore often unclear and remain a major focus in evolutionary biology.

To address this uncertainty, I have focused on the evolutionary radiation of a speciose clade of lizards with an extraordinarily widespread distribution that includes both continents and island archipelagoes. Across their distribution, skinks of the genus *Cryptoblepharus* occur in multiple distinct habitats and habitat specialization has played an important role in their diversification. In this thesis, I explore distinct aspects of the *Cryptoblepharus* radiation and, across all chapters, I shed further light on the ecological and geographic context of diversification with a combined assessment of morphological, biogeographic and phylogenetic patterns. More specifically I evaluate the evolution of habitat specialization, characterize patterns of introgression between phylogenetically and ecologically distinct lineages and quantify whether the probability of dispersal is dependent on habitat preference. However, such processes can only be studied in a phylogenetic context and Ch. I focuses on the challenges associated with coalescent-based species tree estimation using large-scale sequence datasets, which is often required to resolve relationships and interactions among rapidly diversifying lineages.

In collaboration with co-authors, we designed a sequence-capture system based on transcriptomes from closely related skink genera (Singhal 2013). Although sequence-capture methods are costly in comparison to other genome reduction approaches (i.e. ddRAD-seq or transcriptome sequencing), sequence-capture methods are relatively efficient in recovering orthologous loci across highly divergent taxa (Jones and Good 2015). Indeed, with our design we were able to effectively target loci in lizard genera with up to ~40 million years of divergence (Bragg et al. 2016). Our approach was highly successful within *Cryptoblepharus* and we recovered on average over 85% of the targeted loci (Ch. I). In subsequent projects (Ch. III and IV), we modified our original skink capture and also included probes where the design was based on a *C. ruber* transcriptome but this did not result in any major improvement of capture success. Following capture and sequencing, raw sequence reads were processed and assembled with a custom built pipeline.

The correct alignment of sequence data is a major prerequisite for correct phylogenetic inference (Zwickl et al. 2014), but visual inspection of sequence alignments for 1000's of genes is impractical. I therefore developed a bioinformatic workflow that is specifically designed to align and filter exonic sequence data (Appendix A). In comparison to anonymous sequences or introns, reading frames of exon-specific sequences allow additional quality checks on alignments. EAPhy aligns contigs using a traditional aligner, such as MUSCLE (Edgar 2004), and then evaluates each resulting alignment based on a number of user-specified criteria (Blom 2015). More specifically, individual exonic sequences are translated and compared to the alignment consensus. If sequences deviate markedly or contain for example more than one stop codon, alignments are flagged and filtered out. EAPhy provides a high-throughput solution for alignment and alignment filtering, and each exon-capture dataset used in this thesis has been processed accordingly. With increasing volumes of sequence data, it is exceedingly more difficult for empiricists to manually inspect

genetic data. Yet the importance of data processing and quality control does not diminish with datasets of increasing size; in contrast, minor bioinformatic decisions can ultimately have major impact on the biological interpretation of the processed data (Zwickl et al. 2014). EAPhy was specifically developed to prevent incorrect phylogenetic inference due to erroneous alignments and to my knowledge, it is currently the only published method for the processing of exon-capture data for this specific purpose.

The inference of species relationships in clades that include periods of rapid diversification is notoriously difficult due to widespread presence of incomplete lineage sorting (Giarla and Esselstyn 2015). In such instances, coalescent based inference methods are needed to appropriately model coalescent variation while estimating species relationships (Maddison 1997; Degnan and Rosenberg 2009). Yet, full-coalescent based inference, the simultaneous estimation of gene- and species tree, is challenging in datasets that contain large numbers of loci and taxa. Our study on Australian *Cryptoblepharus* for example, includes 29 taxa and more than a thousand genetic markers. This poses a conundrum for empiricists; with datasets that are effectively too large for full-coalescent analyses, but where the evaluation of a concatenated phylogeny suggests that ILS might be a prominent issue and should be accounted for. Summary-coalescent approaches on the other hand are computationally 'cheap', but such methods have been criticized for disregarding uncertainty in gene tree estimation (Springer and Gatesy 2015). This ultimately motivated me to quantify the importance of gene tree estimation error when employing summary-coalescent methods (Ch. I).

Our results highlight that a small number of well-resolved gene trees already converge on the most optimal topology and that the addition of low-resolution gene trees does not introduce phylogenetic noise (i.e. results in well-supported topological

changes). In contrary, the addition of poorly resolved gene trees actually improves support for most bipartitions that had low support with small numbers of loci and erratic changes in support between datasets can highlight potential violations of the structured coalescent model. Our study simultaneously highlights the value of characterizing the heterogeneity in phylogenetic signal between loci in empirical datasets and demonstrates a relatively straightforward procedure how to do so. These findings are in line with other recent studies that have included species tree estimation with summary-coalescent methods and highlights that these approaches are relatively effective in inferring the underlying topology (Mirarab et al. 2014; Ogilvie et al. 2016).

The phylogeny as inferred in Ch. I and Ch. II, illustrates that Australian *Cryptoblepharus* have radiated across the continent in two distinct bouts during the past ~5 Myr. Although *Cryptoblepharus* are known for their high degree of crypticity, hence the marked taxonomic inflation once genetic markers where included (Horner and Adams 2007), the inferred phylogeny illustrates that *Cryptoblepharus* have repeatedly derived similar phenotypes. To examine the ecological context of diversification, I used a phylogeny aware comparative approach in Ch. II, to test whether habitat specialization explains current patterns of phenotypic variation in ecologically relevant traits. Our results strongly suggest that habitat specialization has been a major component in the diversification of Australian *Cryptoblepharus* and we identified two adaptive peaks associated with distinct habitats (rock and arboreal). Moreover, we also recorded repeated peak shifts (5) in isolated regions across the continent, where species converged on the same phenotype as distantly related species that occur in the same habitat. The strong correlation between habitat and ecologically functional traits, repeated peak shifts and rapid diversification, bears strong similarity to the patterns that characterize adaptive radiations in insular systems (Schluter 2000). Yet, in contrast to insular systems where divergent selection

is often the main force promoting diversification, a large number of *Cryptoblepharus* species have also emerged without shifting habitat or changing their morphology. This suggests that uniform selection across isolated populations that occur in the same habitat, has likely been equally important as divergent selection between distinct habitats, for promoting macroevolutionary change. Strong uniform selection could potentially also explain the relatively rapid emergence of species that occur in similar habitat, since speciation due to drift alone should likely take much longer to develop (Gavrilets 2003).

Yet, as highlighted in Ch. III, reproductive isolation between distinct lineages is not complete and occasional introgression might still occur and cannot be predicted based on either phylogenetic distance or habitat preference. In Ch. III, I examined preliminary observations of mitochondrial capture between distinct lineages and found that there is still a strong signal of introgression within the nuclear genome between a pair of non-sister species that diverged approximately 5 Myr ago and which are adapted to distinct habitats (rocks vs trees). By contrast, little or no introgression was observed between two other species pairs that were either more closely related or also occurred in distinct habitats, similar as the 'introgressing pair'. These results were highly surprising and stimulate further research on the evolution of pre- and postzygotic isolation within this genus. Nonetheless, it provides strong evidence that species boundaries can remain semi-permeable for prolonged periods of time (Mallet 2005) and provokes questions regarding the genomic landscape of introgression. Our findings in Ch. II suggest that habitat shifts have resulted in directional selection towards an alternative adaptive peak and have ultimately promoted the evolution of distinct ecomorphs. This implies that individuals with intermediate phenotypes should be relatively less fit, because otherwise we wouldn't expect such a strong distinction in functional trait values between habitats. Alternatively, species could be phenotypically plastic (Goodman et al. 2013) or the ecomorphological traits might be

regulated by a relatively small number of genes that are shielded from gene flow. Recent studies that have assayed genome-wide variation between closely related species with distinct phenotypes, highlight that in some cases gene flow can be basically unrestricted across the majority of the genome except for a few genomic regions (Poelstra et al. 2014). Unfortunately, in the case of *Cryptoblepharus* this will remain unclear for the time being because we are limited in our ability to conduct such analyses given our exon-capture dataset and a lack of a suitable reference genome. Nonetheless, this is an intriguing avenue for future research and can provide important insight on the causes and consequences of long-term interspecific gene flow.

In my final chapter (Ch. IV), I focused on the complete *Cryptoblepharus* genus and explored the biogeographic history of this relatively widespread group and characterized the underlying mechanisms that have shaped their extraordinary distribution. Whereas recent empirical studies around the dispersal-vicariance debate have largely focused on dating results, we wanted to take an alternative approach by focusing on a relatively young group that must have spread via dispersal and identify whether specific environments can increase the probability of such long-distance translocations. We used a probabilistic modeling framework and standard tools of statistical model comparison to identify a model that best fits the data given a range of biogeographic scenarios. Moreover, we modified the DEC+J model (Matzke 2014) so that dispersal probability can be a function of an evolving discrete character, in the case of *Cryptoblepharus*: Habitat preference. Whereas most Indo-Australian lineages occur in inland habitats, species that have dispersed across the Pacific and Indian Ocean mainly occur in littoral habitats. Some species are true littoral specialists and inhabit small coral crevices close to tide pools, others occur in more mesic habitat adjacent to the beach. Nonetheless, they are rarely found away from the coast and much further inland. We inferred the phylogeny using similar methods as in Ch. I and

Ch. II, and asked whether dispersal probability is dependent on ecological state. More specifically, are littoral species more likely to disperse than non-littoral species? The concept of ecology dependent-dispersal is long known (see Sukumaran et al. 2016) but a statistical framework to quantify the importance of ecology was absent until recently (Sukumaran et al. 2016; Ch. IV). In Ch. IV, we provide one of the first empirical examples where we show that a change in habitat has resulted in a significant shift in dispersal probability and, thus, how ecology can modulate the rate of dispersal. A trait dependent dispersal model is significantly better supported than alternative models where dispersal probability is independent of ecological state. Littoral species are six to eight times more likely to disperse than non-littoral species, and this has ultimately resulted in the widespread distribution of *Cryptoblepharus* that ranges from Eastern Africa, to Easter Island and Hawaii. In contrast to the deep divergence between species that occur in the Indo-Australian region, including Wallacea, the littoral species are more closely related and have spread across such a vast distribution within the past 3 million years.

In conclusion, in this thesis I have characterized different evolutionary aspects that have played an important role in the recent diversification of the *Cryptoblepharus* genus. Habitat specialization has been of fundamental importance in the evolutionary radiation of *Cryptoblepharus*; it has promoted adaptive diversification across the Australian continent and simultaneously facilitated widespread dispersal across the Pacific and Indian Ocean. However, while habitat specialization can provide insight on the mechanisms that have promoted initial divergence, the identification of (ancient) introgression between species can inform our understanding of species persistence; an underestimated process that is equally important for the study of lineage diversification (Rosenblum et al. 2012; Dynesius and Jansson 2014). As such, the study of *Cryptoblepharus* has generated an intriguing empirical framework for future studies into the continuous nature between micro- and macroevolutionary change.

References

Blom M.P.K. 2015. EAPhy: A Flexible Tool for High-throughput Quality Filtering of Exon-alignments and Data Processing for Phylogenetic Methods. PLoS Curr.:1–12; doi: 10.1371/currents.tol.75134257bd389c04bc1d26d42aa9089f

Bragg J.G., Potter S., Bi K., Moritz C. 2016. Exon capture phylogenomics: efficacy across scales of divergence. Mol. Ecol. Resour. 16:1059–1068.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Dynesius M., Jansson R. 2014. Persistence of within-species lineages: a neglected control of speciation rates. Evolution. 68:923–934.

Edgar R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Gavrilets S. 2003. Perspective: Models of speciation: What have we learned in 40 years? Evolution. 10:2197-2215.

Giarla T.C., Esselstyn J.A. 2015. The Challenges of Resolving a Rapid, Recent Radiation: Empirical and Simulated Phylogenomics of Philippine Shrews. Syst. Biol. 64:727–740.

Goodman B.A., Schwarzkopf L., Krockenberger A.K. 2013. Phenotypic Integration in Response to Incubation Environment Adaptively Influences Habitat Choice in a Tropical Lizard. Am. Nat. 182:666–673.

Horner P., Adams M. 2007. A Molecular-systematic Assessment of Species Boundaries in Australian Cryptoblepharus (Reptilia: Squamata: Scincidae): A Case Study for the Combined Use of Allozymes and Morphology to Explore Cryptic Biodiversity. The Beagle Suppl. 3:1-20.

Jones M.R., Good J.M. 2015. Targeted capture in evolutionary and ecological genomics. Mol. Ecol. 25:185–202.

Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

Mallet J. 2005. Hybridization as an invasion of the genome. Trends Ecol. Evol. 20:229–237.

Matzke N.J. 2014. Model Selection in Historical Biogeography Reveals that Founder-Event Speciation Is a Crucial Process in Island Clades. Syst. Biol. 63:951–970.

Mirarab S., Bayzid M.S., Warnow T. 2014. Evaluating Summary Methods for Multilocus Species Tree Estimation in the Presence of Incomplete Lineage Sorting. Syst. Biol. 65:366–380.

Ogilvie H.A., Heled J., Xie D., Drummond A.J. 2016. Computational Performance and Statistical Accuracy of *BEAST and Comparisons with Other Methods. Syst. Biol.

65:381–396.

Poelstra J.W., Vijay N., Bossu C.M., Lantz H., Ryll B., Muller I., et al. 2014. The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. Science. 344:1410–1414.

Rosenblum E.B., Sarver B.A.J., Brown J.W., Roches Des S., Hardwick K.M., Hether T.D., et al. 2012. Goldilocks Meets Santa Rosalia: An Ephemeral Speciation Model Explains Patterns of Diversification Across Time Scales. Evol Biol. 39:255–261.

Schluter D. 2000. The ecology of adaptive radiation. Oxford University Press, Oxford.

Singhal S. 2013. De novotranscriptomic analyses for non-model organisms: an evaluation of methods across a multi-species data set. Mol. Ecol. Res. 13:403–416.

Springer M.S., Gatesy J. 2015. The gene tree delusion. Mol. Phylogenet. Evol. 94:1-33.

Sukumaran J., Economo E.P., Lacey Knowles L. 2016. Machine Learning Biogeographic Processes from Biotic Patterns: A New Trait-Dependent Dispersal and Diversification Model with Model Choice By Simulation-Trained Discriminant Analysis. Syst. Biol. 65:525–545.

Zwickl D.J., Stein J.C., Wing R.A., Ware D., Sanderson M.J. 2014. Disentangling methodological and biological sources of gene tree discordance on oryza (poaceae) chromosome 3. Syst. Biol. 63:645–659.

Annex 1

EAPhy: A flexible tool for high-throughput quality filtering of exon-alignments and data processing for phylogenetic methods.

EAPhy: A flexible tool for high-throughput quality filtering of exon-alignments and data processing for phylogenetic methods.

Mozes P.K. Blom

Research School of Biology, The Australian National University, Canberra ACT 0200, Australia

ABSTRACT

Recently developed molecular methods enable geneticists to target and sequence thousands of orthologous loci and infer evolutionary relationships across the tree of life. Large numbers of genetic markers benefit species tree inference but visual inspection of alignment quality, as traditionally conducted, is challenging with thousands of loci. Furthermore, due to the impracticality of repeated visual inspection with alternative filtering criteria, the potential consequences of using datasets with different degrees of missing data remain nominally explored in empirical phylogenomic studies.

In this short communication, I describe a flexible high-throughput pipeline designed to assess alignment quality and filter exonic sequence data for subsequent inference. The stringency criteria for alignment quality and missing data can be adapted based on the expected level of sequence divergence. Each alignment is automatically evaluated based on the stringency criteria specified, significantly reducing the number of alignments that require visual inspection. By developing a rapid method for alignment filtering and quality assessment, the consistency of phylogenetic estimation based on exonic sequence alignments can be further explored across distinct inference methods, while accounting for different degrees of missing data.

INTRODUCTION

High-Throughput Sequencing (HTS) has revolutionised the field of phylogenetics by enabling researchers to question the evolutionary relationships between taxa with large-scale multi-locus datasets (McCormack *et al.* 2013; Yang and Rannala 2012). The development of these methods has been driven by a realisation that the inclusion of many genetic markers helps to account for stochastic coalescent histories of individual genes (Maddison 1997; Edwards 2009; Nakhleh 2013; Liu *et al.* 2014). Species tree inference methods use the multispecies coalescent model to estimate potential gene tree – species tree discordance and large numbers of unlinked loci represent a greater sample of the gene tree distribution underlying the true species tree (Liu *et al.* 2014). However, while phylogenetic estimation might improve by sequencing many loci (Edwards *et al.* 2007; Edwards 2009; Nakhleh 2013; Liu *et al.* 2014), the requirement for high-quality sequence alignments remains unchanged and is fundamental for the correct inference of phylogenetic hypotheses. Existing alignment methods can be extrapolated for use with large-scale multi-locus datasets, but visual inspection of each alignment, the traditional approach for assessing alignment quality, is challenging with thousands of sequenced loci (Lemmon and Lemmon 2013). As a consequence of the impracticality of visual inspection, the impact of missing data in large phylogenomic datasets is often nominally explored and the potential consequences of distinct alignment filtering criteria remain unknown. Nonetheless, contradicting opinions coexist (Roure et al. 2012; Lemmon et al. 2009; Wiens et al. 2011) regarding the effect of missing data on phylogenetic inference and it is therefore advisable to quantify the sensitivity of empirical phylogenetic hypotheses to data filtering choices. Thus we need workflows that automate (as far as possible) the assessment of alignment quality and the consequences (in terms of missing data) of making different choices about filtering criteria. Ideally, such a workflow would facilitate the conversion of individual

contiguous sequences ('contigs') into quality-filtered alignments, and help to minimise the demand for visual inspection.

The need for a high-throughput alignment filtering system emerged with the recent advance in molecular methods to target and sequence large numbers of orthologous loci. Since whole-genome sequencing is still too costly for most research labs that focus on non-model organisms, genome reduction protocols have been developed that isolate large numbers of orthologous loci across the genome of closely related and deeply divergent taxa (McCormack et al. 2013; Lemmon and Lemmon 2013). There are two increasingly popular genome reduction methods that specifically focus on exonic sequence regions and can generate genetic markers suitable for phylogenomic inference. Transcriptome sequencing is a cost-effective method that does not require the a-priori availability of genomic resources. RNA is extracted from the same tissue in different target species and with the expected expression of similar genes, orthologous loci are isolated and sequenced for phylogenetic comparison. An alternative method, exon-capture (Hodges et al. 2007), is a target enrichment approach (Faircloth et al. 2012; Lemmon et al. 2012; Bi et al. 2012; Penalba et al. 2014) that benefits from an increasing number of readily available genomic resources and enables the design of study-specific capture systems. The use of exonic sequence regions for phylogenetic inference, generated by transcriptome sequencing or exon-capture, is promising and has been successfully demonstrated at different levels of divergence across the tree of life (Bi et al. 2013; Ilves et al. 2014; Misof et al. 2014; Hugall et al. 2015). The tremendous increase in the scale of available exonic loci benefits inference methods, but also requires a significant investment in the development of bioinformatic resources to process such data.

Whereas several excellent bioinformatic pipelines have been constructed for processing raw sequence data and conducting sequence assembly (Faircloth et al.

2012; Lemmon et al. 2012; Bi et al. 2012; Penalba et al. 2014; Misof et al. 2014), a bioinformatic scheme is needed for subsequent alignment, alignment quality assessment and alignment filtering. Most published studies still conduct visual inspection of alignment quality and account for missing data by dividing datasets into a limited number of categories manually (i.e. Lemmon et al. 2012; Ilves and López-Fernández 2014) or automated (i.e. Crawford et al. 2014). Recently, Misof et al. developed a method to assess alignment quality in an extensive study that used transcriptomes to infer the phylogeny of insects. They identified potentially erroneous alignments by calculating the BLOSUM62 distance between each amino acid sequence and the best reciprocal hit of a reference taxon. A distance calculation based on a BLOSUM matrix was warranted, due to a significant level of protein divergence between most taxa. The BLOSUM alignment score matrix values the alignment of each amino acid pair differently, representing the likelihood of amino acid substitutions, but lacks resolution when the expected level of protein divergence between two sequences is limited. However, although this has not been tested prior, it can be expected that at shallower levels of divergence subtle misalignments might actually have more significant consequences for phylogenetic estimation than when inferring relationships between distant taxa, stressing the need to identify such misalignments. When assessing alignments with limited levels of sequence divergence, the exact number of clustered amino acid changes is more likely a better indicator of alignment quality than the overall BLOSUM62 distance score. In this short communication, I describe a flexible high-throughput pipeline for quality assessment of exonic sequence alignments and subsequent filtering of missing data. The pipeline is specifically designed to be flexible and process both population and phylogenetic level data, but the method developed by Misof et al. will likely be more effective at deep phylogenetic scales.

EAPhy, exon alignment for phylogenetics, was developed to process exonic sequence data for phylogenetic inference, but is valuable for any type of analysis that requires high-quality filtered alignments (i.e. population genomics or molecular evolution). In this manuscript I will focus on its application for phylogenetic inference. The first objective of the pipeline is to quantify alignment quality and highlight just those loci that require visual inspection. By translating exonic nucleotide alignments into amino acid alignments, EAPhy infers the relative quality of sequences and alignments by assuming that most mutations within exons are silent. In addition, the identification of regions that harbor an excessive cluster of amino acid replacements distinct from a summary reference sequence, is used as a proxy for alignment quality. Simultaneously, insertions and deletions that result in frame shifts and the introduction of multiple stop-codons are unlikely to represent true biological events and such alignments should be addressed. The pipeline can be adapted based on the expected level of divergence between taxa by adjusting the stringency of filtering criteria. The second objective is to provide a user-friendly method to account for missing data. By enabling filtering criteria for missing data to vary, the consistency of phylogenetic estimation can be quantified across different levels of missing data. Lastly, EAPhy was designed to generate alignments of different sorts (haplotype, diplotype and SNP based), in the formats required for most commonly used inference software and facilitate the further exploration of distinct analysis methods. With the development of a high-throughput method for alignment filtering and processing, the overarching aim of this pipeline is to reduce the bioinformatic burden of data analysis involving exonic sequence alignments and ultimately promote further research into the (in)congruences between inference methods, while accounting for different degrees of missing data.

OVERVIEW OF METHODS

EAPhy consists of a collection of scripts that takes as input a set of unaligned sequences for an arbitrary number of species and loci. It will generate multiple sequence alignments using existing aligning software and subsequently filters these alignments for a number of user-specified criteria. The final output consists of quality filtered multiple sequence alignments, allowing different degrees of missing data as preferred, and a list of alignments that still require visual inspection. The output files are automatically exported in the input format of commonly used phylogenetic inference programs. The complete package is freely available at https://github.com/MozesBlom/EAPhy.

EAPhy can be run on most individual computers (i.e. does not require a cluster set-up) and an individual run for a modest dataset can be completed within hours. The pipeline has been used with exon-capture datasets involving tens to hundreds of individuals and thousands of loci, and finished within six hours on a Macintosh desktop computer with a 3.1 GHz Intel i7 processor (2012) and 16 GB of RAM. It is important for the user to adopt filtering criteria suitable for the dataset (level of divergence and data quality) analysed, but if filtering criteria have been carefully reviewed EAPhy should be able to handle larger datasets than currently tested. For EAPhy to function appropriately, I advise to run the pipeline initially with a small subset of the data and replicate the analysis with alternative filtering criteria. If filtering and flagging of alignments works well, then the analysis can be extrapolated for usage with the complete dataset. The importance of specifying appropriate filtering criteria should not be underestimated, since misspecification of filtering criteria will result in a significantly reduced dataset or alternatively a dataset that equals the input data, regardless of potential low-quality alignments.

EAPhy is not designed to identify individual sequencing errors that are often associated with HTS datasets, but will identify sequence regions with excessive non-synonymous substitutions (potential 'low-coverage' sequences) if these have not been filtered out beforehand and appear anomalous in the resulting alignments. Several excellent pipelines have been developed to filter raw sequence data and generate assemblies, and the starting point of this pipeline requires assembled individual contigs that start in first codon frame, for each presumed orthologous locus. A complete overview of the pipeline is outlined in Figure 1 and a general description of the most important components is provided here.

Specification of configuration script

At the onset of each EAPhy run, the system path to an align program executable and all filtering criteria for downstream analysis are specified in a single configuration script. Muscle (Edgar 2004) is the default aligner used by EAPhy, but should be installed by the user independently from downloading EAPhy. Alternative alignment software can be used but requires modifications of several scripts. The EAPhy pipeline is designed as a set of modules that can be executed independently or in consecutive order as a complete analysis (Fig. 1). This provides a straightforward system to reiterate specific components of the pipeline, with alternative filtering criteria for alignment quality or missing data. A complete description of all filtering parameters can be found in the manual and is part of the EAPhy package that can be downloaded from GitHub.

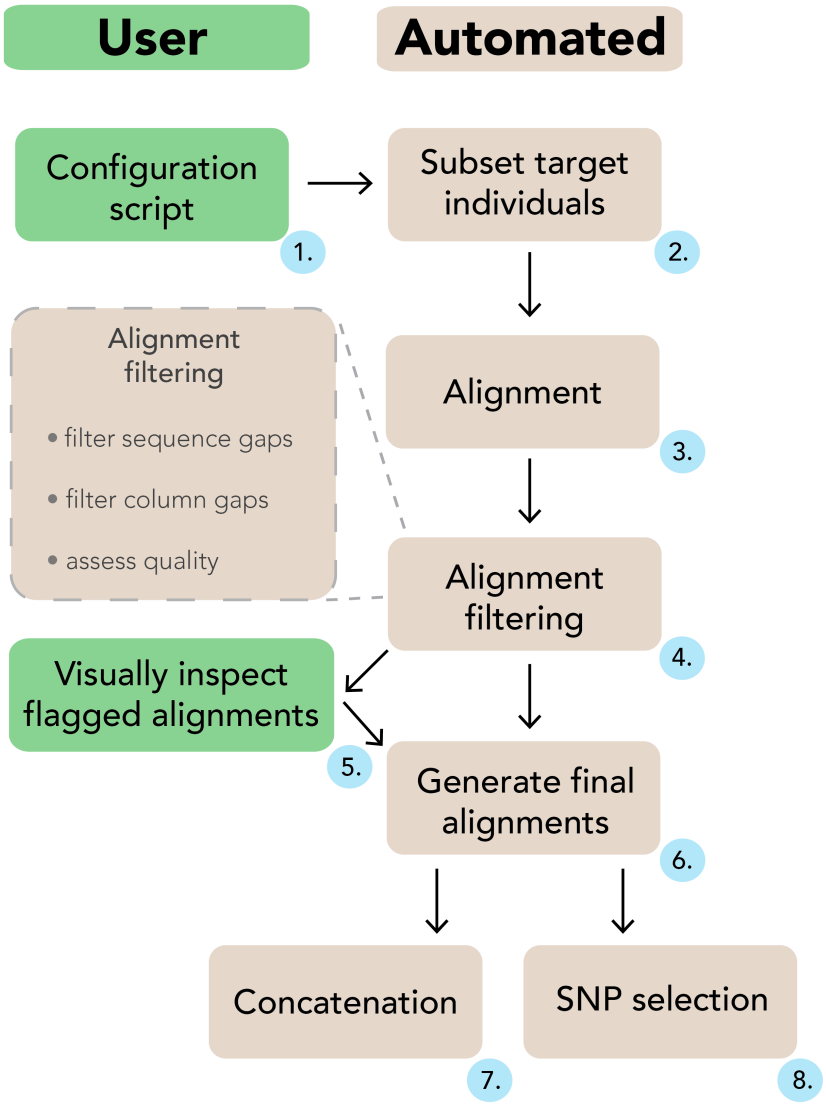


Fig. 1: A schematic overview of the EAPhy workflow. The user specifies filtering instructions in a single configuration script (1). EAPhy will first subset the target individuals from each locus and create new contig files (2). With an existing aligner, new alignments are created for each locus (3) and are subsequently processed and checked (4). Alignments are highlighted that do not fulfill the filtering criteria and can be visually inspected. If deemed appropriate for inclusion, manually checked alignments can be added to the filtered alignment list (5). Alternatively, EAPhy can automatically continue with the alignments that passed filtering and exclude the problematic loci. Once the complete collection of filtered loci has been identified, final alignments are generated (6). If diplotype sequence data was used and heterozygous positions coded according to IUPAC format, concatenated (7) and SNP (8) alignments can be generated if required.

Missing data – within alignments

The effect of missing data on phylogenetic inference is not well understood and contradicting opinions coexist (Lemmon and Lemmon 2009; Roure et al. 2012; Wiens and Morrill 2011). Phylogenetic estimation is likely unbiased with large numbers of loci, if there are no systematic differences in sequence length between individuals for any given locus. However, the maximum-likelihood (ML) estimation might cluster individuals by sequence length rather than sequence similarity for complete positions, when missing data are non-randomly distributed and specific individuals have systematically shorter contigs or are completely missing for specific loci. At sites with missing data, the probability of observing an 'A', 'T', 'C' or 'G', is set to 1 and ML will group taxa together for which there is more signal and less uncertainty (Stamatakis, pers. comm.). Thus the effect of missing data is not limited to small sequence datasets but should also be accounted for and characterized in large-scale datasets. With the development of EAPhy, I do not advocate to discard or include incomplete sites but rather provide the opportunity to account for missing data by generating datasets where different filtering criteria have been enforced.

Missing data within individual sequences are particularly prevalent at the beginning and end of alignments ('jagged edges'), since individual contig sequences often differ in length. Once alignments have been constructed using an existing aligner (e.g. Muscle; Edgar 2004), EAPhy will first address missing data by processing alignments in accordance with stringency criteria specified in the configuration script (Fig. 2). First, potential gaps within individual sequences are removed to yield long consecutive sequences (Fig. 2.1). EAPhy converts all sequence alignments into amino acid alignments and then uses a 'jump-sliding window' approach to assess the presence of potential non-consecutive sequence stretches that are often prevalent at the start/end of individual sequences. A jump-sliding window approach was

developed since a conventional sliding window approach would remove the complete individual sequence if the first window would contain more missing data than allowed. Each window is assessed on the presence of amino acid sequence gaps and if a window contains more gaps than allowed, the complete window is removed for that individual sequence. In-frame gaps (i.e. triplet insertions) are retained if the amount of inserted codon gaps per window does not contain more missing codons than allowed. The window then 'jumps' a sequence distance of half the window size plus one codon and the process reiterates. By converting a nucleotide alignment in its amino acid equivalent, EAPhy specifically takes into account the coding-codon character of exonic sequences. When nucleotide sequence data is removed by codon, the remaining sequence is still in correct frame and codon position can still be inferred for each nucleotide position.

After individual sequences have been trimmed for missing data, EAPhy then assesses missing data between individuals by evaluating the amount of missing data for each amino acid alignment column (Fig. 2.2). The algorithm used is similar to the jump-sliding window approach, but now focuses on the amount of missing data within each amino acid alignment column. The window-length of amino acid columns and the amount of missing data allowed within each column, can be specified in the configuration script. The algorithm evaluates for each amino acid column whether the amount of individuals with missing data exceeds the cut-off specified. If more than half of the columns in a given window have more missing data than allowed, the columns in the first half of the window are removed from the alignment. The window then 'jumps' a sequence distance of half the window size plus one codon and the process reiterates. Amino acid columns at the end of alignments are removed, if they have not been evaluated but the specified window length exceeds the number of remaining columns. When alignments have been filtered for missing data within and between individuals, EAPhy evaluates the presence of single nucleotide insertions, by assessing

the frequency of sequenced individuals for each nucleotide alignment column. If the number of sequenced individuals is below a user specified cut-off, the site is assumed to be a sequencing error and removed from the alignment.

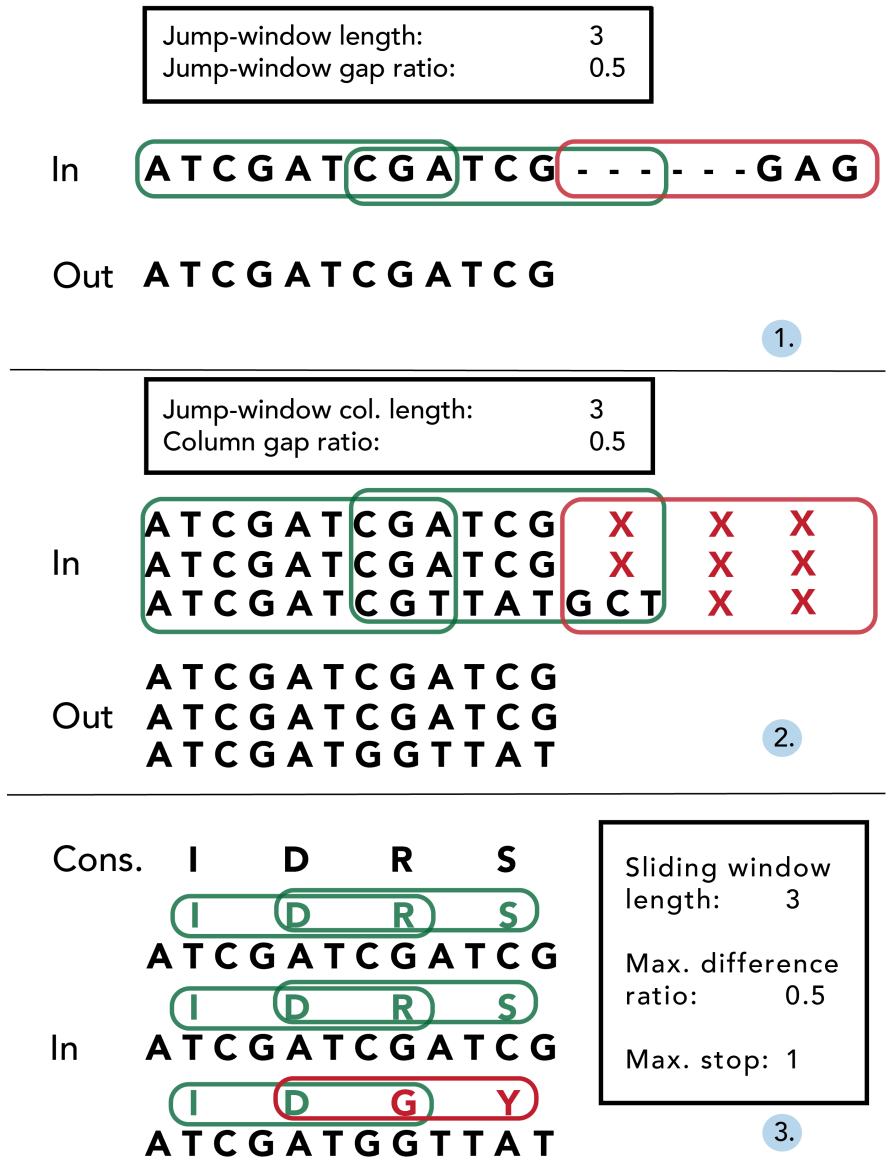


Fig. 2: An exemplary overview of the three main filtering steps conducted during alignment filtering and quality assessment. First, potential gaps within individual sequences are removed to yield long consecutive sequences (1). By converting nucleotide codons into amino-acids, the number of missing amino acids is assessed for each window using a 'jump-sliding-window' approach. If less amino-acids are missing for a given window than the specified 'jump-window gap ratio', the complete corresponding nucleotide stretch is retained (see green frames). If more amino acids are missing for a window, the complete nucleotide stretch is

removed (see red frame). Secondly, the amount of missing data for each amino-acid alignment column in a given window is quantified (2). If more than half of the amino acid columns in a window miss more individuals than the specified 'column gap ratio', all corresponding nucleotide columns are removed (see red frame). Lastly, the quality of the resulting alignment is assessed, by comparing each individual sequence to a consensus sequence (3). Following a common sliding window approach for each individual sequence, the number of amino acids identical to the consensus is quantified. If, for each window, the number of amino acids distinct is less than the specified 'difference ratio', the alignment is retained (see green frames). If for any individual within an alignment, a window would fail this criterion, the alignment is flagged for visual inspection. In addition, for each sequence the number of stop-codons is quantified and if any individual sequence contains more stop-codons than a specified cut-off number (e.g. > 1), the alignment is also flagged for visual inspection.

Alignment quality

Once each alignment has been filtered for missing data, EAPhy then inspects the alignment quality by translating the nucleotide sequences and evaluating the resulting amino acid alignment (Fig. 2.3). First, if the number of stop codons for any individual sequence exceeds a user specified cut-off value (e.g. > 1), the alignment is flagged for visual inspection. Subsequently, a general consensus sequence is estimated for each alignment, and each individual sequence is compared to the consensus sequence in a 'normal' sliding-window approach. The window length is specified in the configuration script and each individual sequence is compared to the consensus sequence by sliding window. For each window, the number of amino acids distinct from the consensus is quantified and if greater than the proportion specified in the configuration script, the alignment is flagged for visual inspection.

Finally, phylogenetic inference is dependent on the comparison of orthologous genetic markers and comparing potential paralogous loci might yield confounded estimates of relationship. EAPhy assumes that the sequenced contigs for each locus are orthologous but has an additional option to potentially identify paralogous loci, by

identifying markers with excessive levels of average individual heterozygosity. The user can inspect the distribution of average individual heterozygosity across all loci and based on this observation make an informed decision whether to exclude a certain percentage of loci with the highest level of average individual heterozygosity.

Concatenation and SNP selection

After visual inspection and filtering of flagged alignments, the collection of final high quality alignments can then be used for a variety of phylogenetic estimation methods. Gene trees can be inferred based on single alignments and a concatenated maximum likelihood tree can be estimated based on all alignments combined. Since all alignment filtering was conducted by codon, each nucleotide can still be assigned its correct codon position. PartitionFinder (Lanfear et al. 2012) estimates the most optimal partitioning scheme across all sequence positions and appropriate substitution model for each partition. A PartitionFinder input file is automatically created with each gene and codon position of the concatenated alignment specified.

In addition to sequence-based alignments, EAPhy will also generate concatenated alignments that include polymorphic sites exclusively. SNAPP (Bryant et al. 2012) is a species tree method that uses unlinked biallelic markers, instead of sequence-based alignments, and EAPhy can generate alignments with a biallelic SNP randomly sampled from each locus. It will verify whether polymorphic sites are biallelic and neglect polymorphic sites with more than two allelic states. Alternatively, SNP alignments can be constructed where every single SNP is considered, regardless of allele count, or with all SNP's across all loci concatenated. If a study is geared towards recovering population structure, such alignments can be used in analyses that model allele frequencies (e.g. Pritchard et al. 2000).

Missing data – number of sequenced individuals

Sequencing success can vary among individual samples. If specific individuals are systematically underrepresented and miss data for many loci, it is possible that the phylogenetic placement of such taxa is ambiguous and the investigator would prefer to exclude these samples. Thus, the potential impact of missing individuals across loci should be accounted for. EAPhy attempts to highlight where this is likely by: a) providing alternative datasets with different numbers of missing individuals allowed and b) providing summary statistic output files quantifying the number of loci sequenced for each individual. This enables the investigator to further explore the potential effects of missing data on phylogenetic inference.

IN SUMMARY

The first objective of developing EAPhy was to provide a flexible and rigorous tool to generate reliable alignments, while minimizing the need for extensive visual inspection. Secondly, EAPhy was designed to allow filtering criteria for missing data to vary and investigate the impact of missing data on phylogenetic estimation. Lastly, EAPhy creates a large number of desired input formats for subsequent analysis, enabling the exploration of distinct inference methods. Negating the effort of manual alignment filtering and processing, EAPhy will hopefully stimulate further research into the potential consequences of applying alternative criteria for missing data and datatype, and how this might ultimately result in (in)congruent estimates of phylogenetic relationships across methods. The simultaneous development of novel molecular approaches to sequence orthologous genetic markers and bioinformatic

methods to analyze such data, will ultimately provide us with the tools to generate a phylogenetic framework for all taxa across the tree of life.

ACKNOWLEDGEMENTS

All scripts are written in Python and throughout the pipeline, many functions benefit from modules that have been developed as part of the excellent BioPython distribution (http://biopython.org/wiki/Main_Page). I would like to thank Craig Moritz and Lisa Schwanz for providing advice and support during the development of EAPhy. I also would like to thank the members of the Moritz' lab but in particular Ana-Catarina Silva for providing helpful suggestions to improve the pipeline and Jason Bragg for his valuable contribution to improve this manuscript.

FUNDING STATEMENT

Mozes Blom is supported by the Australian Research Council with a graduate student fellowship, as part of the ARC Laureate grant (FL110100104) awarded to Craig Moritz.

REFERENCES

Bi K, Linderoth T, Vanderpool D, Good JM, Nielsen R, Moritz C. 2013. Unlocking the vault: next-generation museum population genomics. Mol. Ecol. 22: 6018–6032.

Bi K, Vanderpool D, Singhal S, Linderoth T, Moritz C, Good JM. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics. 13: 403

Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, RoyChoudhury A. 2012. Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. Mol. Biol. Evol. 29: 1917–1932.

Crawford NG, Parham JF, Sellas AB. 2014. A phylogenomic analysis of turtles. Mol. Phylogenet. Evol. 83: 250-257.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32: 1792–1797.

Edwards SV, Liu L, Pearl DK. 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. USA. 104: 5936–5941.

Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? Evolution. 63: 1–19.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61: 717–726.

Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, et al. 2007. Genome-wide in situ exon capture for selective resequencing. Nat. Genet. 39: 1522–1527.

Hugall AF, Ohara TD, Hunjan S, Nilsen R, Moussalli A. 2016. An exon-capture system for the entire class Ophiuroidea. Mol. Biol. Evol. 33: 281–294.

Ilves KL, López-Fernández H. A targeted next-generation sequencing toolkit for exon-based cichlid phylogenomics. Mol. Ecol. Res. 14: 802–811.

Lanfear R, Calcott B, Ho SYW, Guindon S. 2012. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29: 1695–1701.

Lemmon AR, Brown JM, Stanger-Hall K, Moriarty-Lemmon E. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and bayesian inference. Syst. Biol. 58: 130-145

Lemmon AR, Emme SA, Lemmon EM. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61: 727–744.

Lemmon EM, Lemmon AR. 2013. High-throughput genomic data in systematics and phylogenetics. Annu. Rev. Ecol. Evol. Syst. 44: 99–121.

Liu L, Xi Z, Davis C, Edwards SV. 2015. Estimating phylogenetic trees from genome-scale data. Ann. N. Y. Acad. Sci. 1360:36-53

Maddison WP. 1997. Gene trees in species trees. Syst. Biol. 46: 523-536.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol. Phylogenet. Evol. 66: 526-538.

Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science. 346: 763–767.

Nakhleh L. 2013 Computational approaches to species phylogeny inference and gene tree reconciliation. Trends Ecol. Evol. 28: 10.1016/j.tree.2013.09.004.

Penalba JV, Smith LL, Tonione MA, Sass C, Hykin SM, Skipwith PL, et al. 2014. Sequence capture using PCR-generated probes: a cost-effective method of targeted high-throughput sequencing for nonmodel organisms. Mol. Ecol. Res. 14: 1000–1010.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics. 155: 945–959.

Roure B, Baurain D, Philippe H. 2012. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol. Biol. Evol. 30: 197–214.

Wiens JJ, Morrill MC. 2011. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. Syst. Biol. 60: 719-731.

Yang Z, Rannala B. 2012. Molecular phylogenetics: principles and practice. Nature Rev. Genet. 13: 303–314.

Annex 2

Habitat use and new locality records for *Cryptoblepharus poecilopleurus* (Squamata: Scincidae) from French Polynesia

Herpetology Notes, volume 8: 579-582 (2015) (published online on 06 December 2015)

Habitat use and new locality records for *Cryptoblepharus poecilopleurus* (Squamata: Scincidae) from French Polynesia

Mozes P.K. Blom

Cryptoblepharus Wiegmann is the most geographically widespread taxon in the family Scincidae. They occur within the Ethiopian-Malagasy region, on the Indo-Australian continent and on many Pacific islands of both continental and volcanic origin (Ineich and Blanc, 1988; Rocha et al., 2006; Horner, 2007; Hayashi et al., 2009). Although the broad distribution of *Cryptoblepharus* suggests an ecological generalist, Greer (1989) proposed that these small (<55 mm snout-vent length), heliotropic and diurnal lizards, are actually adapted to a narrow set of ecological parameters or microhabitats. Horner (2007) developed this idea further in his comprehensive overview of the genus and distinguished between species occurring on littoral ('beach'), arboreal ('tree') or saxicolous ('rock') substrates.

Littoral species of *Cryptoblepharus* have been characterized as beach-dwelling, intertidal specialists and species have even been observed consuming small crustaceans and polychaetes (Fricke, 1970; Horner, 1984). These coastal lineages are of particular interest since littoral populations might potentially be the source of the vast distribution of the genus and other habitat specialists might derive from such littoral dispersers (i.e. a taxon cycle model (Ricklefs and Bermingham, 2002)). Among all littoral species, *Cryptoblepharus poecilopleurus* has the most extensive distribution, including the majority of islands in the Pacific (Ineich and Blanc, 1988; Horner, 2007). Within the French Polynesian region, these skinks have been recorded on islands that differ in both species richness and habitat complexity (Ineich and Blanc, 1988), providing a promising scenario to examine whether herpetofaunal diversity might induce or reduce alternative habitat use.

Accordingly, I visited eight islands (Fig. 1) across French Polynesia from May 18 until June 19, 2015, to verify species presence and record habitat use of *C. poecilopleurus*. On each island, I walked transects along the beach and where possible further inland, in the morning and afternoon and collected lizards by hand. On transects, I occasionally stripped away long pieces of bark from *Casuarina* trees to spot hiding skinks. This was particularly effective for finding *Lipinia noctua*, but did not reveal many *Cryptoblepharus*.

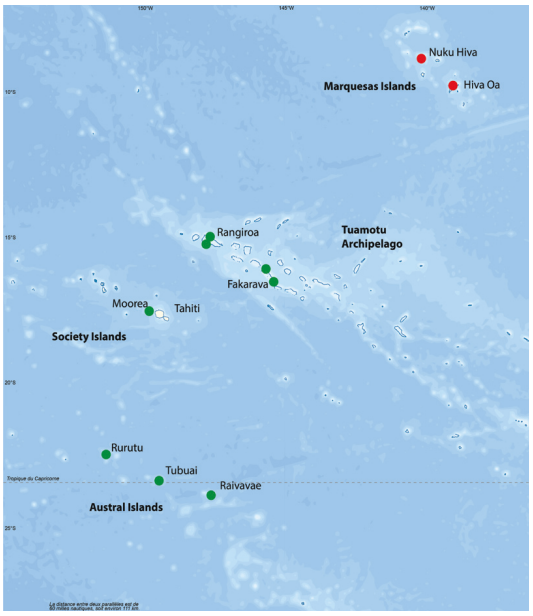


Figure 1. The islands visited in this study are highlighted with a coloured dot. Green dots represent islands where *C. poecilopleurus* have been collected and red dots islands that were visited, but no *C. poecilopleurus* found. The Austral island of Raivavae, is a new locality record, from which *C. poecilopleurus* has not been observed prior.

Research School of Biology, The Australian National University, Canberra ACT 0200, Australia
Correspondence: mozes.blom@gmail.com

Table 1. Main locality and observation records for *C. poecilopleurus* on each visited island. ([1] Two sympatric colour morphs observed, * 1–2 individuals observed in 1 hour search, ** 3–5 individuals observed in 1 hour search, *** more than 5 individuals observed in 1 hour search).

Island	Locality	Abundance	Habitat	Sympatry	Latitude	Longitude
Moorea	Motu Fareone	*	Arboreal	*L. noctua*	-17.491	-149.918
Raivavae	Motu Rani	***	Arboreal	*Emoia sp.*	-23.878	-147.608
Raivavae	Men's rock	**	Saxicolous	*Emoia sp.*	-23.855	-147.66
Tubuai	Mateauta	***	Arboreal	-	-23.344	-149.481
Tubuai	Anua	***	Arboreal	-	-23.377	-149.527
Tubuai	Mateauta harbour	**	Saxicolous	-	-23.344	-149.478
Rurutu	1 km. south airport	***	Arboreal	*Emoia sp.*	-22.459	-151.373
Rurutu	Hidden beach	**	Littoral	*Emoia sp.*	-22.461	-151.373
Rangiroa	Tapuheitini	***	Arboreal	*Emoia sp.*	-14.948	-147.686
Rangiroa	Motu Paati	***	Arboreal	*Emoia sp.*	-15.238	-147.716
Fakarava	Vaiama Village	***	Arboreal	*L. noctua*	-16.114	-145.605
Fakarava[1]	Raimiti	***	Arboreal	-	-16.427	-145.376
Nuku Hiva	-	-	-	-	-	-
Hiva Oa	-	-	-	-	-	-

I recorded *C. poecilopleurus* on six out of eight islands, including a new island record (Raivavae), and they were usually common (<10 individuals spotted within an hour, see Table 1). Whereas littoral habitat use has been extensively described on other islands (Canaris and Murphy, 1965; Canaris, 1973; Fricke, 1970; Horner, 1984), I only observed three populations of *C. poecilopleurus* on a non-arboreal substrate and found none that could be described as beach-dwelling intertidal specialists. Even though populations were always observed less than 50 metres from the coast, most *C. poecilopleurus* individuals were found on trees of the genus *Casuarina* or *Cocos,* adjacent to the beach and not on the beach itself.

Deducing from over 100 observations, the behaviour of these arboreal populations is reminiscent of *Cryptoblepharus* in Australia that occur on the same substrate (pers. obs.). Upon approach, these skinks circumvent the tree and escape by moving upwards. They were active during daytime, if the sun was sufficiently bright (i.e. no cloud cover) and mostly present in patches of full sun light. Areas that had been cleared by human activity, but where large trees remained, yielded high numbers of individuals and each tree sometimes harboured more than one individual. This could have been an observation bias, since it is likely harder to approach and spot lizards in dense vegetation, but these lizards seem to strongly favour brightly illuminated spots, which could be an indication of thermal requirements.

I only recorded three *Cryptoblepharus* populations that used non-arboreal habitats and in these cases, they did not exhibit previously described littoral behaviour. The first observation was on the Austral island of Tubuai, where five individuals were spotted on large dark boulders that were placed as support for the local harbour. The second observation was on a large limestone rock-face (Raivavae), where approximately three individuals scurried between the crevices and ascended, seemingly with ease, a vertical wall. The last observation was on the island of Rurutu, where eight individuals were spotted between vegetation and debris on a pebble beach (Fig. 2a). They remained within close distance to the vegetation adjacent to the beach and were not observed within 20 meters of the shoreline.

Intraspecific aggression was incidentally observed, with two individuals chasing each other across the base of a tree. This behaviour has been recorded prior (Horner, 2007) but it remains unclear whether this is territorial or sex-specific, since *C. poecilopleurus* individuals were also frequently observed in close proximity without any form of aggressive behaviour (Fig. 2b).

Figure 2. A) *C. poecilopleurus* population with (semi-) littoral habitat use. Individual photographed at a pebble beach on the island of Rurutu. B) Two *C. poecilopleurus* individuals observed on arboreal substrate, several *Emoia sp*. were observed on the same tree (not visible on picture).

Three other skink species were observed on these surveys (Table 1). Of these, *Lipinia noctua* tends to occur under bark and be quite secretive. *Emoia cyanura*, uses open habitat and has a high thermal preference seemingly similar to *C. poecilopleurus*. Finally, *Emoia impar* can be found further inland as well, in areas with closed canopy cover and has a lower thermal preference (McElroy, 2014). I did not observe any direct indication for interspecific competition or displacement between skink genera, contrasting previous reports from the Solomon islands (McCoy, 2006) or as observed in Pacific geckos (Case et al., 1994). On the islands surveyed, skinks of the genus *Emoia* tended to be ground-dwelling, but were sometimes also observed on the same trunk as *Cryptoblepharus* without exhibiting any form of belligerent behaviour. Given these observations, displacement does not seem directly evident. Thus it is remarkable that *C. poecilopleurus* populations were only found on trees adjacent to the beach and not further inland (the most inland record was a single juvenile on a *Casuarina* tree 50 m. from the coast). There did not seem to be a clear change in species diversity or richness on trees further inland, except for the absence of *Cryptoblepharus*.

Specialization in littoral beach dwelling as described for other species of *Cryptoblepharus* (Fricke, 1970; Horner, 1984) was not observed on the Polynesian islands visited. Most *C. poecilopleurus* populations recorded, were using an arboreal substrate and often shared trees with other skink species. Nonetheless, all populations were observed on trees in close vicinity to the beach. Given these observations it remains unclear what processes limit the distribution and habitat use of these skinks. If interspecific competition, as recorded on the Solomon islands (McCoy, 2006), does not limit expansion from trees adjacent to the beach to trees further inland, other potential limitations should be considered. One explanation could be that these skinks are potentially restricted in their thermal requirements and require an open habitat with sufficient exposure to bright light. Denser vegetation inland might not provide ample opportunity to optimally thermoregulate, but further research is required.

Acknowledgements. I thank the National Geographic Society, the Mohamed bin Zayed Species Conservation Fund, the Richard B. Gump Station, Jean-Yves Meyer and Craig Moritz for supporting this work. I thank Matthew Fujita and Foteini Spagopoulou for help during the fieldwork and Paul Oliver for helpful suggestions to improve this manuscript.

References

Case, T.J., Bolger, D.T., Petren, K. (1994): Invasions and competitive displacement among house geckos in the tropical Pacific. Ecology 75: 464–477.

Canaris, A.G., Murphy, D.G. (1965): A scincid reptile feeding primarily on marine Crustacea, with a note on its parasites. Journal of the East Africa Natural History Society 25: 129–130.

Canaris, A.G. (1973): Parasites and food habits of a littorial feeding lizard (*Ablepharus*, Scincidae). Copeia 2: 345–346.

Fricke, H.W. (1970): Die ökologische spezialisierung der eidechse *Cryptoblepharus boutonii cognatus* (Boettger) auf das leben in der gezeitenzone (Reptilia, Scincidae). Oecologia 5: 380–391.

Greer, A.E. (1989): The biology and evolution of Australian lizards. Surrey Beatty and Sons: Chipping Norton.

Hayashi, F., Shima, A., Horikoshi, K., Kawakami, K., Segawa, R.D., Aotsuka, T., Suzuki, T. (2009): Limited overwater dispersal and genetic differentiation of the Snake-Eyed Skink (*Cryptoblepharus nigropunctatus*) in the oceanic Ogasawara Islands, Japan. Zoological Science 26: 543–549.

Horner, P. (1984): Notes on the scincid lizard *Cryptoblepharus litoralis* (Mertens 1958) in the Northern Territory. Northern Territory Naturalist 7: 4–7.

Horner, P. (2007): Systematics of the snake-eyed skinks, *Cryptoblepharus* Wiegmann (Reptilia: Squamata: Scincidae) – an Australian-based review. The Beagle Records of the Museums and Art Galleries of the Northern Territory supplement 3: 21–198.

Ineich, I., Blanc, C.P. (1988): Distribution des reptiles terrestres en Polynésie Orientale. Atoll Research Bulletin 318: 1–75.

McCoy, M. (2006): Reptiles of the Solomon Islands, 2nd edition. Pensoft Publishing.

McElroy, M.T. (2014): Countergradient variation in locomotor performance of two sympatric Polynesian skinks (*Emoia impar, Emoia cyanura*). Physiological and Biochemical Zoology 87: 222–230.

Ricklefs, R.E., Bermingham, E. (2002): The concept of taxon cycle in biogeography. Global Ecology & Biogeography 11: 353–361.

Rocha, S., Carretero, M.A., Vences, M., Glaw, F., Harris, D.J. (2006): Deciphering patterns of transoceanic dispersal: the evolutionary origin and biogeography of coastal lizards (*Cryptoblepharus*) in the Western Indian Ocean region. Journal of Biogeography 33: 13–22.

Accepted by Paul Oliver