Title
A Fine Scale Analysis of a Tropical Suture Zone

Author
Singhal, Sonal

Publication Date
2013

Peer reviewed|Thesis/dissertation

A Fine Scale Analysis of a Tropical Suture Zone

by

Sonal Singhal

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Integrative Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Craig Moritz, Chair
Professor Michael Eisen
Professor Rosemary Gillespie
Professor Montgomery Slatkin

Spring 2013

A Fine Scale Analysis of a Tropical Suture Zone

Abstract

A Fine Scale Analysis of a Tropical Suture Zone

by

Sonal Singhal

Doctor of Philosophy in Integrative Biology

University of California, Berkeley

Professor Craig Moritz, Chair

In my dissertation, I use a comparative approach to exploit an outstanding natural experiment, a suture zone in the rainforest of northeast Australia, to answer questions about speciation and hybridization. The suture zone consists of twenty identified contacts between phylogeographic lineages, mostly within morphologically defined species. Although the contacts in the zone likely formed concurrently in response to Holocene expansion from glacial refugia, the lineage-pairs meeting in the contacts exhibit a wide range of genetic divergences. This natural variation enables analysis of the outcomes of secondary contact at different stages of the divergence process. Importantly, although most studies of speciation focus on lineages that show marked phenotypic divergence, I focus on morphologically cryptic lineages, which, though common in nature, have been understudied in this regard. Through my dissertation, I consider contact zones between six lineage-pairs within four morphologically-defined skink species, *Carlia rubrigularis*, *Lampropholis coggeri*, *Saproscincus basiliscus*, and *S. lewisi*.

Through this work, I find support for the reality of cryptic species and argue that the presence of cryptic species can suggest a wider plurality of speciation models than we typically consider (Chapter 1). Indeed, by combining multilocus methods and dense sampling, I find that reproductive isolation between phylogeographic lineages scales tightly with divergence (Chapter 2). These results support the widespread, yet previously unsubstantiated, notion that phylogeographic structure of increasing depth represents a continuum towards complete speciation, even in the absence of overt ecologically-driven divergent selection. I extend these results by looking at introgression across the genome, finding that genome-wide selection, driven by selection against hybrids, structures introgression patterns much more strongly than locus-specific selection histories (Chapter 3). By analyzing the sole lineage-pair in this system that exhibits genealogical discordance, I suggest that geographic stability across time is key to driving divergence (Chapter 4). Further, through combining a fine-scale investigation of a single contact zone with simulations and a meta-analysis, I argue that selection against hybrids, in the form of intrinsic genetic incompatibilities, maintains species boundaries at these contact zones (Chapter

5). Finally, as my work is enabled by emerging genomic technologies for non-model organisms, I summarize my genomics approach, and its associated benefits and challenges, as applied to transcriptome data from these lineages (Chapter 6).

To Ma, Papa and Didi
for being my guiding stars and my biggest fans

lizard tails spin tales:
crypsis hides histories of
evolving species

Contents

List of Figures

List of Tables

Acknowledgments

To my advisor – Craig Moritz, what can I say? The first time we talked, before I was even a graduate student, our conversation was so exciting and so engaging that I was on an intellectual high for a week. How lucky that these conversations continued over the next six years! Thank you for being an ever-constant source of keen questioning, motivation, and enthusiasm. Thank you for suggesting that I do impossible things and then giving me the support that I needed to complete them. But, most of all, thank you for your good example – I am continuously inspired by the kindness, humility and humor with which you live your life.

To my other advisors, both formal and informal – I am grateful for the intelligence and generosity of the faculty who helped train me. I sincerely thank Rauri Bowie, Mike Eisen, Rosie Gillespie and Monty Slatkin for challenging me to read broadly and think deeply during my qualifying exam; your guidance made preparing for the dreaded orals enjoyable. Thank you to Mike, Rosie and Monty for continuing as my thesis committee – your perceptive questions helped me refine my research aims, and your constant support motivated me when I was frustrated. I would especially like to thank those who helped me (even though no paperwork compelled them to!) – Jim Patton inspired me continuously with his humility and unflailing high standards and David Wake challenged me with broad-ranging and quickly moving conversations about evolution, nature, and science & society.

To my teachers before them – your classrooms provided some of my earliest happy memories. A special thanks to Mrs. Worst for teaching me what really matters, Mr. Lavelle for showing me the magic in genetics, Mr. Ballauer for teaching me about the wonder in the scientific world, and Drs. Barbara Schaal and Yu-Chung Chiang for training me to think like a scientist.

To my peers in the MVZ and IB – I am honored to be part of such an amazing community of scholars and friends. The quality of your work inspired me to do better, and your support made the rough days easier and the good days even better. I would like to especially acknowledge those whose guidance and conversation were key to any successes I had as a graduate student – Ke Bi, David Buckley, Ana Carnaval, Ben Carter, Roberta Damasceno, Tom Devitt, Chris DiVittorio, Jon Fong, Matt Fujita, Peter Jorgensen, Shobi Lawalata, Adam Leache, Tyler Linderoth, Matt MacManes, Sean Maher, Jay McEntee, Toni Lyn Morelli, Peter Ralph, Santiago Ramirez, Ricardo Pereira, Sean Rovito, Kevin Rowe, Andrew Rush, Maria Santos, Chodon Sass, Sean Schoville, Skip Skipwith, Adam Smith, Lydia Smith, Maria Tonione, Elaine Vo, Rudi Von May and Guin Wogan. Additional thanks go to those who braved being my office mate – Julie Woodruff, Rachel Walsh, Yu Zeng, Lauren Benedict, Philip Skipwith, Jeremy Crawford, Adam Smith, Mike Holmes. A special thanks to the Endler Reading Group; our challenging conversations changed how I thought about our field and they remain a highlight of my graduate school experience.

To my students – I truly lucked out in getting to work with such motivated and intelligent undergraduates as those in the Ground Squirrel Group. Thank you for your questions that shone light on my ignorance, your humor that kept me in high spirits, and your enthusiasm that spurred me on. I look forward to applying to work for you all some day! And to my students at Patten University at San Quentin – your full-hearted dedication to learning always reminded me why I was in school in the first place.

To all those who helped me in the field – Alice Blackwell, Emily Hoffmann, Craig Moritz, Ben Phillips, and Maria Tonione – your efforts and enthusiasm made field work more enjoyable and productive. A special thanks to all the hard work of the countless researchers in the Moritz and Williams groups who have collected over 10K samples and invaluable ecological and environmental data. My thesis would not have been possible without your willingness to share your data.

To all those who helped along the way – sometimes it seems that dissertations are 10% science and 90% logistics, and there were many people who made my research possible by removing red tape and facilitating connections. Thank you to IB (Mei Greibenow, Susan Gardner), MVZ (Lydia Smith, Anna Ippolito, Stephen Long, John Stenske, Rhanee Guzman-Dungay), QPWS (Keith McDonald, Chris Wegger, Michelle Nissen), SFS (Sigrid Heise, Ian Brennan), JCU (Yvette Williams, Noema Patterson), QB3 & FGL (Minyong Chung, Leath Tonkin, Justin Choi, Karen Lundy), MCB (Fred Khorshidi), and CSIRO (Sandra Kay, Dave Westcott) for all your help.

To my friends, both near and far – thank you for the thought-provoking conversations, home-cooked meals, pun-filled e-mails, weekend adventures, constant teasing, and ever-ready smiles. You all provided me with much needed distractions and reminded me that there was more to life than my research. And a special nod to SMP and AFM; I am so glad we are PhD triplets!

To my family – everything is possible with your love. I am so fortunate to have the best group of siblings around; you all are my built-in laugh factory/hug machine/instant dance party/cheerleading squad. Ma and Papa, I know this isn't what you had in mind when you suggested I become a doctor. Thank you for supporting me, anyways.

And finally, to the lizards – thank you for letting me tell your story.

Chapter 1 was co-authored with Craig Moritz. Chapter 2 was co-authored with Craig Moritz and is currently in review in *Proceedings of the Royal Society B*. Chapter 3 was co-authored with Ke Bi. Chapter 4 was co-authored with Craig Moritz and is re-printed here with permission from *Molecular Ecology*. Chapter 5 was co-authored with Craig Moritz and is re-printed here with permission from *Evolution*. Chapter 6 is re-printed here with permission from *Molecular Ecology Resources.*

Chapter 1

Cryptic species and biodiversity

1.1 Abstract

Cryptic species, or genetic lineages within morphologically-defined species, have been recognized for as long as systematists have collected data on biochemical and genetic polymorphisms. The importance and relevance of these species have been debated, especially as they do not fit the original paradigm of species as morphologically-differentiated units. In this review, we look at the history of cryptic species, summarize findings from the Australian Wet Tropics on the nature of cryptic species, and discuss the consequences of cryptic species for our understanding of speciation and adaptation. We argue for a hypothesis-based approach to delimiting species, in which species boundaries are first outlined using genetic data and then refined with subsequent data on phenotypic divergence of and reproductive isolation between these putative lineages.

1.2 Introduction

From letters from Charles Darwin to Joseph Lee Hooker in which Darwin compares characterizing species as mutable to confessing to murder [73] to recent debates about the validity of using genetic data solely to define species boundaries [192, 29, 108], delimitation of species remains one of the most contentious subjects in biology. The angst surrounding species delimitation can cloud an important reality – for the most part, we make remarkably similar judgements when looking at the bewildering biodiversity that surrounds us and defining the discontinuities we see as species. Perhaps the best evidence for this is the high concordance between lists of "folk species" – taxonomic groupings of regional biota by indigenous communities – and lists of Linnean species – taxonomic groupings of biota by taxonomists [225]. Such work suggests that, although species delimitation is an inherently subjective exercise [148], independent observers often arrive at the same conclusion, and thus, perhaps most species boundaries are not that fuzzy.

That said, the subjectivity of defining species can become quickly apparent in a few cases, none of which are mutually exclusive. First, as noted by Darwin, species are mutable, and diverging units thusly occur on a continuum from population-level differentiation to reproductively-isolated lineages [72, 215]. We lack clear metrics for how much divergence is necessary for two lineages to be defined as two species. Indeed, setting this somewhat arbitrary benchmark undergirds much of the debate around defining operational taxonomic units (OTUs) through DNA barcoding [216]. As an extension, most species are not homogeneous, and they contain some level of polytypy and intergradation within their ranges [213]. Whether this variation is sufficient to name a new species can be unclear, and thus, this variation is instead oft-characterized as subspecies, races, or ecomorphs [80]. Second, we can characterize lineage divergence both through genetic and phenotypic measures. When these two axes of variation portray discordant pictures of differentiation – for example, when phenotypic disparity is high and genetic divergence is low, or when genetic divergence is high and phenotypic disparity is low – the status of these lineages becomes ambiguous. A common example of such a pattern is "cryptic species", or genetic lineages within morphospecies that exhibit little or no morphological divergence [40]. With the ever-increasing ease of obtaining genetic data for diverse taxa across their range, the number of cryptic species recognized continues to grow. Yet, many questions about such species – such as, are they any different from morphospecies? are they just evolutionary ephemera? are they truly cryptic? why have they remained morphologically conserved? – remain unanswered. Further, such cryptic lineages fall throughout the full range of the divergence continuum [142, 228, 124], compounding the first challenge with the second and thus making species limits even more obtuse.

With an eye to these challenges and unanswered questions, we review cryptic speciation in this paper, looking at its historical evolution as a concept and discussing how it affects our understanding of speciation more globally. In doing so, we use an exemplar system, the Australian Wet Tropics, to illustrate approaches for understanding cryptic species and to outline areas for future investigation.

1.3 A Historical Perspective

Early taxonomy – and, by extension, systematics – was based on identifying variation in easily accessible phenotypic data, which typically meant morphological characteristics such as size, shape, and coloration [226]. Even as species concepts moved towards being more process-oriented, morphological data remained an important marker of species differentiation. In fact, Ernst Mayr, who advocated using extent of reproductive isolation to delineate species, recognized that morphological data had a particularly important role to play in understanding species boundaries in allopatric populations. In such populations, he argued, as other researchers have confirmed with empirical data across taxonomic groups [111, 46], that the degree of morphological difference is correlated with the extent of reproductive isolation [226].

However, when systematists started collecting data on biochemical polymorphism, we gained a second axis on which to characterize variation in populations [190], and it quickly became apparent that variation in morphology and genetics were not always correlated. For example, researchers uncovered cases of phenotypic variation with no concomitant genetic differentiation, leading to some lineages being synonymized (*i.e.*, the color morph of the salamander *Plethodon gordoni*; [104]). Importantly, the inverse was true; genetic data also led to the identification of many "sibling species", or morphologically-similar species that showed evidence for reproductive isolation and that were genetically distinct, such as early studies that identified chromosomal races within *Drosophila* species [82, 52]. Re-analyzing phenotypic data in light of these genetic data sometimes showed that these species were pseudo-cryptic, and either these lineages were, in fact, differentiated morphologically, but subtly or in more cryptic morphologies (*i.e.*, genitalia; [366]), or they were differentiated along non-morphological phenotypic axes (*i.e.*, acoustic signals in crickets [270] or flashing patterns in *Photuris* fireflies [21]).

Genetic data became increasingly easier to collect as researchers moved from collecting chromosome data sets to allozyme data sets [12] to mitochondrial data sets, including DNA barcoding studies [182, 141], to multi-locus nuclear data sets [49]. Along the way, helped by ever decreasing prices for collecting genetic data, researchers began assaying patterns of genetic variation across species' ranges, contributing to the emerging field of phylogeography [11]. This confluence led to rapid increases in the number of sibling species – now typically defined as "cryptic species" in recognition of the diversity of evolutionary relationships these lineages exemplified [180] – to the point that a quarter of papers published in Zoological Record Plus mention cryptic species [40]. These cryptic species occur in a diversity of taxa, including species that are considered morphologically simple (fungi, [118]; moss, [322]; parasites; [36]) and those considered complex (vertebrates, [13]; butterflies, [53]). Many of these cryptic species evince as much genetic divergence as morphologically defined species [377, 228, 124], and many show evidence for reproductive isolation, whether they occur in sympatry without hybridizing [36, 123, 171] or there is evidence for hybrid dysfunction [28, 355, 128]. A good number of cryptic species are parapatrically distributed lineages (or, phylolineages) within morphospecies [11], which will be the focus of this review.

1.4 The Australian Wet Tropics: A Case Study in Cryptic Speciation

Parapatrically-distributed phylolineages provide a unique opportunity to investigate cryptic species, because, by characterizing lineage interactions at zones of secondary contact, we can understand what mechanisms, if any, are keeping lineages distinct [145]. Suture zones, or narrow geographic regions in which multiple contact zones overlap [294], can further strengthen inference by allowing comparative studies across a common ecolog-

ical setting and biogeographical history. Here, we review over twenty years of phylogeographic, ecological and morphological data from a suture zone in the Australian Wet Tropics, focusing on phylolineages within three lizard species complexes to understand better the nature and reality of cryptic species.

This suture zone formed as a consequence of repeated glacial cycling during the Neogene [244, 146, 294]; during the cool-dry and warm-wet stages of the glacial cycle, the rainforest contracted into two major and several minor refugia [133, 360]. Rainforest endemics tracked the moving rainforest, leading to divergence between populations restricted to isolated refugia. Since the Last Glacial Maximum, the climate has warmed, the forests have expanded, and long-isolated lineages in a number of taxa (*i.e.*, frogs, beetles, lizards, mammals) came into secondary contact. Our work here focuses on differentiation in three lizard species complexes: *Carlia rubrigularis*/*C. rhomboidalis*, *Lampropholis coggeri*, and *Saproscincus basiliscus*/*S. lewisi*. These species are all fossorial skinks that range in 30 to 70 mm in snout-vent length [376]. They are ecologically-similar – they all reside in the leaf-litter at rainforest edges and gaps [373] – and they are closely related – they likely diverged about 25 mya [331]. Across these five species, there are ten phylolineages (Fig. 1A), which can be paired as seven sets of lineage-pairs (Fig. 1B). Of these seven lineage-pairs, we have identified zones of secondary contact between five; the habitat is not continuous between the remaining two lineage-pairs: *C. rubrigularis* S/*C. rhomboidalis* and *S. basiliscus* C/S.

The phylolineages in these complexes were first identified through broad-range mitochondrial DNA sampling throughout each complex. Sampling revealed multiple, deeply-divergent lineages that were often as genetically distinct as morphospecies in this group (Fig. 2). We identified these major nodal breaks as putative cryptic lineages, all of which but *L. coggeri* N correspond to reciprocally monophyletic lineages. Each of these phylolineages is fractal and shows further, geographically-restricted variation in mtDNA. However, as can be seen in the nuclear network results (Fig. 2), there is little structure in the nuclear genome within these phylolineages, and thus, we focus only on the major clades within each complex. Importantly, for all lineages but *S. basiliscus* S, we see strong concordance between our mtDNA and the nuclear DNA results across geographic patterns of variation, systematic relationships, and divergence depth (Fig. 2; Fig. 3). This high concordance further strengthens the reality of these cryptic lineages. Many cryptic lineages are defined based primarily on plastid genomes (*i.e.*, mitochondrial and chloroplastid genomes; [141, 40]), because plastid markers work across a wide range of taxa and are highly variable [18]. However, theoretical and empirical data show that the unique characteristics of plastid genomes (*i.e.*, their important role in organismal physiology, unilineal inheritance patterns, rapid mutation rate; [18]), along with coalescent variance [365], can lead to discordant patterns between the plastid genome and the rest of the genome. As such, it is important to confirm that lineages identified using single-locus plastid data are robust across other markers [49], as we have done here.

To describe these lineages, we focused on three aspects: their divergence history, morphology, and ecology. First, we fit a standard isolation-with-migration model [280] to

genetic data for each lineage-pair, thus inferring the divergence time and migration rates during divergence. We inferred a 6.2× range in divergence times, scaling from 1.83 mya for the youngest lineage-pair to 11.4 mya for the oldest lineage-pair (Fig. 4). Across all lineage-pairs but for *S. basiliscus* C/S, we inferred very low rates of migration; for *S. basiliscus* C/S, we had trouble precisely defining the number of migrants. However, for this lineage-pair, the number of effective migrants was estimated to be significantly greater than 1, which is considered to be the critical value defining population structure [379, 390]. Morphological variation across phylolineages within each species complex was limited (Fig. 5). Where lineages are different, the differences are generally sex-specific and absolutely limited, such that these differences cannot help discern between lineages in the field. That said, two lineage-pairs are visibly distinct morphologically – the youngest lineage-pair, *C. rubrigularis* S/*C. rhomboidalis*, differ in throat color [160] – and the oldest lineage-pair, *S. basiliscus* N/*S. lewisi*, have differing distributions of the number of para-vertebral scales [65]. Finally, these phylolineages have no obvious differences in their ecology – years of observation have noted no variation in microhabitat use or phenology. Further, although there are broad-scale environmental changes across the AWT [244], this gradient is subtle. Indeed, looking at environmental variation across the ranges of each phylolineage, the extent of variation across lineages overlaps considerably (Fig. 6). Thus, despite the marked difference in divergence times across these lineage-pairs, almost all pairs share remarkable conservation in both morphology and ecology.

What about the "reality" of these lineages, though? Although they are genetically distinct, they could easily be evolutionary ephemera destined to be lost to hybridization as the pairs reunite in secondary contact. An emerging approach to delimiting species uses a coalescent-based models to identify speciation events by determining which lineages are evolving independently [287, 109]. Using such an approach here, every lineage was identified as a species with a high probability (Fig. 1A). However, evolutionary independence in allopatry need not correlate with evolutionary independence in parapatry. Although genetic distance (a proxy for evolutionary independence in allopatry) is correlated with reproductive isolation [67, 312], the strength and form of this relationship is neither perfect nor constant across taxonomic groups [45, 121]. In this system, we see that evolutionary independence in allopatry as measured by coalescent-based models is not synonymous with independence in parapatry, because not all lineage-pairs show evidence for reproductive isolation.

That said, we find that the extent of reproductive isolation evolves in a remarkably predictable fashion [326]. For the lineage-pairs where we could indirectly infer the extent of reproductive isolation, we find evidence for strong isolation between the lineage-pairs *C. rubrigularis* N/C, *L. coggeri* C/S, and *S. basiliscus* N/*S. lewisi*, both at geographical (Fig. 7) and genomic scale (Fig. 8). In fact, we sampled no hybrids where *S. basiliscus* N/*S. lewisi* meet, suggesting that if they are hybridizing, the hybrids fail to reproduce successfully (Fig. 7E). On the other hand, *L. coggeri* N/C and *S. basiliscus* C/S show little evidence for reproductive isolation (Fig. 7 & 8); introgression is rampant throughout their range and their genomes. For the lineage-pairs which are allopatric, behavioral data for

C. rubrigularis S/*C. rhomboidalis* suggest the two lineages mate assortatively [84], which would help maintain genetic independence if the lineages were to meet. For *S. basiliscus* C/S, the status of these lineages remains unclear. On one hand, there appears to have been massive introgression in the past which eroded the nuclear divergence between the lineages [328]. On the other hand, these lineages are quite distinct at the mitochondrial genome. If selection is maintaining mtDNA divergence, then it is quite possible these lineages will continue to evolve as evolutionary independent units in allopatry.

1.5 Implications of Cryptic Species

The data from the AWT and other systems illustrate the ubiquity of cryptic species and suggest, based on observations from where lineages occur in syntopy, that they are evolutionarily independent lineages. Given this, what of cryptic species? Are they somehow distinct from morphospecies, and do they deserve special study? Ernst Mayr, one of the major champions of sibling species, argued no; he believed that morphologically-similar lineages were no different from morphologically-differentiated lineages [225]. In many ways, Mayr is right – there is no reason to assume that the evolutionary forces driving divergence would be different between the two types of species – and, in fact, at a gross level, cryptic species are in no way special. Most empirical data show that cryptic species can both be very young [193, 238] and old [377, 228], and meta-analyses suggest they occur fairly evenly across taxonomic groupings and geographic localities [277]. That said, the mere fact that cryptic species are not morphologically differentiated begs discussion; as Dobzhansky (1946) recognized, "to an ecologist or a geneticist, morphologically similar species are important because their existence proves that morphological differentiation is not an essential, though widespread, concomitant of evolutionary divergence."

How do cryptic species form without accompanying morphological divergence? Currently, much speciation research focuses on the role of divergent natural selection in driving lineage formation ("ecological speciation"; [315, 258]); in such systems, lineages are often markedly different in obvious phenotypic traits, such as floral shape in columbine flowers [368], coloration in desert lizards [301], or body shape and color in *Timema* insect morphs [259]. Under strong divergent natural selection, lineages can evolve pre-zygotic isolation (for example, evolving mate choice based on coloration in cichlid fishes; [318]) and post-zygotic isolation (for example, selecting against hybrids between the different morphs of *Euphydryas* butterflies [227] and evolving Dobzhansky-Muller incompatibilities (DMIs) as a correlated by-product of divergent selection in *Mimulus* [378]). Even in the presence of gene flow [83], divergent natural selection can rapidly drive lineages to different adaptive peaks. However, in cryptic lineages, the handiwork of divergent natural selection appears absent, as there is no change in the most obvious marker of phenotypic divergence – morphology. That does not mean that divergent natural selection fails to act in these systems; instead, divergent natural selection could be acting on more cryptic phenotypes, such as habitat specialization (e.g., *Astraptes* butterflies; [141]),

chemiosensory profiles (e.g., *Drosophila* fruit flies; [66]), structure of reproductive proteins (e.g., the ascidian *Ciona*; [262]), life cycle histories (e.g., *Mordellistena* beetles; [43]), and physiology. However, for most cryptic species, including the AWT lineages, we have little data showing any phenotypic divergence, both because data on these phenotypes are much more time-intensive to collect and because what data we have collected shows no differentiation (*i.e.*, habitat [244] and physiology (Phillips, unpublished)). Thus, although cryptic lineages often show no obvious evidence of adaptation, adaptation is prevalent in speciation [338] – it will just require more time and effort to identify the relevant phenotypes in cryptic lineages.

As another possibility, models in which divergent selection does not play a big role (also termed "non-ecological speciation"; [315]) might explain divergence amongst these cryptic lineages. Such models can be loosely grouped into four types: models involving polyploidy, social and sexual selection models, mutation-order speciation models, and drift-driven models. Though it has driven cryptic divergence in other species complexes [44], polyploidization is unlikely to be a factor here, and we refrain from discussing it futher. In many natural systems, sexual selection interacts with divergent selection to generate spectacular radiations (e.g., the swordfishes *Xiphophorus*; [304]). However, models also suggest that drift and sexual selection might interact to lead to divergence in mating cues between isolated populations [359, 186]. Given that these lineages are fossorial and thus are likely less visually-oriented, it is possible, however, that the species have diverged in chemiosensory profiles. Indeed, given that we have no evidence for hybridization between *S. basiliscus*/*S. lewisi*, we suspect that there is pheromone-mediated pre-mating isolation between these lineages, a hypothesis we are currently testing. Another possibility is that the lineage-pairs diverged in similar habitats, and thus, they were subject to parallel selection. Because different mutations arise in each lineage and are fixed in different orders, these lineages can then still accumulate substantial divergence and post-zygotic isolation via Dobzhansky-Muller incompatibilities [219, 260, 62]. Support for this model is equivocal [295, 78], but, because it predicts phenotypic similarity among populations, it remains a useful framework for explaining genetic divergence in cryptic lineages. Finally, drift remains a "null hypothesis" to explain divergence, whether via traditional population genetic models or updated versions of drift along "holey" adaptive landscapes [113]. In particular, the original verbal models for DMIs proposed that they became fixed in populations through drift [81, 247]. That said, drift can be exceedingly slow at driving divergence [115, 265]. Although models that include a severe bottleneck (e.g., founder-flush models; [56]) can hasten divergence by drift, drift-based models are still likely too slow to appreciably and reliably generate the multitude of cryptic species we see. Further, the likelihood that populations would drift solely along genetic and not morphological axes is unclear; unless some other factor is promoting stasis [364], we would expect drift to drive divergence along both axes. Yet, despite these theoretical challenges and the limited data in support [295, 115, 78], drift-based models remain a starting hypothesis to explain divergence, whether in cryptic or non-cryptic lineages.

Importantly, unlike speciation models involving divergent selection, models without

divergent selection are not robust to gene flow [1, 260, 358], because there is no counterpoint to the homogenizing effects of gene flow. With the development of coalescent-based methods to infer levels of gene flow between diverging lineages [147], we see that most lineage-pairs, including the AWT lineage-pairs (Fig. 4), exchange only a modest number of migrants as they diverge [280]. However, even a modest number of migrants ($Nm < 1$; [19]) can prevent divergence. Further, these models quite possibly underestimate the number of migrants. The actual divergence history of these lineage-pairs is likely much more complex than what we modeled here; biogeographic reconstructions in both this system [133] and other systems affected by Pleistocene cycling [146] show that ranges were dynamic during time and there was likely repeated, brief interludes of connectivity between lineage-pairs during which there could have been introgression. Indeed, introgression during such interglacials is likely what drives the discrepancy between nuclear and mitochondrial genomes in *S. basiliscus* C/S [328]. By fitting our genetic data to a simpler divergence with gene flow model, we likely are underestimating the number of migrants. On one hand, if our lineage-pairs are exchanging appreciable number of migrants as they diverge ($Nm > 0$), speciation hypotheses that include divergent natural selection become more probable. On the other hand, gene flow during these interglacials might be spatially restricted, such that introgression does not extend far beyond the zone of secondary contact [168]. When populations are again sundered, these bouts of interglacial gene flow would leave little evidence behind, except perhaps at loci that experienced selective sweeps. If introgression is limited in both its spatial and genomic extent, then both models that include and exclude divergent natural selection could explain cryptic lineage divergence. Thus, understanding the patterns of gene flow across space, across the genome, and through time in these lineages, when integrated with assays of phenotypic variation along multiple traits, should help us discern what model of speciation best applies to cryptic lineage-pairs.

1.6 Conclusion

Cryptic species challenge traditional notions of species, because the discrepancy between morphological and genetic axes of divergence can make them hard to categorize. Yet, other data suggest that many cryptic species are phenotypically divergent, but on axes of variation that are harder to measure. In cases where we cannot identify phenotypic differences, like the AWT taxa, we can test the reality of these lineages through other means, such as looking at interactions between cryptic lineages in sympatry. Often, these richer, more integrative datasets complement genetic data and show that cryptic lineages are independently evolving units [287]. However, as we also see in the AWT taxa, despite marked genetic differentiation, some cryptic lineages might just be ephemera, destined to be lost to hybridization with sister lineages. These data remind us that proposing species boundaries is a hypothesis [109], an educated guess at the fate of these lineages and a recognition of the ever-evolving nature of species [72, 215, 80].

1.7 Methods

1.7.1 Genetic Analyses

Species Tree

To collect multi-locus nuclear data for these lineages, we employed the anchor-tagged enrichment approach described by [194]. Here, for each lineage, we sampled two individuals, except for *C. rhomboidalis*, *S. basiliscus* S, and *S. lewisi*, for which we sampled one individual. The raw reads from this enrichment experiment were assembled following Lemmon and Lemmon (in prep), resulting in a data set of approximately 480 loci per lineage. To identify haplotypes at each of these loci, we filtered and cleaned the raw reads following the approach outlined in [325], mapped reads back to their corresponding assembly with Bowtie2 [188], and inferred haplotypes using HaplotypeCaller in GATK [77].

These haplotype data were then used to infer species trees via three methods. First, we created three alignments of 50 random loci, for which we estimated the ideal partitioning scheme and model of molecular evolution using PartitionFinder [187]. For each alignment, we then inferred the species tree using STARBEAST, setting the clock model to be "uncorrelated lognormal" and estimating the clock rate to be relative [143]. We ran each alignment 5 times for 1e8 generations, sampling every 1e5 generations, and compared convergence across results with Tracer [290]. STARBEAST can account for uncertainty in estimating gene trees, but because of computational challenges of doing so, can only handle 50 – 100 loci. Thus, to take advantage of our full data set, we used the programs STEAC and STAR, which summarize across gene trees to describe the species tree [204]; importantly, these methods cannot incorporate uncertainty in gene tree estimation and assume gene trees are "true". We divided each locus into three partitions: the anchor region which is slowly-evolving and the two flanking regions which show more variation [194]. We found the ideal partitioning and molecular evolution model using PartitionFinder, and we used the partitioned alignment to infer gene trees with RAxML [340]. We then used STEAC and STAR to infer the species trees given the set of best-scoring gene trees across all the loci.

Bayesian Species Delimitation

An emerging approach in species delimitation is to use coalescent-based approaches to identify lineages that are behaving genetically as species [109]. We apply these approaches to our species complexes using BPP2.2 [292] and the nuclear data described above. For each species complex (*S. basiliscus*/*S.lewisi*, *L. coggeri*, and *C. rubrigularis*/*C. rhomboidalis*) and one outgroup, we created five random alignments of 50 loci each. BPP requires the relationships among terminal taxa to be known; thus, for each complex, we used the relevant subtree of the highly-supported species tree as the guide tree (Fig. 1A). Then, as BPP is sensitive to both priors for τ and θ and has issues with adequate mixing [192, 292], we ran each alignment four times, changing the priors each time and randomly selecting

a starting tree. The prior sets were $\theta \sim G(1,10), \tau \sim G(1,10)$, $\theta \sim G(2,2000), \tau \sim G(2,2000)$, $\theta \sim G(1,10), \tau \sim G(2,2000)$, and $\theta \sim G(2,100), \tau \sim G(2,1000)$. For each of these runs, we used the updated rjMCMC algorithm because [292] showed that it has better mixing than previous versions of this algorithm.

Mitochondrial DNA Phylogeny

To summarize patterns of mitochondrial diversity for each complex, we inferred the gene tree at *ND4* using previously published data [85, 35, 245]. For each data set, we first partitioned the data into the tRNA locus and the three coding positions of the *NADH dehydrogenase* gene. We then co-estimated the ideal partitioning scheme and model of molecular evolution with PartitionFinder. We used MrBayes to infer the gene tree for the partitioned alignment, running the progam twice with four chains (three heated, one cold; default heating parameters) for 50e6 generations with a 6e6 generation burn-in [159].

Nuclear DNA Network

To summarize patterns of nuclear data for each species complex, we used the program POFAD [169], which creates individual-based distance matrices based on multi-locus haplotypic data. Each matrix can then be visualized as a network with SplitsTree [158]. We created nuclear gene networks for *L. coggeri* using data from [35] and for *C. rubrigularis*/*C. rhomboidalis* using data from [85]. The *S. basiliscus*/*S. lewisi* network is reproduced here without modification from [328].

Mitochondrial and Nuclear Divergence

To determine the correlation between nuclear and mitochondrial divergence, we used Arlequin v3.1 to estimate raw D_{xy} and net D_a sequence divergence between the lineage-pairs at both the mitochondrial and nuclear genomes [102], as estimated by the Tamura-Nei model. We calculated the divergence for all lineage-pairs studied here and other lineage-pairs in this complex. These data are presented without modification from [328].

Divergence History Estimation

To infer the divergence history of each lineage-pair, we fit an isolation-with-migration model via three different methods. Because all three methods employ the same basic logic and were parameterized using the same mutation rate and generation time, our results are comparable across methods. For *S. lewisi*/*S. basiliscus* N and *C. rubrigularis* S/*C. rhomboidalis*, we fit previously published data for six to eight nuclear markers using IMa2 [328, 85, 147]. For *S. basiliscus* C/S, we used Approximate Bayesian Computation [70] and sequence data from eight nuclear loci and a single mitochondrial loci to fit a modified isolation-with-migration model, in which migration rates between the nuclear and mitochondrial genomes were allowed to be different. These results were previously published

in [328]. For the remaining four contacts, we used the program dadi to fit an isolation-with-migration model to SNP variation summarized as an unfolded two-dimensional site frequency spectrum (2D-SFS) [135]. These results were previously published in [326].

1.7.2 Morphological Analyses

To look at the extent of crypsis in these complexes, we determined the extent of morphological variation across phylogeographic lineages. The data presented here follow the methods of [326] and add additional, previously-published morphological data for *C. rhomboidalis* from [84]. Briefly, morphological measures for adult lizards (snout vent length, head width, head length, hind limb length) were, where relevant, regressed against elevation and then used in a principal components analysis. We used MANOVAs to determine if variation was significant and conducted follow-up ANOVAs and Tukey HSD tests were relevant. Because these lizards are sexually dimorphic, all analyses were done by sex.

1.7.3 Environmental Analyses

For each species complex, we used more than ten years of sampling data to infer patterns of environmental variation across phylogeographic lineages. To do so, we first rarified each sampling data set such that any points within 5 km of each other were removed (script provided courtesy of S. Maher). We then extracted all 18 Bioclim variables for these points and then used these data to conduct a principal components analysis across lineages [150].

1.7.4 Inferring Reproductive Isolation

To determine the extent of reproductive isolation between sister lineages, we calculated indirect measures of hybridization at all lineage-pairs with clearly-defined contact zones. The data presented here are published without modification from [326]. Briefly, the contact zone between each lineage-pair was densely sampled, and each individual was genotyped at 6–10 nuclear genes and 1 mitochondrial marker. These genotype data were then used to infer clines with Analyse and levels of genomic admixture with Structure [26, 285]. For the four contact zones in which we saw evidence for hybridization, we captured over 3 megabases of sequence using anonymous pooled sequencing and custom exome target arrays. With these data, we inferred an average of 12,300 clines per contact zone using R [288]; data are published here without modification from Singhal and Bi (unpublished).

1.8 Acknowledgements

Funding for this work was provided by the MVZ Writing Year Fellowship and by the Australian National University. We thank R. Damasceno, C. Hoskin, J. McGuire, J. Patton, and D. Wake for helpful discussions.

1.9 Figures

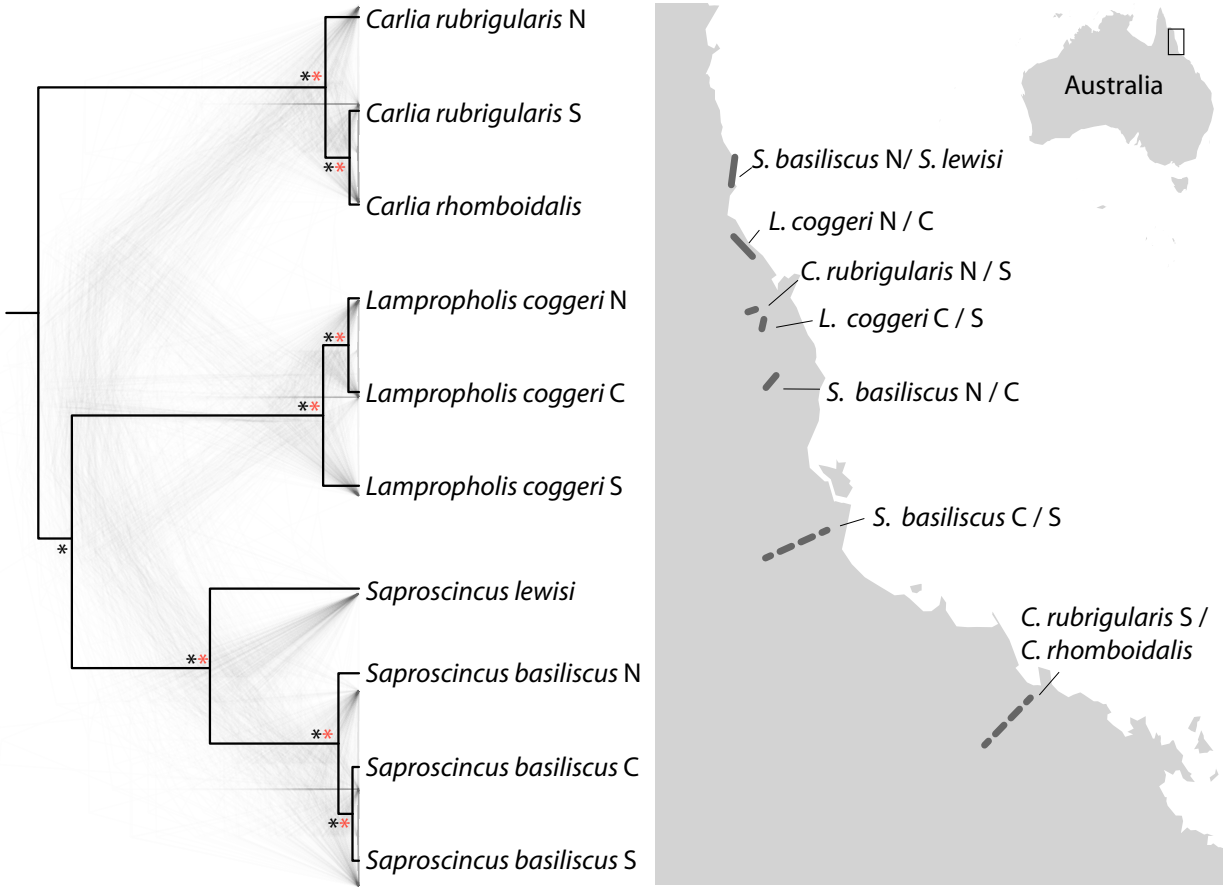

Figure 1.1: A. Species tree of lineages included in this study as inferred with STARBEAST; black asterisks mark nodes with posterior probability ≥ 0.95. Red asterisks indicate nodes with speciation probabilities ≥ 0.95 as inferred with BPP. Gene trees used to infer the species tree shown in the background. B. Map of contact zone localities. Dotted lines indicate midpoints between phylogeographic lineages that are not tightly parapatric.

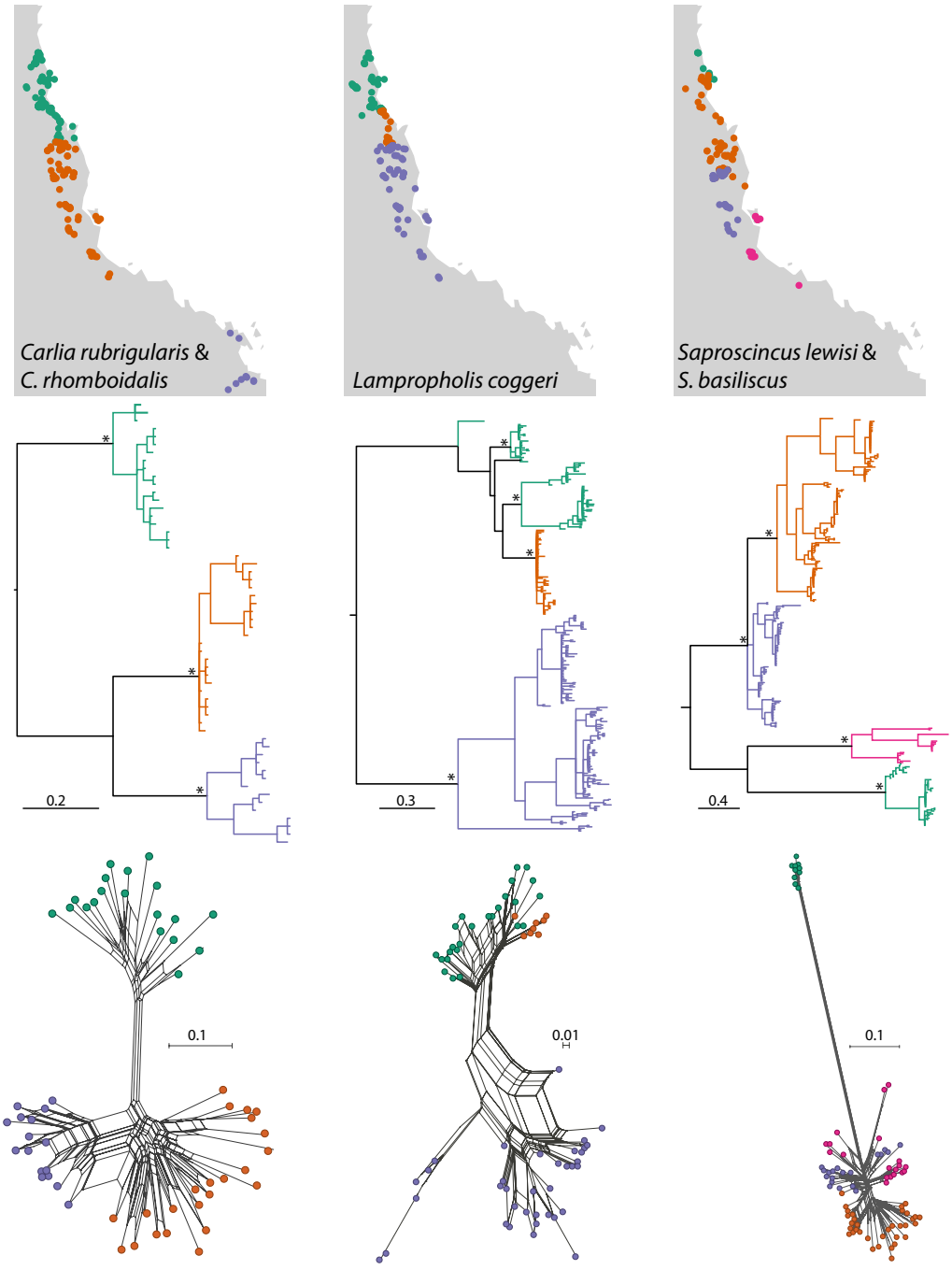

Figure 1.2: Top: Map showing distribution of phylolineages in each species complex; middle: mitochondrial DNA gene tree with with major nodes with posterior probability $\geq$ 0.95 starred; bottom: nuclear network for each species complex as inferred with POFAD.

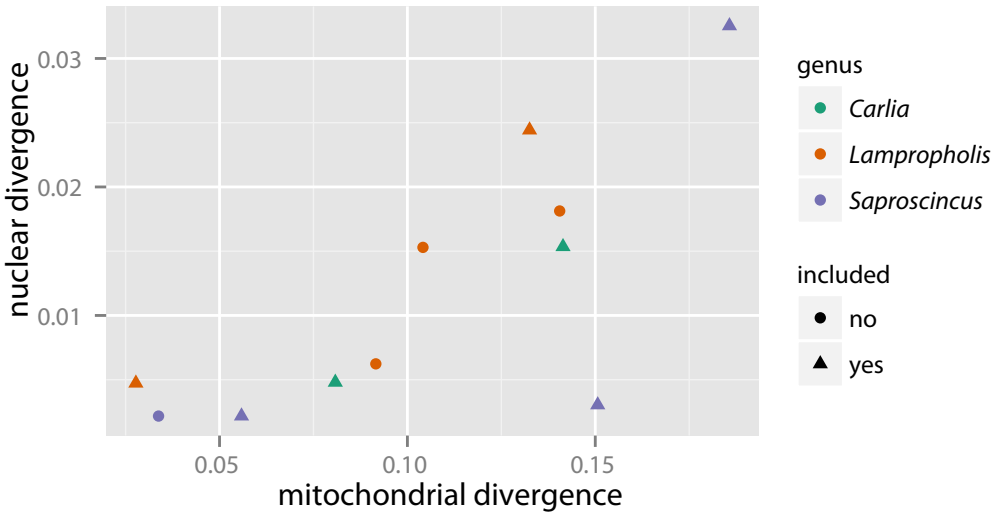

Figure 1.3: Correlation of mitochondrial and nuclear divergence.

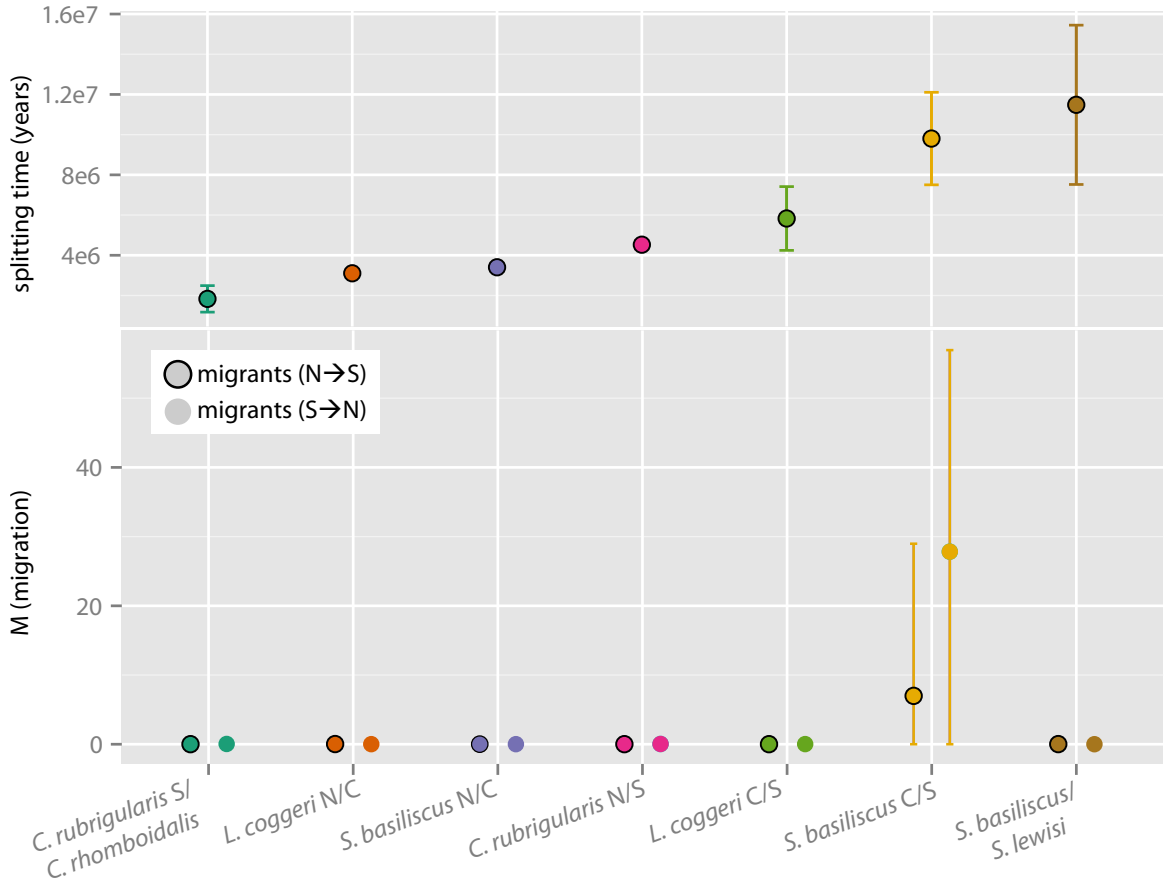

Figure 1.4: Divergence history as inferred from isolation-with-migration analyses between each lineage-pair. Histories inferred using three methods: dadi, IMa2, and Approximate Bayesian Computation; see text for details.

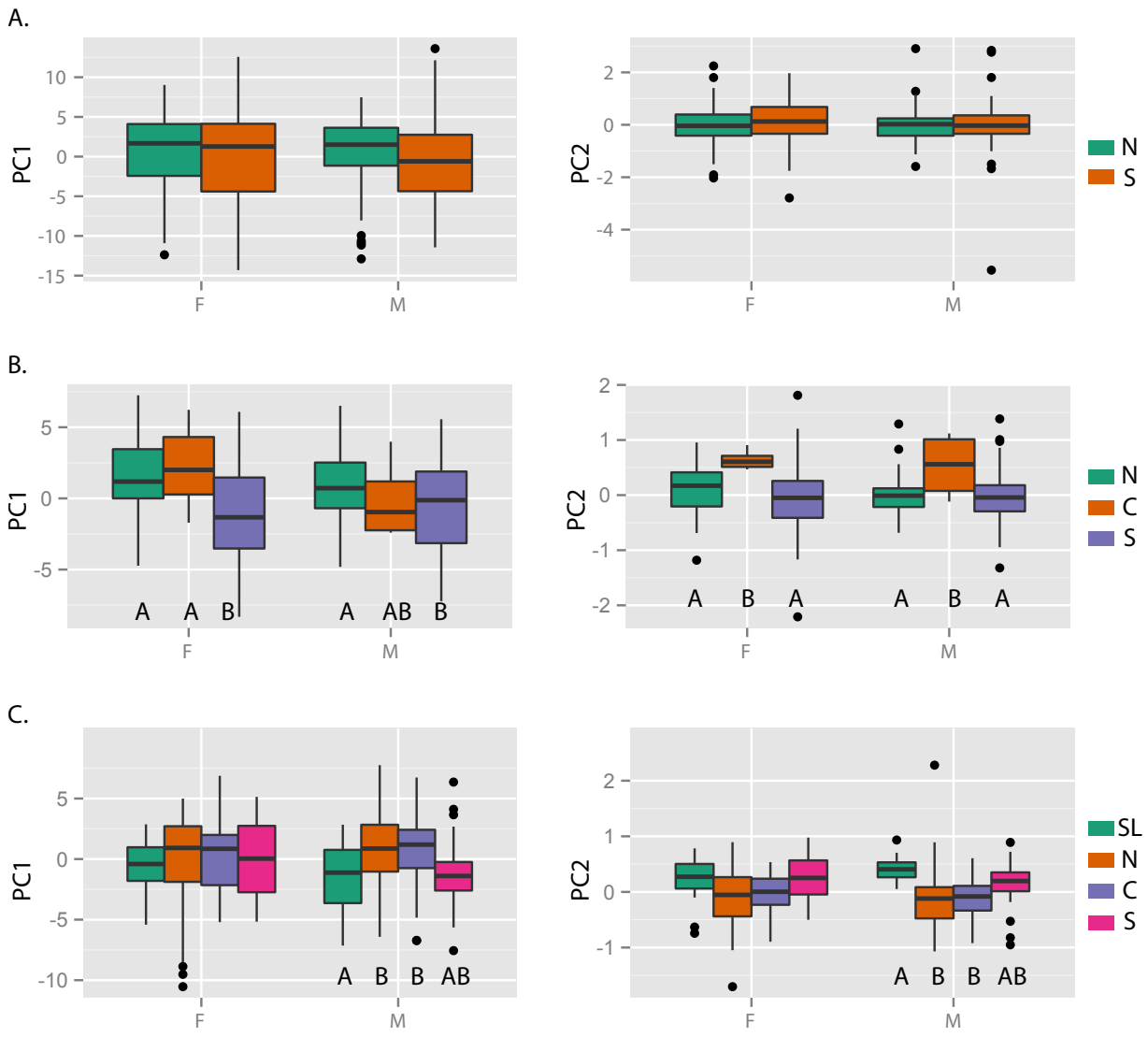

Figure 1.5: Morphological analyses for phylolineages: A. *Carlia rubrigularis*/*C. rhomboidalis*, B. *Lampropholis coggeri* N/C/S, and C. *Saproscincus basiliscus* N/C/S & *S. lewisi*. Significant differences labelled.

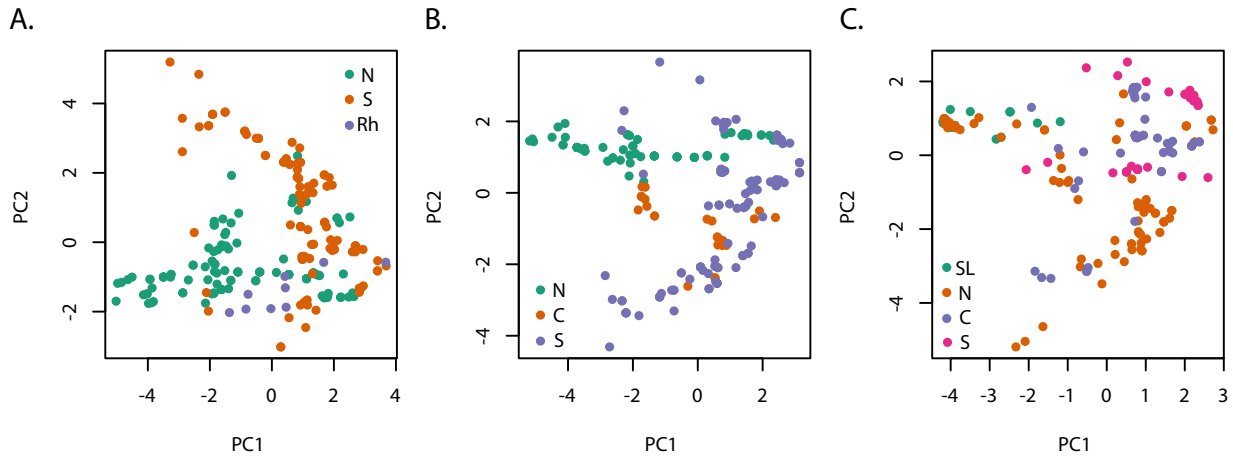

Figure 1.6: Principal components analyses of Bioclim variables across phylogeographic lineages: A. *Carlia rubrigularis*/*C. rhomboidalis*, B. *Lampropholis coggeri* N/C/S, and C. *Saproscincus basiliscus* N/C/S & *S. lewisi*.

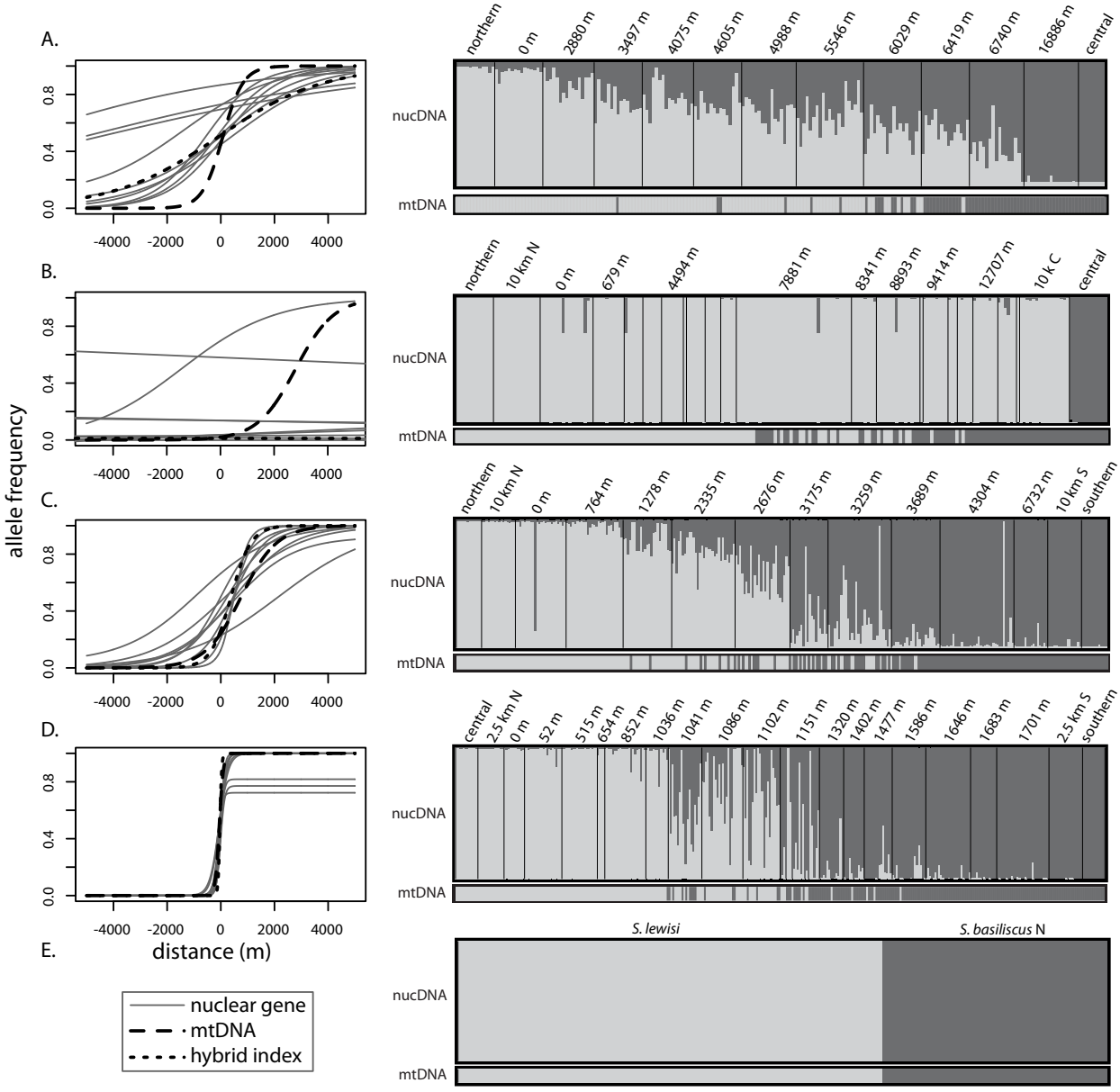

Figure 1.7: Cline fitting (left) and genetic clustering results (right) for contacts in the Australian Wet Tropics suture zone: A. *Lampropholis coggeri* N/C, B. *Saproscincus basiliscus* N/C, C. *Carlia rubrigularis* N/S, D. *L. coggeri* C/S, and E. *S.basiliscus* N/*S.lewisi*. For showing cline fitting results, distances along transects were recalculated so that each hybrid zone center was centered at 0 m. Scale for genetic clustering results differs among contacts.

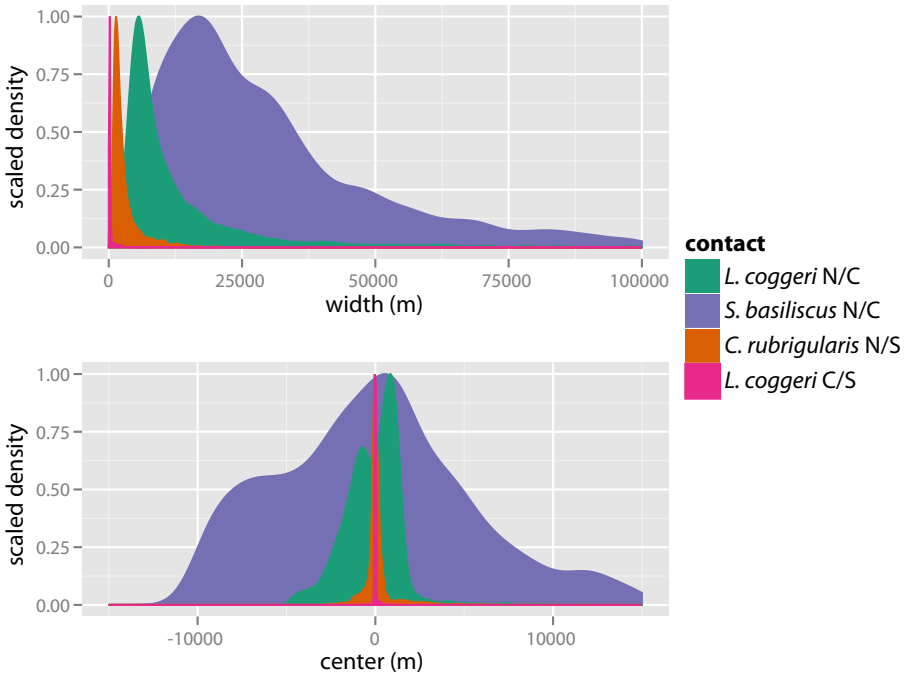

Figure 1.8: Distributions for A. cline width and B. cline center for the four hybrid zones.

Chapter 2

Reproductive isolation scales with divergence

2.1 Abstract

Phylogeographic studies often reveal multiple morphologically-cryptic lineages within species. What is yet unclear is whether such lineages represent nascent species or evolutionary ephemera. To address this question, we compare five contact zones, each of which occurs between eco-morphologically cryptic lineages of rainforest skinks from the rainforests of the Australian Wet Tropics. Although the contacts likely formed concurrently in response to Holocene expansion from glacial refugia, we estimate that the divergence times (τ) of the lineage-pairs range from 3.1 to 11.5 Myr. Multilocus analyses of the contact zones yielded estimates of reproductive isolation that are tightly correlated with divergence time and, for longer-diverged lineages ($\tau > 5$ Myr), substantial. These results show that phylogeographic splits of increasing depth can represent stages along the speciation continuum, even in the absence of overt change in ecologically relevant morphology.

2.2 Introduction

There is now abundant evidence for deep phylogeographic divisions within traditionally described taxa, suggesting that morphologically cryptic species are common [40]. Indeed, deep phylogeographic structure based on mitochondrial DNA (mtDNA), and confirmed by multilocus nuclear DNA (nDNA), is increasingly used as an initial step in species delimitation via integrative taxonomy [266]. As we grow better able to identify evolutionarily independent genetic lineages within morphologically-defined species, what is often missing is both an understanding of the forces leading to this diversity and evaluation of whether these lineages are more than ephemera [303]. One way to start addressing these issues is to test for reproductive isolation (RI) among such morphologically cryptic lin-

eages. Is there substantial RI between cryptic lineages, and if so, how does this scale with divergence time and historical gene flow? Answers to these questions will both inform modern systematics and contribute to our understanding of speciation processes.

It has long been supposed that phylogeographic lineages represent a step in the continuum from population divergence to speciation [14]. More generally, speciation theory posits that RI, especially post-zygotic RI, increases with divergence time, with the tempo and form of the relationship depending on the genetic architecture of Dobzhansky-Muller incompatibilities (DMIs) and the interaction of selection, drift, and gene flow [113, 130, 264]. But, given this, the heterogeneity of divergent selection is expected to blur the relationship between RI and divergence [113]. A growing body of evidence supports a general increase in RI with divergence time; however, with few exceptions [243], these results derive from analyses of phenotypically distinct species pairs [45, 312]. As phylogeographic lineages within morphologically defined species are often parapatrically distributed, comparative analyses of RI indices in secondary contact zones could provide a unique window into the dynamics of eco-phenotypically cryptic speciation [139]. Such studies have the added advantage of addressing the evolution of RI in nature, in the organisms' ecological context, rather than laboratory crosses, as is more common in the literature.

To investigate the evolution of RI in nature, we exploit a system characterized by climate-driven fluctuations in habitat extent and connectivity during the Neogene – the rainforests of the Australian Wet Tropics (AWT; Fig. 1). Extended periods of retraction of rainforests to mesic mountain tops has resulted in pronounced phylogeographic structure within endemic faunal species, but with variable levels of sequence divergence and divergence time among intraspecific lineages [244, 35]. Where tested, these mtDNA lineages are generally corroborated by multilocus nuclear gene analysis [244] (but see [328]); however, eco-morphological divergence is subtle or absent [35, 154]. Following Holocene rainforest expansions, over twenty contact zones involving pairs of morphologically cryptic lineages formed between the historic refugia [244]. These contact zones provide a natural experiment with which to test the hypothesis that RI increases with divergence time among cryptic lineages. Previous studies of contact zones have revealed outcomes ranging from negligible to strong RI [244], including one case of speciation by reinforcement [155]. However, the cases studied to date are taxonomically and ecologically heterogeneous.

Here, we use comparative analysis of RI within five contact zones involving lineage-pairs from a closely-related and ecologically-similar clade of terrestrial rainforest skinks, across which there are varying levels of sequence divergence among component clades: (*Carlia rubrigularis* N/S, *Lampropholis coggeri* N/C, *L. coggeri* C/S, *Saproscincus basiliscus* N/C, and *S. basiliscus* N/*S. lewisi*; Fig. 1). We combine genome-scale analyses of divergence history between allopatric populations with multilocus analysis of intensively sampled contact zones to test for increasing RI with divergence time. We assume that per-generation dispersal rates are similar across lineages, which indirect dispersal estimates from this clade support [278, 327]. Because the focal lineages are ecologically-similar,

rainforest-edge species that likely tracked the expanding rainforest front closely [373], we further assume that the contact zones formed concurrently. Following from these assumptions, we predict that RI scales closely with divergence, especially given the limited ecomorphological divergence of these lineage-pairs [35] and, for at least one lineage-pair, apparent absence of mate choice [84]. More specifically, we predict that, as divergence time increases, cline widths should narrow, clines should exhibit less variance in cline width, disequilibrium – both within- and between-loci – should increase, and frequency of hybrids within the hybrid zone should decrease [27].

2.3 Methods

2.3.1 Sampling

We sampled five contact zones in the AWT from 2008 to 2011 (Table S1): *Carlia rubrigularis* N/S, *Lampropholis coggeri* N/C, *L. coggeri* C/S, *Saproscincus basiliscus* N/C, and *S. basiliscus* N/*S. lewisi* (Fig. 1, Fig. S1). For each contact, we first identified the location of the contact zone by genotyping individuals at the mitochondrial genome (Fig. 1B). Then, we collected samples from populations geographically-isolated from the contact zone, which we used to infer demographic history and to develop markers. For four of the five contacts, we sampled individuals non-destructively along a linear transect through the contact zone (Fig. S1). For *S. basiliscus* N/*S. lewisi*, we sampled opportunistically because initial data suggested the lineages were not hybridizing (Fig. S1). Data from the *L. coggeri* C/S hybrid zone were previously published in [327], and we expanded the *C. rubrigularis* N/S data set collected by [278], by genotyping new genes, increasing sample sizes, and adding new populations.

2.3.2 Morphological Analyses

To test for phenotypic divergence across the major phylogeographic lineages within each traditionally defined species, we measured adult lizards outside of the contact zones at four standard characters for lizards: snout-vent length, head width, head length, and hind limb length. Because these species show evidence of sexual dimorphism, we analyzed these data by spliting each species data set by sex. Further, where deemed relevant by MANCOVA, we removed the effect of elevation by taking the unnormalized residuals of morphological characters against elevation. Using the first two major orthogonal axes from a scaled principal components analysis, we tested for morphological differentiation across phylogeographic lineages using MANOVAs, conducted follow-up ANOVAs on results that were significant, and followed significant ANOVAs with Tukey HSD tests [288].

2.3.3 Genetic Data Collection

We collected two types of genetic data: (1) transcriptomic data from populations isolated from the contact zone to infer demographic history and to develop markers and (2) genotypic data from populations located in the transition zone between lineages to infer the extent of reproductive isolation. We collected transcriptomic data for five individuals per lineage; with these data, we created pseudo-reference assemblies and called individual genotypes as described in [325]. This variant information was then used to infer demographic history as described below. Using variant data, we also designed PCR-RFLP markers for genotyping lizards from the contact zones. We selected variants that were located in untranslated regions (UTRs) of genes, diagnostic for the two lineages, and easily resolved by robust and inexpensive restriction enzymes. Marker details, including primers, annealing temperatures, and corresponding restriction enzymes, can be found in Table S2. In total, individuals in the *S. basiliscus* N/*S. lewisi* contact zone were genotyped at 6 nuclear markers and mtDNA, and individuals in the other four zones at 10 nuclear markers and their mtDNA (Table S1).

2.3.4 Analysis

To analyze the data, we characterized the demographic history of the lineages and calculated several indirect measures of RI.

Fitting divergence histories

Many comparative studies use genetic distance as a proxy for time since divergence [312, 246]; however, genetic distance might be decoupled from divergence time, especially if migration rates are high or ancestral population sizes are large [252]. Accordingly, we inferred divergence time and other demographic parameters for each lineage-pair by fitting an isolation-with-migration model to genetic data. For *S. basiliscus* N and its sister species *S. lewisi*, because we did not have genomic data for *S. lewisi*, we used previously published genetic data for eight loci [328] to infer model parameters with IMa2 [147].

For the remaining contacts, we used the transcriptomic data to fit an isolation-with-migration model using dadi [135]. dadi uses a diffusion approximation to fit a likelihood model for demographic history to the two-dimensional site frequency spectrum (2D-SFS), which summarizes the distribution of allele frequencies for shared and private alleles. We inferred the unfolded 2D-SFS using ANGSD, which is able to infer a population's SFS without calling individual genotypes [255]. Here, we only used UTR sequence because UTRs are more likely to evolve neutrally than coding sequence [374], and we restricted our analysis to high-coverage regions ($\geq$20$\times$) where we had greater confidence in genotype calling [255]. To construct the unfolded SFS, we polarized SNPs with sequence data from other lineages in the clade. Inferred demographic parameters were converted from coalescent units to real-time units by using estimates of the nuclear mutation rate, assum-

ing a molecular clock, and accounting for differences in total sequenced length across contacts [135]. Our estimate of the nuclear mutation rate (9×10^{-10} $\frac{substitutions}{bp\times generation}$) is derived from fossil-calibrated estimates of the mitochondrial mutation rate in this broader clade of lizards [48] and estimates of the nuclear-mitochondrial substitution rate scalar as inferred from IMa2 results and [328].

Measuring reproductive isolation

To infer the strength of RI and to correlate its evolution with divergence time, we calculated six indices – average nuclear cline width, mitochondrial cline width, coefficient of variance of nuclear cline width, Hardy-Weinberg disequilibrium (F_{IS}), linkage disequilibrium (R_{ij}), and percent hybrids – based on the genotypic data from the contact zones. Importantly, we note that these indices are independent measures of isolation, though some would show correlated responses under certain conditions, such as under a tension zone model [27]. We first collapsed adjoining sampling localities into geographic populations based on their Euclidean distance to their nearest neighbor. Then, we fit clines to our data using the program Analyse [26]. Second, using Analyse, we calculated multilocus measures of Hardy-Weinberg and linkage disequilibrium for all geographic localities across all contacts. Third, we used Structure to estimate each individual's hybrid index [285] and NewHybrids to calculate the number of hybrids in the contact zone [7].

We used our data to contrast different models for how RI accumulates through time, including *linear* and *quadratic* models for the accumulation of total RI and *linear*, *snowball*, and *slowdown* models for the accumulation of DMIs through time (c.f. [130]). We restricted our analyses to three indices of RI – F_{IS}, R_{ij}, and percent hybrids – as starting values for these three indices could be predicted. We did all model fitting using the least-squares approach implemented in R [288], and we chose the best-fitting models by calculating the relative weight of each model based on AIC score.

2.4 Results

2.4.1 Eco-morphological divergence

We used morphological data, including ecologically relevant traits such as body size, limb length and head dimensions, to test for phenotypic divergence across the major phylogeographic lineages within each traditionally defined species (mean sample size, $\bar{N}$=155). We summarized the data as two principal component axes that explained over 97% of the variation and, using these axes, we found little significant morphological variation across phylogeographic lineage-pairs (Fig. S2). Where there is significant variation (Fig. S2), the differences are sex-specific and of small magnitude – e.g., mean body size for female *L. coggeri* varies from 35.7$\pm$2.7 to 38.5$\pm$3.3 mm by lineage.

2.4.2 Fitting divergence histories

Inferring the demographic history of these lineage pairs gave two primary results. First, divergence times vary 3.7×, from 3.1 mya to 11.5 mya. Second, estimates of nuclear divergence and divergence times are tightly correlated (r^2=0.98; p-value = 0.01), an unsurprising result given the low estimates for migration during divergence ($M = 4\times10^{-3}$ to 3.5×10^{-2} $\frac{migrants}{generation}$; Fig. 2B). Full model parameters are available in Table S4.

2.4.3 Measuring reproductive isolation

We used fine-scale spatial sampling and multilocus estimates of hybridization and introgression to infer the strength of RI. We sampled densely through each contact zone, averaging 20 individuals for each of 12 populations per contact ($N = 55 - 406$; Table S1), with the geographic scale of sampling determined by preliminary data on the respective mtDNA transition (range of transect length $= 2 - 16$ km). Based on genotypic data, we then calculated six indirect indices of RI (average nuclear cline width, mitochondrial cline width, coefficient of variance of nuclear cline width, Hardy-Weinberg disequilibrium (F_{IS}), linkage disequilibrium (R_{ij}), and percent hybrids) for each lineage-pair.

We first describe general patterns at each contact zone before summarizing across all the zones, stepping from the least to most divergent contact (Fig. 3). In the *L. coggeri* N/C contact, we see widespread introgression that extends throughout the sampled transect and evidence for two general patterns of introgression in nuclear loci: clines whose center and width is similar to the mtDNA cline and clines which show broad introgression of the Central (C) alleles into the Northern (N) lineage (Fig. 3A). In the *S. basiliscus* N/C contact, we were unable to infer clines at all but one of the nuclear loci; it appears that northern alleles have almost completely introgressed into the Central lineage (Fig. 3B). The asymmetric hybridization in both the *L. coggeri* N/C and *S. basiliscus* N/C contacts could stem from stochastic, demographic or selective processes; disentangling the causes of asymmetry is not possible here so we focus on consensus patterns. Both the *C. rubrigularis* N/S and *L. coggeri* C/S show similar clines across all loci, and both show limited introgression beyond the contact zone (Fig. 3C and 3D). Finally, there is no evidence for hybridization between *S. basiliscus* N and its ecomorphologically similar sister species *S. lewisi*, even when sampled in sympatry (Fig. 3E).

Examining the correlation of divergence time with the six indices of RI, we see significant and strong correlations for all indices but mitochondrial cline width (Fig. 4). As predicted with increasing divergence time, we see decreased cline width and variance in cline width, fewer hybrids, and increased between- and within-loci disequilibria. These results are robust to our estimates of splitting times; using pairwise nuclear divergence gives quantitatively similar results (Fig. S4). Note that not all indices of RI could be estimated for all contacts. We did not infer cline indices for the *S. basiliscus* N/*S. lewisi* contact zone because of insufficient sampling, and we did not estimate either disequilibrium or

hybridization measures for *S. basiliscus* N/C as most nuclear loci were nearly monoallelic throughout the sampled contact zone.

To make our data more broadly comparable to other published data sets, we fit *linear* and *quadratic* models to the increase of RI (as measured by F_{IS}, R_{ij}, and percent hybrids) through time [45]. Although these models have no formal theoretical basis, they reflect the speed and accumulation at which total RI accumulates. Using relative weights from AIC scores, we determined that total RI, as measured by each of these three indices, best fits a model of quadratic growth with time (Fig. S5A-C). We then used our data to contrast three models for the accumulation of DMIs and found that our data fit the *slowdown* model better than the *linear* or *snowball* models, suggesting the rate at which DMIs accumulated slowed down with time ([130]; Fig. S5D-F).

2.5 Discussion

By looking across five contacts in a clade of closely-related and ecologically-similar skinks in the Australian Wet Tropics, we find strong support for the prediction of increasing RI with divergence time. To our knowledge, this is the first comparative study of the strength of lineage boundaries across eco-morphologically similar lineages. These data support the view that phylogeographic splits of increasing depth can represent stages along the speciation continuum – including genetically-cohesive lineages with long-term potential for persistence.

Interestingly, most other data sets comparing RI with divergence time show significantly more noise than ours [45, 312], even thogh these data sets were collected in controlled laboratory settings. In comparison, the strength of our correlations is unexpected, especially given that stochastic processes often influence hybrid zone structure and dynamics significantly [243, 91]. We speculate that the close fit between RI and divergence time in our study stems from the lack of overt divergent selection on eco-morphology, as varying strengths and forms of selection would be expected to introduce rate heterogeneity [113]. In fact, the relationships between different indices of RI and divergence time are so strong that we can use them as a general metric for predicting the progress of speciation in this group. As has been suggested by [178], populations achieve "species status" when linkage disequilibrium (measured here as R_{ij}) is 0.5. Using this relationship, we find that our lineage-pairs are predicted to show $R_{ij} = 0.5$ at 7 Myr or 8.1 Myr after divergence (and at 0.79 or 0.81% nDNA divergence), under a linear versus quadratic model for accumulation of RI, respectively. Although this specific calibration is contingent on our divergence time calibration and unlikely to be generalizable to other taxa, it does provide a yardstick for the tempo of speciation in this group.

Given this rate of speciation – and noting that *C. rubrigularis* N/S and *L. coggeri* C/S show significant but incomplete isolation with even less time – we suggest that the accumulation of RI here is rapid relative to a purely drift-driven model of the evolution of intrinsic RI via DMIs. A simplistic model of drift-driven accumulation of DMIs in allopatry

suggests that the waiting time to speciation is approximately the number of substitutions needed for RI divided by the substitution rate [113], which, given the skinks' estimated mutation rate, could be on the order of hundreds of millions of years. Thus, accumulating substantial RI with minimal phenotypic divergence suggests (1) rapid drift-driven divergence along "holey adaptive landscapes" [113], (2) parallel selection driving mutation-order speciation [260], and/or (3) natural or sexual selection acting on more cryptic phenotypes, like chemiosensory production and perception. As yet, we lack the fine-scale ecological data necessary to characterize the barriers to gene flow acting in this group and to determine which of these hypothetical drivers of divergence are relevant [338]. However, these data do confirm cryptic speciation among phylogeographic lineages and suggest that this could be common, in contrast to the present focus on speciation driven by divergent selection [315].

Looking beyond the velocity of RI accumulation to its acceleration, we find that the total strength of RI increases exponentially through time in this clade (Fig. S5A-C). The pattern of exponentially increasing RI emerges when individual barriers to gene flow combine multiplicatively rather than additively [130], and it is occasionally recovered in other studies [232]. This result suggests that as barriers to gene flow start to evolve, the cumulative effect of these barriers can grow quickly. Thus, species formation can be thought of as an accelerating process – particularly as RI decreases gene flow, which typically further promotes divergence [100]. Further, although our data suggest the rate at which DMIs accumulate might decrease through time (Fig. S5D-F), we refrain from over-interpreting these results because our indices of RI potentially include both pre-zygotic and post-zygotic factors and few data have addressed the model's assumptions of equal and multiplicative fitness effects [130].

2.5.1 Implications for cryptic speciation

Because we can quickly and cheaply query geographic variation within species using mtDNA, deep mtDNA divergence is often used as an initial hypothesis for species delimitation. However, deep mtDNA divergence is neither necessary nor sufficient to delimit species [149], and it is often discordant with other genetic or phenotypic measures of divergence [356]. In this work, we provide additional evidence that mtDNA often presents an idiosyncratic perspective on historical dynamics, finding that mitochondrial cline width is the sole index of RI that had a non-significant correlation with divergence time. Given our and others' findings of mtDNA's idiosyncracy [18], we concur that the observation of deeply divergent mtDNA phylogroups is a useful start of taxonomic and phylogeographic studies, but it certainly should not be the end.

Further, since researchers have begun cataloguing diversity within species, most species have been found to show geographically-restricted variation [11]. Why do some of these intraspecific lineages continuously diverge and exhibit reproductive isolation from sister-lineages, whereas others collapse [317]? As suggested by many, geography is a powerful determinant [11]. But, as we see here, RI can take millions of years to accumu-

late, particularly in the absence of strong ecologically-driven divergence, suggesting that isolation must be sustained across millions of years to lead to genetically independent lineages. In the absence of historical stability, geography can be insufficient, leading to the extensive introgression and discordance often reported in other systems [229, 174]. Thus, as a working hypothesis, we suggest that climatically and geomorphologically stable regions, such as the major refugia of the AWT, are more likely to accumulate such cryptic diversity than are more spatio-environmentally dynamic regions. Indeed, broad-range introgression and discordance are exceedingly rare in the AWT, except amongst lineages endemic to the relatively unstable southern rainforest isolates [328, 34].

Finally, these data add to a growing body of literature that support a Darwinian perspective on species formation [215] and extend this perspective to cryptic diversification. Whether from lineages that likely diverged with gene flow [218] or lineages that diverged in allopatry [274], these data sets show that the accumulation of RI is often a gradual process and that species are not static entities. Indeed, divergence is a continuous, reversible process [317]. For these lineage-pairs, we find that the evolution of RI followed a predictable timeline during their divergence in allopatry (Fig. 4; Fig. S3). Now that lineages have expanded following the Last Glacial Maximum and the original barrier to gene flow (*i.e.*, geography) has disappeared, these lineage-pairs will move again along the continuum. For *L. coggeri* N/C and *S. basiliscus* N/C, this initial divergence will likely be reversed, because the lineage-pairs are hybridizing freely in the apparent absence of RI. However, *L. coggeri* C/S and *C. rubrigularis* N/S appear to be more strongly isolated, because the scale and extent of hybridization and introgression between these lineage-pairs are very limited. Given the limited eco-morphological differentation of these lineages, we hypothesis this RI is intrinsic and not environmentally-dependent and, thus, likely to maintain lineage boundaries even in changing environments. As such, these lineage pairs will likely continue to diverge. That said, RI between these lineages is not complete, but these "leaky" species boundaries can serve as a source of novelty, whether through the evolution of reinforcement or through the selective introgression of adaptive alleles [155, 6].

2.6 Acknowledgements

For advice and discussions, we gratefully acknowledge C. Ellison, T. Linderoth, R. Pereira, and members of the Moritz Lab, and for comments on previous versions of this manuscript, we thank R. Damasceno, C. DiVittorio, J. Patton, R. Von May, and D. Wake. Additionally, T. Linderoth provided scripts used to infer the 2D-SFS from ANGSD output. For assistance in the field, we thank A. Blackwell, E. Hoffmann, C. Hoskin, B. Phillips, M. Tonione, and S. Williams. Support for this work was provided by Museum of Vertebrate Zoology Annie Alexander Fund, National Geographic Society, NSF-GRFP and NSF-DDIG, and the SSE Rosemary Grant Award. Supercomputing resources used in this study were provided by the grid resources at the Texas Advanced Computing Center and

Pittsburgh Supercomputing Center.

2.7 Figures

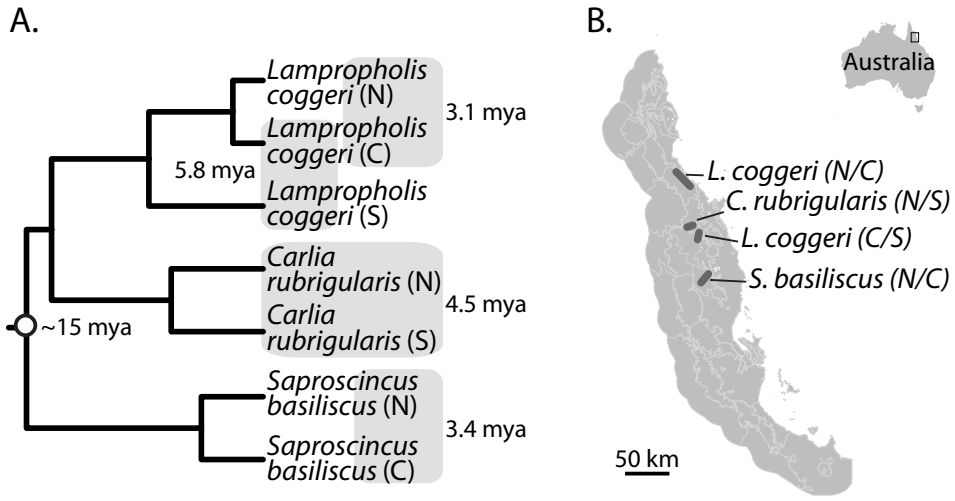

Figure 2.1: A. Phylogeny showing relationships among focal lineages (see SI); boxes outline contact zones. Boxes are labeled with pairwise mitochondrial (top) and nuclear (bottom) divergence. B. A map of the Australian Wet Tropics, labeled with contact zones. A more detailed map of each contact zone is available in Fig. S1.

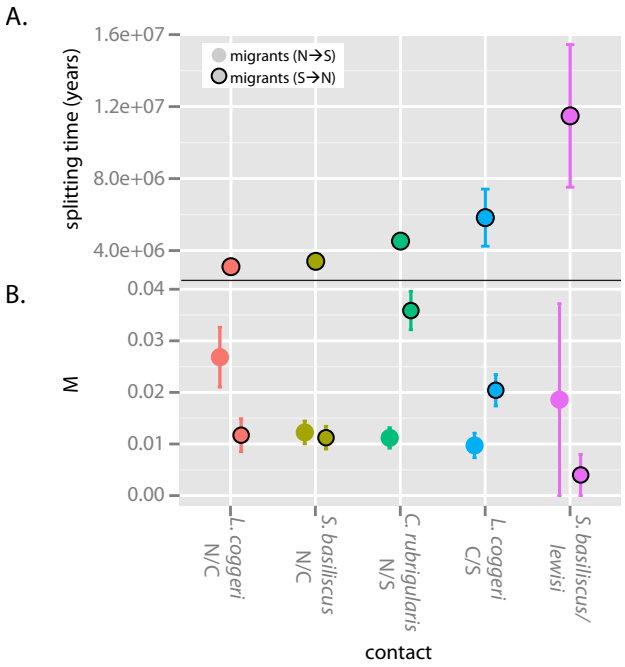

Figure 2.2: A. Divergence times and B. effective number of migrants for the lineage-pairs in this study, as inferred from the two-dimensional site frequency spectrum by dadi and from IMa2. Error bars reflect (as relevant) standard deviation or 95% limits of posterior distribution. For *Saproscincus basiliscus* N/*S. lewisi*, *S. lewisi* is the northern lineage and *S. basiliscus* N is southern.

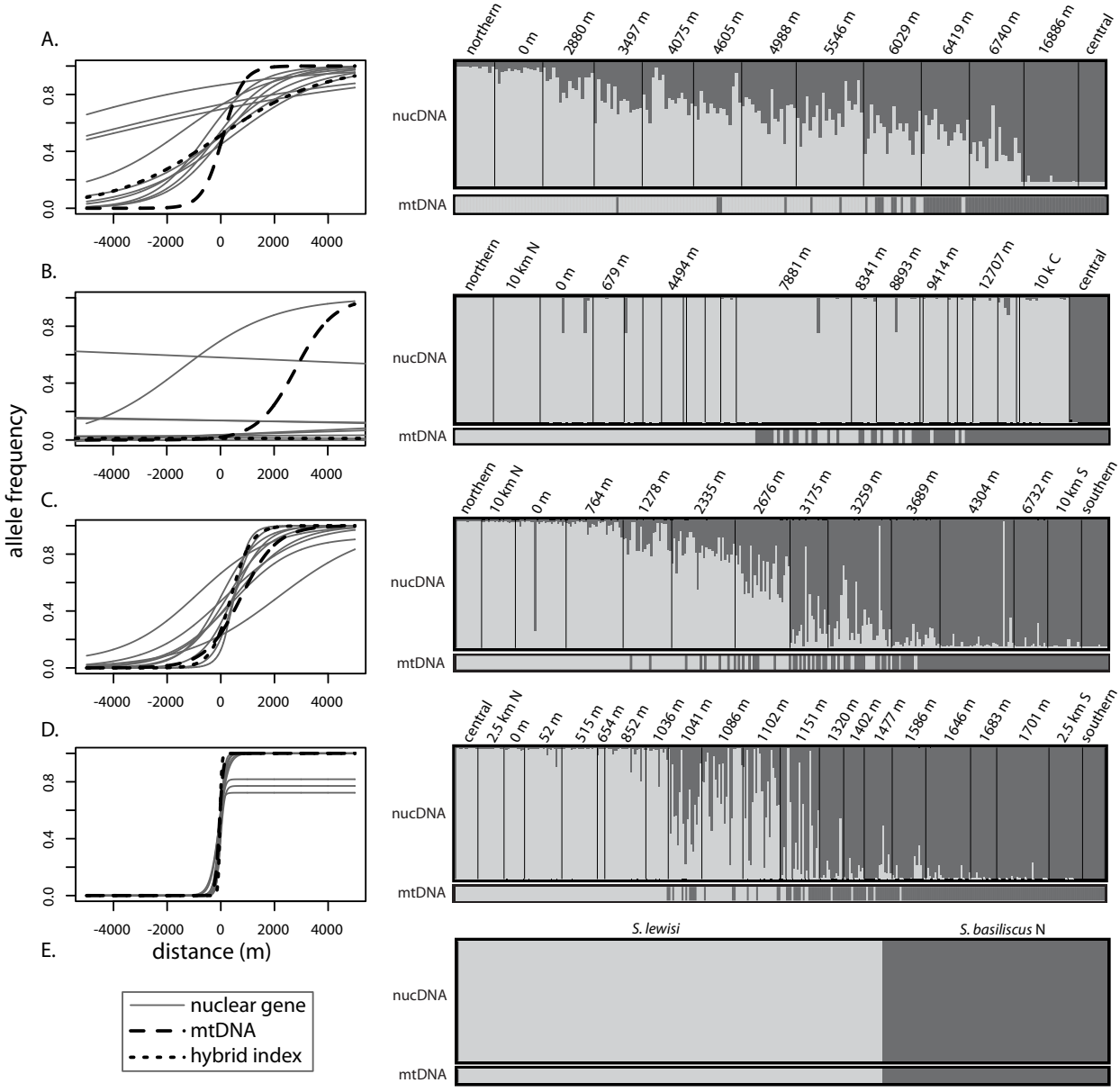

Figure 2.3: Cline fitting (left) and genetic clustering results (right) for contacts in the Australian Wet Tropics suture zone: A. *Lampropholis coggeri* N/C, B. *Saproscincus basiliscus* N/C, C. *Carlia rubrigularis* N/S, D. *L. coggeri* C/S, and E.*S.basiliscus* N/*S.lewisi*. For showing cline fitting results, distances along transects were recalculated so that each hybrid zone center was centered at 0 m. Scale for genetic clustering results differs among contacts.

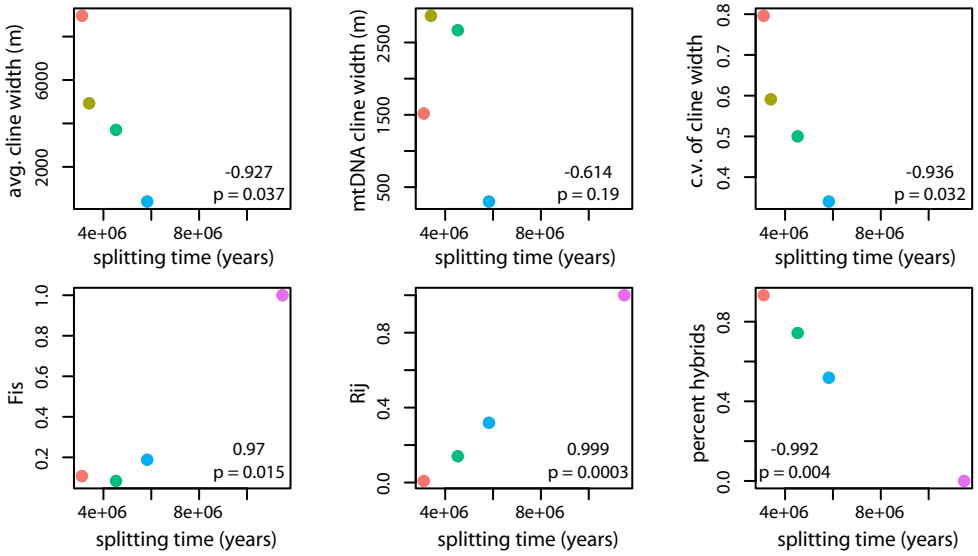

Figure 2.4: Comparative results showing the correlation between divergence time (as measured in years) and different indices of reproductive isolation: average nuclear cline width, mitochondrial cline width, coefficient of variance in nuclear cline width, Hardy-Weinberg disequilibrium (F_{IS}), linkage disequilibrium (R_{ij}), and percent of hybrids in the contact zone. Graphs are labeled with correlation coefficients. Colors follow Figure 2.

Chapter 3

History cleans up messes

3.1 Abstract

Hybrid zones provide an excellent arena in which to address questions about genomic divergence during lineage divergence, because they vividly illustrate the duality of potential interactions between lineages. On one hand, hybrid zones show evidence for barriers between lineages due to low fitness of hybrids and the evolution of pre-zygotic isolation. On the other hand, hybrid zones can be viewed as selective filters, with adaptive alleles introgressing rapidly from one lineage to another, neutral chromosome segments diffusing, and divergently selected segments stacking up at the zone center. These views illustrate the richness of predictions that can be tested in appropriate empirical systems, especially with new genomic approaches. Thusly motivated, we analyze genome-wide introgression data from four contact zones in tropical lizards found in the Australian Wet Tropics. These contact zones all formed between morphologically cryptic lineage-pairs within morphologically defined species, and their divergence times span from 3.1 million years ago to 5.8 million years ago. By fitting geographic clines to these data and inferring the selection history of these same loci, we test two predictions: first, in more highly divergent lineages, the extent of introgression, both in terms of proportion of genome and spatial range, will be more limited than in less divergent lineages. Second, introgression patterns should reflect selection history, such that genes under positive selection will show outlier behavior in introgression extent compared to genes evolving neutrally. Thus, through this work, we simultaneously compare introgression across lineage-pairs with a common biogeographic setting yet different demographic histories, and across genes with variable selection histories within lineage-pairs. We find that divergence history is an important predictor of introgression patterns, whereas locus-specific selection history does not strongly structure introgression patterns.

3.2 Introduction

As lineages diverge, mutation, selection, drift, recombination and gene flow interact to shape patterns of genomic divergence. This rich interplay of processes leads to varied outcomes in natural systems. At one extreme, there are lineages which exhibit extremely different phenotypes but for which genetic divergence is limited to a few regions of the genome, like the benthic and limnetic forms of the widespread stickleback fish [170]. On the other extreme, there are lineages which exhibit conserved phenotypes but show marked differentiation across the genome, like the cryptic lineages of the West African gecko [192]. Further, each of these cases is dynamic, and, with time, the patterns of genomic divergence change. Data from comparative systems show that, just as lineage divergence is an ever iterative process [72, 215, 225], genomic divergence is too [301, 144, 318, 236].

Understanding how genomes diverge as lineages diverge can be explored through two primary, non-mutually exclusive approaches: we can (1) investigate patterns of genomic variation of lineages falling along the speciation continuum and (2) determine what happens when differentiated lineages and genomes interact. In this study, we focus on understanding the genomic consequences of hybridization between lineages meeting in secondary contact. Theory predicts this interaction will be structured by the balance between selection and recombination [25]. During hybridization, two differentiated genomes meet, and the resulting admixed genome is subject to selection, whether due to extrinsic, environmentally-dependent selection against resulting hybrid phenotypes [314] or intrinsic selection against genetic incompatibilities [81, 247]. When selection is strong, disequilibrium remains high across the genome, and loci are trapped from introgressing even if they have no phenotypic effect. However, when selection is weak, recombination dissociates disequilibrium between loci, thus allowing loci to introgress at a rate and extent that reflect the selective effects of that locus and closely-linked loci [16]. In such a case, the hybrid zone functions as a sieve [224], with adaptive alleles introgressing rapidly from one lineage to another, neutral chromosome segments diffusing, and divergently selected segments stacking up at the zone center. As such, hybrid zones both represent the build-up of reproductive isolation between lineages and a potential source of evolutionary novelty through selective introgression, and exploring these two outcomes leads to a richness of predictions that can be tested in appropriate empirical systems, especially with new genomic approaches [272].

We apply this approach to a natural laboratory for comparative analyses of speciation and hybridization: the suture zone of the Australian Wet Tropics. This suture zone, or a geographically-restricted area consisting of multiple overlapping contact zones [294], occurs in a narrow strip of rainforest in northeastern Australia [244]. The suture zone formed due to the repeated glacial cycles of the Pliocene and Pleistocene, and it is comprised of over twenty contacts that formed between long-isolated phylogeographic lineages of rainforest endemics [244]. Almost all the lineage-pairs are morphologically cryptic; however, the lineages span a wide range of genetic divergences. This natural variation

in divergence history allows us to address questions about speciation along the speciation continuum. Here, we focus on four contact zones within a clade of closely-related and ecologically-similar skinks: *Lampropholis coggeri* N/C, *Saproscincus basiliscus* N/C, *Carlia rubrigularis* N/S, and *L. coggeri* C/S. Previous work showed that these lineages have divergence times that span from 3.1 million years ago (mya) to 5.8 mya (Fig. 1A) and that hybrid zones have formed between all these lineage-pairs [326].

Here, we collect a genome-wide data set for introgression across these four hybrid zones. By inferring the selection history of the loci sampled and fitting geographic clines to variation in these same loci, we test two predictions: first, in more highly divergent lineage-pairs, selection against hybrids will be greater, and thus, the extent of introgression, both in terms of proportion of genome and spatial range, will be more limited than in less divergent lineage-pairs [27, 25]. Further, in highly divergent versus less divergent lineage-pairs, the size of introgressing blocks will be smaller [22, 282]. Importantly, in making these predictions, we assume that these contact zones formed concurrently and that recombination rates are similar across the lineages, both reasonable assumptions given the lineages' shared history and similar biologies [244, 373]. Second, given recombination has broken up associations between loci, introgression patterns should reflect selection history, such that genes under positive selection will show outlier behavior in introgression extent compared to genes evolving neutrally [106, 100, 333]. By testing these two predictions, we simultaneously compare introgression across lineage-pairs with a common biogeographic setting yet different demographic histories, and across genes with variable selection histories within lineage-pairs.

3.3 Methods

3.3.1 Sampling

For this study, we sampled four hybrid zones, previously characterized in [326]: *Lampropholis coggeri* N/C, *Saproscincus basiliscus* N/C, *Carlia rubrigularis* N/S, and *L. coggeri* N/C (Fig. S1; Table S1). For each hybrid zone, we defined nine populations across each transect. We identified six populations that occurred approximately 10, 2.5 and 1 km north and south of the hybrid zone center, two populations at the 20% and 80% tails of the average cline in the hybrid zone, and a final population at the average cline center. Here, average cline was estimated using previously published data for ten nuclear clines [326]. In total, we sampled an average of 133 individuals per contact zone, and each population consisted of an average of 14.8 individuals ($N = 8 - 17$). But for individuals comprising the the 10 km populations for the *L. coggeri* C/S contact zone, all other individuals were included in previous studies [327, 278, 326].

3.3.2 Design of exome capture array

In order to capture a uniform subset of the genome across populations, we designed chip-based exome capture arrays, with each array being specific for a single contact zone. Following [39], we used blast [5] to annotate gene identities and exonerate [332] to define exon-intron boundaries of each lineage's *de novo* transcriptome assembly [325]. These transcriptomes used were sequenced from populations geographically-isolated from the contact zone and are unlikely to contain recently introgressed alleles. All annotation was done with *Anolis carolinensis* (AnoCar2.0+), the reference genome most closely-related to our clade. From all possible exons, we identified exons for which (1) we could identify orthologs in both lineages of the lineage pair, (2) GC-content was between 30% and 80% as suggested by [39], (3) multi-individual transcriptome data showed evidence for variation at the locus, whether fixed or segregating [325], and (4) exon length was at least 200 base pairs (bp). To this, we added (1) full-length transcripts for genes ($N = 95 - 99$) that showed evidence of positive and divergent selection, as identified by calculating $\frac{divergence}{polymorphism}$, $\frac{dN}{dS}$ and F_{ST} values [385], (2) full-length transcripts for genes ($N = 92$) with metabolic or reproductive function, because these genes possibly contribute to cryptic phenotypes with relevance for species boundaries, (3) two mitochondrial loci, *NADH dehydrogenase subunit 4* and *16s ribosomal RNA*, and (4) the 5' and 3' untranslated regions (UTR) for regions previously genotyped in these populations ($N = 10$; [326]), which we used to evaluate the efficacy of our pooled strategy. Importantly, because this study's goal was to infer allele frequencies of variation in targeted loci, we tried to minimize bias in capture efficiency by printing orthologs from each lineage for each targeted exon.

After selecting the initial set of targets, we processed the targets further to ensure that none contained repetitive sequence – such targets are likely to be captured at excessive coverage and can thus reduce the overall efficiency of an array. To do so, we removed targets that (1) matched to repeats in the RepeatMasker database [336], (2) were highly similar to other targets on the same array, and (3) contained *k*-mers that were disproportionally common in the *A. carolinensis* genome.

In total, we targeted an average of 3082 unique loci, representing 1.83 Mbp of sequence. Across all four exome capture arrays, there were 1120 loci in common. Although we did introduce some bias in how we selected these loci, the patterns of evolution at our targeted loci mimic those for all known loci (Fig. S3). The targets were then printed at 2bp tiling on Agilent 10M eArrays. The scripts used to design the exome capture array are available at https://github.com/singhal/probeDesign.

3.3.3 Library preparation, exome capture, and sequencing

Making uniquely barcoded libraries for each individual would have been prohibitively expensive and time-consuming; thus, we employed an anonymous pooling strategy instead [391]. We first extracted high-quality DNA from each sample using a high-salt extraction method [4], measured each sample's DNA concentration using a Nanodrop, di-

luted samples to approximately 100 $\frac{ng}{uL}$, and then measured DNA concentration using a broad sensitivity kit with Qbit. Using these estimates of DNA concentration, we pooled equimolar amounts of DNA per individual for each population. We then sonicated the pooled DNA samples to 200 bp – 400 bp using a Covaris shearer and used the sheared DNA to prepare uniquely barcoded libraries following [234]. After measuring the concentrations of the libraries using a Qbit, we pooled libraries by contact to obtain a total of 20μg for exome capture. For each contact, we then captured targeted exons using the custom arrays, following the protocol published in [152] with modifications by [39]. To improve capture efficiency, we isolated COT1 DNA from *L. coggeri*, following the approach outlined in [357], and used a 50/50 mix of *L. coggeri* and commercially-purchased chicken COT1 DNA as our blocking reagent for all exome capture experiments (Hybloc; Applied Genetics Lab). We validated the success of the capture experiment by using qPCR to calculate the difference in T_m for pre-capture and post-capture libraries at target and non-target loci. The size distribution, quality, and quantity of final libraries were checked using a Bioanalyzer and Qbit and then sequenced on an Illumina HiSeq 2000 at the Vincent Coates Genome Sequencing Laboratory at University of California, Berkeley. Each contact was sequenced to one lane depth to get the necessary coverage for accurate allele frequency estimation (see Supplemental Information).

3.3.4 Data filtration, assembly, and variant discovery

The raw data were first cleaned and trimmed for quality, using cutadapt and trimmomatic to remove adaptors, trimmomatic to remove low-quality sequence via a sliding window approach, and cope and flash to merge overlapping paired-end data into single reads [223, 205, 203, 212]. Then, although we already had target sequences that could be used as a reference genome, we assembled the cleaned reads to form a new reference assembly, to extend target lengths and thus reduce edge effects [39]. For all contacts, we used ABYSS to generate *de novo* assemblies at varying *k*-mer lengths (8 assemblies from $k = 21 - 91$). Because the resulting assemblies are highly redundant, we then merged the *de novo* assemblies using our custom merge script, which uses blat and cd-hit-est to identify overlapping clusters of contigs and cap3 to assemble these clusters [176, 201, 156]. Following assembly, we identified contigs that matched to targeted exons using a reciprocal-BLAST approach [176]; this set of contigs became our reference assembly.

We then aligned our cleaned sequencing reads to the reference assembly to identify variable sites. First, we identified variation that was segregating far from the hybrid zone. To do so, we used the reads from our transcriptome sequencing [325]; these reads are derived from individuals from populations geographically-isolated from the contact zone. Second, we identified variation that was segregating in the hybrid zones, by aligning reads from our pooled libraries. For both steps, we used bowtie2 to iteratively map reads globally and then locally, using paired-end information where possible [189]. Importantly, although duplicate reads resulting from PCR over-amplification of libraries are typically removed at this stage, we chose not to do so here, because our high coverage

likely led to many "accidental" duplicates. Variable sites were then called using samtools mpileup [198], modifying default settings to discard indel variation, to turn off the filter removing closely-occurring mutations, and to ignore the Hardy-Weinberg filter. Through this two-step process, we identified a putative set of variable sites. For these variable sites, we used samtools to call genotypes for each individual from the geographically-isolated populations and to calculate allele frequencies for each population in the hybrid zone transect [198].

The pipeline outlined here is available at https://github.com/singhal/exomeCapture and is summarized in Fig. S2.

3.3.5 Analysis

Our analysis consisted of three primary objectives.

Evaluating Success of the Experiment

We measured the efficacy of our anonymous pooling strategy and our overall exome capture experiment. To do so, we compared known allele frequencies to estimated allele frequencies, looked at correlation of allele frequencies across SNPs within non-recombining molecules (*i.e.*, mitochondrial DNA), and calculated standard measures of exome capture efficacy [268]. Further details are available in Supplemental Information.

Inferring Patterns of Introgression

We used the variation data to infer introgression patterns at each SNP and to identify SNPs that showed outlier patterns of introgression. We first filtered our allele frequency data by restricting analysis to SNPs which had greater than $50\times$ coverage in every transect population (see Supplementary Information for rationale) and removing any SNPs for which we were missing allele frequency information at any transect population. For those remaining SNPs where the difference between the lowest and greatest allele frequencies across all populations was $p \geq 0.50$, we fit sigmoidal clines to the transition in allele frequencies through the contact zone. We only used allele frequency data from the seven central populations, because the 10-km populations were typically off the linear hybrid zone transect. Clines were fit using standard cline functions [27] and the nls function in R [288]. We did not explore more complex models of cline shapes, such as those allowing asymmetry or exponential decay in the tails of the cline [27], because these models require denser sampling than we have. For those SNPs where the difference in allele frequency across all the populations was $p < 0.5$, we determined if any of these SNPs showed introgression patterns indicative of a sweep. Introgression at a SNP was categorized as a "sweep" if the allele frequency difference between the ancestral populations was $p \geq 0.5$ and allele frequencies at all populations in the contact zone – including the 10 km populations – were uniformly within $p = 0$ or $p = 1$.

Having defined introgression patterns at each SNP, we then identified those SNPs with outlier patterns of introgression. Ideally, to identify genomic outliers, one would generate a neutral distribution for the given metric under a certain demographic history [253]. Those loci that deviate substantially from the neutral distribution would then be identified as outliers. While characterizing the demographic history of diverging populations has become standard (c.f. [354, 147]), doing so for hybrid zones has yet to be done and is beyond the scope of this work. Thus, we choose to identify outliers by using simple cutoff – the clines with widths in the bottom 5% of the distribution were considered to be "narrow" and those in the top 5% "wide" (Fig. S14, S16).

Testing Predictions

To test our predictions, we conducted several analyses. First, we profiled general patterns of cline center and width and tested if any of the hybrid zones show evidence for asymmetric introgression based on displacement of individual cline centers from the consensus center. Second, because we lacked the individual genotypic data necessary to infer linkage disequilibrium across the genome, we calculated Moran's I, a spatial auto-correlation measure, as a proxy. In systems where selection overwhelms recombination, clinal patterns will be constant over genomic space and will lead to high Moran's I. Where selection is weak, recombination will quickly dissociate patterns between neighboring SNPs, leading to low Moran's I. We calculated Moran's I for cline width following [267], focusing only on those contigs for which we characterized introgression at multiple SNPs. Third, we investigated the connection between locus-specific selection histories and patterns of introgression. To do this, we first characterized patterns of molecular evolution at each contig, as determined by variation at the populations geographically-isolated from the contact zone (Fig. S1). Using those variants, we calculated several indices of evolution: $\frac{divergence}{polymorphism}$, F_{ST}, and $\frac{dN}{dS}$ [385]. For $\frac{divergence}{polymorphism}$ and F_{ST}, loci that had the highest 5% values were identified as outliers, and for $\frac{dN}{dS}$, loci with $\frac{dN}{dS} > 1$. We then (1) calculated correlations between cline metrics and these indices, (2) compared these indices across the different introgression categories ("normal", "narrow", "wide, "sweep"), and (3) compared patterns of molecular evolution across loci with "normal" and "outlier" patterns of introgression.

3.4 Results

3.4.1 Efficacy of Exome Capture Experiment

The exome capture experiments for each contact zone were successful; briefly, we acquired high-coverage and high-quality data for our targets, extended our in-target assembly by 60% by assembling our reads *de novo*, and recovered high and consistent speci-

ficity (≈65%). Further details on and explanations of how we evaluated the experiments are available in the Supplementary Material.

We also evaluated the success of our anonymous pooling strategy two ways. First, we compared estimated allele frequencies to known allele frequencies from previous genotyping studies (Fig. S12; [326]). We found substantial and significant correlation between estimated and known allele frequencies (average $r^2 = 0.97$), suggesting that sampling drift due to anonymous pooling was minimal. Second, we estimated variance in estimated allele frequencies at fixed SNPs in the mitochondrial genome. Because the mitochondrial genome is non-recombining, all fixed SNPs should have the same allele frequency in a population. As seen in Figure S13, the variance in estimated mtDNA allele frequencies was minimal for almost all populations and contacts.

The approach effectively discovered variation for downstream analyses; we identified an average of 57K SNPs after filtering (Table S4), and we were able to fit clines at anywhere from 2.6K to 14.2K of these (Table S5).

3.4.2 Testing Biological Predictions

Across Contacts

We tested the influence of divergence history on introgression patterns by comparing results across contacts. The number and type of introgression patterns we found varied across contacts (Fig. S14); with increasing divergence time between lineage pairs, we were able to fit more clines because more SNPs were highly differentiated. That said, we were unable to fit clines at many of the highly-differentiated SNPs in *L. coggeri* N/C, because the patterns of allele frequencies across the hybrid zone were erratic. Additionally, we see that more SNPs in the *S. basiliscus* N/C contact zone show evidence of a sweep pattern, a result that follows from previously-collected data that suggest there has been massive unidirectional introgression in that zone.

To further describe the hybrid zones, we looked at patterns of asymmetry in cline center. Shifts in distribution might reflect demographic forces that structure the hybrid zone, such as differences in population density between lineages that lead to biased introgression [74]. Such demographic processes would inform our ability to generalize these results. Here, we see significant evidence of skew in cline center (Fig. S18); this asymmetry provides important context for any interesting locus-specific patterns.

Summarizing patterns of cline width across contact zones (Fig. 2) recapitulates the pattern we found in previous work of decreasing cline width as divergence time between lineage-pairs increases [326]. What is clearer with these data, however, is that both the average spatial extent of and the variance in introgression are reduced as divergence time increases (Fig. 2). Indeed, comparing the distributions of cline width and center across contact zones, we can see that they become noticeably tighter between older lineage-pairs.

Because we could not calculate linkage disequilibrium in these contact zones, we used Moran's I, a spatial auto-correlation measure, to infer how quickly patterns of cline width

change over genomic space. Here, we find that the two less divergent lineage-pairs (*L. coggeri* N/C and *S. basiliscus* N/C) have almost no evidence of spatial autocorrelation beyond 100 base pairs (Fig. 3). The two more divergent lineage-pairs (*C. rubrigularis* N/S and *L. coggeri* C/S) have more extensive autocorrelation that extends for at least 1 kB, with only moderate declines over distance.

Across the Genome

We tested whether locus-specific selection histories influence introgression patterns at that locus, such that loci with outlier patterns of evolution also show outlier patterns of introgression. For each contact, we calculated correlations between indices of molecular evolution at a locus and cline width (Fig. 4). Almost all the correlations were statistically insignificant, but for widths from *C. rubrigularis* N/S and contig and SNP F_{ST}. However, even in these cases, the absolute magnitude of the correlation was weak.

Further, comparing patterns of molecular evolution across cline types indicated some significant differences for patterns at contig F_{ST} and divergence (Fig. 5). Here, narrow clines showed evidence of occurring in loci with greater F_{ST} and divergence, whereas wide and sweep clines had lower F_{ST} and divergence. However, although statistically significant, the differences among cline types are absolutely limited, suggesting that the impact of locus-specific selection on introgression patterns is minimal. If we instead look at cline widths conditioned on whether a locus shows outlier patterns of molecular evolution, we see significant differences for all contacts when considering F_{ST}. However, the patterns are inconsistent across contacts. We also see significant differences in cline width given $\frac{divergence}{polymorphism}$ outliers for *C. rubrigularis* N/S and *S. basiliscus* N/C, but again the patterns are inconsistent between the two contacts.

3.5 Discussion

In this study, we employed a genomic perspective to look at comparative patterns of introgression across a suture zone. By sampling across the genome and across contacts, we characterized patterns across different selection and divergence histories. Through our data sets, we find that divergence history is an important predictor of introgression patterns, whereas locus-specific selection history does not strongly structure introgression patterns.

3.5.1 Overall Patterns

These genome-scale results broadly recapitulate the results we uncovered in previous work. In that and this work, we broadly see that lineage-pairs become more divergent, the extent of introgression becomes more limited. Our work builds on these results in a few novel ways. First, earlier work from the *L. coggeri* C/S hybrid zone found that

clines between this lineage-pair were all exceptionally narrow and largely concordant [327]. Theory predicts that, unless selection is extremely strong, some clines will diffuse neutrally and thus should be wider and non-concordant [25]. Because we failed to recover this pattern, we originally hypothesized that this hybrid zone might not be at equilibrium. In a system with neutral diffusion, cline width is given as $w = \sqrt{2\pi}\sigma\sqrt{t}$, where σ is the per-generation dispersal length and t is generations since secondary contact [100]. Our previous estimates suggest $\sigma \approx 80\frac{m}{\sqrt{gen}}$ and $t \approx 80$ generations, which would give neutral clines of width 10,000 meters. In this expanded data set, we found ≈5% (946 of 18,081) of all clines were this wide or wider, suggesting neutral diffusion is occurring, albeit rarely. Thus, in contrast to our proposed hypothesis, these data suggest that this hybrid zone is at tension zone equilibrium; however, most of the genome is subject to direct or correlated selection, leading to narrow and coincident clines. Second, the clines are much narrower for *L. coggeri* C/S than *C. rubrigularis* N/S (Fig. 2), and given the two species have similar dispersal rates [278, 327], this difference suggests selection against hybrids is stronger for *L. coggeri* C/S. However, *C. rubrigularis* N/S exhibits a much higher level of Moran's I (Fig. 3), a result that suggests recombination is less effective in breaking up associations between loci in *C. rubrigularis* N/S [16]. While this pattern could certainly emerge due to stochastic demographic forces, it also could result from differences in the genomic structure of segments under selection, such that the individual locus effect is stronger in *C. rubrigularis* N/S, leading to more extensive spatial auto-correlation (or, linkage disequilibrium) than in *L. coggeri* C/S. Third, earlier work showed that both the *S. basiliscus* N/C and *L. coggeri* N/C contact zones had evidence for asymmetric hybridization. For most variation in *S. basiliscus* N/C, the Northern allele had completely introgressed into the Central lineage [326], and for *L. coggeri* N/C, a small portion of loci had introgressed largely from the Central lineage into the Northern lineage. For both of these hybrid zones, it was unclear if the patterns we recovered were from stochastic or selective processes. However, with a larger data set, we find that a substantial portion of loci show asymmetric patterns (Fig. S18), suggesting that the asymmetry is likely due to demographic effects, such as those outlined in [71, 91]. Interestingly, this asymmetry might help us better understand genomic differentiation as it relates to speciation in these lineages. Because asymmetric introgression has swept away so much of the genomic divergence that had accumulated in allopatry, the divergence that remains might reside in those genomic regions that contribute to species-specific phenotypic differences.

3.5.2 Implications for Speciation and Adaptation

In this work, we explore the dual nature of selection in defining introgression patterns. In testing the effect of divergence history on introgression, we are implicitly testing the strength of selection on hybrids, which acts at the level of the individual and thus can affect a large portion of the genome. Strong total selection against hybrids leads to extensive linkage disequilibrium in hybrid zones, preventing introgression even at loci that

neither have, nor are linked to loci with, selective effect. We see this pattern in the highly divergent lineage-pairs *L. coggeri* C/S and *C. rubrigularis* N/S, both as the narrow and limited range of introgression (Fig. 2) and as the high spatial auto-correlation in introgression across genomic space (Fig. 3). These results follow nicely from population genetic descriptions of speciation as the accumulation of linkage disequilibrium [105, 178]. In the lesser divergent lineage-pairs *L. coggeri* N/C and *S. basiliscus* N/C, selection against hybrids appears to be weaker, and accordingly, the extent of introgression is broader and spatial correlation is limited. This leads to the observation that history cleans up messes, or that time, and the divergence that typically accumulates with time, leads to patterns across the genome congealing such that two divergent genomes eventually act as completely isolated units [381]. Selection against hybrids could be due to either a mismatch of the hybrid and its environment (*i.e.*, extrinsic effects) or due to internal genic incompatibilities (*i.e.*, intrinsic effects) [68]; here, given the limited ecological and morphological differentiation between lineages [326, 244, 373], we suggest intrinsic effects are more likely. This work is mirrored by studies in ecologically distinct lineages – such as multiple pairs of ecomorphs in benthic/limnetic sticklebacks [144], white-sand/dark-soil lizards [301], blue/red cichlids [318], which have shown increased ecological differentiation, along with geographic isolation, leads to ever increasing genome-wide differentiation. Thus, extending a Darwinian perspective on species as ever evolving lineages [72, 215], we see that genomic divergence during lineage divergence is also ever evolving, with the genome differentiating into isolated units through the interaction of geography, selection, and time.

We also tested how selection varies across the genome and how that structures introgression, finding some evidence that the selection history under which a locus evolved influences the rate and extent of its introgression (Fig. 4 & 5). However, the effects, while significant, were modest, particularly given the striking differences in introgression extent across contact zones. Some empirical studies have similarly found little correlation [261, 126], whereas others have shown that locus-specific histories can do a good job predicting introgression patterns at a locus [267, 296]. Given this, why does selection history have such a limited role in explaining introgression patterns in this system? First, our power to detect such patterns is eroded by demographic processes – such as genetic drift across hybrid zone populations [281], differences in recombination rate across loci [16, 248], and variance in historical gene flow across loci [286, 384] – which can swamp the signal from an individual locus. Second, we only sampled a small portion of the genome (ca. 15% of coding regions; 0.1% of the total genome), and although we biased our genomic sampling to include those loci with outlier patterns of molecular evolution, we could easily have failed to sample the loci that are the outliers in this system. Third, our hypothesis that selection history influences introgression assumes that highly-differentiated loci are adaptive. The genic basis of adaptation is still little understood, but thus far, the data are equivocal that genetic patterns suggestive of adaptive differentiation actually reflect a locus-specific effect on organismal fitness [308, 289, 250]. Fourth, if the genetic basis of the traits under selection in these lineage-pairs is complex, such that there are many loci

of small effect [299], then the selective effect of a given locus would be small. As width is related to the inverse square root of selection [333], small differences in selection will only have modest impacts on cline width. Sixth, and perhaps most importantly, our analysis considers the selection history of each exon as a whole and attempts to connect locus-wide patterns of selection to introgression patterns at a single SNP within that locus. However, our Moran's I results show minimal spatial auto-correlation in *S. basiliscus* N/C and *L. coggeri* N/C, suggesting that, within a locus with the same selection history, SNPs have a variety of introgression behaviors. Perhaps the relevant unit of analysis is the selection history of a given SNP (*i.e.*, the QTN; [299]) and not the locus in which the SNP resides.

3.5.3 Study Limitations

This study has a number of limitations, which influence the reach of the biological inference we can make and are therefore useful to consider as context. First, we do not have a good approximation for genetic distance between variants in these lineages. Genetic linkage maps provide useful data on recombination rates across the genome, which empirical data from other systems show can vary greatly [248] and which can strongly influence patterns of introgression [25, 16, 184]. Thus, this work does not appropriately control for recombination in evaluating locus-specific patterns; in future work, it will be important to understand how recombination varies across the genome and interacts with selection to define introgression extent. Second, in studies aiming to identify outlier patterns of variation across the genome, having an expectation for the neutral distribution of variation is important [253]. Demography can introduce stochasticity, such as the asymmetry in introgression we see in *L. coggeri* N/C and *S. basiliscus* N/C, which affects the neutral distribution of variation, and thus, how one defines outliers. However, characterizing the demography of a hybrid zone is challenging, because no formal analytical models exist to describe how time since contact, population sizes and growth rates, deme structure, migration rates, dispersal lengths, and asymmetry in any of these variables influence introgression. Simulation studies suggest these factors can have profound effects on hybrid zone structure and introgression patterns [281, 235, 120], which would obscure our ability to disentangle the relative roles of demography and selection in defining introgression outliers. Third, further compounding this challenge, we were only able to run one transect through each hybrid zone because of habitat availability. Other studies have shown that, while global and general patterns of introgression are similar across transects, locus-specific patterns often vary greatly [163, 257], for the reasons outlined above. Thus, one way to counteract this intrazone variability is to take consensus patterns across zones, which we are unable to do in this work.

3.5.4 Conclusions

Most studies investigating genomic divergence through lineage divergence have focused on lineages exchanging large number of migrants every generation and between which

there is marked ecological differentiation [258, 170, 261]. This study focuses on lineages that diverged with minimal gene flow and that are ecologically and morphologically similar [326]; such divergences are an important, and in some ways understudied, component of Earth's biodiversity [40]. This work finds that the evolution of isolation across the genome is an iterative and heterogeneous process, as theory predicts [381], and in particular, our work underscores the role of time in defining this process. Given the current emphasis on how ecology drives lineage divergence [258, 315], this work reminds us of the pivotal role history has to play in cleaning up messes and focusing patterns.

3.6 Acknowledgements

We thank D.B. Wake for his thoughtful dialogue, during which he shared the idea that "history cleans up messes", which served as this paper's inspiration. Additionally, for advice, we gratefully acknowledge G. Coop, C. Moritz, T. Linderoth, J. Novembre, R. Pereira, J. Schraiber, M. Slatkin, and F. Viera, and for technical support, M. Chung and L. Smith. Funding was generously provided by a NSF Graduate Research Fellowship, Museum of Vertebrate Zoology Koford & Albert Preston Fund, and NSF DDIG. The Texas Advanced Computing Center (TACC) at The University of Texas at Austin provided grid resources that contributed to the research results reported within this paper.

3.7 Data Accessibility

Data are available at the following locations:

1. Scripts used for exome capture array design are available at `https://github.com/singhal/probeDesign`.
2. Scripts used for exome capture bioinformatic analysis are available at `https://github.com/singhal/exomeCapture`.
3. Scripts used for exome capture biology analysis are available at `https://github.com/singhal/exomeAnalysis`.

3.8 Figures

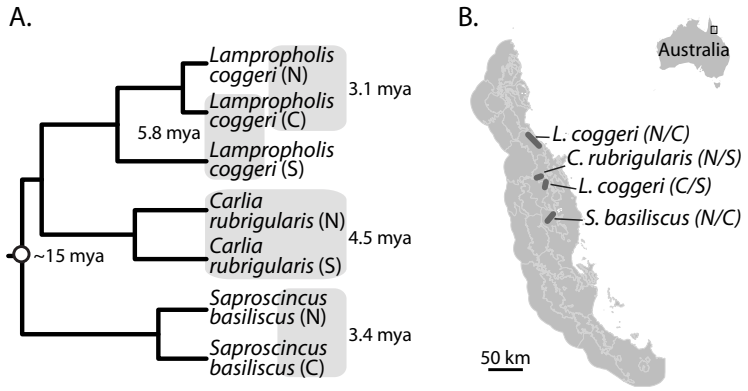

Figure 3.1: A. Phylogeny of lineages used in this study; gray boxes indicate lineage-pairs meeting in hybrid zones. Boxes are labelled with divergence time estimates for the lineage-pair [326]. B. Map of the Australian Wet Tropics, labelled with contact zone locations.

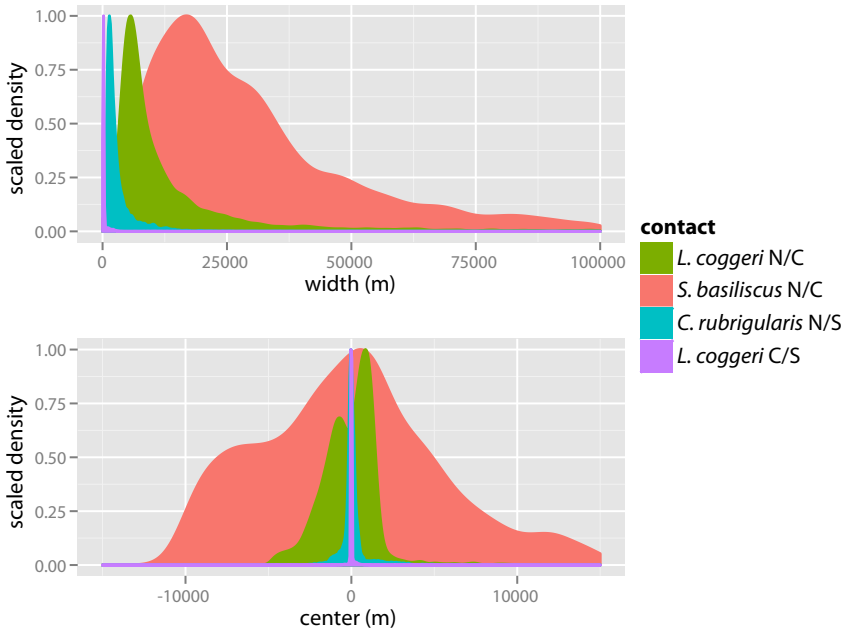

Figure 3.2: Distributions for A. cline width and B. cline center for the four contact zones.

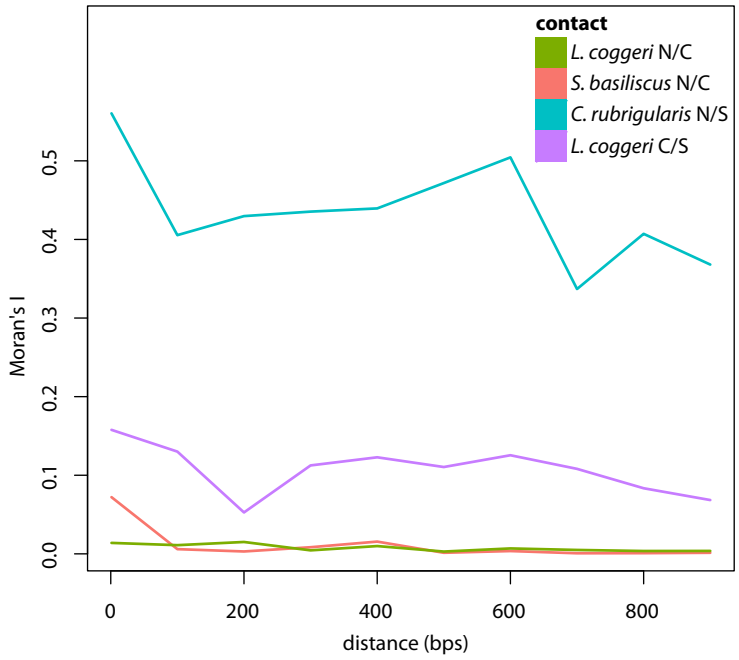

Figure 3.3: Moran's I, a measure of spatial auto-correlation applied to genomic distance, for cline width at each of the four contacts.

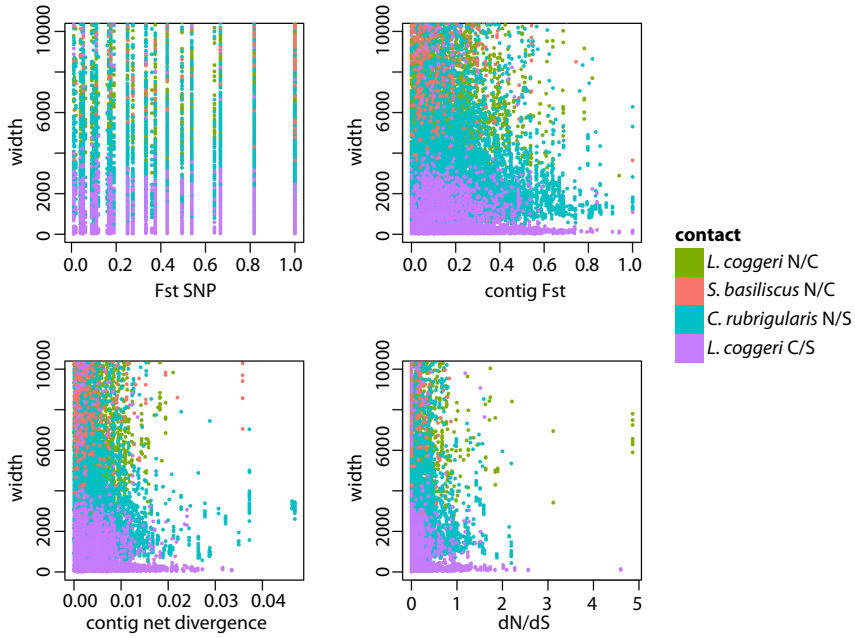

Figure 3.4: Correlation between locus-specific summary statistics measuring the rate of molecular evolution and cline widths.

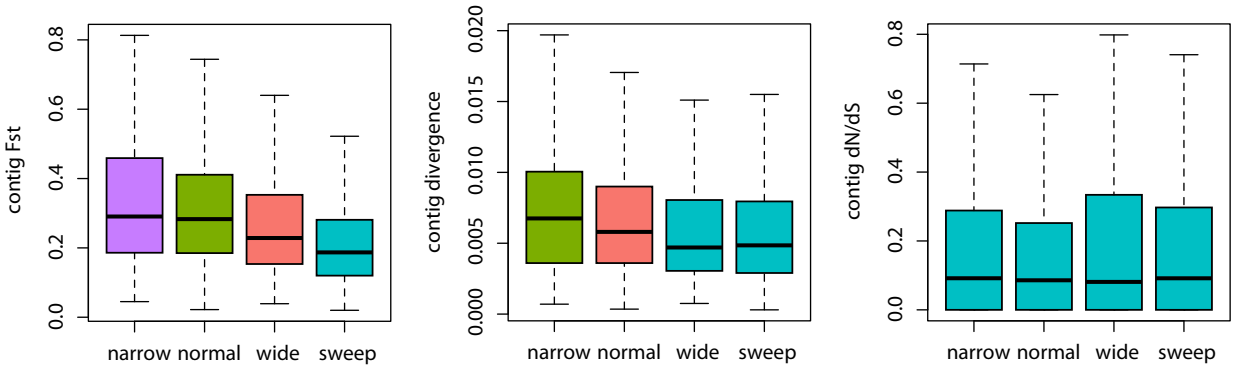

Figure 3.5: Differences in locus-specific summary statistics across cline types. Fill color reflects statistically significant groupings.

Chapter 4

Genealogical discordance in a rainforest lizard

4.0.1 Abstract

Genealogical discordance, or when different genes tell distinct stories although they evolved under a shared history, often emerges from either coalescent stochasticity or introgression. In this study, we present a strong case of mito-nuclear genealogical discordance in the Australian rainforest lizard species complex of *Saproscincus basiliscus* and *S. lewisi*. One of the lineages that comprises this complex, the Southern *S. basiliscus* lineage, is deeply divergent at the mitochondrial genome but shows markedly less divergence at the nuclear genome. By placing our results in a comparative context and reconstructing the lineages' demography via multi-locus and coalescent-based Approximate Bayesian Computation (ABC) methods, we test hypotheses for how coalescent variance and introgression contribute to this pattern. These analyses suggest that the observed genealogical discordance likely results from introgression. Further, to generate such strong discordance, introgression probably acted in concert with other factors promoting asymmetric gene flow between the mitochondrial and nuclear genomes, such as selection or sex-biased dispersal. This study offers a framework for testing sources of genealogical discordance and suggests that historical introgression can be an important force in shaping the genetic diversity of species and their populations.

4.1 Introduction

Genealogical discordance is a common phenomenon in natural systems [229, 249, 230, 293], yet the causes and consequences of discordance are often unclear. Under genealogical discordance, not all loci appear to tell the same story, even though the genes evolved under a common demographic history. This discordance can take several forms; most notably, topologies and branch lengths among organismal lineages can vary across

loci [165, 96]. Of note is discordance between organelle (chloroplastic and mitochondrial genomes) and nuclear loci, which appears throughout the natural world [58, 276]. Studies often point to the special characteristics of the organelle genome – *i.e.*, its smaller effective population size, uniparental inheritance, lack of recombination, key role in organismal metabolism, and, in the case of the mitochondrial genome, increased mutation rate [18] – to explain this discordance. However, it is unclear if any of the special characteristics of cytoplasmic genomes need to be invoked to explain cytonuclear discordance as compared to general genealogical discordance [71].

Whether in the form of discrepancies in topology or branch lengths, genealogical discordance typically arises from three, non-exclusive processes: coalescent variance ("incomplete lineage sorting"), introgression, and gene duplication [211]. Here, we focus on coalescent variance and introgression. First, the coalescent, or the process by which alleles in a population find a common ancestor, is inherently stochastic [365]. Thus, theory predicts that any genealogical reconstruction should exhibit some heterogeneity across loci – not only because the coalescent is a sampling process – but also because we rely on the distribution of mutations, another stochastic process, to estimate coalescent histories [365]. How much heterogeneity is expected is unclear; for a subset of population histories, researchers have derived analytical expectations for the variance in coalescent times and inferred genealogical relationships [352, 334]. However, this work is generally limited to simple splitting histories, and more complex patterns of divergence could possibly increase this variance [195]. A second powerful source of genealogical discordance is introgression, or the movement of an allele from one gene pool to another [6, 185]. Particularly when introgression acts in concert with other forces such as locus-specific selection or sex-biased dispersal, genealogical discordance can increase further [221].

Both coalescent variance and introgression are often invoked by researchers trying to explain patterns of genealogical discordance [249]. Although not always easy [293, 231], it is important to determine how these forces interact in the context of a species' history to create genealogical discordance. After all, discordance is useful for inferring key parameters about the divergence process [97]; in particular, discordance increases with larger ancestral population sizes [147]. Further, because it can result from introgression, discordance can inform us about both historical and ongoing hybridization in the lineages of interest [249]. Thus, although genealogical discordance has sometimes been regarded as a complication [76], it actually can provide an informative window into species' histories.

In this study, we present a compelling case of genealogical discordance in the rainforest lizard species complex, *Saproscincus basiliscus* and *S. lewisi*. Endemic to the Australian Wet Tropics (AWT), a narrow strip of rainforest in northeastern Australia, this complex consists of four major, highly divergent mitochondrial lineages: *S. lewisi* in the far north, and the Northern, Central and Southern lineages of *S. basiliscus* (Fig. 1A,B; [245]). Throughout most of their history, paleo-modeling suggests that the lineages likely diverged in isolated glacial-period refugia, with brief opportunities for gene flow during interglacial periods [360, 245]. Here, to further explore the divergence history of these regional populations, we collect a species-wide and multi-locus data set. In doing so,

we find significant branch length heterogeneity between the nuclear and mitochondrial genomes for the Southern-most populations – 50-fold greater divergence of mitochondrial DNA than nuclear DNA. To test hypotheses about how coalescent variance and introgression contribute to this discordance, we place the data in a comparative context, and we exploit the region's well-understood biogeography to reconstruct the lineages' demographic history via Approximate Bayesian Computation (ABC) analyses.

4.2 Methods

4.2.1 Sampling and Genetic Data

Our sampling covers the known distribution of the sister species *Saproscincus basiliscus* and *S. lewisi* throughout the AWT [245]. These species are leaf-litter skinks that, while generally found in association with rainforest, can also extend into adjacent wet sclerophyll forests [65]. The two species are ecologically and morphologically similar; they are delineated based on a minor morphological feature – differing number of paravertebral scales [65].

In this study, we added to the mitochondrial data set collected by Moussalli *et al.*, (2009) to generate a 291-individual data set. Our expanded sampling focused on sequencing individuals located in geographic gaps between previously-defined phylogeographic lineages; habitat is contiguous between all lineages, except for the Central and Southern lineages (Fig. 1A). For a subset of these individuals (N=86; Table S1), we sequenced eight nuclear loci. Because the relevant unit of analysis in this study is the phylogeographic lineage, these individuals were sampled roughly proportional to the prevalence of each mitochondrial clade, and we ensured representation of the full geographic range of the species.

We extracted DNA from preserved tail tissue using a high-salt DNA extraction [4]. To assay mitochondrial variation, we sequenced the ND4 locus [9]; to assay nuclear variation, we sequenced eight loci, including six published previously: β-globin intron, C-mos exon, R35 exon, Rhodopsin intron, and TPI and RPS8 introns [86, 191, 311, 35]. To this, we added two additional intronic loci (CRISP and LGMN), designed for the closely-related lizard *Lampropholis coggeri* [35]. All PCRs were conducted in standard conditions in 12 μL volumes, using a touchdown protocol of 14 cycles of decremental and 22 cycles of stable annealing temperatures (details available in Table S2). Following PCR amplification, we visualized products on an agarose gel, cleaned PCR products via ExoSAP-IT (USB), and sequenced products using BigDye v3.1 on a ABI3730 (Applied Biosystems). The majority of reads were assembled and edited using Geneious [88]; to resolve assemblies with heterozygous indels, we used CodonCode Aligner's heterozygoteIndel feature (CodonCode Co.).

For the mitochondrial locus, final assemblies were aligned with the published alignment from Moussalli *et al.* (2009) using MUSCLE [95]. For the nuclear loci, we inferred

haplotypes from our diplotypes computationally using PHASE2.1 [341], running the algorithm 100 times and assuming a constant recombination rate. We used the most probable haplotype resolution to determine haplotypes; here, all heterozygous sites were resolved to greater than 95% probability. Final nuclear alignments were made with MUSCLE and checked manually in Geneious. The final nuclear data set was 94% complete by locus with a combined length of 3.98 Kb.

4.2.2 Tree-based and multi-locus analyses

To infer genealogical relationships among our mitochondrial haplotypes, we used the Bayesian phylogenetic approach implemented in MrBayes v3.1.2 [159]. We included sequences from the species *Saproscincus czechurai* and *S. tetradactylus* as outgroups ([245], GenBank IDs: FJ195325.1,FJ195291.1). We partitioned the alignment into the genic and tRNA regions and assigned to each partition the most appropriate model for nucleotide substitution using MrModelTest [263]. The partitioned alignment was run twice, each with four chains (three heated, one cold, default heating parameters), for 20,000,000 generations with a 6,000,000 generation burn-in. MCMC chain convergence was assessed by calculating ESS values using Tracer, and the posterior distribution of trees was summarized using TreeAnnotator [87].

To describe heterogeneity in topologies and branch lengths among loci, we inferred gene trees for each of our nuclear loci using the maximum-likelihood approach implemented in RAxML [340]. Each alignment was run unpartitioned, under the substitution model inferred to be most probable by MrModelTest [263]. If the model selected by MrModelTest was simpler than those implemented in RAxML, we chose the simplest model RAxML provides (GTRGAMMA). For each locus, we found the best-scoring maximum-likelihood tree and conducted 1000 rapid bootstrap analyses to determine support for the tree.

We employed two approaches to summarize and visualize patterns of variation across our multi-locus results. First, to identify population clusters, we used the program Structure v2.3.2, which identifies populations (K) by minimizing linkage and Hardy-Weinberg disequilibrium within a cluster [285]. We ran Structure 20 times with our phased nuclear data (10,000,000 steps with 1,000,000 burn-in) under the 'admixture' model for each of 12 K values (ranging from 1 to 12), determined the best-supported K value following Evanno *et al.* (2005) as implemented in StructureHarvester [94], and summarized and plotted results using Clumpp and Distruct [162, 300]. Second, we used POFAD to construct a network of individual similarity based on phased nuclear data. POFAD is a distance-based method that explicitly accounts for haplotypic variation within individuals [169]. We inferred Tamura-Nei corrected distance matrices for each locus using PAUP [348], calculated a final individual-based distance matrix with POFAD, and visualized the results as unrooted networks using SplitsTree [158].

4.2.3 Between-lineage Diversity

Discrepant branch lengths for mitochondrial and nuclear loci could simply be due to differences in mutation rates between the two genomes. If mutational variance is a minor factor in this system, we would expect sequence divergence at mitochondrial and nuclear genomes to be correlated and to reflect the difference in substitution rate between the genomes. To explore this possibility, we used Arlequin v3.1 to estimate raw D_{xy} and net D_a sequence divergence between the lineages of the *S. basiliscus* species complex (Fig. 1B) at both the mitochondrial and nuclear genomes, as estimated by the Tamura-Nei model. To compare patterns of nuclear-mitochondrial divergence more generally, we expanded our analysis of divergence between lineages to five other closely-related and co-distributed species of lizards: *Carlia rubrigularis* and *C. rhomboidalis* (7 nuclear loci; [85]), *Lampropholis robertsi* (8 nuclear loci; [35]), *L. coggeri* (6 nuclear loci; [35]), and *Gnypetoscincus queenslandiae* (2 nuclear loci; Singhal, unpublished). For these species, many of the nuclear loci sequenced were also used in this study, and the same mitochondrial marker was sequenced across all species. In each of these species complexes, we identified major lineages based on the mitochondrial genealogy and then determined sequence divergence between sister phylogeographic lineages at both mitochondrial and phased nuclear data.

4.2.4 Divergence and demographic analyses

To determine if coalescent variance or introgression could explain the observed, strong genealogical discordance, we inferred the most likely demographic history using Approximate Bayesian Computation (ABC). For many demographic scenarios, both defining and calculating the likelihood function for the model can be challenging; the crude approximation afforded by ABC, however, can estimate the likelihood function by simulation, and thus, can allow researchers to test a much wider range of biologically-relevant models [31]. That being said, recent research has suggested that model choice via ABC can give slightly biased results for un-nested models, particularly when summary statistics fail to capture the full complexity of the raw data [298]. However, this and subsequent research also indicate, by using summary statistics with differing distributions under alternative models and by validating the model choice procedure itself [220], ABC can remain a powerful tool for exploring and testing the fit of different models to data.

With these caveats in mind, we tested the fit of two major classes of models to our empirical data: models without introgression, designed to test how coalescent variance can contribute to generating genealogical discordance, and those with introgression, designed to test how introgression, together with coalescent variance, can contribute to generating genealogical discordance. Our choice of models is motivated by our knowledge of biogeographic history of the *S. basiliscus* species complex and its rainforest habitat [245, 360]. We focus on modeling the sister lineages that exhibit strong genealogical discordance, the Central and Southern lineages of *S. basiliscus*. To simplify the models, we model only the Spec Uplands population of the Southern lineage; the Hinchinbrook Island and Elliot

Upland populations have too low of sample sizes to model accurately and also introduce additional spatial complexity beyond the scope of this study. As predicted by paleomodels (Fig. 2A), the Central and Southern lineages likely evolved largely in allopatry through glacial cycles, and thus we consider multiple variations on a basic allopatric model in our simulations (Fig. 3). These models are:

- Models with no introgression
    1. A simple model of population splitting, in which an ancestral population splits into the Central and Southern lineages with no post-divergence gene flow (Fig. 3A);
    2. An extension of the simple model ("peripatric divergence") in which the Southern lineage is initially very small when it splits from the ancestral lineage (Fig. 3B);
    3. A model in which there is ancestral population structure, such that the ancestral population consists of multiple populations with limited gene flow, after which it splits into the Central and Southern lineages (Fig. 3C).
- Models with introgression
    4. A model in which there is a pulse of post-divergence gene flow, in which the Central lineage expands and exchanges migrants with the Southern lineage for a brief period of time in the past (Fig. 3D);
    5. A model in which there is a pulse of post-divergence gene flow, in which the mitochondrial gene flow between the Central and Southern lineages is greater than nuclear gene flow (Fig. 3E). This model would allow the relictual Southern mitochondrial genome to introgess into the invading Central lineage (Fig. 2B);
    6. A model in which there is a pulse of post-divergence gene flow, in which nuclear gene flow between the Central and Southern lineages is greater than mitochondrial gene flow (Fig. 3F). This model would allow the invading Central nuclear genome to completely introgress the relictual Southern population (Fig. 2C).

In particular, we include models 2 and 3 because such histories can substantially affect the coalescent [166, 334], and they are plausible in context of the biogeographic history of these lineages [244]. The introgression models (models 4 - 6) reflect our knowledge of this system's history; a model in which there is constant gene flow throughout divergence is unlikely to describe this system. Reconstruction of the AWT rainforest during the Pleistocene suggests suggests that the Central and Southern lineages were isolated for most of their history – including at present [245] – with brief periods of increased connectivity during cool-wet periods, most recently in a brief period during the early Holocene (Fig. 2; [360]). Further, as has been suggested in numerous other systems of cytonuclear discordance [129, 18], models 5 and 6 explore the possibility of differential rates of gene flow

at the mitochondrial and nuclear genomes if, for example, there is allele surfing [71], the mitochondrial genome is under selection (*i.e.*, cytonuclear incompatibilities), or there is sex-biased gene flow (see **Discussion**). Importantly, model 5 differs from models 4 and 6 in the structure of the simulation (Fig. 3D-F), because, although the geography of all three scenarios is the same, the relative movement of mitochondrial and nuclear genomes between the Central and Southern lineages differs across models (Fig. 2B,C).

We used the program msABC to simulate and generate summary statistics for each of these models, modifying the program via Perl scripts to both simulate nuclear and mitochondrial data and to calculate six additional summary statistics ($D_{a,nuc}$, $D_{a,mito}$, $D_{xy,nuc}$, $D_{xy,mito}$) and the corresponding cytonuclear divergence ratios ($\frac{D_{a,mito}}{D_{a,nuc}}$, $\frac{D_{xy,mito}}{D_{xy,nuc}}$). For loci lengths, mutation rate, and recombination rate, we used well-circumscribed priors defined by our empirical data and data from other studies of lizards [48, 302]; for all other parameters, we used broad, uninformative priors (Table S3).

We generated an initial set of 10,000 simulations under each model and used the results from these simulations to evaluate which summary statistics differed the most between our two major model classes (c.f. [298]) and to determine which statistics, if any, were significantly correlated and, thus, unlikely to provide additional information. Through this approach, we defined three summary statistics ($D_{a,nuc}$, $D_{a,mt}$, $\frac{D_{a,mito}}{D_{a,nuc}}$), and we used these along with summary statistics more generally useful for inferring demography ($\theta_{w,nuc}$ and $\theta_{w,mito}$ for both lineages). We then simulated larger data sets of 1 million simulations for each model to use in model choice. Using the R package abc [70], we conducted all downstream inference. Primarily, we used a weighted multinomial logistic regression to estimate the posterior probabilities of our models' fit to our data. Following a rejection step across all models (tolerance rate; γ=0.01), regression was performed on the retained simulations, where the model is treated as a categorical response variable and the summary statistics are the independent variables [32]. Further, while our primary objective here is model choice rather than model fitting, we used model fitting to test the accuracy of model choice [119]. Thus, we then inferred the posterior distributions for each parameter in each demographic model using a local linear regression-corrected rejection scheme (γ=0.01) and log-transforming parameters prior to fitting to ensure the posterior distributions fell within prior ranges [138].

Finally, we evaluated the performance of our model choice procedure via two methods. First, we generated pseudo-observed data sets, for which we randomly selected a simulated data set, defined which model supported the data best using the same model-choice procedure described above, and then calculated how often these data sets were mis-classified. Second, we generated posterior predictive distributions under the inferred posterior distributions for each model and then compared our empirical summary statistics to these simulated distributions; if model-choice is accurate, our empirical summary statistics should lie within the simulated distributions. [354].

4.3 Results

4.3.1 Mitochondrial and Multi-locus Phylogeography

The mitochondrial gene tree recovered the same lineages as described by Moussalli *et al.*, (2009) (Fig. 1B). Here, we refer to these mitochondrial lineages as the *S. lewisi*, Northern, Central, and Southern lineages. Each major mitochondrial lineage consists of several, well-supported subclades, each of which is geographically restricted (Fig. 1B). Our improved sampling located areas of sympatry between the *S. lewisi* and Northern lineages and between the Northern and Central lineages. As described earlier, *S. lewisi* and Southern *S. basiliscus* are each highly divergent from the rest of the clade; net corrected sequence divergence ranges from 15 - 18% between these lineages. Further, the Southern lineage is highly structured; populations in the currently isolated Spec Uplands, Elliot Uplands, and Hinchinbrook Islands are 3-8% divergent from each other (Fig. 1A-B).

The primary result found from analyses of the eight-loci nuclear data is marked genealogical discordance between the mitochondrial and nuclear data for the *S. basiliscus* Southern lineage. As shown by the multi-locus nuclear gene network generated via POFAD (Fig. 1D), *S. lewisi* is quite divergent from the rest of the species complex ($D_a = 0.032$). However, the *S. basiliscus* Southern lineage shows an order of magnitude less nuclear divergence from the Central and Northern lineages ($D_a = 0.003$), even though its mitochondrial divergence is nearly as great as that of *S. lewisi*.

Besides this instance of genealogical discordance, the data otherwise show broad-scale concordance among markers. Although individual gene trees all show differences in topology and branch length for most of these lineages (Fig. S1), a consensual history emerges from multi-locus analyses. First, the major clades and subclades identified by mitochondrial sequencing are all recovered by Structure-clustering of nuclear genotypes (Fig. 1C). Using the Evanno method [101], we determined that nine clusters provided the best fit to the data, each of which correspond to a mitochondrial clade. Examination of the nuclear data shows, however, that only the *S. lewisi* lineage is recovered as a distinct clade, even though the main mitochondrial lineages are largely separated in the distance-based network (Fig. 1D). Second, geographic concordance between the two marker types is strong; there are only two cases where an individual belongs to two different major clusters (*S. basilicus* N, C, S lineages, and *S. lewisi*) for mitochondrial and nuclear data (Fig. 1C). Both cases trace to individuals sampled at the parapatric boundaries between the Northern and Central lineages.

4.3.2 Between-lineage Diversity

To place our finding of incongruent branch lengths between the nuclear and mitochondrial genomes for the Southern lineage in context, we compared sequence divergence (D_a) in mitochondrial and nuclear genomes between major phylogeographic lineages in seven co-distributed species of lizards. As shown in Fig. 4, divergence levels at the two genomes

are highly correlated (r^2=0.91; p <0.005), with an average divergence ratio of 11.2. As expected, this divergence ratio reflects the estimate of the nuclear-mitochondrial mutation scalar for lizards (≈14; [48]). The mito-nuclear divergence ratio between the *S. basiliscus* Southern and Central lineages (50.4; as identified by the arrow in Fig. 4) is a noticeable outlier in this group. Including this datum substantially weakens the strength of the correlation between nuclear and mitochondrial divergence (r^2=0.76; p <0.05). Thus, the discordance we see in branch lengths between the nuclear and mitochondrial genomes for the Southern lineage is unlikely due to variance in mutation rates.

4.3.3 Divergence and Demographic Analyses

To determine if coalescent stochasticity or introgression better explains the genealogical discordance we see, we used an ABC approach to test the fit of six different models to our empirical data. First, we used a small number of simulations to test the utility of a wide-range of summary statistics to distinguish between our models. Perhaps because our models explore a large parameter space–portions of which lead to competing models becoming nearly identical–many of the tested summary statistics showed little difference in distributions between the models. That said, we identified seven summary statistics (($D_{a,nuc}$, $D_{a,mt}$, $\frac{D_{a,mito}}{D_{a,nuc}}$), $\theta_{w,nuc}$ and $\theta_{w,mito}$ for both lineages), which were not strongly correlated with each other (r^2 <0.2) and that showed differing distributions between the models (Fig. S2).

We extended these initial simulations to conduct model choice by calculating posterior probabilities of our differing models. The general class of models exploring coalescent stochasticity (models 1 through 3) were supported with low posterior probability (summed p=0.0483), and models invoking a pulse of gene flow were strongly supported (summed p=0.951; Table 1; Bayes Factor=19.68). In particular, within models with introgression, a model that allowed for asymmetric gene flow between the two genomes (summed p=0.922) was very strongly supported compared to a model with equal gene flow between the two genomes (p=0.0291; Bayes Factor=31.68). Further, there is some support that a model with more mitochondrial gene flow (model 5; p=0.763) is more likely than a model with more nuclear gene flow (model 6; p=0.159; Bayes Factor=4.79). Given the results of Robert *et al.*, 2011, we refrain from over-interpreting the posterior probabilities reported here, but we do suggest that these results strongly support a demographic history of pulsed gene flow that is heterogenous between the nuclear and mitochondrial genomes.

We estimated the posterior distributions of the parameters for our best-fitting model (model 5; pulsed gene flow with greater mitochondrial than nuclear gene flow). Our results, shown in Fig. S3, are plausible given our knowledge of the species' ecology and the region's biogeographic history. Model-fitting showed very little power to estimate key parameters of the pulse of gene flow, *i.e.*, when it started, how long it lasted, number of migrants in either direction, or the magnitude of asymmetry in gene flow between the

two genomes. Other studies looking at pulsed gene flow have shown similarly limited power [386, 199]; fortunately, our conclusions do not depend on precise estimates of these parameters.

To evaluate the performance of our model choice, we generated pseudo-observed data sets and looked at the frequency of mis-classification. As seen in Fig. S4 and Table S4, the frequency of false positives and negatives is high across most of the models. However, for most of these mis-classified models, the posterior probability of the best-supported model was low (p <0.5). Following Fagundes *et al.*, 2007, we computed the probability that the best-supported model is the correct model, given the observed posterior probability. Comparing models with and without introgression, we computed the probability of observing our posterior probability (p=0.951) in error as 0. Comparing models with and without heterogeneous introgression, we computed the probability of observing our posterior probability (p=0.922) in error as 0. Comparing a model with more mitochondrial gene flow than nuclear gene flow, we computed the probability of observing our posterior probability (p=0.763) in error as 0.122. We further evaluated our model-choice procedure by comparing the empirical value for our summary statistics to the posterior predictive distributions inferred for each model. For our best-supported model (model 5), each of our observed values is within the posterior predictive distributions (Fig. 5); the probability of recovering the same or more extreme value for a given summary statistic ranged from 0.129 to 0.724. For the other 5 models, this is only also true for model 6 (probabilities ranged from 0.074 to 0.823; Fig. S5). These evaluations of our model choice procedure strongly support model 5 or model 6 as the best fit our data and support the hypothesis that historical introgression, with asymmetry in gene flow between the mitochondrial and nuclear genomes, is likely the source of the discordance seen in *S. basiliscus* and *S. lewisi*.

4.4 Discussion

By collating a multi-locus data set in the species complex *S. basiliscus* and *S. lewisi*, we uncovered a striking example of genealogical discordance: the populations representing *S. basiliscus* Southern mitochondrial lineage, which are 15% divergent from the rest of *S. basiliscus*, are an order of magnitude less divergent at the nuclear genome than expected. That *S. lewisi*, a lineage with similar levels of mitochondrial divergence, exhibits genealogical concordance across loci (Fig. 1; Fig. S1) and is reproductively isolated from the rest of the clade based on multi-locus data from sympatric populations (Singhal and Moritz, unpublished), further underlines how anomalous this result is.

Perhaps the most parsimonious explanation for reduced divergence in the nuclear genome when compared to the mitochondrial genome is that the nuclear genome has a lower substitution rate. Calculating divergence between phylogeographic lineages in seven species of closely-related AWT skinks shows that divergence at the mitochondrial genome is otherwise tightly correlated with nuclear divergence (Fig. 4). Although the

substitution rates between the two genomes are different – the mito-nuclear divergence rate is approximately 11:1 – our discordant clade has a divergence ratio of 50, far beyond what could be explained by genome-specific substitution profiles.

4.4.1 Source of genealogical discordance

What, then, can explain this level of genealogical discordance? Genealogical discordance is most often attributed to either the stochasticity of the coalescent or introgression across lineage boundaries [177]. With respect to coalescent variance, because the mitochondrial genome acts as a single locus [18], it might be capturing an errant view of history, relative to the rest of the genome. If coalescent variance is underpinning genealogical discordance, we might expect that the population diverged under a history that promoted increased variance in genealogical patterns. Our ABC simulations, which explored both a broad set of divergence histories and parameter space, suggest under certain population histories, genealogical discordance (as measured here) can increase. However, the increase in variance is limited compared to the magnitude of the discrepancy we see (Fig. S2). Thus, as supported by the low posterior probabilities for these models, coalescent variance is unlikely the source of this system's genealogical discordance.

Our divergence history reconstruction suggests the genealogical discordance likely results from historical introgression between the Central and Southern lineages. Ongoing introgression between the Central and Southern lineages is unlikely; niche models and field surveys suggest the two lineages do not currently meet [245], and nuclear data show no evidence for recent admixture between these lineages (Fig. 1C). However, niche models for the preferred habitat of the species complex (*e.g.*, wet sclerophyll/rainforest) through time support a model of divergence in isolation during glacial periods with transient connectivity and, thus, opportunities for introgression in the early Holocene [360]. Habitat during the cold-dry portion of the glacial cycle was predicted to be restricted to two major refugia in the north and south of the AWT (corresponding to the Northern and Central lineages respectively), with smaller refugia in the Spec Uplands and on Hinchinbrook Island and limited interconnectivity between the refugia (Fig. 2A). During the cool-wet and warm-wet portions of the glacial cycles, the forest between the Central and Southern lineages was predicted to be contiguous. As such, the Central lineage likely expanded out of its refugium and invaded the southern AWT, allowing for gene flow between the Central and Southern lineages (Fig. 2B,C). Following gene flow between the two lineages, the ancestral Southern nuclear genome was replaced by the invading Central nuclear genome, but the highly-divergent Southern mitochondrial genome persisted. In a sense, the only evidence for the Southern population that persisted through time in the small southern refugia is its highly divergent mitochondrial genome, present as distinct subclades in the Spec and Elliot Uplands and Hinchinbrook Island. Interestingly, the same phenomenon may have occurred in a rainforest frog, *Litoria nannotis*, distributed across the same region that shows a similar pattern of cyto-nuclear discordance [34].

4.4.2 Why is introgression heterogeneous?

If this is the history explaining the pattern of genealogical discordance in this system, why is the pattern of discordance marker-specific? Heterogeneous introgression, particularly when the markers being compared are cytoplasmic and nuclear, has several possible root causes: neutral and stochastic effects, selection, or sex-biased processes. First, heterogeneous introgression can arise because of stochastic, neutral effects [185], like differences in drift among markers or the confounding effects of introgression and demography. Researchers have explained cytonuclear genealogical discordance by invoking the differential rates of drift and fixation in cytoplasmic and nuclear markers [375]. However, the balance between migration-drift is the same for both genomes because the higher drift in the mitochondrial genome is counteracted by a lower effective number of migrants [380]. Thus, drift is unlikely to explain this pattern. Yet, introgression does introduce stochasticity [185], and as the mitochondrial genome is just one marker, it might capture an extreme end of this variance. But, as our ABC-based demographic reconstructions suggest, such effects are still unlikely to lead to the sort of discordance seen here. Further, stochastic effects can be compounded by demographic events, such as changes in ranges or population growth [307, 71]. In particular, in the model of "allele surfing", alleles from a resident population introgress readily into low-density populations at the edge of an expanding population [71]. Loci that experience higher levels of drift – here, loci linked to the low-dispersing sex like the mitochondrial genome – are more likely to introgress quickly [276, 230]. However, this work also predicts that even with low rates of admixture, complete replacement is expected at all loci, whether nuclear or mitochondrial [71]. As such, we might to see swamping of variation in the expanding population at both the nuclear and mitochondrial genomes. As we do not see this, we think allele surfing is unlikely to explain the patterns we see here.

Indeed, as shown by our demographic reconstruction, in order for introgression to commonly lead to strong genealogical discordance, gene flow levels must differ significantly between the two marker types – and selection and sex-biased processes are two factors that can lead to differential gene flow. Here, either a situation in which gene flow is elevated at the mitochondrial genome (model 5, Fig. 2B) or in which gene flow is depressed at the mitochondrial genome could explain the pattern we see (model 6; Fig. 2C). These two situations have the same geography, but the patterns of introgression differ and they reflect two very different biological realities. However, the genetic patterns, particularly when introgression is historical and introgressed alleles have been fixed, are hard to distinguish, and based on our demographic reconstructions, we cannot fully reject one or the other.

With respect to selection, rates of introgression and fixation are determined largely by the balance between migration and selection [333]. Selection can lead to loci having more limited or increased introgression compared to the background rate. Negative selection can take several forms – for example, the mitochondrial genome could be adapted to local bioclimatic niches and thus would introgress less readily [18, 61]. This seems

unlikely here, as the Southern and Central lineages do not occupy markedly different bioclimatic space (Fig. S6). Alternately, cytonuclear incompatibilities (*e.g.*, *Tigriopus californicus*; [99]) could limit introgression at the mitochondrial genome and a subset of the nuclear genome, while the rest of the nuclear genome would introgress freely [351]. Cytonuclear incompatibilities can evolve due to selection or due to drift [115], and they are a plausible explanation for this discordance. In other studies, selective sweeps have lead to the cytoplasmic genome introgressing rapidly into other lineages [127]. However, the Southern mitochondrial lineage is deeply structured among southern isolates, and thus, does not have the typical signature of a selective sweep (*i.e.*, low but rare diversity among haplotypes).

Finally, sex-specific processes could also explain these patterns. For example, if there is female philopatry, then the maternally-inherited mitochondrial genome will experience less gene flow, which could lead to less introgression upon secondary contact (but see Petit and Excoffier 2009 for an alternative perspective). This is a common explanation for discrepancies between the mitochondrial and nuclear genomes [38], and as we have evidence for male-biased dispersal in related species [90], this might explain our results as well. Further, patterns of mating, whether due to differences in population density [157] or active mate choice, can lead to rapid introgression of the maternal mitochondrial genome across lineage boundaries [58]. We have no evidence either against or in support of this hypothesis, but it certainly could be a factor. In sum, it appears that the genealogical discordance we see in this system results from historical introgression, and it is quite possible this introgression acted in concert with either selection or sex-biased processes. Knowing how these selective or sex-biased processes are contributing and interacting to lead to this pattern cannot be determined from genetic data alone, but it is amenable to future study through field observation and experiments.

4.4.3 Significance

Examples of genealogical discordance are many, and cytonuclear discordance accounts for a significant number of these cases [58, 71]. Stochastic effects, whether arising in the presence or absence of gene flow, could certainly explain many instances of discordance. However, many other cases, like ours, likely arise because of the unique biology of the mitochondria – *i.e.*, selection limits introgression because of cytonuclear incompatibilities [99] or mate choice patterns promote introgression in sex-linked markers [157, 58]. Further, most examples of differential introgression across mitochondrial and nuclear genomes are cases where the introgressed marker has not yet reached fixation [191, 230]. In this respect, this system stands alongside a few other cases, most notably the arctic charr, North American chipmunks, polar bear, and two species of temperate hares [375, 293, 231, 137] in which distinct evolutionary units have seemingly become fixed for discordant mitochondrial DNA. Both this system and these other species share a history of small populations; the increased efficacy of drift in such populations might have quickened replacement of introgressing alleles. In other systems, where introgression often

occurs between large populations that are expanding from glacial refugia [146], insufficient time might have passed to allow introgressed alleles to reach fixation.

Finally, our and others' results suggest that introgression is likely manifest in the natural world. Because very little hybridization, or mating between different lineages, is necessary to spur introgression [6], the frequency of introgression might say very little about the frequency of hybridization. However, it does hint at interesting population histories of infrequent but dynamic changing hybridization, and it certainly suggests that introgression could be a pervasive and powerful force shaping the diversity of species and their populations [214]. Whether introgression erodes genetic divergence as we see here or leads to adaptive change [6, 339], it is an undeniably important evolutionary process.

4.5 Acknowledgements

We gratefully acknowledge all those in the Moritz and Williams laboratories who collected samples for this work, in particular A. Moussalli, S. Williams and C. Hoskin. Assistance in the lab was provided by K. Whittaker and A. Xu and advice by members of the Moritz Lab, M. Hickerson, J. McEntee, and M. Tonione. We thank A. Davis Rabosky, M. Hickerson, A. Leaché, J. McGuire, R. Pereira, M. Slatkin and two anonymous reviewers for thoughtful comments on earlier versions of this manuscript. Financial support was provided by the Museum of Vertebrate Zoology Koford Fund, a NSF Graduate Research Fellowship to SS, and NSF DEB0416250 to CM. The Texas Advanced Computing Center (TACC) at The University of Texas at Austin provided grid resources that contributed to the research results reported within this paper.

4.6 Data Accessibility

Data are available at the following locations:

1. DNA sequences are available as on Genbank, accessions JX313797 – JX315247
2. Final alignments and scripts used for simulations are available on DRYAD, entry `doi:10.5061/dryad.dn0m4`

4.7 Figures

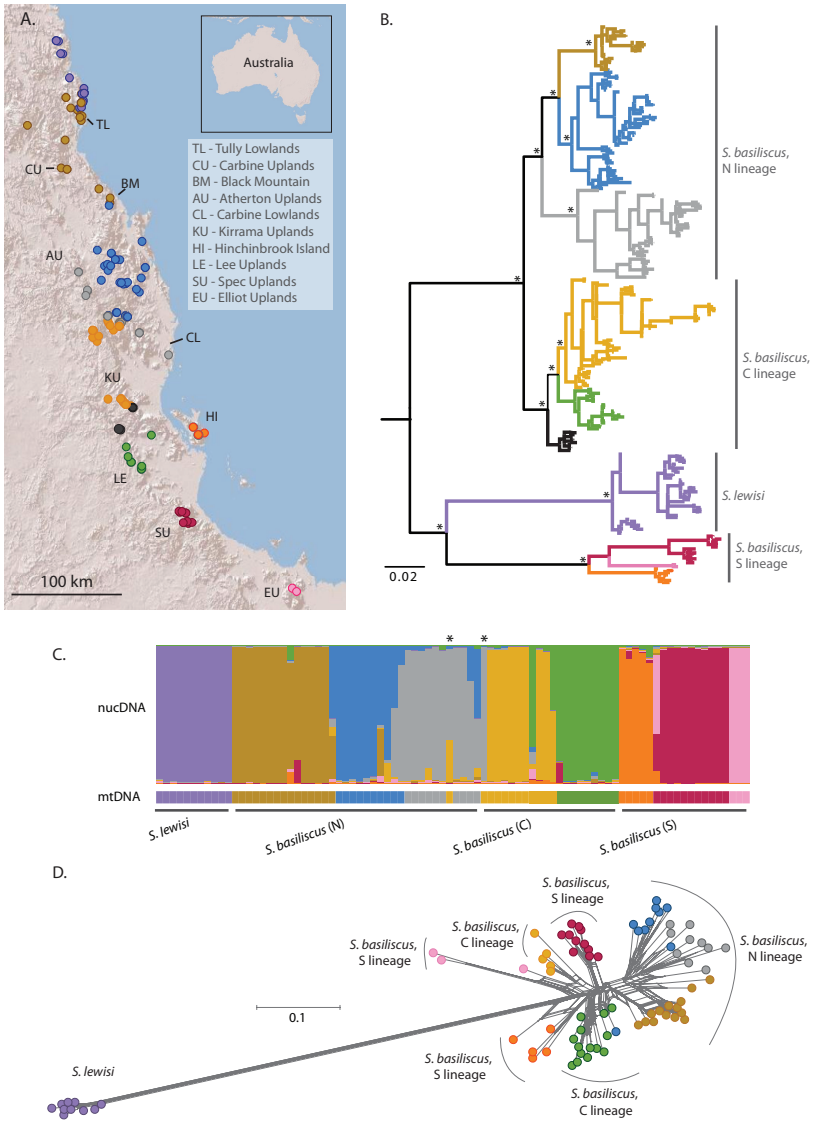

Figure 4.1: A. Map of Australian Wet Tropics showing sampled points for *S. basiliscus* and *S. lewisi* and identifying bioregions [372]. B. Mitochondrial gene tree as inferred by Bayesian analysis for *S. basiliscus* and *S. lewisi*. Major clades with posterior probability >0.95 are marked with an asterisk. C. Structure results based on haplotypes at eight nuclear loci, with mitochondrial identity for each individual shown. Individuals are ordered by location from north to south. Instances of mito-nuclear discordance are identified by asterisks. D. Gene network based on haplotypes at eight nuclear loci. Scale is in a standardized, non-unit based measurement given by POFAD.

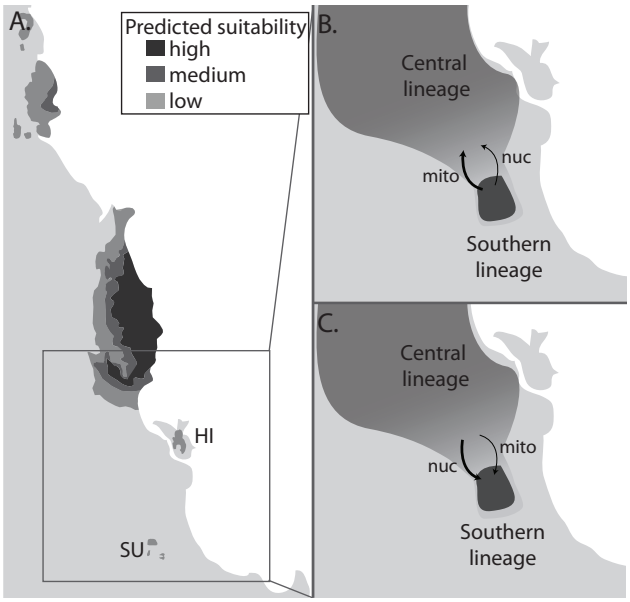

Figure 4.2: A. A suitability map for wet sclerophyll-rainforest in the Australian Wet Tropics showing isolated glacial refugia during the cold-dry stage of the glacial cycle (18,000 ybp), modified from Vanderwal *et al.*, 2009. Cartoon depictions of the possible historical introgression event during a cool-wet period of higher connectivity, showing the dynamics of mitochondrial and nuclear introgression between the Central and Southern lineages of *S. basiliscus* for B. model 5, of more mitochondrial than nuclear gene flow and C. model 6, of more nuclear than mitochondrial gene flow.

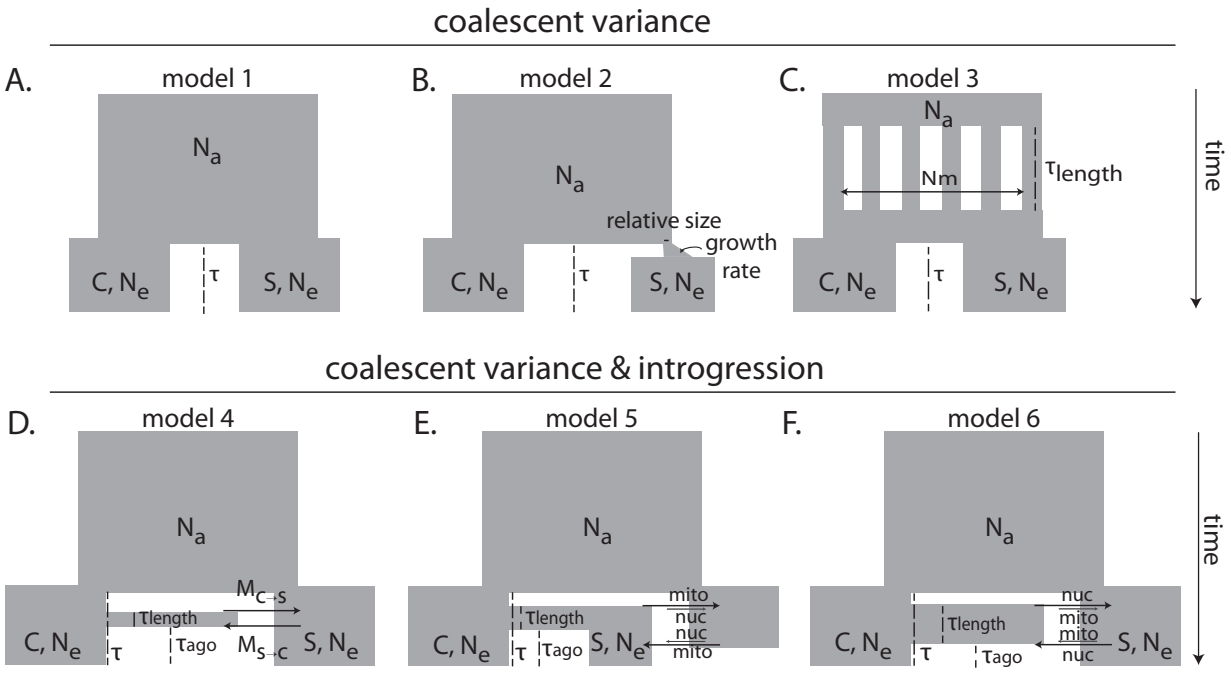

Figure 4.3: Cartoon depictions of the six models used in the ABC analysis to infer the divergence history of the Central and Southern lineages of *S. basiliscus*: A. simple splitting, B. a "peripatric" splitting model, C. a splitting model with ancestral population structure, D. a model with pulsed, post-divergence gene flow, E. a model with pulsed gene flow in which mitochondrial gene flow is greater than nuclear gene flow, and F. a model with pulsed gene flow in which nuclear gene flow is greater than mitochondrial gene flow

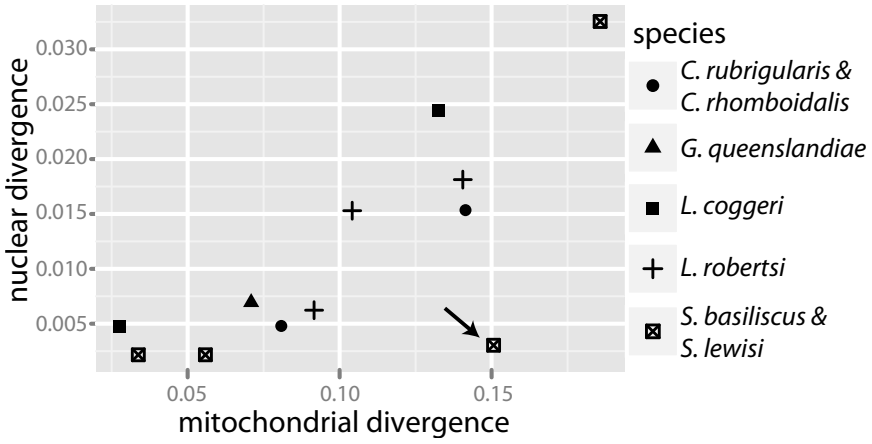

Figure 4.4: Correlation between mitochondrial and nuclear divergence between phylogeographic lineages in Australian Wet Tropics lizards; points are labelled according to the species in which the contact is found. The arrow identifies the contact of interest: *S. basiliscus* Central/Southern lineages.

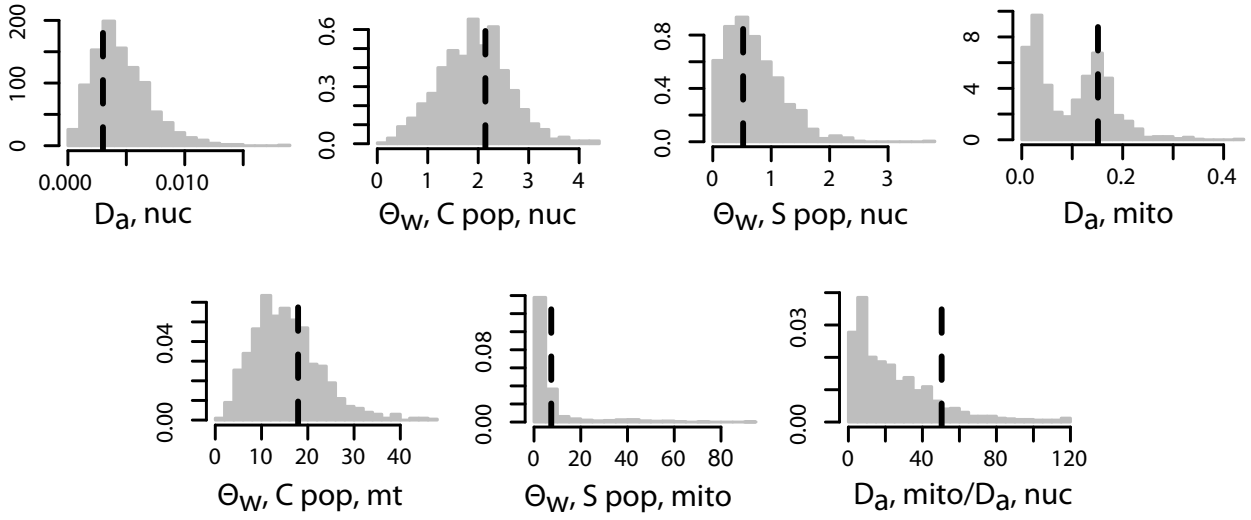

Figure 4.5: Following ABC analysis of *S. basiliscus* Central and Southern lineages, posterior predictive results for the most highly-supported model (model 5: greater, pulsed gene flow at the mitochondrial genome) across all seven summary statistics. Dashed black lines reflect true value of summary statistic for the empirical data.

4.8 Tables

model	posterior probability
simple split (model 1)	0.0263
peripatric split (model 2)	0.0220
ancestral structure split (model 3)	0.000
pulsed gene flow (model 4)	0.0291
more mitochondrial gene flow (model 5)	0.763
more nuclear gene flow (model 6)	0.159

Table 4.1: Models, numbered as in text, with posterior probabilities as inferred from ABC analyses.

Chapter 5

Strong selection in a narrow contact zone

5.0.1 Abstract

Phenotypically cryptic lineages comprise an important yet understudied part of biodiversity; in particular, we have much to learn about how these lineages are formed and maintained. To better understand the evolutionary significance of such lineages, we studied a hybrid zone between two morphologically-cryptic phylogeographic lineages in the rainforest lizard, *Lampropholis coggeri*. Analyzing a multilocus genetic dataset through cline inference, individual-based methods and population measures of disequilibrium and using simulations to explore our genetic results in context of theoretical expectations, we inferred the processes maintaining this hybrid zone. We find that these lineages meet in a hybrid zone that is narrow (≈400 m) relative to inferred dispersal rate. Further, the hybrid zone exhibits substantial genetic disequilibrium and sharply coincident and largely concordant clines. Based on our knowledge about the region's biogeography, the species' natural history, and our simulation results, we suggest that strong selection against hybrids structures this system. As all clines show a relatively narrow range of introgression, we posit that this hybrid zone might not yet be in equilibrium. Nonetheless, our results clearly show that phylogeographic lineages can evolve substantial reproductive isolation without concomitant morphological diversification, suggesting that such lineages can constitute a significant component of evolutionary diversity.

5.1 Introduction

The growth of phylogeography, or the study of geographic variation in genetic diversity within a species, has shown that many species consist of multiple, highly-divergent genetic lineages. These lineages often exhibit levels of genetic divergence equal to or greater than those between morphologically-defined species [11] yet have no or limited phenotypic differentiation. These morphologically cryptic lineages are common yet understudied; in particular, understanding how these lineages form and what maintains boundaries

between these lineages in the absence of overt phenotypic differentiation are open questions [40]. Contact zones, or geographic regions in which previously isolated lineages meet and interact, are an excellent tool for exploring and understanding the forces that maintain lineages' identities [139]. Although such studies traditionally focus on contacts between morphologically differentiated species, contact zones can be particularly useful in evaluating the significance of morphologically-cryptic phylogeographic lineages as a component of biodiversity. After all, contact zones exhibit a continuum of outcomes that reflect how the lineages meeting diverged; lineages can exchange genes freely in the absence of reproductive isolation [319], gene flow can be spatially limited if barriers to gene flow (*i.e.*, assortative mating or selection against hybrids) have evolved [140], or lineages can show complete reproductive isolation and remain phenotypically and genetically distinct at contact [367]. Looking at contact zones between cryptic lineages can help us start to answer these questions; already, studies have shown that such lineages can evolve substantial barriers to gene flow [155, 278, 173]. This initial work suggests that cryptic lineages might not just be evolutionary ephemera; rather, they can be regarded as nascent species [15]. However, these studies are still in their infancy, and we have much more to learn about cryptic biodiversity.

The Australian Wet Tropics (AWT) suture zone offers a set of replicated natural experiments with which we can address these questions. This suture zone reflects effects of late Quaternary climate change on rainforest and the species endemic to it [244, 146]. Palynological data and ecoclimatic models show that the forest contracted into two major refugia during the glacial cycles of the Late Quaternary [256, 133, 360]. From 3 to 8 Kya, the rainforest expanded rapidly from these refugia, allowing diverse rainforest-specialist fauna to form spatially clustered contact zones [244]. Genetic divergence at these contacts varies widely, yet lineage-pairs are morphologically similar and occur in the same ecological settings. Previous analyses of contacts in frogs, lizards and mammals have shown several evolutionary outcomes: speciation by reinforcement, post-zygotic isolation without assortative mating, and no reproductive isolation [155, 278, 84, 283]. These results show that lineages can maintain their genetic integrity at secondary contact, even without concomitant morphological diversification. However, studies from this suture zone are still few, and analysis of additional contacts in lineage-pairs with different levels of divergence is needed to test predictions such as the correlation between genetic divergence and extent of reproductive isolation [367].

We add to this emerging portrait of secondary contact outcomes in a suture zone by characterizing a contact in the AWT-endemic rainforest skink, *Lampropholis coggeri*. *L. coggeri* (family: *Scincidae*) is a small, semi-fossorial lizard (SVL $\approx$ 45 mm) that is often found in sun patches within or at the edge of forests [373]. Previous sampling and genetic analysis of *L. coggeri* identified two major lineages in this species, which we name Northern and Southern (Fig. 1A; [35]). These lineages are 9.4% divergent at mitochondrial DNA and 1.1% divergent at nuclear introns, yet do not differ in any morphological traits (*i.e.*, size, shape, scale counts, coloration) [35]. In this study, we locate and characterize the contact zone between the Northern and Southern lineages by collecting a multi-locus genetic data

set and conducting individual-based, population-level, and cline-based analyses. Then, motivated by results from the genetic analyses, we use simulations to explore alternative hypotheses about the forces governing this hybrid zone, and to extend the relevance of our results, we compare our findings to those found for other hybrid zone systems. Given previous results from similar studies in this system [278] and the expectation that post-zygotic isolation scales with genetic divergence [67], we predict substantial levels of genetic disequilibrium and narrow genetic clines in the present system.

5.2 Methods

5.2.1 Sampling

Based on the initial characterization [35] and subsequent sampling, we identified the contact zone between the Northern and Southern lineages. From 2008 to 2010, we sampled the contact extensively along a linear transect running through the contact zone (the Gillies Transect) and, to provide some level of replication, opportunistically around a nearby lake (Lake Barrine) bisected by the contact zone (Figure 1; Table S1). Although each of these regions are forested, the area between them was cleared in the early twentieth century, preventing further sampling. Animals were captured by hand, sampled for tail tissue, measured and sexed, geo-referenced, and then released at the site of capture. For the linear transect, we sampled 17 localities (average of 17 individuals per locality) over 2.5-km. At Lake Barrine, we sampled 58 animals at 28 unique locations. Additionally, we sampled two localities (17-18 individuals per locality) at 2.5 km on either side of the hybrid zone center to determine whether any of the alleles showed introgression outside of the central hybrid zone. Finally, from each lineage, we sampled one locality (12 individuals) about 40 km away from the hybrid zone center. These latter individuals were sampled for liver tissue and were collected as voucher specimens to be accessioned at the Museum of Vertebrate Zoology, Berkeley, California. As these localities are geographically isolated (hereafter, "allopatric") from the hybrid zone center, they are unlikely to contain alleles introgressed from the hybrid zone center, and thus, we used them for marker development.

5.2.2 Marker Development

To assay hybridization and introgression in the contact zone, we used one mitochondrial locus and ten nuclear loci. We used previously published primers to amplify the mitochondrial locus ND4 [9] and three nuclear loci: β-globin [86], LC5, and LC17 [35]. To design additional markers, we developed markers based on data from a high-throughput sequencing run. Briefly, we extracted total RNA from five individuals from each of our Northern and Southern allopatric localities, pooled equimolar amounts of individual RNA for each locality, isolated mRNA, and prepared a sequencing library as di-

rected by Illumina [37]. Each of the two resulting libraries was sequenced at one lane on an Illumina Genome Analyser II [37]. Resulting reads were trimmed for quality and for adapter sequence and assembled using the *de novo* assembler ABySS; contigs were annotated using a custom vertebrate gene database compiled from Ensembl [42, 41]. To find fixed differences between the two lineages, we mapped the trimmed reads from the Northern and Southern localities to the Northern reference assembly using bwa [197]. Resulting single nucleotide polymorphism (SNP) calls were parsed using samtools [198], and we identified SNPs that were fixed in the lineages for different alleles. We then identified a subset of SNPs that were in annotated genes, that were either non-coding (*i.e.*, located in the 5′ or 3′ untranslated region) or resulted in synonymous amino acid changes, and that could be resolved using commonly available and robust restriction enzymes. We used Primer3 [306] to design primers for 12 of these SNPs and Sanger-sequenced these alleles in a larger sample from the Northern and Southern allopatric localities (12 individuals each) to confirm these SNPs were fixed or nearly-fixed between the localities. We successfully sequenced 11 loci and selected for further analysis the seven most robust loci (*ABHD5*, *AUTO*, *NDST2*, *LEMD2*, *PCBD1*, *RTN3*, *SAR1*). The bioinformatics pipeline to develop these markers was written in Perl, available at `https://github.com/singhal/transcriptomic`.

5.2.3 Collection of Genetic Data

We extracted genomic DNA from tail tissue using a high-salt method [4] and confirmed DNA quality and quantity using a Nanodrop. To amplify DNA, we used standard PCR conditions in a 15 μL reaction. Each of the 11 markers contained a diagnostic SNP that was fixed or nearly-fixed between the two lineages and that could be genotyped using PCR-RFLP. To genotype loci, we digested 10 μL of the amplified product in a 25 μL reaction with one unit of the appropriate enzyme, following manufacturer's suggestions for use (NEB). We visualized the digested products on a 1.5% agarose gel stained with ethidium bromide and scored genotypes manually. Details on primer sequences, annealing temperatures, enzymes used, and restriction patterns can be found in Table S2.

5.2.4 Analyses

We conducted four types of analysis. First, we determined hybrid composition using individual-based methods, as the genetic make-up of hybrids can reflect the nature of selection against different hybrid classes [371]. Second, we fit clines to our data to measure introgression extent and to determine whether clines were concordant and coincident, as we expect greater introgression extent and lack of coincidence and concordance if there are no barriers to gene flow. Third, we estimated population parameters of disequilibrium, as disequilibrium only persists after secondary contact if there is selection against hybrids or strong assortative mating [27]. Fourth, we combined estimates of cline width and linkage disequilibrium to infer dispersal and selection in the hybrid zone given a

tension-zone model. We were unable to obtain sufficient samples at Lake Barrine to enable cline-based analyses; thus, these samples were only included in the individual-based analyses and disequilibrium measures.

First, to determine the composition and type of hybrids in the contact zone, we used two programs, Structure [285] and NewHybrids [7]. For both programs, we used genotypic data from all ten nuclear loci for all 406 individuals sampled. Using a Bayesian approach, Structure estimates the probability that an individual belongs to a genetic cluster by minimizing linkage and Hardy-Weinberg disequilibrium within a cluster (K). We ran Structure 10 times under the 'admixture' model for each of nine K values (ranging from 2 to 10), recording posterior probability distributions for admixture proportion. We then determined the best-supported K value following [101] as implemented in StructureHarvester [94], summarized results across that K value using Clumpp [162], and plotted results using Distruct [300]. NewHybrids also implements a Bayesian approach to determine the probability that an individual belongs to one of the six genotypic classes that results from the first two generations of crossing (*i.e.*, either parental form, F1 hybrid, F2 hybrid, or first-generation backcross). We ran NewHybrids five times and summarized runs by averaging probabilities across runs.

To fit clines to our data, we first pooled our unique sampling points along the transect into localities. On average, the area around pooled points (as measured by minimum convex polygons) was 297 m^2, with areas ranging from 3 to 1498 m^2. We then calculated the location of the pooled points along the transect by collapsing the points to a one-dimensional transect; because our sampling regime followed a linear transect, the results were nearly identical to the original sampling points. We then fit three models of clines to our data via the maximum likelihood framework implemented in Analyse [26]. First, we fit a basic two-parameter sigmoidal model (Sig) which describes the transition in allele frequency through space (p) with respect to cline center (c) and width (w) as:

$$p = \frac{1 + \tanh[\frac{2(x-c)}{w}]}{2} \tag{5.1}$$

Here, cline width is the inverse of the maximum slope of the curve. Sigmoidal clines do not necessarily evoke a specific selection model, and thus, can be used to describe a frequency change in any trait. Second, we fit a four-parameter stepped model (Step) in which the center of the cline is described by the sigmoidal model and the tails of the cline are described by the parameter B and the exponential decay function [349]:

$$p \propto \exp(\frac{-4x\theta^{\frac{1}{2}}}{w}) \tag{5.2}$$

The stepped model is appropriate for multi-trait data; it allows for a sharp change in frequency at the center of cline, as might be seen due to epistatic interactions. In the tails of the cline, where recombination has broken down epistasis and selection is accordingly weaker, introgression occurs more quickly and clines are shallower [25]. The parameter

θ reflects the strength of selection against the character outside of the cline center, and B describes the size of the tails, or the proportion of alleles that are introgressing. Finally, we fit a special case of the stepped model, the six-parameter asymmetric stepped model (Astep), in which either side of the cline has different introgression extent (*i.e.*, θ and B are fit separately to either side of the cline). This model describes scenarios in which introgression is greater into one lineage than the other. In each of these models, we allowed p_{min} and p_{max} to vary at either end of the cline. Each of these models is nested within each other; thus, to determine whether more complex models fit the data better, we calculated twice the difference of log likelihoods and found significance by comparing to the critical value. Clines were fit to each of the 11 loci and a composite multilocus hybrid index obtained from the Structure results. After fitting the clines, we determined whether clines were coincident (*i.e.*, sharing the same center) and concordant (*i.e.*, sharing the same width). We would expect coincidence if the hybrid zone is recent or if selection is strong and concordance if selection strength is uniform across loci. Following Phillips *et al.* 2006 , we constructed log-likelihood profiles for each locus over a range of center (c) and width (w) values. We then calculated two likelihood values: (1) for the non-coincident model, we summed the locus-specific maximum-likelihood values and (2) for the coincident model, we summed the c log-likelihood profiles over all loci to find the maximum likelihood value for the loci's shared center. To determine whether non-coincidence fit the data better than coincidence, we determined if the difference in likelihood between the two models was significant using a chi-square distribution (df=10, or one less than the number of loci). The same approach with cline widths was used to determine cline concordance.

To calculate within locus and between locus disequilibrium, we also used Analyse. Within locus disequilibrium, or Hardy-Weinberg disequilibrium, results when the proportion of heterozygotes at a locus deviates from the expected proportion under random mating. We calculated Hardy-Weinberg disequilibrium by estimating maximum likelihood values for F_{IS} across all nuclear loci and across all sites. Analyse calculates between-locus disequilibria, or linkage disequilibrium (LD), as:

$$R_{ij} = \frac{D_{ij}}{\sqrt{p_i q_i p_j q_j}} \tag{5.3}$$

Because the magnitude of LD relies, in part, on the allele frequencies at the loci under consideration, this method reduces this dependency by dividing the estimate of linkage disequilibrium by the square root of the product of the allele frequencies at the loci [27]. To calculate LD, two challenges arise: first, how to account for within locus disequilibrium as it can affect measures of LD, and second, how to estimate multilocus disequilibrium properly when pairwise measures of disequilibria are not independent. Analyse addresses the first issue by accounting for within locus disequilibria by downsizing the effective sample size of a population, as each allele sampled does not reflect a unique data point when there is within-locus disequilibrium. This method does not, however, address the

second issue; it assumes that pairwise LD estimates are independent. Other multilocus LD methods, which allow dependency, are not appropriate here. Barton's (2000) method for estimating multilocus disequilibria cannot handle a data set of this size, and Barton and Gale's (1993) method for estimating disequilibria by hybrid index assumes no within-locus disequilibrium.

Assuming a tension zone model (*i.e.*, that all clines are in migration-selection equilibrium), estimates of cline width and linkage disequilibrium can be used to estimate dispersal in the hybrid zone as represented by σ, the variance in position between offspring and parents [27]. If selection in the hybrid zone is weak, σ is given by:

$$R_{ij} = \frac{4\sigma^2}{w^2 r} \tag{5.4}$$

where r is the recombination rate between loci; we assume no linkage between markers and thus take r to be 0.5. Here, R_{ij} is calculated at the center of the zone, where it is predicted to be the largest. We calculated average LD at each of the four center localities in the hybrid zone and averaged these estimates to derive a value for R_{ij}. As our clines were not concordant, we calculated σ for the range of observed widths we saw, the mean width, and the width of the composite cline based on hybrid index. Once σ has been estimated, we can estimate selection against heterozygotes by using the equation:

$$s^* = 8(\frac{\sigma}{w})^2 \tag{5.5}$$

where s^* is a measure of effective selection on a locus, which reflects both selection acting directly on the locus and on loci in disequilibrium with the locus [27].

For these analyses, we used R and the package ggplot2 to conduct all mathematical and statistical operations and to make all graphs [288, 369].

5.2.5 Simulations

Similar patterns of cline width and genetic disequilibrium in hybrid zones can emerge from very different biological realities; in fact, it can be notoriously challenging to distinguish between competing hypotheses for hybrid zone maintenance [140]. Thus, to place our genetic results in context of hybrid zone models, we simulated secondary contact between two isolated lineages using the forward-time simulation program simuPOP [273]. While models for hybrid zones are abundant in the literature and have contributed greatly to our understanding of hybrid zone dynamics [16, 92, 184], few incorporate assortative mating and selection against hybrids in a multi-locus framework as we do here. For this model, we implement a one-dimensional chain of sixty populations; the two lineages occupy either side of the chain at time zero (Fig. S4). At time zero, the populations begin exchanging migrants under a stepping-stone model of migration. Under a one-dimensional stepping stone model, the proportion of migrants is related to σ by $\sigma^2 = m\epsilon^2$ where ϵ

is the distance between demes. Cline width is thus a dimensionless value measured in deme number; converting to distance to compare to empirical systems necessitates an estimate of ϵ. Selection against hybrids was defined by a multiplicative selection model, where fitness was dependent on the number of loci at which an individual was heterozygous. We allowed strength of assortative mating to vary from random mating to nearly complete associative mating. In the model of assortative mating used here (a multilocus model based on Felsenstein's (1981) "group-based model") , some proportion of individuals (α) mate preferentially with any individuals who share at least 90% or more of their ancestry, and the remaining individuals (1-α) mate randomly. We simulated clines under a range of scenarios, including varying the number of loci under selection (2, 5, 10), the strength of selection (0, 0.2, 0.4, 0.6, 0.8, 0.9, 0.95, 0.99), the strength of assortative mating (0, 0.2, 0.4, 0.6, 0.8), and migration rates (0.1, 0.3, 0.5). Population sizes ($N = 1000$) and recombination rates ($r = 0.5$) were kept constant across simulations, and in each simulation, we modeled ten neutral loci to look at introgression patterns under neutrality. Simulations were run for 1000 generations as preliminary runs suggested that, where relevant, this was sufficient for runs to reach equilibrium. We recorded data from the simulations every 100 generations, which included estimates of cline width and center (as calculated by Kruuk *et al.*, 1999) , per-locus Hardy-Weinberg disequilibrium, linkage disequilibrium, and hybrid indices. We simulated 1000 data sets for each parameter combination. Supplementary information contains more information about model choice and parameterization (Table S4 and Fig. S4).

5.3 Results

5.3.1 Identification of the contact zone and development of marker loci

We located the zone of secondary contact at two places, along a road running through both lineages along the Gillies Range (Gillies Transect) and around Lake Barrine. The habitat is continuous through the contact zone, with no major transitions in habitat types, and *L. coggeri* are abundant throughout. This contact zone is near several other contact zones in the AWT suture zone; in particular, Lake Barrine is home to an admixed population of two frog species *Austrochaperina robusta/fryi*, and the previously described contact in the skink *Carlia rubrigularis* is located 15 km northward [244, 278].

High-throughput sequencing of the pooled libraries from the allopatric localities resulted in the identification of over 20,000 putatively-fixed and annotated SNPs between the two lineages. Sanger-sequencing data of a subset of these SNPs and their surrounding regions from allopatric (c. 40 km distant samples; N=12 each) localities confirmed that all were fixed or nearly-fixed, although other variants close to the target SNPs ($\pm$200 bp) showed incomplete lineage sorting.

Our mitochondrial locus and seven of the ten nuclear SNPs were completely fixed between allopatric localities of the two lineages. SNPs located in three loci (*PCBD1*, *LC17*,

and *ABHD5*) showed incomplete fixation when comparing the Northern and Southern allopatric localities, with F_{ST} estimates of 0.92, 0.75, and 0.96 respectively. Although this incomplete fixation could be due to introgression from the contact zone, this is unlikely as these localities are geographically disjunct from the hybrid zone.

In total for the contact zone, we genotyped all ten nuclear and one mitochondrial markers in 406 lizards, 348 from the Gillies transect and 58 from Lake Barrine. These data are available at Dryad at `http://datadryad.org/handle/10255/dryad.36371`.

5.3.2 Estimates of hybrid frequency and composition

Results from the individual-based Structure analyses showed that the individuals at the contact zone best fit a two-population model, where each parental lineage is a genetic cluster and hybrids are the result of admixture between these clusters. At the Gillies Transect, population assignment tests showed that the majority of hybrids (45 out of 63; 71%) were limited mostly to just four geographically close localities, spanning from 1041 to 1151 m along the transect. Here, hybrids are defined as any individual that has an admixture proportion of ≥ 0.10, and for whom the posterior probability of admixture proportion does not include 0 or 1. Of the 81 individuals in these four central localities, 45 (or 55%) were hybrids, of which 65% and 35% had Northern and Southern mitochondrial types, respectively. As shown in Fig. 2B, the admixture proportion (*i.e.*, hybrid index) in the hybrid zone center (localities at 1041 to 1151 m) spans a wide range, a pattern that has been described as flatly uniform [167], as opposed to unimodal or bimodal patterns. Outside of these four localities, only 18 additional lizards (5% of genotyped lizards) were identified as hybrids, and the two localities located 2.5-km from the hybrid zone center showed no sign of introgression (Fig. 2A). Thus, introgression beyond the hybrid zone center, as estimated from these markers, appears very rare.

Using NewHybrids to assign individuals to hybrid class showed that the hybrid zone center contained no F1s, 16 parents from the Northern lineage (20%), and 12 parents from the Southern lineage (15%) (Fig. S1). Delineating between F2s, first-generation backcrosses, and older backcrosses confidently is challenging [275]; thus, we do not categorize the 53 hybrids further.

The pattern seen along the Gillies transect was matched at Lake Barrine (Fig. 1B). Of the 58 individuals genotyped there, 16 (27%) were hybrids, and none could be confidently classified as F1s. These hybrids were all found in two narrow bands on either side of the lake, each measuring about 300 m in width (Fig. S2).

5.3.3 Estimates of cline shape

All clines were best fit by the sigmoidal model; while some markers showed a marginally better fit under the stepped or asymmetric stepped model, the improvement over the sigmoidal model was insignificant. We report and discuss the results from the sigmoidal clines only. Clines were exceptionally narrow – average cline width was 403 m, and

widths ranged from 280 m (*LC5*) to 695 m (*LEMD2*) (Fig. 3; Table 1). The mitochondrial cline and the composite hybrid index cline were both narrower than average; they had widths of 300 m and 370 m respectively. For most clines, p_{min} and p_{max} were 0 and 1 respectively; however, the three loci for which fixation between allopatric samples was incomplete had, as expected, non-zero p_{min} (Table 1).

Visual inspection of the clines showed that they had coincident centers; all centers were within 100 meters of each other. A formal test of cline coincidence confirmed this observation (χ^2=7.26, df=10, p=0.70); the best-fitting center was located 1.18 km from the start of the transect, at the southern edge of the four localities at the core of the hybrid zone. Yet, the clines were not concordant; allowing cline width to vary across loci fit the data significantly better than constraining clines to the same width (χ^2=184.8, df=10, p¡0.05). Although the clines are not concordant, their widths fall within a narrow range. Most cline widths are within 400$\pm$100 m, which, given that $\sigma \approx 80\frac{m}{\sqrt{gen}}$ (see below), is a relatively small deviation from concordance.

5.3.4 Estimates of disequilibrium

Estimates of F_{IS} and R_{ij} showed that both within-locus and between-loci disequilibrium is substantial in the hybrid zone. In most localities, power to measure within-locus disequilibrium was limited as many localities only had one allele at each locus. However, six localities showed significant departures from Hardy-Weinberg Equilibrium (HWE) at one or more loci. The four localities at the center of the hybrid zone had across-loci F_{IS} values ranging from 0.239 to 0.453; per locus measures showed that almost all loci had significantly non-zero F_{IS} values (Fig. 4A). Two localities away from the center of the hybrid zone showed deviations from HWE; per locus results indicated that departures from HWE at two of the incompletely-sorted loci, *PCBD1* and *LC17*, drove this pattern. Disequilibrium where the two lineages meet at Lake Barrine was similarly substantial; across-locus F_{IS} was estimated to be 0.408 (0.266-0.540).

Similarly, power to estimate linkage disequilibrium was limited in many localities as most localities consisted of individuals homozygous at all loci. However, six localities had significant R_{ij}, all in or near the center of the hybrid zone (Fig. 4B). The four localities at the center of the hybrid zone all had significant, positive LD at nearly all between-locus comparisons; average LD across all loci at these localities ranged from 0.184 to 0.446. Again, disequilibrium at Lake Barrine was strong; almost all locus pairs had significant, positive LD and the multi-locus estimate of linkage disequilibrium was 0.490 (0.440-0.523).

5.3.5 Estimates of dispersal rate and selection

Assuming a tension zone model and that the hybrid zone is at migration-selection equilibrium, measures of cline width and linkage disequilibrium can be used to estimate dispersal rate in the hybrid zone and selection against hybrids. These measures do not ex-

plicitly account for variation in cline width among characters; thus, we report here point estimates based on the width of our multilocus cline, as well as the range of values corresponding to the range of cline widths. Dispersal rate (here measured as σ) was estimated as 80 $\frac{m}{\sqrt{gen}}$(40 - 160 $\frac{m}{\sqrt{gen}}$). Although estimating dispersal rate is fraught with assumptions, in particular that the system conforms to a tension zone model, our estimates correspond well to those from two closely-related species. In the rainforest skink *Carlia rubrigularis* [278], σ was estimated to range from 90 to 133 $\frac{m}{\sqrt{gen}}$ by using both the method described here and a F_{ST} based measure [305], and in the rainforest skink *Gnypetoscincus queenslandiae*, σ, as estimated via the Rousset (1997) method, was 29 $\frac{m}{\sqrt{gen}}$ [344]. Using our estimate of σ for *L. coggeri*, we estimate average effective selection (s^*) at a locus as 0.403 (0.106 to 0.653). The wide range of possible values for selection strength reflects both the challenge in measuring the various parameters of this composite measure and the variation of selection strength across loci. However, our estimates suggest that, if the equilibrium tension model is appropriate for this system, then selection against hybrids is substantial. If our system does not fit a tension zone model, then we are likely over-estimating σ, in which case average effective selection would be less than we suggest here. However, we think that our estimate of σ is reasonable, as it overlaps with that for *C. rubrigularis*, a species with similar natural history and habitat use as *L. coggeri* [373].

5.3.6 Results from hybrid zone simulations

Motivated by the finding of strong and ubiquitous genetic disequilibrium and narrow clines in this hybrid zone, we used simulations to explore how reproductive isolation (both pre-zygotic and post-zygotic) versus neutral diffusion affects hybrid zone dynamics. Under models that ranged from random mating to nearly complete assortative mating and from neutral diffusion to nearly complete selection against hybrids, we simulated expected patterns of (1) hybrid composition, (2) cline width and concordance, and (3) disequilibrium. For ease of presentation, we discuss and show results from just one parameter set (migration rate = 0.3 and 10 loci under selection); results from other parameter combinations are qualitatively similar (Fig. S3). Several key conclusions emerge from these simulations. First, strength of assortative mating has no significant effect on any of the measured parameters; while assortative mating slows initial introgression and the decay of disequilibrium, the equilibrium outcomes are the same under strong assortative mating and random mating (Fig. S4). Second, under neutral diffusion (e.g., no selection against hybrids) and very soon after secondary contact (¡50 generations under a range of demographic parameters), the system exhibits many of the same patterns one would see in a tension zone model – *i.e.*, narrow clines, high disequilibrium, and flatly uniform distribution of hybrid indices (Fig. 5). Beyond this point, disequilibrium is low and clines increasingly broad with gradually increasing variance in cline width. Third, patterns of disequilibrium and cline shape diverge between markers under selection and unlinked neutral markers rapidly after secondary contact (<50 generations) and only re-

main concordant when total selection against hybrids is strong (>90%; Fig. 5). In particular, assuming independent assortment between neutral and selected loci, maintaining narrow clines at neutral loci requires total selection against hybrids to be nearly complete (>95%). Fourth, continuing this theme, variance in cline width among neutral, unlinked loci increases rapidly after secondary contact, even under strong selection against hybrids. Although average cline width remains narrow at neutral loci when selection against hybrids is strong, some loci show stochastic behavior and wide clines, increasing variance. Seeing this variance, however, requires a large number of loci to be sampled. Finally, hybrid index tends towards unimodality, unless contact is recent or there is strong selection against hybrids (>90%), in which case it is bimodal to flatly uniform. Further, in all these cases, finding F1s in the hybrid zone is rare. These results suggest that our pattern of narrow, fairly concordant clines and high disequilibrium is likely the result of very recent secondary contact (e.g., <50 generations) or strong selection against hybrids.

5.4 Discussion

In this paper, we show that the major lineages of *L. coggeri* meet in an extremely narrow hybrid zone, which we can describe further as:

- evincing clines of average width 403 m in a species with dispersal rate of approximately 80 $\frac{m}{\sqrt{gen}}$,
- consisting of coincident clines, which while not concordant, show low variance in cline width,
- showing substantial LD and HW disequilibrium at the center of the zone at nearly every marker,
- having a flatly uniform distribution of hybrid index with no F1 hybrids,
- and, exhibiting the same general patterns of limited introgression and extensive disequilibrium at a second independent sampling site, Lake Barrine.

Narrow hybrid zones can emerge due to several processes [239]. Here, we describe what we consider the four dominant causes, noting that these are not mutually exclusive. First, as a null hypothesis, a narrow hybrid zone can result from neutral diffusion after secondary contact between previously-isolated lineages [100]. If there are no barriers to gene flow between the lineages, the two lineages will eventually become genetically and phenotypically indistinguishable. Prior to this equilibrium, however, the system will exhibit clines, the width of which are a function of dispersal length and time since contact [27]. Second, narrow clines can result from selection against heterozygotes or, more generally, hybrids under the environmentally-independent "tension zone" model [30, 349]. Here,

clines are stable at the equilibrium between parental dispersal into the hybrid zone and selection against resulting hybrids, and selection is independent of the environment though the clines often cluster in areas of low environment suitability [349]. Third, narrow clines can result due to assortative mating between parental forms, whether due to active mate choice (*e.g.*, *Heliconius* butterflies [217]) or habitat selection (*e.g.*, in *Bombina* toads [209]). Finally, narrow clines can form at the edge of an ecotone between two distinct environments, when populations on either side of the ecotone are differentially adapted to these conditions [100]. Here, we explore these possible explanations for the *L. coggeri* hybrid zone by considering their predictions in light of our data and simulation results.

Although environment-dependent selection is certainly important in shaping numerous hybrid zones (*e.g.*, in *Iris* flowers [69] and in *Colaptes* birds [241]), it likely not a factor contributing to this hybrid zone. Bioclimatic analysis of the suture zone relative to adjacent refugial (source) areas indicated the former has relatively low suitability, but also that the parental lineages are from analogous rainforest habitats [244]. As there is no noticeable difference between parental habitats or in eco-phenotypes [35] and this region of low suitability is much broader and more subtle than one would expect for such a narrow hybrid zone, we think environmental selection is unlikely to contribute significantly here. Distinguishing between the remaining three explanations is notoriously difficult [10], but it is key if we want to use hybrid zones to understand speciation better. After all, by understanding what maintains a hybrid zone, we can understand what barriers have evolved to gene flow, and thus, what factors are contributing to lineage maintenance.

If the system is described by neutral diffusion, then cline width is given by $w = \sqrt{2\pi}\sigma\sqrt{t}$, where w is cline width and t is the time (in generations) since secondary contact [100]. We do not have direct estimates of t for this system, but we use what we know about the system to evaluate the plausibility of this non-equilibrium model. First, paleomodels suggest the rainforest expanded from glacial refugia sometime 3 to 8 Kya [360], and as *L. coggeri* is found at rainforest edges and gaps, it likely tracked this expansion closely. Assuming we estimated σ correctly and assuming a conservative 3 $years/gen$, then under neutral diffusion, we would expect clines to be (at minimum) 6.3 km wide, nearly 15 times wider than our average cline width. To get clines as narrow as those measured here given this estimate of time since secondary contact, σ would need to be just 14 $\frac{m}{\sqrt{gen}}$. This estimate of dispersal rate is nearly an order of magnitude lower than that estimated in the closely-related and ecologically-similar species, *C. rubrigularis* [278]. Further, *L.coggeri* is a gap-edge species such that population densities are naturally dynamic in space, and thus based on natural history, this estimate is too low to be realistic [373]. Moreover, under this scenario, our simulations show that we would not expect to see extensive disequilibrium and cline coincidence as we do. However, it is possible that we have wrongly inferred time since secondary contact, and secondary contact is more recent. Given the width of our multilocus and mitochondrial clines and using a range of dispersal rates for a closely-related lizard species (*C. rubrigularis*; 90 - 133 $\frac{m}{\sqrt{gen}}$), we estimate time since secondary contact as 0.81 to 2.70 generations, or 2 to 8 years given a conservative generation time

of three years. Our simulations suggest that in the early stages of neutral diffusion (<50 generations), we can expect to recover all the patterns of hybridization seen in this contact zone (*i.e.*, high disequilibrium, sharply narrow and concordant clines, a flatly uniform hybrid distribution; Fig. 5). However, it seems unlikely that the lineages met for the first time in just the last 10 years. It is more likely that the lineages met in the past, and environmental change, whether natural or human-induced, has affected the hybrid zone such that the clock for time since secondary contact has been reset [271]. However, even in this scenario, the lineages would have been previously in secondary contact, during which there would have been ample time for broad-scale introgression, which is noticeably absent. In conclusion, while neutral diffusion can result in the patterns of introgression and disequilibrium we see here, the timing of secondary contact and dispersal rate necessary to generate such patterns are unrealistic based on our knowledge about this system.

Tension zones can produce the same patterns as well (Fig. 5), and most systems that exhibit similar patterns as the *L. coggeri* zone tend to be defined as tension zones (e.g., the *(Vandiemenella* grasshoppers [173] and the *Carlia* lizards [278]). However, these and our systems do not fit the tension zone model in one important way; even when hybrids are under strong selection, introgression patterns are expected to be uneven across loci. Introgression extent is inversely proportional to effective selection strength, which is both a function of direct selection on a locus and selection on any loci in linkage disequilibrium with the focal locus [25]. For neutral loci, which experience no direct selection, theory suggests that indirect selection can retard introgression [25]. But, as theory further suggests and as our simulations show, recombination breaks down linkage disequilibrium between neutral loci and loci under selection, over time increasing extent of introgression and variance in cline width ([24],Fig. 5). The speed at which this occurs depends on the strength of selection relative to the rate of recombination between the neutral and selected loci [16]. Under equilibrium, we would thus expect to see a wide range of cline widths; however, in our study, much like most other studies that have measured cline width, the range of cline widths is fairly limited (Table 2).

In the context of the tension zone model, this pattern of limited introgression has three possible roots. First, it is possible all our markers are under strong direct selection. Others have argued this is likely when researchers chose to assay introgression extent using diagnostic markers, which have possibly become fixed due to divergent natural selection [388]. However, in this system, we see that even our markers with incomplete lineage sorting have narrow clines. Second, as theory and our simulations suggest [114], if total selection against hybrids is great (>95%), then introgression at neutral alleles, even if unlinked to loci under selection, will be greatly reduced. Although we measured strong effective selection at individual loci (average ¿40%) in this system, we cannot easily convert this to total selection strength as this measure is dependent on the number of loci under selection and the recombination rates between these loci. Finally, it is possible that selection is strong and so widely dispersed over the genome that every marker, even if neutral, is closely linked to multiple markers under selection [25]. In such systems, multi-locus clines can be narrower than predicted by the direct selection each locus is

experiencing, and neutral loci can be slow to introgress past the cline [16]. Testing pervasive selection as a possible mechanism for narrow clines at neutral loci would require investigating introgression at more loci, ideally in context of their genomic location.

Finally, assortative mating has been invoked in several hybrid zones as maintaining lineage boundaries [136, 310]. Theoretical studies have shown that assortative mating can limit introgression at neutral loci [116, 235, 240], but this result depends on the model used for assortative mating as all these models used a one-locus or two-locus model for mating traits and preference (for a counter-example where assortative mating has little effect on outcomes, see [309]). When using a multi-locus model (appropriate for quantitative traits) and group-based mating as done here, assortative mating does little to limit introgression further. Thus, although assortative mating could potentially solidify lineage boundaries through reinforcement (e.g., the *Litoria serrata* frog [155]), we think assortative mating is unlikely to be an important force structuring this hybrid zone. However, to test this conclusively, we would need to do mate choice experiments for the lineages meeting in the hybrid zone.

Of all the possible forces structuring this hybrid zone, it seems most likely that there is strong selection against hybrids. It is, of course, possible both that secondary contact has been recent and that selection is strong, such that this hybrid zone is not yet at equilibrium. In the case of non-equilibrium, using cline width to infer selection will overestimate selection strength. In particular, when selection is pervasive across a genome, approach to equilibrium can be slowed down by multi-locus effects [16]. If so, we would expect to see a limited range of cline widths as we see here (Table 1). Ultimately, to disentangle how recency of contact and selection against hybrids are contributing to hybrid zone dynamics, we need to (1) collect data on hybrid viability and fitness based on experimental crosses or field-based studies and/or (2) look at distributions of the length of the introgressed blocks to estimate the age of the contact [282]. Upon introgressing into the foreign population, chromosomes will be whole, but with time, recombination will break up the introgressed chromosome into small blocks of introgression. As such, even with strong selection, blocks should be much smaller in older contacts than in recent contacts [16] and, given estimates of the population recombination rate, should enable us to determine the age of these contacts.

Data from this hybrid zone suggests that selection against hybrids is key in maintaining narrow clines. Very few studies have investigated contact zones between cryptic lineages; most analyzed hybrid zones are between phenotypically-distinct lineages or chromosomally-distinct races (Table 2). The few other studies of contact zones between morphologically-cryptic lineages have shown a range of outcomes – from neutral diffusion (*Chioglossa lusitanica* salamander; [319]) to narrow and stable clines (*Carlia rubrigularis* lizard; [278] and *Lacerta schreiberi* lizard; [342]). Such studies are still too few to draw broad conclusions about the nature of hybrid zones between phylogeographic isolates. Our results are striking, both in comparison to these studies and other systems (Table 2), in how narrow the clines are, and certainly suggest that reproductive isolation can evolve without overt morphological differentiation. The evolution of substantial reproductive

isolation is particularly fascinating as these lineages likely came contact during previous interglacials, during which they could have merged [146, 112]. Yet, these lineages have remained distinct at all assayed loci, suggesting that substantial reproductive isolation can evolve quickly despite opportunity for cyclic introgression and without detectable morphological or ecological divergence.

With this work, *L. coggeri* joins an ever-growing list of systems in which researchers have identified evidence for deep phylogeographic structure with little or no overt morphological or ecological divergence (for example, *Heteronotia binoei* lizard [110], *Hemidactylus fasciatus* lizard [192], *Aptostichus atomarius* spider [47], *Aneides flavipunctatus* salamander [297], *Mielichhoferia elongata* moss [321]). As yet, there are very few cases where the extent of reproductive isolation has been estimated from either experimental crosses or hybrid zone analyses. However, like the *L. coggeri* system, these cases often demonstrate strong barriers to gene flow (e.g., the *Litoria myola/serrata* frog [155], *C. rubrigularis* lizard [278], *Astraptes fulgerator* butterfly [141], *Brachionus plicatilis* rotifers [124], *Draba sp.* plants [134]). This work suggests that, in contrast to the current emphasis on divergent ecologically-based selection as a driver of speciation, divergence in less obvious niche dimensions (e.g., shifts in physiology or in chemiosensory cues [164, 335]), divergence due to parallel adaptation [219, 260], or drift-driven, non-adaptive divergence [363]) might be more important in species formation than generally recognized. Indeed, this growing body of work suggests that phylogeographic lineages are, like species, "merely one stage in progressive modification", and thus, important in their own right (Grinnell, as cited in [345]).

5.5 Acknowledgements

For funding, we acknowledge support from the David and Marvalee Wake Fund at the Museum of Vertebrate Zoology, the National Science Foundation Graduate Research Fellowship, the National Geographic Committee for Research and Exploration, and the Center for Integrative Genomics at UC-Berkeley. The Texas Advanced Computing Center (TACC) at The University of Texas at Austin provided grid resources that have contributed to the research results reported within this paper. For help in the field, we thank Rayna Bell, Emily Hoffmann, Conrad Hoskin, Jim McGuire, Ben Phillips, and Maria Tonione. For help in the lab, we thank Tarini Ullal. For logistical support, we thank CSIRO (Sandra Kay and Dave Westcott), JCU (Noema Patterson and Yvette Williams), SFS (Ian Brennan and Sigrid Heise), and QPWS (Keith McDonald, Michelle Nissen, and Chris Wegger), and for help with simuPOP, Bo Peng and for help with Analyse, Stuart Baird. For advice and discussions, we gratefully acknowledge Tom Devitt, Jay McEntee, Montgomery Slatkin, and David Wake, and for thoughtful comments on previous versions of this manuscript, Stuart Baird, Jim Patton, Ben Phillips, and Ricardo Pereira.

5.6 Data Accessibility

Data are available at the following locations:

1. Genotypes collected are available on DRYAD, entry `doi:10.5061/dryad.4gh6hf5g`
2. Base script used for hybrid zone simulations available on DRYAD, entry `doi:10.5061/dryad.4gh6hf5g`
3. Assembled and annotated transcriptome is available on DRYAD, entry `doi:10.5061/dryad.4gh6hf5g`
4. Scripts used is available at `https://github.com/singhal/transcriptomic`

5.7 Figures

Figure 5.1: A. Range of *L. coggeri* with major lineages identified; localities used for marker development (≈40 km away from the hybrid zone center) are shown by stars. B. Close-up of contact zone, showing sampling points and present land cover.

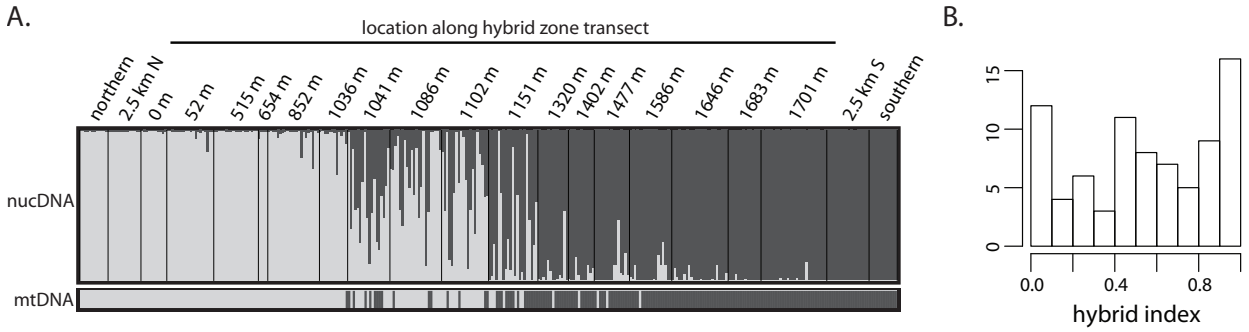

Figure 5.2: A. Structure results for all individuals in the linear Gillies transect, showing the point estimate for admixture coefficient. Individuals are arranged north to south. B. Histogram of point estimates of hybrid index for individuals at the center of the hybrid zone.

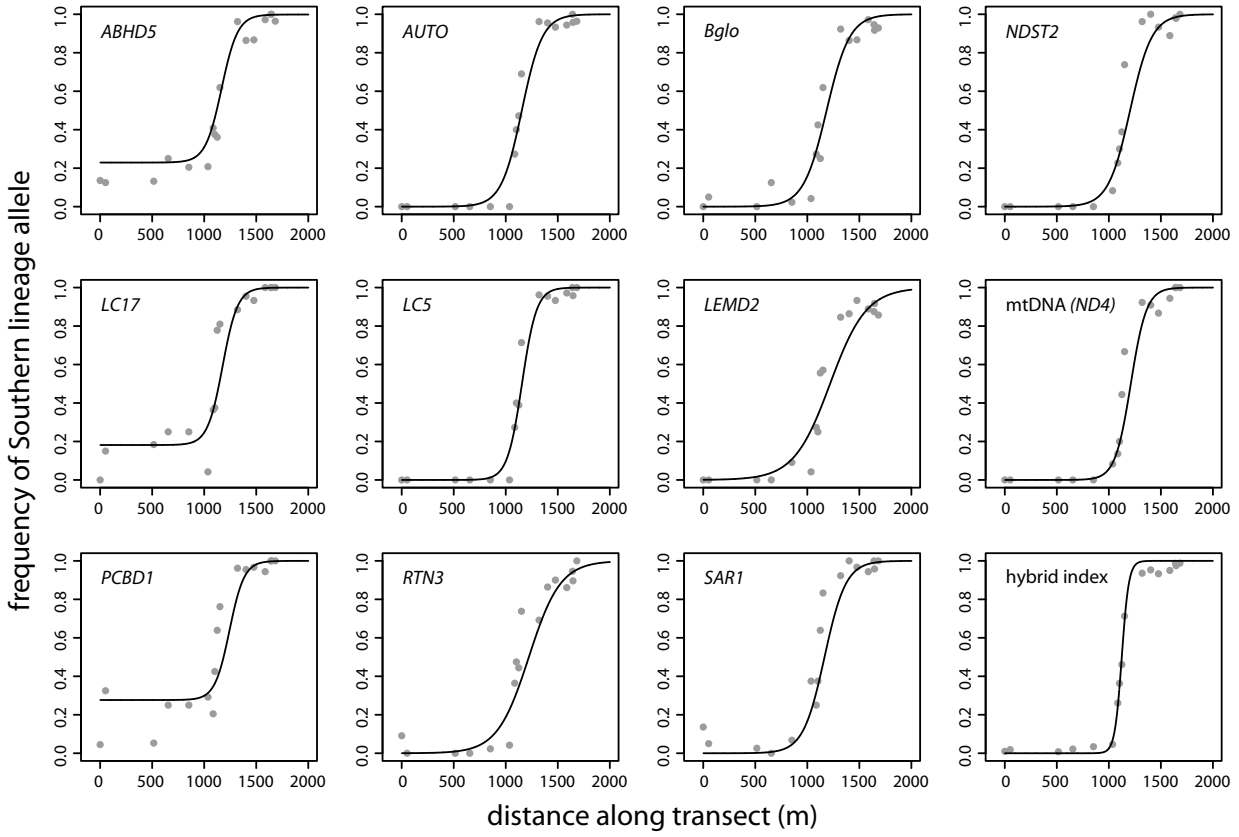

Figure 5.3: Maximum likelihood estimates for cline shape and location for 11 markers studied in this hybrid zone and the cline for our composite measure, hybrid index.

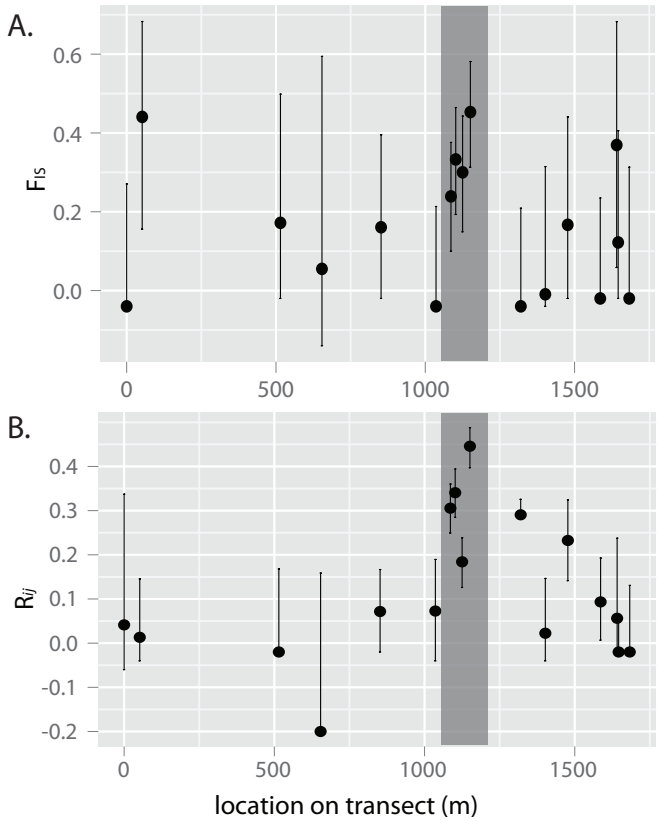

Figure 5.4: A. Hardy-Weinberg disequilibrium F_{IS} measures and B. Linkage disequilibrium R_{ij} in localities along the linear transect. Darker gray box outlines the localities in the hybrid zone center.

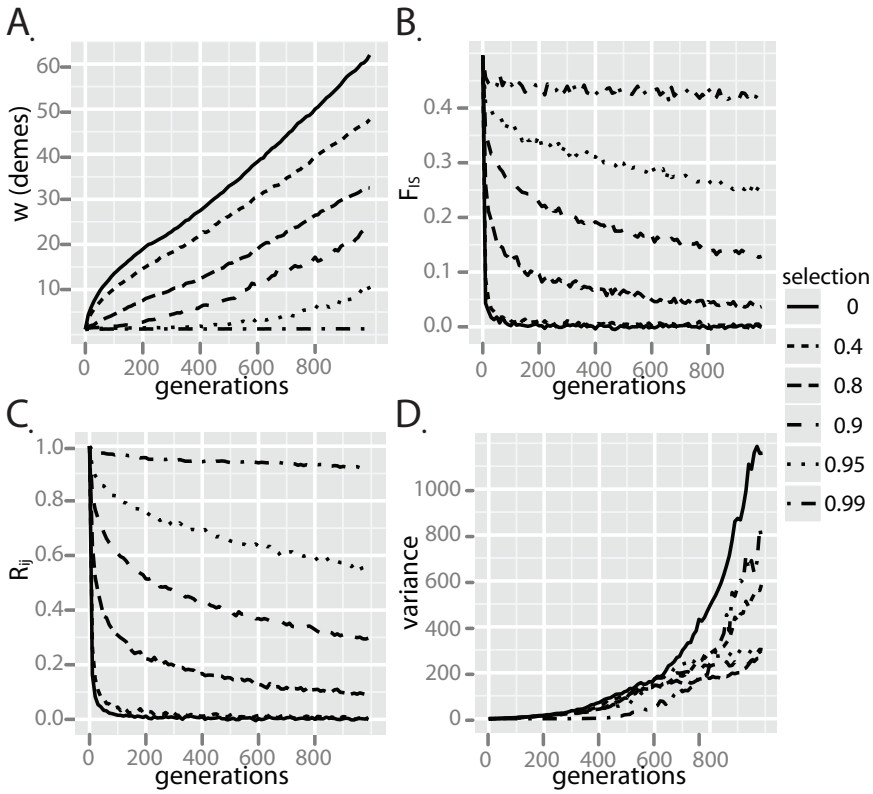

Figure 5.5: Results from the simulations of secondary contact, A. Cline width (in demes), B. F_{IS}, and C. R_{ij} at neutral loci for a range of values for selection against hybrids. Shown for migration rate of 0.3, with ten loci under selection, and random mating.

5.8 Tables

locus	LnL	p_{min}	p_{max}	width (w) in meters	center (c) in meters
mtDNA (*ND4*)	-4.469	0	1	300 (223-416)	1211 (1165-1266)
hybrid index	-3.21	0	1	370 (310-456)	1150 (1104-1216)
ABHD5	-1.952	0.23	1	311 (3-621)	1167 (1036-1313)
AUTO	-8.842	0	1	395 (234-464)	1155 (1118-1227)
Bglo	-4.07	0	1	444 (334-610)	1185.5 (1114-1249)
LC17	-0.786	0.18	1	288 (14-488)	1176.5 (1091-1312)
LC5	-2.764	0	1	280 (175-436)	1156 (1104-1215)
LEMD2	-3.624	0	1	696 (519-980)	1219 (1123-1290)
NDST2	-8.799	0	1	419 (312-562)	1204 (1143-1262)
PCBD1	-7.705	0.28	1	282 (2-372)	1243 (1094-1317)
RTN3	-2.619	0	1	607 (460-826)	1222 (1144-1293)
SAR1	-3.471	0	1	411 (316-573)	1165.5 (1106-1229)

Table 5.1: Summary of estimates of cline parameters with two-unit support limits shown in parentheses.

citation	species	taxa	minimum width (km)	maximum width (km)	coefficient of variation	number of loci	s*
Dasmahapatra *et al.*, 2002	*Anaratia fatima* and *amathea*	butterfly	26	28	0.06	4	-
Sites *et al.*, 1995	*Sceloperous grammicus* (chromosomal races)	lizard	0.87	1.29	0.26	3	0.3
Porter *et al.*, 1994	*Pontia daplidice* and *edusa*	butterfly	18.95	25.15	0.29	4	~0.5
Szymura and Barton 1991	*Bombina bombina* and *variegata*	toad	5	7.92	0.3	6	0.22
Alexandrino *et al.*, 2005	*E. eschscholtzii xanthoptica* and *E. e. platensis*	salamander	0.5	0.9	0.43	9	0.46-0.75
Brumfield *et al.*, 2001	*Manacus candei* and *vitellinus*	bird	3.9	11	0.56	4	-
Buno *et al.*, 1994	*C.p. erythropus* and *C.p.parallelus*	grasshopper	33.02	55.82	0.6	3	-
Yanchukov *et al.*, 2006	*Bombina bombina* and *variegata*	toad	0.86	4.25	0.66	8	-
this study	***L. coggeri* (C and S lineages)**	**lizard**	**0.28**	**0.696**	**0.73**	**11**	**0.11-0.65**
Machalon *et al.*, 2007	*Mus musculus* and *domesticus*	mouse	6.43	18.07	0.84	6	0.056-0.090
Carling and Brumfield 2008	*Passerina cyanea* and *amoena*	bird	175	523	0.87	7	-
Gay *et al.*, 2008	*Larus glaucescens* and *occidentalis*	bird	400	1140	0.93	7	-
Mettler and Spellman 2009	*Pheucticus melanocephalus* and *ludovicianus*	bird	82	356	0.99	4	-
Carling and Brumfield 2009	*Passerina cyanea* and *amoena*	bird	2.8	584	1.11	10	-
Phillips *et al.*, 2004	*Carlia rubrigularis* (N and S lineages)	lizard	0.45	2.24	1.15	4	0.50-0.70
Kawakami *et al.*, 2008	*Vandiemenella viatica* (chromosomal races)	grasshopper	0.093	0.347	1.89	12	0.197
Teeter *et al.*, 2008	*Mus musculus* and *domesticus*	mouse	6.5	341	2.41	38	-
Dufkova *et al.*, 2011	*Mus musculus* and *domesticus*	mouse	0.23	16.76	2.95	13	0.25

Table 5.2: A summary of minimum and maximum widths for studies measuring clines. All studies found by doing a search for ("hybrid zone" OR hybridization) AND clin*) in Web of Knowledge on 19 April 2011. Only studies that measured clines using a sigmoidal or stepped model and that assayed 3 or more loci were included. Coefficient of variation is taken for the square of the widths, as $s^* \propto w^2$.

Chapter 6

Genomic analyses for non-model organisms

6.1 Abstract

High-throughput sequencing (HTS) is revolutionizing biological research by enabling scientists to quickly and cheaply query variation at a genomic scale. Despite the increasing ease of obtaining such data, using these data effectively still poses notable challenges, especially for those working with organisms without a high-quality reference genome. For every stage of analysis – from assembly to annotation to variant discovery – researchers have to distinguish technical artifacts from the biological realities of their data before they can make inference. In this work, I explore these challenges by generating a large *de novo* comparative transcriptomic dataset data for a clade of lizards and constructing a pipeline to analyze these data. Then, using a combination of novel metrics and an externally validated variant data set, I test the efficacy of my approach, identify areas of improvement, and propose ways to minimize these errors. I find that with careful data curation, HTS can be a powerful tool for generating genomic data for non-model organisms.

6.2 Introduction

High-throughput sequencing (HTS) is poised to revolutionize the field of evolutionary genetics by enabling researchers to assay thousands of loci for organisms across the tree of life. Already, HTS data sets have facilitated a wide range of studies, including identification of genes under natural selection [387], reconstructions of demographic history [206], and broad scale inference of phylogeny [337]. Daily, sequencing technologies and the corresponding bioinformatics tools improve, making these approaches even more accessible to a wide range of researchers. Still, acquiring HTS data for non-model organisms is non-trivial, especially as most applications were designed and tested using data for organisms with high-quality reference genomes. Assembly, annotation, variant discovery,

and homolog identification are challenging propositions in any genomics study [17, 254]; doing the same *de novo* for non-model organisms adds an additional layer of complexity. Already, many studies have collected HTS data sets for organisms of evolutionary and ecological interest [153, 175, 98] and have developed associated pipelines. Some have published these pipelines to share with other researchers [57, 151, 75]; such programs make HTS more accessible to a wider audience and serve as an excellent launching pad for beginning data analysis. However, because each HTS data set likely poses its own challenges and idiosyncrasies, researchers must evaluate the efficacy and accuracy of any pipeline for their data sets before they are used for biological inference. Evaluating pipeline success is easier for model organisms, where reference genomes and single nucleotide polymorphism (SNP) sets are more common; however, for most non-model organisms, we often lack easy metrics for gauging pipeline efficacy.

In this study, I generate a large HTS data set for five individuals each from seven phylogeographic lineages in three species of Australian skinks (family: *Scincidae*; Fig. S2), for which the closest assembled genome (*Anolis carolinesis*) is highly divergent (most recent common ancestor [MRCA], 150 million years ago [Mya], [3]). These seven lineages are closely related; they shared a MRCA about 25 Mya [331]. This clade is the focus of a set of studies looking at introgression across lineage boundaries [327], and to set the foundation for this work, I generate and analyze transcriptomic data for lineages meeting in four of these contact zones, two of which are between sister-lineages exhibiting deep divergence (*Carlia rubrigularis* N/S, *Lampropholis coggeri* C/S) and two which show shallow divergence (*Saproscincus basiliscus* C/S, *Lampropholis coggeri* N/C) (Fig. S2). I use these data to develop a bioinformatics pipeline to assemble and annotate contigs, and then, to define variants within and between lineages and identify homologs between lineages. Using both novel and existing metrics and an externally validated SNP data set, I am able to test the effectiveness of this pipeline across all seven lineages. In doing so, I refine my pipeline, identify remaining challenges, and evaluate the consequences of these challenges for downstream inferences. My work makes suggestions to other researchers conducting genomics research with non-model organisms, offers ideas on how to evaluate the efficacy of pipelines, and discusses how the technical aspects of HTS sequencing can affect biological inference.

6.3 Methods

All bioinformatic pipelines are available as Perl scripts on `https://github.com/singhal/transcriptomic`, and they are summarized graphically in Figs. 1A and S1. I have also shared R scripts [288] that use ggplot2 to do the statistical analyses and graphing presented in this paper [369].

6.3.1 Library Preparation and Sequencing

Even though costs of sequencing continue to drop and assembly methods improve [122, 313], whole-genome *de novo* sequencing remains inaccessible for researchers interested in organisms with large genomes (*i.e.*, over 500 Mb) and for researchers who wish to sample variation at the population level. Thus, most *de novo* sequencing projects must still use some form of complexity reduction (*i.e.*, target-based capture or restriction-based approaches) in order to interrogate a manageable portion of the genome. Here, I chose to sequence the transcriptome, because it is appropriately sized to ensure high coverage and successful *de novo* assembly, I will surely obtain homologous contigs across taxa, I can capture both functional and non-coding variation, and assembly can be validated by comparing to known protein-coding genes.

Liver and, where appropriate, testes samples were collected from adult male and female lizards during a field trip to Australia in fall 2010 (Table S1); tissues and specimens are accessioned at the Museum of Vertebrate Zoology, UC-Berkeley. I extracted total RNA from RNA-later preserved liver tissues using the Promega Total RNA SV Isolation kit. After checking RNA quality and quantity with a Bioanalyzer, I used the Illumina mRNA TruSeq kit to prepare individually barcoded cDNA libraries. Final libraries were quantified using qPCR, pooled at equimolar concentrations, and sequenced using four lanes of 100bp paired-end technology on the Illumina HiSeq2000.

6.3.2 Data Quality and Filtration

I evaluated raw data quality by using the FastQC v0.10.0 module [8] and in-house Perl scripts that calculate sequencing error rate. Sequencing error rates for Illumina reads have been reported to be as high as 1% [237]; such high rates can both lead to poor assembly quality and false positive calls for SNPs. To compare to these reported values, I derived an empirical estimate of sequencing error rate. To do so, I aligned a random subsample of overlapping forward-reverse reads (N=100,000) using the local aligner blat v34 [176], identified mismatches and gaps, and calculated error rates as the total number of errors divided by double the length of aligned regions. Data were then cleaned: exact duplicates due to PCR amplification were removed, low-complexity reads (e.g., reads that consisted of homopolymer tracts or more than 20% 'N's) were removed, reads were trimmed for adaptor sequence and for quality using a sliding window approach implemented in Trimmomatic v0.16 [205], reads matching contaminant sources (e.g., ribosomal RNA and human and bacterial sources) were removed via alignment to reference genomes with Bowtie2 v2.0.0-beta5 using default settings [188], and overlapping paired reads were merged using Flash v1.0.2 [212]. Following data filtration but prior to read merging, I again estimated sequencing error rates using the method described above.

6.3.3 *de novo* Assembly

Determining what kmer, or nucmer length, to use is key in *de novo* assembly of genomic data [93]. In assembling data with even coverage, researchers typically use just one kmer [93]; however, with transcriptome data, contigs have uneven coverage because of gene expression differences [222]. Thus, some have shown the ideal strategy for transcriptomes is to assemble data at multiple kmers and then assemble across the assemblies to reduce redundancy [346]. To assemble across assemblies, I first identify similar contigs using clustering algorithms (cd-hit-est v4.5.7; [201]) and local alignments (blat v34; [176]) and then assemble similar contigs using a light-weight *de novo* assembler (cap3; [156]). I used this custom multi-kmer approach along with other existing approaches, including:

- A single kmer approach implemented in the program Trinity r2012-01-25 (a *de novo* RNA transcript assembler, after which I used my clustering script [132]);
- A single kmer approach implemented in ABySS v1.3.2 (a *de novo* genomic assembler; [324]), Velvet v1.1 (a *de novo* genomic assembler; [389]), and SOAPdenovo-Trans v1.01 (a *de novo* RNA transcript assembler; [208]), which I implemented as a multi-kmer approach using my custom multi-kmer script
- A multi-kmer approach implemented in the program OASES v0.2 [316]

I explore a wide-range of assembly methods because generating a high-quality and complete assembly is key for almost all downstream applications. Particularly with genome assembly, which is both an art and a science, researchers should try multiple approaches and evaluate their efficacy before further analyses [93]. However, without a reference genome, evaluating the quality of a *de novo* assembly is challenging. Here, I implement novel metrics for evaluating *de novo* transcriptome assemblies. In addition to existing metrics in the literature (N50, mean contig length, total assembly length) [222], I determined which proportion of reads were used in the assembly, measured putative levels of chimerism in transcripts due to misassemblies, determined the proportion of assembled transcripts that could be annotated and the accuracy of these transcripts (as determined by the number of nonsense mutations, or premature stop codons), and calculated the completeness and contiguity of the assembly [222].

Here, I assembled across all individuals in a lineage rather than assembling each individual separately. Although this introduced additional polymorphism into the data which can reduce assembly efficiency [362], previous work suggests the additional data lead to more complete assemblies (Singhal, unpublished).

6.3.4 Annotation

Following evaluation of my final assemblies, I chose the best assembly (here, Trinity-generated assemblies) for annotation to protein databases. Determining the most appropriate database for annotation is important, so I tested multiple options, including using

a single-species database, whether from a distantly-related but well-annotated genome or closely-related but poorly-annotated genome, using a multi-species database, or using a curated protein set, such as UniRef90 [347]. For one randomly selected lineage, I tested the efficiency and accuracy of five different reference databases:

- the non-redundant Ensembl protein database [107] for the lizard *Anolis carolinensis*; with a most-recent common ancestor to my lineages of about ≈150 mya, it is the closest available genome [3],
- the non-redundant Ensembl protein data set for *Gallus gallus*, whose genome is higher quality than the *Anolis* genome but is more distantly related (≈250 mya),
- a non-redundant, curated data set (UniRef90) of proteins from a wide range of organisms, whose genes have been clustered at 90% similarity,
- a highly-redundant Ensembl protein data set for eight vertebrates sequenced to high quality (human, dog, rat, mouse, platypus, opossum, dog, chicken),
- a highly-redundant Ensembl protein data set for the 54 vertebrates whose genomes have been annotated.

I evaluated the number of matching contigs, and for the non-redundant data sets, the number of uniquely matching contigs. Distinguishing between contigs that match and contigs that match uniquely is important, as despite my clustering during assembly, many contigs in the assembly appear redundant. These highly similar contigs likely result from misassemblies, allelic variants, alternative splicing isoforms, or recently duplicated paralogs. Parsing these categories is challenging without a reference genome and when expected coverage across contigs is uneven. Especially for projects interested in functional genomics, annotation of redundant contigs remains an important and unresolved issue. Here, I try to mitigate these errors by using reciprocal BLAST best matching to annotate contigs and selecting the best match. In doing so, I likely failed to annotate recently evolved paralogs, but I should not have multiple copies of the same gene in my downstream analyses.

Once I determined the best database both with respect to efficacy and efficiency, I used a custom script to annotate the contigs using a reciprocal best-match strategy via BLASTx v2.2.24 and tBLASTx with an e-value cutoff of 1e-20 [5] and defined the untranslated regions and coding sequence of the transcript using Exonerate v2.1 [332]. Further, initial tests of the annotation pipeline uncovered two challenges: first, many contigs were chimeric and consisted of multiple, combined transcripts, and second, many of the predicted open reading frames (ORFs) had nonsense mutations, largely due to frameshift mutations. To correct for chimeric contigs, I identified contigs that had two or more non-overlapping and high-quality matches to different genes using BLASTx and split these contigs accordingly. Further, I used the program FrameDP v1.2 to identify and correct for frameshift mutations [131].

Finally, I searched unannotated contigs against the NCBI 'nr' database using BLASTn to determine these contigs' identity. As described in the Results, these unannotated contigs largely went unidentified. Thus, although some of these unannotated transcripts have viable open reading frames and/or had homologs in other lineages, and therefore, might be genes, I will be conservative and only use annotated transcripts in all downstream analyses.

Finally, to describe the putative biological functions of my annotated contigs, I determined gene ontology using Blast2Go [64].

6.3.5 Alignment

The first step in identifying variants or estimating gene expression levels is to align the sequencing reads to one's reference genome. Here, I use my annotated transcripts as a pseudo-reference genome [370], thus aligning the reads used to generate the assembly to the assembly itself. Here, I tested seven different aligners (bowtie v0.12.7, bowtie2 v2.0.0-beta5, bwa v0.6.1, novoalign v2.07.07, smalt v0.5.8, SOAPaligner v2.21, stampy v1.0.14; [189, 188, 197, 207, 200]) to determine their efficacy and accuracy. These programs run the gamut of being fast but less sensitive to being slower and more sensitive. Here, sensitivity is defined as the aligner's ability to align reads with multiple mismatches. Previous results have shown [196] that alignment error is a common cause of miscalled SNPs, particularly alignment errors around indel sites. To evaluate these programs, I inferred genotypes from the alignments with SAMtools v0.1.18 [198]. I then compared these genotypes to a small data set of known genotypes from one of the contact zones, *C. rubrigularis* N/S. In another study, I had Sanger sequenced 200-400 bp of sequence from 10 to 15 genes for the same individuals sequenced here (Singhal, unpublished). Importantly, all these genes were represented at high coverage ($\geq$20$\times$) in this data set; thus, coverage is sufficiently great to ensure accurate genotype calling [255]. I used these validated genotypes to determine the number of false positives (or variation called at a non-polymorphic site) and negatives (or variation not called at a polymorphic site) in my inferred genotypes. Further, I evaluated these programs based on the proportion of reads and read pairs they aligned and the concordance of SNP calls across data sets.

6.3.6 Variant discovery

Two major types of variant discovery are SNP identification and genotype calling. Many researchers are interested only in identifying SNPs, or determining which nucleotide positions are variable in a sample of individuals. SNP-containing regions are then resequenced or genotyped for further analysis [370]. Increasingly, researchers are both identifying variable sites, and then, summarizing variation at these sites using the site frequency spectrum (SFS) or calling genotype likelihoods for each individual for subsequent population genomics analyses. SNP identification has become an easier exercise as sequencing costs dropped and coverage has increased. However, genotype calling remains

a challenging proposition, particularly in diploid and polyploid individuals, as distinguishing heterozygosity, homozygosity, and sequencing errors at variable sites is difficult unless there is high coverage ($\geq$20$\times$, [255]). Thus, I focus on genotype calling and its use in characterizing variation for population genomics analyses. Importantly, I assume in my approach and discussion that both alleles are expressed in each individual. Although there are some data to suggest that expression can be allele-biased, properly controlling and testing for this issue requires having previously identified variants or genomic data [330].

My results indicated that Bowtie2 was the most effective and efficient aligner (see *Results*); thus, I used it for all downstream analyses. When identifying variants from alignment data, there are several approaches:

1. brute strength methods, in which the read counts for given alleles at a site are calculated, and variants are determined by an arbitrary cut-off [383]

2. maximum likelihood (VarScan v2.2) and Bayesian methods (SAMtools v0.1.18) [181, 198], in which algorithms consider strand bias, alignment quality, base quality, and depth to call genotype likelihoods for individuals. These methods have been developed further to account for Hardy-Weinberg disequilibrium and linkage disequilibrium in calling and filtering variants [198, 77], to use machine learning with a set of validated SNPs to improve algorithms [77], and to re-align reads near indel areas to ensure inaccurate alignments do not lead to false SNPs.

3. Bayesian methods (ANGSD v0.3) which infer the site frequency spectrum for all the variants in the data set, which is, in turn, used as a prior to estimate genotype likelihoods for individuals [255]. This method is particularly useful for data sets with large population samples.

Here, I test these three general types of SNP and genotype discovery, using read counting, VarScan, SAMtools, and ANGSD in two sister lineage-pairs for which I have validated genotypes (*C. rubrigularis* N/S and *L. coggeri* N/C). I both looked at concordance of SNP and genotype calls across methods and calculated the number of false positives and negatives.

6.3.7 Homolog discovery

Homologs between lineages must be identified for any comparative genomics analyses. In this study, my lineages are all closely-related, so homology identification is less challenging than in many other comparative studies. However, ensuring I am identifying orthologs across lineages and not paralogs is challenging, particularly as my annotation pipeline could not conclusively distinguish orthologs and paralogs in the absence of a reference genome. With that caveat, I test three different methods for identifying homology:

1. defining homologs by their annotation; *i.e.*, contigs that share the same annotation are assumed to be homologs,

2. defining homologs by reciprocal best-hit BLAST, as is most commonly done in other studies [242],

3. the SNP method, or defining homologs by mapping reads from one lineage to the other lineages' assembly, identifying variants, and thus determining homologous sequence.

I evaluated these methods by the number of homologs found, the percent of aligned sequence between homologs, and the raw number of differences between homologous sequence. I looked at homology discovery both between sister lineages and non-sister lineages, as I expect discovery across non-sister lineages will be harder.

6.3.8 Biological inference

Finally, I determined how robust biological inference is to the analysis method used. First, to determine how genotype calling affects downstream inference, I inferred the site frequency spectrum and associated summary statistics (Tajima's D, θ, π) for one lineage across different genotype calling methods and different coverage levels using dadi 1.6.2 [135]. Second, to determine how homology identification affects downstream inference; I determined *dN/dS* ratios using PAML 4.4 [385] and raw sequence divergence for each gene across different methods of homology.

6.4 Results

6.4.1 Data Quality and Filtration

Library preparation and sequencing were successful for all individuals. On average, I generated 3.5 ± 0.5 Gb per individual. Duplication rates, low-complexity sequences, and contamination levels were low (Table S2). However, aggressive filtering and merging significantly reduced the raw data set; I lost $27.1 \pm 3.8\%$ of raw base pairs per individual. As seen in Figure S3, this strategy significantly improved the per-base quality of my data. Indeed, I was able to reduce sequencing error rates in my final data set five-fold (initial error rates: $0.3 \pm 0.1\%$, final error rates: $0.06 \pm 0.01\%$). These error rates are likely overestimates, because I used the lower-quality portion of the read (the tail end) to identify sequencing errors. Despite this reduction in error rates, profiling of mismatches across the reads showed that both the head and tail of the read still harbor a higher number of mismatches compared to the rest of the read. This pattern persisted even when the first and last five base pairs of each read were trimmed prior to alignment (Fig. S4). Possibly, as others have found residual adaptor sequence in their data sets despite using rigorous

adaptor trimming (Bi, unpublished), these heightened error rates could be due to adaptor sequences leading to misalignments and spurious SNPs.

6.4.2 *de novo* Assembly

To assemble my data, I tested five different programs, which employed different strategies (e.g., single k-mer, built-in multi-kmer approach, my custom multi-kmer approach). I evaluated the assemblies on many metrics; here, I show data for four of these metrics. With respect to the percentage of paired reads that aligned to the assembly, SOAPdenovo and Trinity performed far better than the rest of the assemblers (Fig. 2A), suggesting their assemblies were more contiguous. The same two assemblers and Velvet also recovered the greatest number of annotated transcripts, measured here by the number of core eukaryotic genes found in these assemblies (CEGMA; [269]; Fig. 2B). OASES and Trinity appeared to be the most accurate, as they contained the fewest number of nonsense mutations in annotated ORFs (Fig. 2C). Finally, OASES, Trinity and SOAPdenovo assemblies had the fewest number of putative chimeric transcripts (Fig. 2D). Looking across all these metrics, Trinity emerges as the best assembler. Further, Trinity did a good job assembling most of the data; on average, just $8.1 \pm 4.3\%$ of contigs from other assemblies were unique to that assembly compared to Trinity. As such, I used Trinity assemblies for all downstream analyses. As seen in Table 1, the basic metrics of these assemblies (e.g., number of contigs, total length of assembly, and N50) were fairly constant across all lineages. Unlike other studies [63], I find no correlation between contig length and coverage, suggesting my assembly is not data-limited (Fig. S5). I do find a weak but significant negative correlation between polymorphism levels and contig length ($r^2 = -0.169$, $p - value < 0.05$; Fig. S6), suggesting that, for more variable contigs, combining across individuals negatively impacts assembly contiguity.

6.4.3 Annotation

After assembling the data, I annotated the assemblies in order to identify uniquely annotated contigs for downstream analyses and to refine the assemblies further. First, because my focal lineages are evolutionarily distant from the nearest genome (MRCA ≈150 mya to *Anolis carolinensis*), I wanted to test the efficacy of different databases to annotate my contigs. While more complete databases did lead more annotated contigs (Table S3), the increase was marginal. Further, larger databases consume significantly more computing time; here, annotating to the UniProt90 database took nearly 100 times the processor hours as annotating to *A. carolinensis*. Thus, I used the *A. carolinensis* database for all further annotations. Importantly, I could annotate these genomes to more distant relatives (*G. gallus* and *T. guttata*; MRCA ≈300 mya), without seeing a significant decrease in annotation success (Table S3). This result suggests such an annotation approach could work for organisms in even more genomically depauperate clades.

While annotating contigs, I identified a low percentage of chimeric contigs (≈4%), which I resolved by splitting these contigs into individual genes (Table S4). Inspecting alignments of sequencing reads to these chimeric contigs suggested that these contigs form during assembly and not due to technical errors during library preparation, as chimeric junctions generally had significantly reduced coverage. Further, a small portion of the predicted open reading frames (ORFs) of annotated contigs (≈3%) had premature stop codons. Although it is possible that these ORFs are pseudogenes [172], it seems more likely that they are due to assembly errors, as these contigs were generally highly expressed. Using FrameDP, I was able to identify and fix many of these likely frameshift errors (Table S4).

Through this pipeline, I annotated an average of 23360 contigs per lineage, of which, which matched to an average of 11366 unique genes in the *A. carolinensis* genome (Table 1). I also recovered the full coding sequence for many genes; 67% of unique annotated contigs encompassed the entire coding sequence for a gene, including portions of the 5′ and 3′ UTRs. These numbers appear reasonable – the annotation for the *A. carolinensis* genome currently includes 19K proteins, and liver tissue does not express all genes at a sufficiently high level to be represented here [291]. These genes contribute to a diversity of biological processes and serve a wide range of molecular functions, suggesting I assayed a varied portion of the transcriptome (Fig. S7).

Further, my pipeline appears to be robust; almost all unannotated contigs failed to find a good match in the NCBI 'nr' database (Fig. S8). Approximately 9% of unannotated contigs matched to genes; however, further analysis of these matches showed that almost all of them matched with such low-quality to prevent annotation.

Additionally, by annotating contigs rigorously to limit the number of putative duplicate contigs, I significantly reduced the redundancy of my data set. When I aligned sequencing reads to my initial, unannotated assembly, I found that ≈10% of mapped reads aligned to multiple places in the assembly. Some of these multiple alignments might be because of biological redundancy – perhaps these reads are aligning across recently duplicated genes or across common motifs in genes – but it is likely a good portion of them are aligning multiply because the initial assembly had many redundant contigs (≈50% of annotated contigs were not unique). After annotating the genome and removing redundant contigs, I reduced the percentage of mapped reads aligning non-uniquely to ≈2%. However, removing redundant contigs also lead to an average 8% decline in overall mapping efficiency. Thus, it seems likely these redundant contigs are "biologically real", but we do not yet have the tools to parse such contigs properly [361].

6.4.4 Alignment

Identifying variants and quantifying gene expression first require that sequencing reads are aligned to the reference genome. Here, I tested the efficacy of seven different alignment programs, which employ different algorithms over a range of sensitivity and speed. I evaluated these programs in three ways. First, I used my externally validated set of

genotypes to see how many genotypes were inferred correctly. Almost all of the aligners performed well and led to the correct genotype at $\geq$90% of the sites. Although the false negative rate was moderately high ($\approx$5% for most aligners), the false positive rate was low (Table 2). Bowtie2 clearly outperformed the rest of the aligners and was thus used for all downstream analyses. Second, I evaluated how many read pairs and reads the programs could align. Although Novoalign, smalt and stampy are generally considered to be more sensitive aligners, I found little variation in the percentage of reads aligned across programs (Fig. 3). Bowtie2 and stampy were able to align the most paired reads, which is useful as aligning paired reads reduces the likelihood of errant matches and non-unique matches [20]. Finally, I looked at overlap in SNPs inferred across programs. Problematically, although all programs were fed the same reference genome and sequencing reads, I saw only moderate overlap – on average, only 77$\pm$9% of SNPs were shared. Checking the raw alignments suggested these discrepancies often arose from differences in alignment rather than differences in SNP inference post-alignment. These results suggest that alignment is likely a major source of error in *de novo* HTS analyses, as has been suggested by other studies [196, 202, 179]. Further, although the common set of SNPs found across these programs is likely to be high-quality, considering only these SNPs is likely to lead to many false negatives. That said, when the same SNPs were called across programs, genotype inference was highly concordant; 94$\pm$2% of genotype calls were the same across alignment methods, and inferred allele frequency at these SNPs was highly correlated (r=0.94$\pm$0.01).

6.4.5 Variant Discovery

After alignment, programs for variant inference are used to call SNPs and genotypes. In the previous tests, I used the variant discovery program SAMtools for all analyses; here, I test a few approaches: a brute strength approach, in which I call SNPs and genotypes based solely on count data, two probabilistic methods (SAMtools and VarScan), and a probabilistic method that uses the allele frequency spectrum (ANGSD). I first assessed accuracy of genotype calls by using my externally validated genotype set. In general, I found that all methods performed fairly well – particularly, when a SNP was identified, all programs inferred the correct genotype with high accuracy ($\geq$98%; Table 3). However, the count method of identifying variation led to many false positives, an unsurprising result given its failure to account for sequence error or alignment score. ANGSD had a high false negative rate, the reason for which is unclear, though is possibly due to the small sample sizes used here. But, as shown by other work, ANGSD is best suited for correctly inferring the shape of the site frequency spectrum [255]. Comparing across all SNPs found across all programs, I found that concordance across all SNPs was moderate, similar to my comparative alignment results. On average, only 83% of SNP calls are shared across programs; this lack of concordance was largely driven by SNPs inferred from count data. More promisingly, when a site is inferred as a SNP, 98% of the genotype calls are shared

across programs. Overall, these results suggested SAMtools performed the best, so I used it for all downstream analyses.

Upon defining SNPs and then genotypes for each individual, I explored how different variant discovery methods affect biological inference by constructing the SFS. Despite the only moderate levels of concordance in SNP calls, I find that the SFS is nearly identical across all the different approaches but VarScan (Fig. 4). Importantly, this result only holds true when I restrict analysis to higher-coverage contigs ($\geq 10\times$); low-coverage contigs show aberrant patterns. Although the SFS is similar across all approaches, estimates of key population genetic summary statistics (*i.e.*, θ_w, π) vary depending on the approach – an unsurprising result given that the total number of SNPs inferred differs across approaches. Thus, prior to using these data for population genetic analyses, ascertainment bias must be factored into any downstream inference [251]. Finally, to look at these SNPs in greater detail, I annotated the SNPs I found in two sister-lineages, with respect to how they are segregating, their location relative to the gene, and their coding type (Fig. S10). Not only are the patterns of polymorphism and non-synonymous/synonymous mutations reasonable [33], but there are many types of variants (*i.e.*, coding vs. non-coding, non-synonymous vs. synonymous, fixed vs. polymorphic), which will allow the data to be used to identify adaptive signatures of molecular evolution, infer demographic history, and develop markers.

6.4.6 Homolog discovery

To identify homologs between lineages, I tested three different methods and then evaluated their effectiveness. All three methods performed well, identifying more than 8000 homologous pairs between lineages within-genera and between-genera for a significant portion of the contig length (Fig. 5). However, with the SNP method for homology, alignment efficiency dropped off significantly in between-genera comparisons, leading to identified homologs being shorter. I chose to use reciprocal BLAST matching to identify homologs for all downstream analyses as it was able to identify more homologs than the two other methods and it worked well across evolutionary distances (Fig. 5). This approach identified 8800 homologous contigs across all seven lineages for use in comparative analyses.

Estimation of the summary statistics (sequence divergence and dN/dS ratios between homologs from lineage-pairs) is affected by how homologs are defined (Fig. S11). Defining homologs via annotation or via reciprocal BLAST matching gives very similar results for both sequence divergence and dN/dS. However, using SNPs to reconstruct the homolog results in a fuzzier pattern. When I restrict the analysis to homologs with higher coverage ($>10\times$) for which there is greater confidence in SNP inference (see *Results: Variant Discovery*), all three methods are highly correlated. Thus, this method for homolog identification should account for differences in coverage, where appropriate.

6.5 Discussion

In creating and implementing a pipeline for high-throughput sequence data, I noted several possible sources of error (Fig. 1B):

1. Errors introduced during library preparation, which can include human contamination, errors introduced during PCR amplification of the library, and contamination between samples
2. Errors introduced during sequencing, the frequency and type of which are dependent on the chemistry of sequencing platform, and subsequent de-multiplexing
3. Errors introduced during assembly [17], such as misassembly of reads to create chimeric contigs
4. Errors due to misalignment of reads to assembly during variant discovery, particularly caused by indels in alignments and reads that map to multiple locations
5. Errors in SNP and genotype calling, such as not sampling both alleles and thus mistakenly calling a homozygote

To this, I add two additional sources of uncertainty that every study in evolutionary genomics faces – have contigs been annotated correctly and have orthologs between compared genomes been identified correctly [59]? Errors can arise at any stage in the process; such errors percolate through subsequent steps, likely affecting all downstream inference [361, 202, 179]. Whether using their own pipeline or a pre-existing pipeline, researchers will want to incorporate some of the checks suggested here to ensure that the pipeline is working well for their data and that incidence of errors is low. Moving forward, the questions become how to limit these errors and how to mitigate their effects.

All these sources of error are non-trivial, but with careful data checking and willingness to discard low-quality data, it is possible to mitigate the effects of these errors. First, as has now become standard, scrubbing reads for low-quality bases and adaptors is a must – as shown here, read cleaning can reduce error rates noticeably. When possible, merging reads from paired-end reads can further decrease error rates and will lead to more accurate estimates of coverage for expression studies [212]. Second, having a high-quality assembly is crucial both for accurate annotation and variant discovery. Inferring the quality of *de novo* assemblies is challenging, as there are no clear metrics or comparisons to use [222]. However, I propose a few metrics, which can be used with transcriptome data – primarily, looking for assemblies that minimize chimerism and non-sense mutations, that are contiguous, and that capture a significant portion of known key genes. Undoubtably, errors remain in the final assemblies, but these metrics helped me select the most accurate assembly for downstream analyses. Additionally, contig redundancy in final assemblies remains a pressing challenge. By using a strict reciprocal-BLAST annotation strategy, I removed many of these apparently redundant contigs. However, this approach certainly removed some biologically real contigs that were recent duplicates and

alternative splicing isoforms of interest to those interested in expression differences between biological groups [361]. Researchers should continue to explore better methods to identify orthologs and paralogs. Until better methods are developed, using strict criteria for identifying non-redundant gene sets is a must, as most biological inference programs assume that each contig offered for analysis is a unique evolutionary unit.

Alignment and variant discovery remain notable challenges. In part, a poor-quality assembly genome truly can affect variant discovery – alignments across misassemblies can led to errant SNP calls, particularly when misassemblies introduce indels [196]. Further, unless some sort of redundancy reduction is used, many contigs will be nearly identical in an assembly, leading to a high rate of non-unique alignments and miscalled SNPs. I was able to remove most redundant contigs, and thus, I reduced the proportion of non-unique alignments. I still see evidence for errors in alignment as (1) discrepancies between our externally-validated SNP set and genotype calls from these alignments and (2) the only moderate level of congruence between different approaches fueled by the same data. The same patterns hold for SNP inference after alignment. Some of these errors are likely driven by the quality of the assembly – by removing alternative splicing isoforms and recently duplicated genes, some of the reads likely misaligned to retained contigs although they were derived from another, rejected contig. However, many of these errors disappear at higher coverage, thus, given these data, the best approach is to rely on contigs with higher coverage – 10 to 20$\times$, at least – and to account for this ascertainment bias in any biological inference. Importantly, however, by relying on high-coverage contigs in transcriptome analyses one is biased to more slowly-evolving genes, as there is a strong negative correlation between expression levels and rate of molecular evolution [89].

Further, to ensure the vagaries of variant discovery do not unduly influence our biological inference, we should use the genotype likelihoods and not genotype calls for downstream work. Ideally, researchers would conduct subsequent inference that use the SFS or genotype likelihoods as input, such as BAMOVA [125] or dadi [135], thus ensuring uncertainty in SNP and genotype calling is incorporated into model fitting. However, many analyses, particularly those used by most biodiversity researchers (*i.e.*, coalescent-based demography and phylogeny programs), require known genotypes or haplotypes. Until uncertainty is incorporated into such programs, researchers will have to arbitrarily chose cutoffs to determine most likely genotypes. In such cases, researchers might want to restrict their analyses to regions with high coverage, where calls are likely more certain [255].

Moving forward, how can we reduce the sources of errors stemming from alignment errors and genotype inference? Improved assemblies, facilitated by new long-read sequencing technologies, will certainly help. As researchers collect externally validated SNP data sets, they can use programs like GATK to recalibrate variant calling and to realign around indels [77]. Researchers will also increasingly sequence more individuals in a population, which will better take advantage of multi-sample methods like SAMtools and ANGSD [198, 255]. Finally, programs like Cortex, which assemble across individuals to provide both a reference assembly and individual assemblies, are promising [161]. Simu-

lations suggest that this method can also better handle data with indel polymorphism.

Finally, homolog discovery is a challenge in any genome project [59], and this project was no exception. All three methods I tested for homolog discovery worked well, but I recommend only using a SNP-based approach between lineages that are closely-related and for contigs with high coverage. Moving forward, as we acquire more comparative genomic data across the tree of life, homolog discovery should become an easier problem, as fueled by comparative clustering programs like OrthoMCL [60].

Given this, other researchers should carefully consider the benefits and challenges of working with transcriptomic data before embarking on a similar studies. For researchers interested in obtaining variation data for non-model organisms and who do not require expression data, they might consider using restriction-based methods like RADtags or reduced-representation libraries [153] or collecting target-based capture data [39]. These methods are cheaper than transcriptome methods, and they do not require that genetic samples have been preserved to maintain RNA quality. However, finding homologous contigs across phylogenetic depths can be challenging, and such contigs typically cannot be annotated. Target-based capture methods can be used with low-quality DNA and have the same benefits of transcriptome data (*i.e.*, homologous contigs can be identified across phylogenetic depths and contigs can be annotated) without its disadvantages (*i.e.*, coverage is expected to even across contigs and redundancy in assemblies can be more easily handled) [39]. However, exome-capture is more expensive than restriction-based methods and designing probes requires previously acquired genomic data. Thus, determining which approach is ideal for a given study depends on the number and quality of samples to be assayed, the amount of money available, and the phylogenetic span of the samples.

Despite the challenges of HTS data and transcriptome data, through this work I collated a large data set of over 12K annotated contigs, spanning a wide-range of biological functions, and over 100K SNPs between lineage-pairs, spanning a wide-range of locations and coding types. Notably, I was able to do all of these analyses using existing, open-source software and, but for assembly, by using a low-end desktop machine. Genomic analyses are not just for those working with humans or mice anymore. With careful and thoughtful data curation, HTS can enable researchers to use genomic approaches to explore all the branches in the tree of life.

6.6 Acknowledgements

I gratefully acknowledge M. Chung, J. Penalba, and L. Smith for technical support, and the Seqanswers.com community for providing timely and thoughtful advice. Three anonymous reviewers, R. Bell, K. Bi, J. Bragg, C.A. Buerkle, T. Linderoth, M. MacManes, C. Moritz, R. Nielsen, S. Ramirez, F. Zapata, and members of the Moritz Lab Group provided comments and suggestions during this work and on this manuscript that greatly improved its quality. Financial support for this work was provided by National Sci-

ence Foundation (Graduate Research Fellowship and Doctoral Dissertation Improvement Grant), the Museum of Vertebrate Zoology Wolff Fund, and a Rosemary Grant Award from the Society of the Study of Evolution. This work was made possible by the supercomputing resources provided by NSF XSEDE, in particular the clusters at Texas Advanced Computing Center and Pittsburgh Supercomputing Center.

6.7 Data Accessibility

Data are available at the following locations:

1. Specimens and tissues used in this study are accessioned at the Museum of Vertebrate Zoology, UC-Berkeley, catalog numbers 269023-269105
2. Original Illumina reads are available on SRA, entry SRA062739 under BioProject PRJNA183544
3. Final assemblies are available on DRYAD, entry `doi:10.5061/dryad.7c99f`
4. Scripts used are available on DRYAD, entry `doi:10.5061/dryad.7c99f` and at `https://github.com/singhal/transcriptomic`

6.8 Figures

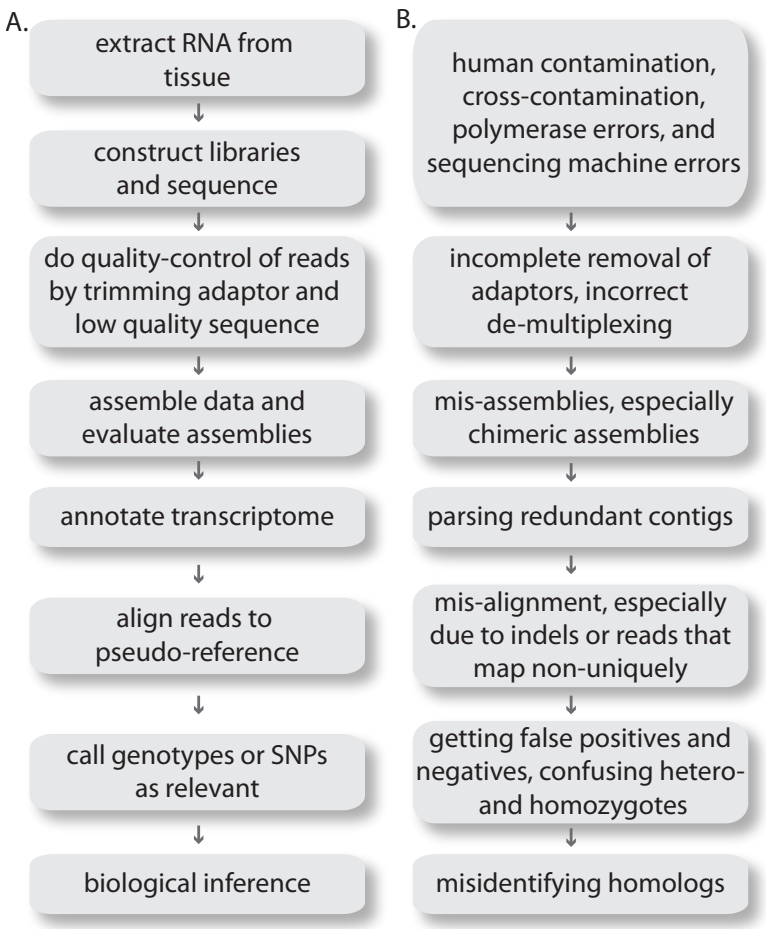

Figure 6.1: A. Pipeline for handling transcriptome data for *de novo* population genomic analyses, as presented in this study. B. Errors introduced at each stage in the pipeline.

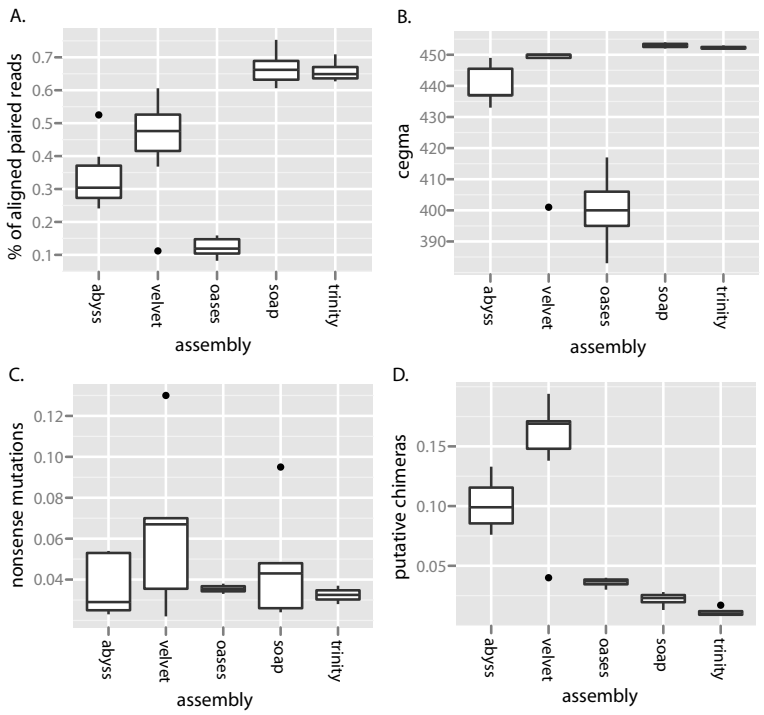

Figure 6.2: Evaluation of assemblies across the seven sequenced lineages according to A. percentage of paired reads that aligned to reference, B. number of CEGMA genes that are found in assembly, C. percentage of annotated coding sequences that had nonsense mutations, and D. percentage of contigs that were putative chimeras.

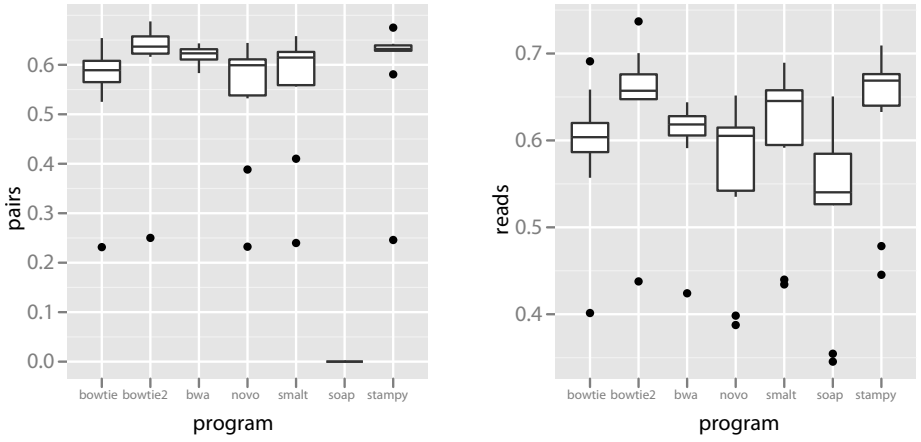

Figure 6.3: Evaluation of different alignment software across three randomly selected lineages with respect to two metrics, A. number of paired reads aligned and B. number of reads aligned.

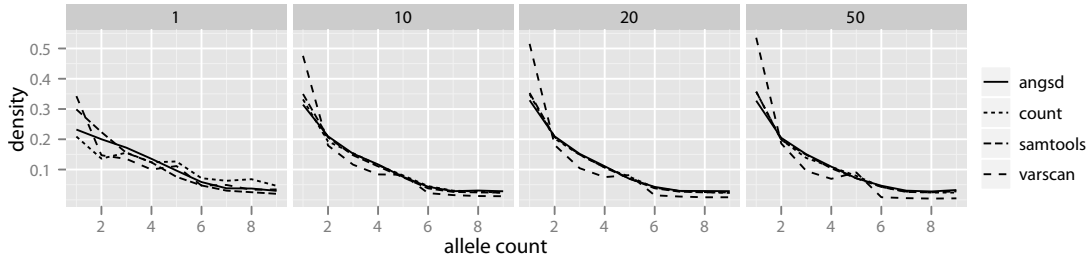

Figure 6.4: Unfolded allele frequency spectrum for variants within a randomly selected lineage for sites represented at 1x, 10x, 20x, and 50x coverage per individual, across different methods for genotype inference.

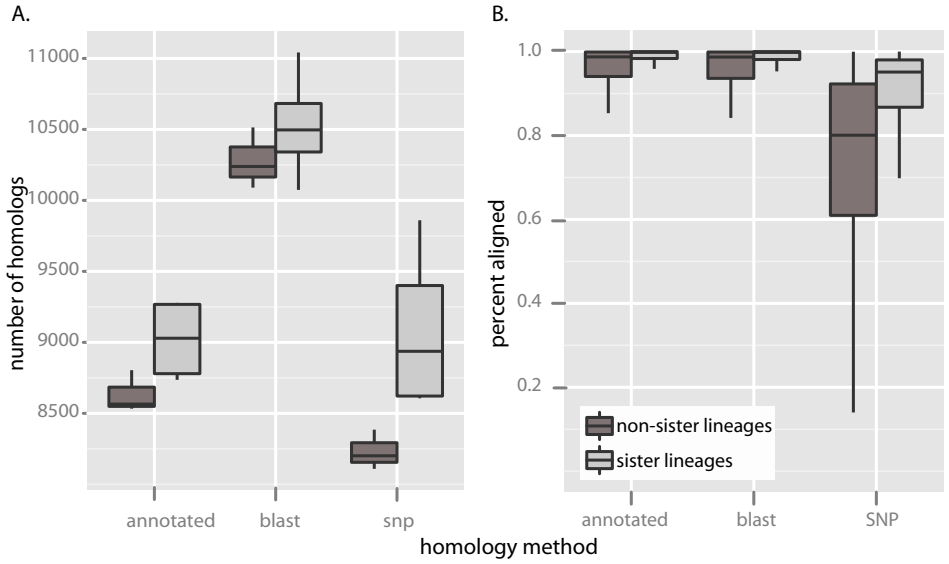

Figure 6.5: Summary of different methods for homolog discovery between all lineage comparisons of interest, considering A. number of homologs for which 75% of sequence was aligned and B. percent of homolog aligned.

6.9 Tables

assembly	number contigs	total length	n50	annotated contigs	annotated contigs (unique)	complete annotated contigs
C. rubrigularis, N	104648	89.1e6	1806	25198	12063	8179
C. rubrigularis, S	98280	84.3e6	1780	24323	11558	7697
L. coggeri, N	96798	87.5e6	1972	22760	11457	7344
L. coggeri, C	106937	92.7e6	1845	23852	10894	7796
L. coggeri, S	112935	89.6e6	1549	23774	11029	7258
S. basiliscus, C	84756	77.7e6	1951	21584	11221	7586
S. basiliscus, S	98685	83.5e6	1749	22031	11340	7696

Table 6.1: Summary of assemblies and their annotation. Complete annotated contigs are those with some 5′ and 3′ UTR sequence, as well as the full coding sequence.

genotype	bowtie	bowtie2	bwa	novoalign	smalt	SOAPaligner	stampy
right genotype	379 (89.8%)	419 (99.2%)	381 (90.3%)	383 (90.8%)	393 (93.1%)	207 (49.0%)	391 (92.7%)
wrong genotype	29 (6.9%)	3 (0.7%)	7 (1.7%)	9 (2.1%)	6 (1.4%)	52 (12.3%)	8 (1.9%)
false negative	12 (2.8%)	0 (0%)	34 (8.1%)	30 (7.1%)	23 (5.5%)	163 (38.6%)	23 (5.5%)
false positive	3	1	1	1	1	1	5

Table 6.2: Accuracy of genotype inference following the use of different programs for alignment; all genotypes were inferred using samtools post-alignment. Parenthetical percentages show the relative proportions of genotype types.

Genotype	ANGSD	count data	SAMtools	VarScan
right genotype	520 (68.4%)	745 (98.0%)	750 (98.7%)	745 (98.0%)
wrong genotype	3 (0.3%)	15 (2.0%)	10 (1.3%)	15 (2.0%)
false negative	230 (30.2%)	0 (0%)	0 (0%)	0 (0%)
false positive	6	134	1	12

Table 6.3: Accuracy of genotype inference across different programs for genotype inference; for all, Bowtie2 was used for alignment. Parenthetical percentages show the relative proportions of genotype types.

Bibliography

[1] A.F. Agrawal, J.L. Feder, and P. Nosil. "Ecological divergence and the origins of intrinsic postmating isolation with gene flow". In: *International Journal of Ecology* 2011 (2011).

[2] J. Alexandrino et al. "Strong selection against hybrids at a hybrid zone in the *Ensatina* ring species complex and its evolutionary implications". In: *Evol.* 59 (2005), pp. 1334–1347.

[3] J. Alföldi et al. "The genome of the green anole lizard and a comparative analysis with birds and mammals". In: *Nature* 477 (2011), pp. 587–591.

[4] S.M. Aljanabi and I. Martinez. "Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques". In: *Nucl. Acids Res.* 25 (1997), pp. 4692–4693.

[5] S.F. Altschul et al. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". In: *Nucleic Acids Research* 25 (1997), pp. 3389–3402.

[6] E. Anderson. *Introgressive hybridization*. New York: Wiley, 1949.

[7] E.C. Anderson and E.A. Thompson. "A model-based method for identifying species hybrids by using multilocus genetic data". In: *Genetics* 160 (2002), pp. 1217–1229.

[8] S. Andrews. *FastQC*. 2012. URL: `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`.

[9] E. Arevalo, S.K. Davis, and J.W. Sites. "Mitochondrial DNA sequence divergence and phylogenetic relationships among eight chromosome races of the *Sceloporus grammicus* complex (*Phyrnosomatidae*) in Central Mexico". In: *Syst. Biol.* 43 (1994), pp. 387–418.

[10] M.L. Arnold. *Natural hybridization and evolution*. Oxford, UK: Oxford University Press, 1997.

[11] J.C. Avise. *Phylogeography: the history and formation of species*. Cambridge, MA: Harvard University, 2000.

[12] J.C. Avise. "Systematic value of electrophoretic data". In: *Systematic Biology* 23.4 (1974), pp. 465–481.

[13] J.C. Avise and D. Walker. "Species realities and numbers in sexual vertebrates: perspectives from an asexually transmitted genome". In: *Proceedings of the National Academy of Sciences* 96.3 (1999), pp. 992–995.

[14] J.C. Avise, D. Walker, and G.C. Johns. "Speciation durations and Pleistocene effects on vertebrate phylogeography". In: *Proc. Roy. Soc. B* 265 (1998), pp. 1707–1712.

[15] J.C. Avise and K. Wollenberg. "Phylogenetics and the origin of species". In: *PNAS* 94 (1997), pp. 7748–7755.

[16] S.J.E. Baird. "A simulation study of multilocus clines". In: *Evol.* 49 (1995), pp. 1038–1045.

[17] M. Baker. "*de novo* genome assembly: what every biologist should know". In: *Nature Methods* 9 (2012), pp. 333–337.

[18] J.W. Ballard and M.C. Whitlock. "The incomplete natural history of mitochondria". In: *Mol. Ecol.* 13 (2004), pp. 729–744.

[19] C. Bank, R. Bürger, and J. Hermisson. "The Limits to Parapatric Speciation: Dobzhansky–Muller Incompatibilities in a Continent–Island Model". In: *Genetics* 191.3 (2012), pp. 845–863.

[20] S. Bao et al. "Evaluation of next-generation sequencing software in mapping and assembly". In: *Journal of Human Genetics* 56 (2011), pp. 406–414.

[21] H.S. Barber and F.A. McDermott. *North American fireflies of the genus Photuris*. Vol. 117. Smithsonian Institution, 1951.

[22] N. Barton and B.O. Bengtsson. "The barrier to genetic exchange between hybridizing populations". In: *Heredity* 57 (1986), pp. 357–376.

[23] N.H. Barton. "Estimating multilocus linkage disequilibria". In: *Heredity* 84 (2000), pp. 373–389.

[24] N.H. Barton. "Gene flow past a cline". In: *Heredity* 43 (1979), pp. 333–339.

[25] N.H. Barton. "Multilocus clines". In: *Evol.* 37 (1983), pp. 454–471.

[26] N.H. Barton and S.J.E. Baird. *Analyse 1.10*. 1998. URL: `http://helios.bto.ed.ac.uk/evolgen/Mac/Analyse/`.

[27] N.H. Barton and K.S. Gale. "Genetic analysis of Hybrid zones". In: *Hybrid zones and the evolutionary process*. Oxford, UK: Oxford University Press, 1993, pp. 13–45.

[28] N.H. Barton and G.M. Hewitt. "The genetic basis of hybrid inviability in the grasshopper Podisma pedestris". In: *Heredity* 47.3 (1981), pp. 367–383.

[29] A.M. Bauer et al. "Availability of new Bayesian-delimited gecko names and the importance of character-based species descriptions". In: *Proceedings of the Royal Society B: Biological Sciences* 278.1705 (2011), pp. 490–492.

[30] A.D. Bazykin. "A hypothetical mechanism of speciation". In: *Evol.* 23 (1969), pp. 685–687.

[31] M.A. Beaumont. "Approximate Bayesian Computation in evolution and ecology". In: *Ann. Rev. Ecol. Evol. Syst.* 41 (2010), pp. 379–406.

[32] M.A. Beaumont. "Joint determination of topology, divergence time and immigration in population tree". In: *Simulation, genetics and human prehistory*. Ed. by S. Matsumura and P. Forster. Cambridge, UK: University of Cambridge Press, 2008, pp. 135–154.

[33] D.J. Begun et al. "Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*". In: *PLoS Biology* 5 (2011), p. 310.

[34] R.C. Bell et al. "Comparative multi-locus phylogeography confirms multiple vicariance events in co-distributed rainforest frogs". In: *Proceedings of the Royal Society B* 279 (2012), pp. 991–999.

[35] R.C. Bell et al. "Patterns of persistence and isolation indicate resilience to climate change in montane rainforest lizards". In: *Mol. Ecol.* 19 (2010), pp. 2531–2544.

[36] S. Bensch et al. "Linkage between nuclear and mitochondrial DNA sequences in avian malaria parasites: Multiple cases of cryptic speciation?" In: *Evolution* 58.7 (2004), pp. 1617–1621.

[37] D.R. Bentley, S. Balasubramanian, et al. "Accurate whole genome sequencing using reversible terminator chemistry". In: *Nature* 456 (2006), pp. 53–59.

[38] P. Berthier, L. Excoffier, and M. Ruedi. "Recurrent replacement of mtDNA and cryptic hybridization between two sibling species *Myotis myotis* and *Myotis blythii*". In: *Proc. Roy. Soc. B* 273 (2006), pp. 3101–3109.

[39] K. Bi et al. "Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales". In: *BMC Genomics* 13 (2012), p. 403.

[40] D. Bickford et al. "Cryptic species as a window on diversity and conservation". In: *TREE* 22 (2006), pp. 148–155.

[41] E. Birney et al. "Ensembl: a genome infrastructure". In: *Cold Spring Harb. Symp. Quant. Biol.* 68 (2003), pp. 213–216.

[42] I. Biron et al. "De novo transcriptome assembly with ABySS". In: *Bioinformatics* 25 (2009), pp. 2872–2877.

[43] C.P. Blair et al. "Cryptic speciation and host-race formation in a purportedly generalist tumbling flower beetle". In: *Evolution* 59.2 (2005), pp. 304–316.

[44] J.P. Bogart and M. Tandy. "Polyploid amphibians: three more diploid-tetraploid cryptic species of frogs". In: *Science* 193.4250 (1976), pp. 334–335.

[45] D.I. Bolnick and T.J. Near. "Tempo of hybrid inviability in Centrarchid Fishes (Teleostei: Centrarchidae)". In: *Evol.* 59 (2005), pp. 1754–1767.

[46] D.I. Bolnick, T.J. Near, and P.C. Wainwright. "Body size divergence promotes post-zygotic reproductive isolation in centrarchids". In: *Evolutionary Ecology Research* 8.5 (2006), pp. 903–913.

[47] J.E. Bond and A.K. Stockman. "An integrative method for delimiting cohesion species: finding the population-species interface in a group of Californian trap-door spiders with extreme genetic divergence and geographic structuring". In: *Syst. Biol.* 57 (2008), pp. 628–646.

[48] M.C. Brandley et al. "Accommodating locus-specific heterogeneity in molecular dating methods: an example using inter-continental dispersal of *Plestiodon* (*Eumeces*) lizards". In: *Sys. Bio.* 60 (2011), pp. 3–15.

[49] P.H. Brito and S.V. Edwards. "Multilocus phylogeography and phylogenetics using sequence-based markers". In: *Genetica* 135.3 (2009), pp. 439–455.

[50] R.T. Brumfield et al. "Evolutionary implications of divergent clines in an avian (*Manacus*,Aves) hybrid zone". In: *Evol.* 55 (2001), pp. 2070–2087.

[51] I. Buno et al. "A hybrid zone between two subspecies of the grasshopper *Chorthippus parallelus* along the Pyrenees - the west end". In: *Heredity* 73 (1994), pp. 625–634.

[52] H. Burla et al. "The willistoni group of sibling species of Drosophila". In: *Evolution* (1949), pp. 300–314.

[53] J.M. Burns et al. "DNA barcodes and cryptic species of skipper butterflies in the genus Perichares in Area de Conservacion Guanacaste, Costa Rica". In: *Proceedings of the National Academy of Sciences* 105.17 (2008), pp. 6350–6355.

[54] M.D. Carling and R.T. Brumfield. "Haldane's rule in an avian system: using cline theory and divergence population genetics to test for differential introgression of mitochondrial, autosomal and sex-linked loci across the *Passerina* bunting hybrid zone". In: *Evol.* 62 (2008), pp. 2600–2615.

[55] M.D. Carling and R.T. Brumfield. "Speciation in *Passerina* buntings: introgression patterns of sex-linked loci identify a candidate gene region for reproductive isolation". In: *Mol. Ecol.* 18 (2009), pp. 834–847.

[56] H.L. Carson and A.R. Templeton. "Genetic revolutions in relation to speciation phenomena: the founding of new populations". In: *Annual Review of Ecology and Systematics* 15 (1984), pp. 97–131.

[57] J.M. Catchen et al. "Stacks: building and genotyping loci *de novo* from short read sequences". In: *Genes, Genomes, Genetics* 1 (2011), pp. 171–182.

[58] K.M. Chan and S.A. Levin. "Leaky prezygotic isolation and porous genomes: rapid introgression of maternally inherited DNA". In: *Evol.* 59 (2005), pp. 720–9.

[59] F. Chen et al. "Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes". In: *PLoS ONE* 2 (2007), p. 383.

[60] F. Chen et al. "OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups". In: *Nucleic Acids Research* 34 (2005), pp. 363–368.

[61] Z.A. Cheviron and R.T. Brumfield. "Migration-selection balance and local adaptation of mitochondrial haplotypes in Rufous-collared Sparrows (*Zonotrichia capensis*) along an elevational gradient". In: *Evol.* 63 (2009), pp. 1593–1605.

[62] F.M. Cohan and A.A. Hoffmann. "Uniform selection as a diversifying force in evolution: evidence from Drosophila". In: *American Naturalist* (1989), pp. 613–637.

[63] A.A. Comeault et al. "*De novo* characterization of the *Timema cristinae* transcriptome facilitates marker discovery and inference of genetic divergence". In: *Molecular Ecology Resources* 12 (2012), pp. 549–61.

[64] A. Conesa et al. "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research". In: *Bioinformatics* 15 (2005), pp. 3674–6.

[65] P.J. Couper and L.D. Keim. "Two new species of *Saproscincus* (*Reptilia: Scincidae*) from Queensland". In: *Memoirs of the Queensland Museum* 42 (1998), pp. 465–473.

[66] J.A. Coyne. "Genetics of a difference in male cuticular hydrocarbons between two sibling species, Drosophila simulans and D. sechellia". In: *Genetics* 143.4 (1996), pp. 1689–1698.

[67] J.A. Coyne and H.A. Orr. "Patterns of speciation in *Drosophila* revisited". In: *Evol.* 51 (1997), pp. 295–303.

[68] J.A. Coyne and H.A. Orr. *Speciation*. Sinauer Associates Sunderland, MA, 2004.

[69] M.B. Cruzan and M.L. Arnold. "Assortative mating and natural selection in an *Iris* hybrid zone". In: *Evol.* 48 (1994), pp. 1946–1958.

[70] K. Csilléry, O. Francois, and M.G.B. Blum. "abc: an R package for approximate Bayesian computation (ABC)". in press. 2012.

[71] M. Currat et al. "The hidden side of invasions: massive introgression by local genes". In: *Evol.* 62 (2008), pp. 1908–1920.

[72] C. Darwin. *On the origin of species*. London, UK: John Murray, 1859.

[73] DarwinCorrespondenceDatabase. *Darwin Correspondence Database*. 2013. URL: `http://www.darwinproject.ac.uk/entry-729`.

[74] K.K. Dasmahapatra et al. "Inferences from a rapidly moving hybrid zone". In: *Evol.* 56 (2002), pp. 741–753.

[75] P. de Wit et al. "The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis". In: *Molecular Ecology Resources* 12 (2012), pp. 1058–1067.

[76] J.H. Degnan and N.A. Rosenberg. "Gene tree discordance, phylogenetic inference, and the multispecies coalescent". In: *TREE* 24 (2009), pp. 332–340.

[77] M.A. DePristo et al. "A framework for variation discovery and genotyping using next-generation DNA sequencing data". In: *Nature Genetics* 43 (2011), pp. 491–8.

[78] J.R. Dettman et al. "Incipient speciation by divergent adaptation and antagonistic epistasis in yeast". In: *Nature* 447.7144 (2007), pp. 585–588.

[79] T. Dobzhansky. "Complete reproductive isolation between two morphologically similar species of Drosophila". In: *Ecology* (1946), pp. 205–211.

[80] T. Dobzhansky. "Speciation as a stage in evolutionary divergence". In: *American Naturalist* (1940), pp. 312–321.

[81] T. Dobzhansky. "Studies on Hybrid Sterility. I. Spermatogenesis in pure and hybrid *Drosophila pseudoobscura*". In: *Zeitschrift fur Zellforschung und mikroskopische Anatomie* 21 (1934), pp. 169–221.

[82] T. Dobzhansky, H. Burla, and A.B. da Cunha. "A comparative study of chromosomal polymorphism in sibling species of the willistoni group of Drosophila". In: *American Naturalist* (1950), pp. 229–246.

[83] M. Doebeli and U. Dieckmann. "Speciation along environmental gradients". In: *Nature* 16 (2003), pp. 259–64.

[84] G. Dolman. "Evidence for differential assortative female preference in association with refugial isolation of rainbow skinks in Australia's tropical rainforests". In: *PLoS ONE* 3 (2009), p. 3499.

[85] G. Dolman and C. Moritz. "A multilocus perspective on refugial isolation and divergence in rainforest skinks (*Carlia*)". In: *Evol.* 60 (2006), pp. 573–582.

[86] G. Dolman and B. Phillips. "Single copy nuclear DNA markers characterized for comparative phylogeography in Australian Wet Tropics rainforest skinks". In: *Mol. Ecol. Notes* 4 (2004), pp. 185–187.

[87] A.J. Drummond and A. Rambaut. "BEAST: Bayesian evolutionary analysis by sampling trees". In: *BMC Evo. Bio.* 7 (2007), p. 214.

[88] A.J. Drummond et al. *Geneious v5.3*. `http://www.geneious.com`. 2010.

[89] D.A. Drummond and C.O. Wilke. "Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution". In: *Cell* 25 (2008), pp. 341–52.

[90] S. Dubey and R. Shine. "Restricted dispersal and genetic diversity in populations of an endangered montane lizard (*Eulamprus leuraensis*, *Scincidae*)". In: *Mol. Ecol.* 19 (2010), pp. 886–897.

[91] P. Dufkova, M. Macholan, and J. Pialek. "Inference of selection and stochastic effects in the house mouse hybrid zone". In: *Evol.* 65 (2011), pp. 993–1010.

[92] R. Durrett, L. Buttel, and R. Harrison. "Spatial models for hybrid zones". In: *Heredity* 84 (2000), pp. 9–19.

[93] D. Earl et al. "Assemblathon 1: a competitive assessment of *de novo* short read assembly methods". In: *Genome Research* 21 (2011), pp. 2224–41.

[94] D.A. Earl. *Structure Harvester v0.6.6*. 2011. URL: `http://taylor0.biology.ucla.edu/struct\_harvest`.

[95] R.C. Edgar. "MUSCLE: multiple sequence alignment with high accuracy and high throughput". In: *Nucl. Acids Res.* 32 (2004), pp. 1792–97.

[96] S.V. Edwards. "Is a new and general theory of molecular systematics emerging?" In: *Evol.* 63 (2009), pp. 1–19.

[97] S.V. Edwards and P. Beerli. "Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies". In: *Evol.* 54 (2000), pp. 1839–1854.

[98] H. Ellegren et al. "The genomic landscape of species divergence in *Ficedula* flycatchers". In: *Nature* in press (2012).

[99] C.K. Ellison and R.S. Burton. "Genotype-dependent variation of mitochondrial transcriptional profiles in interpopulation hybrids". In: *PNAS* 105 (2008), pp. 15831–15836.

[100] J.A. Endler. *Geographic variation, speciation and clines*. Princeton, NJ: Princeton, 1977.

[101] G. Evanno, S. Regnaut, and J. Goudet. "Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study". In: *Mol. Ecol.* 14 (2005), pp. 2611–20.

[102] L. Excoffier and H.E.L. Lischer. "Arlequin suite v. 3.5: A new series of programs to perform population genetics analyses under Linux and Windows". In: *Molecular Ecology Resources* 10 (2010), pp. 564–567.

[103] N.J.R. Fagundes et al. "Statistical evaluation of alternative models of human evolution". In: *PNAS* 104 (2007), pp. 17614–17619.

[104] J.H. Feder, G.Z. Wurst, and D.B. Wake. "Genetic variation in western salamanders of the genus Plethodon, and the status of Plethodon gordoni". In: *Herpetologica* (1978), pp. 64–69.

[105] J. Felsenstein. "Skepticism towards Santa Rosalia, or why are there so few kinds of animals". In: *Evol.* 35 (1981), pp. 124–138.

[106] R.A. Fisher. "Gene frequencies in a cline determined by selection and diffusion." In: *Biometrics* 6 (1950), pp. 353–61.

[107] P. Flicek et al. "Ensembl 2012". In: *Nucleic Acids Research* 40 (2012), pp. 84–90.

[108] M.K. Fujita and A.D. Leaché. "A coalescent perspective on delimiting and naming species: a reply to Bauer et al." In: *Proceedings of the Royal Society B: Biological Sciences* 278.1705 (2011), pp. 493–495.

[109] M.K. Fujita et al. "Coalescent-based species delimitation in an integrative taxonomy". In: *TREE* 27 (2012), pp. 480–488.

[110] M.K. Fujita et al. "Diversification and persistence at the arid-monsoonal interface: Australia-wide biogeography of the Bynoe's Gecko (*Heteronotia binoei*,Gekkonidae". In: *Evol.* 64 (2010), pp. 2293–2314.

[111] D.J. Funk, P. Nosil, and W.J. Etges. "Ecological divergence exhibits consistently positive associations with reproductive isolation across disparate taxa". In: *Proceedings of the National Academy of Sciences of the United States of America* 103.9 (2006), pp. 3209–3213.

[112] D.J. Futuyma. "Evolutionary constraints and ecological consequences". In: *Evol.* 64 (2010), pp. 1865–1884.

[113] S. Gavrilets. *Fitness Landscapes and the Origin of Species*. Cambridge, MA: Princeton University Press, 2004.

[114] S. Gavrilets. "Hybrid zones with Dobzhansky-type epistatic selection". In: *Evol.* 51 (1997), pp. 1027–1035.

[115] S. Gavrilets. "Models of speciation: what have we learned in 40 years?" In: *Evol.* 57 (2003), pp. 2197–2215.

[116] S. Gavrilets and M.B. Cruzan. "Neutral gene flow across single locus clines". In: *Evol.* 52 (1998), pp. 1277–1284.

[117] L. Gay et al. "Comparing clines on molecular and phenotypic traits in hybrid zones: a window on tension zone models". In: *Evol.* 18 (2008), pp. 1–18.

[118] D.M. Geiser, J.I. Pitt, and J.W. Taylor. "Cryptic speciation and recombination in the aflatoxin-producing fungus Aspergillus flavus". In: *Proceedings of the National Academy of Sciences* 95.1 (1998), pp. 388–393.

[119] A. Gelman et al. *Bayesian Data Analysis*. London, UK: Chapman & Hall, CRC, 2004.

[120] M.D. Gimenez et al. "Understanding the basis of diminished gene flow between hybridizing chromosomal races of the house mouse". In: *Evolution* 67 (2013), pp. 1446–1462.

[121] T. Giraud and S. Gourbière. "The tempo and modes of evolution of reproductive isolation in fungi". In: *Heredity* 109.4 (2012), pp. 204–214.

[122] T.C. Glenn. "Field guide to next-generation DNA sequencers". In: *Molecular Ecology Resources* 11 (2011), pp. 759–769.

[123] A. Gomez et al. "Mating trials validate the use of DNA barcoding to reveal cryptic speciation of a marine bryozoan taxon". In: *Proceedings of the Royal Society B: Biological Sciences* 274.1607 (2007), pp. 199–207.

[124] A. Gómez et al. "Speciation in ancient cryptic species complexes: evidence from the molecular phylogeny of Brachionus plicatilis (Rotifera)". In: *Evolution* 56.7 (2002), pp. 1431–1444.

[125] Z. Gompert and C.A. Buerkle. "A Hierarchical Bayesian Model for Next-Generation Population Genomics". In: *Genetics* 187 (2011), pp. 903–917.

[126] Z. Gompert et al. "Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species". In: *Evolution* 66 (2012), pp. 2167–2181.

[127] Z. Gompert et al. "Widespread mito-nuclear discordance with evidence for introgressive hybridization and selective sweeps in *Lycaeides*". In: *Mol. Ecol.* 17 (2008), pp. 5231–5244.

[128] D.A. Good. "Hybridization and cryptic species in Dicamptodon (Caudata: Dicamptodontidae)". In: *Evolution* (1989), pp. 728–744.

[129] J.M. Good et al. "Ancient hybridization and mitochondrial capture between two species of chipmunks". In: *Mol. Ecol.* 17 (2008), pp. 1313–1327.

[130] S. Gourbiere and J. Mallet. "Are species real? the shape of the species boundary with exponential failure, reinforcement and the "Missing Snowball"". In: *Evol.* 64 (2009), pp. 1–24.

[131] J. Gouzy, S. Careere, and T. Schiex. "FrameDP: sensitive peptide detection on noisy matured sequences". In: *Bioinformatics* 25 (2009), pp. 670–671.

[132] M.G. Grabherr et al. "Full-length transcriptome assembly from RNA-Seq data without a reference genome". In: *Nature Biotechnology* 15 (2011), pp. 644–652.

[133] C.H. Graham, C. Moritz, and S.E. Williams. "Habitat history improves prediction of biodiversity in rainforest fauna". In: *PNAS* 103 (2006), pp. 632–636.

[134] H.H. Grundt et al. "High biological species diversity in the artic flora". In: *PNAS* 103 (2006), pp. 972–975.

[135] R.N. Gutenkunst et al. "Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data". In: *PLoS Genetics* 5 (2009), p. 1000695.

[136] F. Haas, J. Knape, and A. Brodin. "Habitat preferences and positive assortative mating in an avian hybrid zone". In: *J. of Avian Bio.* 41 (2010), pp. 237–247.

[137] F. Hailer et al. "Nuclear genomic sequences reveal that polar bears are an old and distinct bear lineage". In: *Science* 336 (2012), pp. 344–347.

[138] G. Hamilton, M. Stoneking, and L. Excoffier. "Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations". In: *PNAS* 102 (2005), pp. 7476–7480.

[139] R.G. Harrison. "Hybridization and hybrid zones: historical perspective". In: *Hybrid zones and the evolutionary process*. Oxford, UK: Oxford University Press, 1993, pp. 3–12.

[140] R.G. Harrison. "Pattern and process in a narrow hybrid zone". In: *Heredity* 56 (1986), pp. 337–349.

[141] P.D.N. Hebert et al. "Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator". In: *Proceedings of the National Academy of Sciences of the United States of America* 101.41 (2004), pp. 14812–14817.

[142] D. Hedgecock and F.J. Ayala. "Evolutionary divergence in the genus Taricha (Salamandridae)". In: *Copeia* (1974), pp. 738–747.

[143] J. Heled and A.J. Drummond. "Bayesian inference of species trees from multilocus data". In: *Molecular biology and evolution* 27.3 (2010), pp. 570–580.

[144] A.P. Hendry et al. "Along the speciation continuum in sticklebacks". In: *J. of Fish Biology* 75 (2009), pp. 2000–2036.

[145] G.M. Hewitt. "Hybrid zones-natural laboratories for evolutionary studies". In: *Trends in Ecology & Evolution* 3.7 (1988), pp. 158–167.

[146] G.M. Hewitt. "Quaternary phylogeography: the roots of hybrid zones". In: *Genetica* 139 (2011), pp. 617–638.

[147] J. Hey. "Isolation with migration models for more than two populations". In: *Mol. Biol. Evol.* 27 (2010), pp. 905–20.

[148] J. Hey. "On the arbitrary identification of real species". In: *Speciation and Patterns of Diversit (eds Butlin RK, Bridle J, Schluter D)* (2009), pp. 15–28.

[149] M.J. Hickerson, C.P. Meyer, and C. Moritz. "DNA barcoding will often fail to discover new animal species over broad parameter space". In: *Syst. Biol.* 55 (2006), pp. 729–39.

[150] R.J. Hijmans et al. "Very high resolution interpolated climate surfaces for global land areas". In: *International journal of climatology* 25.15 (2005), pp. 1965–1978.

[151] S.M. Hird, R.T. Brumfield, and B.C. Carstens. "PRGmatic: an efficient pipeline for collating genome-enriched second-generation sequencing data using a 'provisional reference genome'". In: *Molecular Ecology Resources* 11 (2011), pp. 743–748.

[152] E. Hodges et al. "Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing." In: *Nat. Protoc.* 4 (2009), pp. 960–974.

[153] P.A. Hohenlohe et al. "Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags". In: *PLoS Genetics* 6 (2010), p. 1000862.

[154] C. Hoskin et al. "Persistence in peripheral refugia promotes phenotypic divergence and speciation in a rainforest frog". In: *American Naturalist* 178 (2011), pp. 561–578.

[155] C.J. Hoskin et al. "Reinforcement drives rapid allopatric speciation". In: *Nature* 437 (2005), pp. 1353–1356.

[156] X. Huang and A. Madan. "CAP3: a DNA sequence assembly program". In: *Genome Research* 9 (1999), pp. 868–77.

[157] C.L. Hubbs. "Hybridization between fish species in nature". In: *Syst. Zool.* 4 (1955), pp. 1–20.

[158] D.H. Hudson and D. Bryant. "Application of Phylogenetic Networks in Evolutionary Studies". In: *Mol. Biol. Evol.* 23 (2006), pp. 254–267.

[159] J.P. Huelsenbeck and F. Ronquist. "MRBAYES: Bayesian inference of phylogeny". In: *Bioinformatics* 17 (2001), pp. 754–755.

[160] G. Ingram and J. Covacevich. "Revision of the genus Carlia (Reptilia, Scincidae) in Australia with comments on Carlia bicarinata of New Guinea". In: *Memoirs of the Queensland Museum* 27.2 (1989), pp. 443–490.

[161] Z. Iqbal et al. "*De novo* assembly and genotyping of variants using colored de Bruijn graphs". In: *Nature Genetics* 44 (2012), pp. 226–232.

[162] M. Jakobsson and N.A. Rosenberg. "CLUMPP: a cluster matching and permutation program for dealing with multimodality in analysis of population structure". In: *Bioinformatics* 23 (2007), pp. 1801–1806.

[163] V. Janousek et al. "Genome-wide architecture of reproductive isolation in a naturally occurring hybrid zone between *Mus musculus musculus* and *M. m. domesticus*". In: *Mol. Ecol.* 21 (2012), pp. 3032–3047.

[164] D.H. Janzen. "Why mountain passes are higher in the tropics". In: *Am. Nat.* 101 (1967), pp. 233–249.

[165] W.B. Jennings and S.V. Edwards. "Speciational history of Australian grass finches (*Poephila*) inferred from thirty gene trees". In: *Evol.* 59 (2005), pp. 2033–2047.

[166] F.F. Jesus, J. Wilkins, and J. Wakeley. "Expected coalescence times and segregating sites in a model of glacial cycles". In: *Genetics and Molecular Research* 5 (2006), pp. 466–474.

[167] C.D. Jiggins and J. Mallet. "Bimodal hybrid zones and speciation". In: *TREE* 15 (2000), pp. 250–255.

[168] E.L. Jockusch and D.B. Wake. "Falling apart and merging: diversification of slender salamanders (Plethodontidae: Batrachoseps) in the American West". In: *Biological Journal of the Linnean Society* 76.3 (2002), pp. 361–391.

[169] S. Joly and A. Bruneau. "Incorporating allelic variation for reconstructing the evolutionary history of organisms from multiple genes: an example from *Rosa* in North America". In: *Syst. Biol.* 55 (2006), pp. 623–636.

[170] F.C. Jones et al. "The genomic basis of adaptive evolution in threespine sticklebacks". In: *Nature* 484 (2012), pp. 55–61.

[171] G. Jones and S.M. Van Parijs. "Bimodal echolocation in pipistrelle bats: are cryptic species present?" In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 251.1331 (1993), pp. 119–125.

[172] S. Kalyana-Sundaram et al. "Expressed pseudogenes in the transcriptional landscape of human cancers". In: *Cell* 149 (2012), pp. 1622–1634.

[173] T. Kawakami et al. "Genetic analysis of a chromosomal hybrid zone in the Australian grasshoppers (*Vandiemenella, viatica* species group)". In: *Evol.* 63 (2009), pp. 139–152.

[174] B.P. Keck and T.J. Near. "Geographic and temporal aspects of mitochondrial replacement in Nothonotus darters (*Teleostei: Percidae: Etheostomatinae*)". In: *Evol.* 64 (2010), pp. 1410–1428.

[175] I. Keller et al. "Population genomic signatures of divergent adaptation, geneflow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes". In: *Molecular Ecology* in press (2012).

[176] W.J. Kent. "BLAT–the BLAST-like alignment tool". In: *Genome Research* 12 (2002), pp. 656–64.

[177] J.F.C. Kingman. "The coalescent". In: *Stochastic processes and their applications* 13 (1982), pp. 235–248.

[178] M. Kirkpatrick and V. Ravigne. "Speciation by natural and sexual selection: models and experiments". In: *Am. Nat.* 159 (2002), S22–35.

[179] C.L. Kleinman and J. Majewski. "Comment on "Widespread RNA and DNA sequence differences in the human transcriptome"". In: *Science* 335 (2012), p. 1302.

[180] N. Knowlton. "Cryptic and sibling species among the decapod Crustacea". In: *Journal of crustacean biology* (1986), pp. 356–363.

[181] D.C. Koboldt et al. "VarScan: variant detection in massively parallel sequencing of individual and pooled samples". In: *Bioinformatics* 25 (2009), pp. 2283–5.

[182] T.D. Kocher et al. "Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers". In: *Proceedings of the National Academy of Sciences* 86.16 (1989), pp. 6196–6200.

[183] M.R. Kronforst et al. "Linkage of butterfly mate preference and wing color preference cue at the genomic location of *wingless*". In: *PNAS* 103 (2006), pp. 6575–6580.

[184] L.E.B. Kruuk et al. "A comparison of multilocus clines maintained by environmental selection or by selection against hybrids". In: *Genetics* 153 (1999), pp. 1959–1971.

[185] C.H. Kuo and J.C. Avise. "Phylogeographic breaks in low-dispersal species: the emergence of concordance across gene trees". In: *Genetica* 124 (2005), pp. 179–186.

[186] R. Lande. "Models of speciation by sexual selection on polygenic traits". In: *Proceedings of the National Academy of Sciences* 78.6 (1981), pp. 3721–3725.

[187] R. Lanfear et al. "PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses". In: *Molecular biology and evolution* 29.6 (2012), pp. 1695–1701.

[188] B. Langmead and S.L. Salzberg. "Fast gapped-read alignment with Bowtie 2". In: *Nature Methods* 9 (2012), pp. 357–359.

[189] B. Langmead et al. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome". In: *Genome Biology* 10 (2009), p. 25.

[190] A. Larson. "The relationship between speciation and morphological evolution". In: (1989).

[191] A.D. Leaché. "Species tree discordance traces to phylogeographic clade boundaries in North American fence lizards (*Sceloporus*)". In: *Syst. Biol.* 58 (2009), pp. 547–559.

[192] A.D. Leache and M.K. Fujita. "Bayesian species delimitation in West African forest geckos (*Hemidactylus fasciatus*)". In: *Proc. Roy. Soc. B* 277 (2010), pp. 3071–3077.

[193] C.E. Lee. "Global phylogeography of a cryptic copepod species complex and reproductive isolation between genetically proximate populations". In: *Evolution* 54.6 (2000), pp. 2014–2027.

[194] A.R. Lemmon, S.A. Emme, and E.M. Lemmon. "Anchored hybrid enrichment for massively high-throughput phylogenomics". In: *Systematic biology* 61.5 (2012), pp. 727–744.

[195] T. Lenormand, D. Roze, and F. Rousset. "Stochasticity in evolution". In: *TREE* 24 (2009), pp. 157–165.

[196] H. Li. "A statistical framework for SNP calling, mutation discovery,association mapping and population genetical parameter estimation from sequencing data". In: *Bioinformatics* 27 (2011), pp. 2987–2993.

[197] H. Li and R. Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform". In: *Bioinformatics* 25 (2009), pp. 1754–1760.

[198] H. Li et al. "The sequence alignment/map (SAM) format and SAMtools". In: *Bioinformatics* 25 (2009), pp. 2078–2079.

[199] J. Li et al. "Rejecting strictly allopatric speciation on a continental island: prolonged post-divergence gene flow between Taiwan (*Leucodioptron taewanus*, *Passeriformes*, *Timaliidae*) and Chinese (*L. canorum canorum*) hwameis". In: *Mol. Ecol.* 19 (2010), pp. 494–507.

[200] R. Li et al. "SOAP: short oligonucleotide alignment program". In: *Bioinformatics* 24 (2008), pp. 713–714.

[201] W. Li and A. Godzik. "CD-Hit: a fast program for clustering and coomparing large sets of protein or nucleotide sequences". In: *Bioinformatics* 22 (2006), pp. 1658–59.

[202] W. Lin et al. "Comment on "Widespread RNA and DNA sequence differences in the human transcriptome"". In: *Science* 335 (2012), p. 1302.

[203] B. Liu et al. "COPE: An accurate k-mer based pair-end reads connection tool to facilitate genome assembly". In: *Bioinformatics* 28 (2012), pp. 2870–2874.

[204] L. Liu et al. "Estimating species phylogenies using coalescence times among sequences". In: *Systematic Biology* 58.5 (2009), pp. 468–477.

[205] M. Lohse et al. "RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics". In: *Nucleic Acids Research* 40 (2012), pp. 622–7.

[206] F. Luca et al. "A reduced representation approach to population genetic analyses and applications to human evolution." In: *Genome Research* 21 (2011), pp. 1087–1098.

[207] G. Lunter and M. Goodson. "Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads". In: *Genome Research* 21 (2011), pp. 936–939.

[208] R. Luo et al. "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler". In: *GigaScience* 1.1 (2012), pp. 1–6.

[209] C.J. Maccallum, B. Nurenberger, and N.H. Barton. "Experimental evidence for habitat dependent selection in a *Bombina* hybrid zone". In: *Proc. Roy. Soc. B* 260 (1995), pp. 257–264.

[210] M. Macholan et al. "Genetic analysis of autosomal and X-linked markers across a mouse hybrid zone". In: *Evol.* 61 (2007), pp. 746–771.

[211] W.P. Maddison. "Gene trees in species trees". In: *Syst. Biol.* 46 (1997), pp. 523–526.

[212] T. Magoč and S.L. Salzberg. "FLASH: fast length adjustment of short reads to improve genome assemblies". In: *Bioinformatics* 27 (2011), pp. 2957–63.

[213] J. Mallet. "A species definition for the modern synthesis". In: *Trends in Ecology & Evolution* 10.7 (1995), pp. 294–299.

[214] J. Mallet. "Hybridization as an invasion of the genome". In: *TREE* 20 (2005), pp. 229–37.

[215] J. Mallet. "Why was Darwin's view of species rejected by 20th century biologists?" In: *Biology and Philosophy* 25 (2010), pp. 497–527.

[216] J. Mallet and K. Willmott. "Taxonomy: renaissance or Tower of Babel?" In: *Trends in Ecology & Evolution* 18.2 (2003), pp. 57–59.

[217] J. Mallet et al. "Estimates of selection and gene flow from measures of cline width and linkage disequilibrium in *Heliconius* hybrid zones". In: *Genetics* 124 (1990), pp. 921–936.

[218] J. Mallet et al. "Natural hybridization in heliconiine butterflies: the species boundary as a continuum". In: *BMC Evol. Bio.* 7 (2008), p. 28.

[219] G.S. Mani and B.C. Clarke. "Mutational order: a major stochastic process in evolution". In: *Proc. Roy. Soc. B* 240 (1990), pp. 29–37.

[220] J.M. Marin et al. "Relevant statistics for Bayesian model choice". available on arXiv: 1110.4700v1. 2011.

[221] L.S. Maroja, J.A. Andres, and R.G. Harrison. "Genealogical discordance and patterns of introgression and selection across a cricket hybrid zone". In: *Evol.* 63 (2009), pp. 2999–3015.

[222] J.A. Martin and Z. Wang. "Next-generation transcriptome assembly". In: *Nature Reviews Genetics* 12 (2011), pp. 671–682.

[223] M. Martin. *Cutadapt removes adapter sequences from high-throughput sequencing reads.* 2011. URL: `http://journal.embnet.org/index.php/embnetjournal/article/view/200`.

[224] G.D. Martinsen et al. "Hybrid populations selectively filter gene introgression between species". In: *Evolution* 55 (2001), pp. 1325–1335.

[225] E. Mayr. *Animal species and evolution*. Cambridge, MA: Belknap Press, 1963.

[226] E. Mayr. *Systematics and the origin of species, from the viewpoint of a zoologist*. 13. Harvard University Press, 1942.

[227] C.S. McBride and M.C. Singer. "Field studies reveal strong postmating isolation between ecologically divergent butterfly populations". In: *PLoS biology* 8.10 (2010), e1000529.

[228] S.F. McDaniel and J.A. Shaw. "Phylogeographic structure and cryptic speciation in the trans-antarctic moss Pyrrhobryum mnioides". In: *Evolution* 57.2 (2003), pp. 205–215.

[229] J.A. McGuire et al. "Mitochondrial introgression and incomplete lineage sorting through space and time: phylogenetics of Crotaphytid lizards". In: *Evol.* 61 (2007), pp. 2878–2897.

[230] J. Melo-Ferreira et al. "Interspecific X-chromosome and mitochondrial DNA introgression in the Iberian hare: selection or allele surfing". In: *Evol.* 65 (2011), pp. 1956–1968.

[231] J. Melo-Ferreira et al. "Recurrent Introgression of Mitochondrial DNA Among Hares (*Lepus* spp.) Revealed by Species-Tree Inference and Coalescent Simulations". in press. 2012.

[232] T.C. Mendelson, B.D. Inouye, and M.D. Rausher. "Quantifying patterns in the evolution of reproductive isolation". In: *Evol.* 58 (2004), pp. 1424–1433.

[233] R.D. Mettler and G.M. Spellman. "A hybrid zone revisited : molecular and morphological analysis of the maintenance, movement, and evolution of a Great Plains avian (*Cardinalidae*: *Pheucticus*) hybrid zone". In: *Mol. Ecol.* 18 (2009), pp. 3256–3267.

[234] M. Meyer and M. Kircher. "Illumina sequencing library preparation for highly multiplexed target capture and sequencing." In: *Cold Spring Harb. Protoc.* 6 (2010), pdb.prot5448.

[235] L.K. M'Gonigle and R.G. FitzJohn. "Assortative mating and spatial structure in hybrid zones". In: *Evol.* 64 (2009), pp. 444–455.

[236] A.P. Michel et al. "Widespread genomic divergence during sympatric speciation". In: *PNAS* 107 (2010), pp. 9724–9729.

[237] A.E. Minoche, J.C. Dohm, and H. Himmelbauer. "Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems". In: *Genome Biology* 12 (2011), p. 112.

[238] D. Molbo et al. "Cryptic species of fig-pollinating wasps: Implications for the evolution of the fig–wasp mutualism, sex allocation, and precision of adaptation". In: *Proceedings of the National Academy of Sciences* 100.10 (2003), pp. 5867–5872.

[239] W.S. Moore. "An evaluation of narrow hybrid zones in vertebrates". In: *Quat. Rev. Bio.* 52 (1977), pp. 263–277.

[240] W.S. Moore. "Assortative mating genes selected along a gradient". In: *Heredity* 25 (2008), pp. 2241–2246.

[241] W.S. Moore and J.T. Price. "Nature of selection in the northern flicker hybrid zone and its implications for speciation theory". In: *Hybrid zones and the evolutionary process*. Oxford, UK: Oxford University Press, 1993, pp. 196–225.

[242] G. Moreno-Hagelsieb and K. Latimer. "Choosing BLAST options for better detection of orthologs as reciprocal best hits". In: *Bioinformatics* 24 (2008), pp. 319–324.

[243] M. Morgan-Richards and G.P. Wallis. "A comparison of five hybrid zones of the weta *Hemideina thoracica* (Orthoptera: Anostomatidae): degree of cytogenetic differentiation fails to predict zone width". In: *Evol.* 57 (2003), pp. 849–861.

[244] C. Moritz et al. "Identification and dynamics of a cryptic suture zone in a tropical rainforest". In: *Proc. Roy. Soc. B* 276 (2009), pp. 1235–1244.

[245] A. Moussalli et al. "Variable responses of skinks to a common history of rainforest fluctuation: Concordance between phyleogeography & paleo-distribution models". In: *Molecular Ecology* 18 (2009), pp. 483–499.

[246] L.C. Moyle and T. Nakazato. "Hybrid Incompatibility "snowballs" Between *Solanum* Species". In: *Science* 329 (2010), pp. 1521–1523.

[247] H.J. Muller. "Isolation mechanisms, evolution and temperature". In: *Biology Symposium* 6 (1942), pp. 71–125.

[248] M.W. Nachman and B.A. Payseur. "Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice". In: *Phil. Trans. R. Soc. Lond B* 367 (2012), pp. 409–421.

[249] T.J. Near et al. "Phylogeny and temporal diversification of darters (*Percidae: Etheostomatinae*)". In: *Syst. Biol.* 60 (2011), pp. 565–595.

[250] R. Nielsen. "Molecular signatures of natural selection". In: *Ann. Rev. Genet.* 39 (2005), pp. 197–218.

[251] R. Nielsen. "Population genetic analysis of ascertained SNP data". In: *Human Genomics* 1 (2004), pp. 218–224.

[252] R. Nielsen and J. Wakeley. "Distinguishing Migration From Isolation: A Markov Chain Monte Carlo Approach". In: *Genetics* 158 (2001), pp. 885–896.

[253] R. Nielsen et al. "Darwinian and demographic forces affecting human protein coding genes". In: *Genome Research* 19 (2009), pp. 838–849.

[254] R. Nielsen et al. "Genotype and SNP calling from next-generation sequencing data". In: *Nature Review Genetics* 12 (2011), pp. 443–51.

[255] R. Nielsen et al. "SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data". In: *PLoS One* 7 (2012), p. 37558.

[256] H.A. Nix. "Biogeography: patterns and process". In: *Rainforest animals: atlas of vertebrates endemic to Australia's Wet Tropics*. Canberra: Australian Nature Conservation Agency, 1991, pp. 11–40.

[257] A.W. Nolte, Z. Gompert, and C.A. Buerkle. "Variable patterns of introgression in two sculpin hybrid zones suggest that genomic isolation differs among populations". In: *Mol. Ecol.* 18 (2009), pp. 2615–2627.

[258] P. Nosil. *Ecological Speciation*. Oxford, UK: Oxford University Press, 2012.

[259] P. Nosil and B.J. Crespi. "Experimental evidence that predation promotes divergence in adaptive radiation". In: *Proceedings of the National Academy of Sciences* 103.24 (2006), pp. 9090–9095.

[260] P. Nosil and S. Flaxman. "Conditions for mutation order speciation". In: *Proc. Roy. Soc. B* 278 (2011), pp. 399–407.

[261] P. Nosil et al. "Do highly divergent loci reside in genomic regions affecting reproductive isolation? A test using next-generation sequence data in Timema stick insects". In: *BMC Evo. Bio.* 12 (2012), p. 164.

[262] M.L. Nydam and R.G. Harrison. "Reproductive protein evolution in two cryptic species of marine chordate". In: *BMC evolutionary biology* 11.1 (2011), p. 18.

[263] J.A.A. Nylander. "MrModeltest v2. Program distributed by the author". In: *Evolutionary Biology Center Uppsala University* 2 (2004), pp. 1–2.

[264] H.A. Orr. "The population genetics of speciation: the evolution of hybrid incompatibilities". In: *Genetics* 139 (1995), pp. 1805–1813.

[265] H.A. Orr and M. Turelli. "The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities". In: *Evolution* 55.6 (2001), pp. 1085–1094.

[266] J.M. Padial et al. "The integrative future of taxonomy". In: *Frontiers in Zool.* 7 (2010), p. 16.

[267] T.L. Parchman et al. "The genomic consequences of adaptive divergence and reproductive isolation between species of manakins". In: *Molecular Ecology* (2013).

[268] J.S. Parla et al. "A comparative analysis of exome capture". In: *Genome Biology* 12 (2011), R97.

[269] G. Parra, K. Bradnam, and I. Korf. "CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes". In: *Bioinformatics* 23 (2007), pp. 1061–67.

[270] Y.M. Parsons and K.L. Shaw. "Species boundaries and genetic diversity among Hawaiian crickets of the genus Laupala identified using amplified fragment length polymorphism". In: *Molecular ecology* 10.7 (2001), pp. 1765–1772.

[271] J.L. Patton. "Hybridization and hybrid zones in pocket gophers (Rodentia, Geomyidae)". In: *Hybrid zones and the evolutionary process*. Oxford, UK: Oxford University Press, 1993, pp. 290–308.

[272] B.A. Payseur. "Using differential introgression in hybrid zones to identify genomic regions involved in speciation". In: *Mol. Ecol. Res.* 10 (2010), pp. 806–820.

[273] B. Peng and M. Kimmal. "simuPOP: a forward-time population genetics simulation environment". In: *Bioinformatics* 21 (2005), pp. 3686–3687.

[274] R.J. Pereira, W.B. Monahan, and D.B. Wake. "Predictors for reproductive isolation in a ring species complex following genetic and ecological divergence". In: *BMC Evol. Bio.* 11 (2011), p. 194.

[275] R.J. Pereira and D.B. Wake. "Genetic leakage after adaptive and nonadaptive divergence in the *Ensatina eschscholtzii* ring species". In: *Evol.* 63 (2009), pp. 2288–2301.

[276] R.J. Petit and L. Excoffier. "Gene flow and species delimitation". In: *TREE* 24 (2009), pp. 386–393.

[277] M. Pfenninger and K. Schwenk. "Cryptic animal species are homogeneously distributed among taxa and biogeographical regions". In: *BMC Evol. Biol.* 7 (2007), p. 121.

[278] B.L. Phillips, S.J.E. Baird, and C. Moritz. "When vicars meet: a narrow contact zone between morphologically cryptic phylogeographic lineages of the rainforest skink, *Carlia rubrigularis*". In: *Evol.* 58 (2004), pp. 1536–1548.

[279] J. Pinheiro et al. *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-102. 2011.

[280] C. Pinho and J. Hey. "Divergence with gene flow: models and data". In: *Annual Review of Ecology, Evolution, and Systematics* 41 (2010), pp. 215–230.

[281] J. Polechova and N. Barton. "Genetic Drift Widens the Expected Cline but Narrows the Expected Cline Width". In: *Genetics* 189 (2011), pp. 227–235.

[282] J.E. Pool and R. Nielsen. "Inference of historical changes in migration rate from the lengths of migrant tracts". In: *Genetics* 181 (2009), pp. 711–719.

[283] L.C. Pope, A. Estoup, and C. Moritz. "Phylogeography and population structure of an ecotonal marsupial, *Bettongia tropica*, determined using mtDNA and microsatellites". In: *Mol. Ecol.* 9 (2000), pp. 2041–2053.

[284] A.H. Porter et al. "The *Pontia daplidice-edusa* hybrid zone in northwestern Italy". In: *Evol.* 52 (1997), pp. 1561–1573.

[285] J.K. Pritchard, M. Stephens, and P. Donnelly. "Inference of population structure using multilocus genotype data". In: *Genetics* 145 (2000), pp. 945–959.

[286] A.S. Putnam, J.M. Scriber, and P. Andolfatto. "Discordant divergence times among Z chromosome regions between two ecologically distinct swallowtail butterfly species". In: *Evol.* 61 (2007), pp. 912–927.

[287] K. De Queiroz. "Species concepts and species delimitation". In: *Systematic Biology* 56.6 (2007), pp. 879–886.

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

[289] J. Radwan and W. Babik. "The genomics of adaptation". In: *Proc. R. Soc. B* 279 (2012), pp. 5024–5028.

[290] A. Rambaut and A.J. Drummond. *Tracer v1. 4*. 2007.

[291] D. Ramsköld et al. "An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data". In: *PLoS Computational Biology* 5 (2009), p. 1000598.

[292] B. Rannala and Z. Yang. "Improved Reversible Jump Algorithms for Bayesian Species Delimitation". In: *Genetics* (2013).

[293] N. Reid, J. R. Demboski, and J. Sullivan. "Phylogeny estimation of the radiation of western American chipmunk (*Tamias*) in the face of introgression using reproductive protein genes". In: *Systematic Biology* 61 (2012), pp. 44–62.

[294] C.L. Remington. "Suture zones of hybrid interaction between recently joined biotas". In: *Evolutionary Biology*. Ed. by T. Dobzhansky, M.K. Hecht, and W.C. Steere. New York: Plenum, 1968, pp. 321–428.

[295] W.R. Rice and E.E. Hostert. "Laboratory experiments on speciation: what have we learned in 40 years?" In: *Evolution* (1993), pp. 1637–1653.

[296] L.H. Rieseberg, J. Whitton, and K. Gardner. "Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species". In: *Genetics* 152 (1999), pp. 713–727.

[297] L.J. Rissler and J.J. Apodaca. "Adding more ecology into species delimitation: ecological niche models and phylogeography help define cryptic species in the black salamander (*Aneides flavipunctatus*)". In: *Syst. Biol.* 57 (2007), pp. 924–942.

[298] C.P. Robert et al. "Lack of confidence in approximate Bayesian computation model choice". In: *PNAS* 108 (2011), pp. 15112–15117.

[299] M.V. Rockman. "The QTN program and the alleles that matter for evolution: all that's gold does not glitter". In: *Evolution* 66 (2012), pp. 1–17.

[300] N.A. Rosenberg. "DISTRUCT: a program for the graphical display of population structure". In: *Mol. Ecol. Res.* 4 (2004), pp. 137–138.

[301] E.B. Rosenblum and L.J. Harmon. "Same same but different: replicated ecological speciation at White Sands". In: *Evolution* 65 (2011), pp. 949–60.

[302] E.B. Rosenblum, H.E. Hoekstra, and M.W. Nachman. "Adaptive reptile color variation and the evolution of the *MC1R* gene". In: *Evol.* 58 (2004), pp. 1794–1808.

[303] E.B. Rosenblum et al. "Goldilocks Meets Santa Rosalia: An Ephemeral Speciation Model Explains Patterns of Diversification Across Time Scales". In: *Evol. Biol.* 39 (2012), pp. 255–261.

[304] G.G. Rosenthal and C.S. Evans. "Female preference for swords in Xiphophorus helleri reflects a bias for large apparent size". In: *Proceedings of the National Academy of Sciences* 95.8 (1998), pp. 4431–4436.

[305] F. Rousset. "Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance". In: *Genetics* 145 (1997), pp. 1219–1228.

[306] S. Rozen and H. Skaletsky. "Primer3 on the WWW for general users and for biologist programmers". In: *Bioinformatics methods and methods in molecular biology*. Totowa, NJ: Humana Press, 2000, pp. 365–386.

[307] M. Ruedi, M.F. Smith, and J.L. Patton. "Phylogenetic evidence of mitochondrial DNA introgression among pocket gophers in New Mexico, family: *Geomyidae*". In: *Mol. Ecol.* 6 (1997), pp. 453–462.

[308] P.C. Sabeti et al. "Positive natural selection in the human lineage". In: *Science* 312 (2006), pp. 1614–1620.

[309] S. Sadedin and M.J. Littlejohn. "A spatially explicit individual-based model of reinforcement in hybrid zones". In: *Evol.* 57 (2003), pp. 962–970.

[310] G. Saetre et al. "A sexually selected character displacement in flycatchers reinforces pre-mating isolation". In: *Nature* 387 (1997), pp. 589–592.

[311] K.M. Saint et al. "C-mos, a nuclear marker useful for squamate phylogenetic analysis". In: *Mol. Phy. Evol.* 10 (1998), pp. 259–263.

[312] M.M. Sasa, P.T. Chippindale, and N.A. Johnson. "Patterns of postzygotic reproductive isolation in frogs". In: *Evol.* 52 (1998), pp. 1811–1820.

[313] M.C. Schatz, A.L. Delcher, and S.L. Salzberg. "Assembly of large genomes using second-generation sequencing". In: *Genome Research* 20 (2010), pp. 1165–73.

[314] D. Schluter. "Ecology and the evolution of the species". In: *TREE* 16 (2001), pp. 372–380.

[315] D. Schluter. "Evidence for Ecological Speciation and Its Alternative". In: *Science* 323 (2009), pp. 737–741.

[316] M.H. Schulz et al. "*Oases*: robust *de novo* RNA-seq assembly across the dynamic range of expression levels". In: *Bioinformatics* 28 (2012), pp. 1086–1092.

[317] O. Seehausen. "Conservation: losing biodiversity by reverse speciation". In: *Curr. Biol.* 16 (2006), pp. 334–7.

[318] O. Seehausen et al. "Speciation through sensory drive in cichlid fish". In: *Nature* 455 (2008), pp. 620–627.

[319] F. Sequeira et al. "Genetic exchange across a hybrid zone within the Iberian endemic golden-striped salamander, *Chiloglossa lusitanica*". In: *Mol. Ecol.* 14 (2005), pp. 245–254.

[320] M.R. Servedio. "The evolution of premating isolation: local adaptation and natural and sexual selection against hybrids". In: *Evol.* 58 (2004), pp. 913–924.

[321] A.J. Shaw. "Molecular phylogeography and cryptic speciation in the mosses *Mielichhoferia elongata* and *M. mielicchoferiana* (*Bryaceae*)". In: *Mol. Ecol.* 9 (2000), pp. 595–608.

[322] J. Shaw. "Biogeographic patterns and cryptic speciation in bryophytes". In: *Journal of Biogeography* 28.2 (2001), pp. 253–261.

[323] K.L. Shaw and S.C. Lesnick. "Genomic linkage of male song and female acoustic preference QTL underlying a rapid species radiation". In: *PNAS* 106 (2009), pp. 9737–9742.

[324] J.T. Simpson et al. "ABySS: a parallel assembler for short read sequence data". In: *Genome Research* 19 (2009), pp. 1117–1123.

[325] S. Singhal. "De novo transcriptomic analyses for non-model organisms: an evaluation of methods across a multi-species data set". In: *Mol. Ecol. Res.* 13 (2013), pp. 403–16.

[326] S. Singhal and C. Moritz. "Reproductive isolation between phylogeographic lineages scales with divergence". In: *arXiv preprint arXiv:1301.4276* (2013).

[327] S. Singhal and C. Moritz. "Strong selection maintains a narrow hybrid zone between morphologically cryptic lineages in a rainforest lizard". In: *Evolution* 66 (2012), pp. 1474–1489.

[328] S. Singhal and C. Moritz. "Testing hypotheses of genealogical discordance in a rainforest lizard". In: *Mol. Ecol.* 21 (2012), pp. 5059–5072.

[329] J.W. Sites, N.H. Barton, and K.M. Reed. "The genetic structure of a hybrid zone between two chromosome races of the Sceloporus grammicus complex (Sauria, *Phrynosomatidae*) in central Mexico". In: *Evol.* 49 (1995), pp. 9–36.

[330] D.A. Skelly et al. "A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data". In: *Genome Research* 21 (2011), pp. 1728–1737.

[331] A. Skinner, A.F. Hugall, and A.N. Hutchinson. "Lygosomine phylogeny and the origins of Australian scincid lizards". In: *Journal of Biogeography* 38 (2011), pp. 1044–1058.

[332] G.S.C. Slater and E. Birney. "Automated generation of heuristics for biological sequence comparison". In: *BMC Bioinformatics* 6 (2005), p. 31.

[333] M. Slatkin. "Gene flow and selection in a cline". In: *Genetics* 75 (1973), pp. 733–756.

[334] M. Slatkin and J.L. Pollack. "Subdivision in an ancestral species creates asymmetry in gene trees". In: *Mol. Biol. Evol.* 25 (2008), pp. 2241–2246.

[335] C. Smadja and R.K. Butlin. "On the scent of speciation: the chemosensory system and its role in pre-mating isolation". In: *Heredity* 102 (2009), pp. 77–97.

[336] A.F.A. Smit, R. Hubley, and P. Green. *RepeatMasker Open-3.0*. 2013. URL: `www.repeatmasker.org`.

[337] S.A. Smith et al. "Resolving the evolutionary relationships of molluscs with phylogenomic tools". In: *Nature* 480 (2011), pp. 364–367.

[338] J.M. Sobel et al. "The biology of speciation". In: *Evol.* 54 (2009), pp. 295–315.

[339] Y. Song et al. "Adaptive introgression of anticoagulant rodent poison resistance by hybridization between Old World mice". In: *Current Biology* 21 (2011), pp. 1296–1301.

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

[341] M. Stephens, N. Smith, and P. Donnelly. "A new statistical method for haplotype reconstruction from population data". In: *Amer. J. of Hum. Gen.* 68 (2001), pp. 978–989.

[342] D. Stuart-Fox et al. "Variation in phenotype, parasite load and male competitive ability across a cryptic hybrid zone". In: *PLoS ONE* 4 (2009), p. 5677.

[343] A. Sulonen et al. "Comparison of solution-based exome capture methods for next generation sequencing". In: *Genome biology* 12.9 (2011), R94.

[344] J. Sumner et al. "Neighbourhood size, dispersal and density estimates of the prickly forest skink (*Gnypetoscincus queenslandiae*) usng individual genetic and demographic methods". In: *Mol. Ecol.* 10 (2001), pp. 1917–1927.

[345] M. Sunderland. *Teaching Natural History at the Museum of Vertebrate Zoology*. in press.

[346] Y. Surget-Groba and J.I. Montoya-Burgos. "Optimization of *de novo* transcriptome assembly from next-generation sequencing data". In: *Genome Research* 20 (2010), pp. 1432–40.

[347] B.E. Suzek et al. "UniRef: comprehensive and non-redundant UniProt reference clusters". In: *Bioinformatics* 23 (2007), pp. 1282–8.

[348] D.L. Swofford. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Sunderland, MA: Sinauer Associates, 2002.

[349] J.M. Szymura and N.H. Barton. "Genetic analysis of a hybrid zone between the fire-bellied toads *Bombina bombina* and *B. variegata*, near Cracow in Southern Poland". In: *Evol.* 40 (1986), pp. 1141–1159.

[350] J.M. Szymura and N.H. Barton. "The genetic structure of the hybrid zone between the fire-bellied toads *Bombina bombina* and *B. variegata*: comparisons between transects and between loci". In: *Evol.* 45 (1991), pp. 237–261.

[351] N. Takahata and M. Slatkin. "Mitochondrial gene flow". In: *PNAS* 81 (1984), pp. 1764–1767.

[352] S. Tavaré et al. "Inferring coalescence times from DNA sequence data". In: *Genetics* 145 (1997), pp. 505–518.

[353] K.C. Teeter et al. "Genome-wide patterns of gene flow across a house mouse hybrid zone". In: *Genome Research* 18 (2007), pp. 1–10.

[354] K. Thornton and P. Andolfatto. "Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*". In: *Genetics* 172 (2006), pp. 1607–1619.

[355] D.L. Toews and D.E. Irwin. "Cryptic speciation in a Holarctic passerine revealed by genetic and bioacoustic analyses". In: *Molecular Ecology* 17.11 (2008), pp. 2691–2705.

[356] D.P.L. Toews and A. Brelsford. "The biogeography of mitochondrial and nuclear discordance in animals". In: *Mol. Ecol.* 21 (2012), pp. 3907–3930.

[357] V.A. Trifonov, N.N. Vorobieva, and W. Rens. "FISH with and without COT1 DNA". In: *Fluorescence in situ hybridization FISH*. Berlin, Germany: Springer Berlin, 2009, pp. 99–109.

[358] R.L. Unckless and H.A. Orr. "Dobzhansky–Muller incompatibilities and adaptation to a shared environment". In: *Heredity* 102.3 (2009), pp. 214–217.

[359] J.C. Uyeda et al. "Drift promotes speciation by sexual selection". In: *Evolution* 63.3 (2009), pp. 583–594.

[360] J. VanDerWal, L.P. Shoo, and S.E. Williams. "New approaches to understanding late Quaternary climate fluctuations and refugial dynamics in Australian tropical rain forests". In: *J. Biogeo.* 36 (2009), pp. 291–301.

[361] N. Vijay et al. "Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments". In: *Molecular Ecology* (in press).

[362] J.P. Vinson et al. "Assembly of polymorphic genomes: Algorithms and application to *Ciona savignyi*". In: *Genome Research* 15 (2005), pp. 1127–1135.

[363] D.B. Wake. "Problems with species: patterns and processes of species formation in salamanders". In: *Ann. Missouri Bot. Gard.* 93 (2006), pp. 8–23.

[364] D.B. Wake, G. Roth, and M.H. Wake. "On the problem of stasis in organismal evolution". In: *Journal of Theoretical Biology* 101.2 (1983), pp. 211–224.

[365] J. Wakeley. *Coalescent theory: an introduction*. Greenwood Village, Colorado: Roberts and Company, 2008.

[366] T.J. Walker. "Cryptic species among sound-producing ensiferan Orthoptera (Gryllidae and Tettigoniidae)". In: *Quarterly Review of Biology* (1964), pp. 345–355.

[367] J.T. Weir and T.D. Price. "Limits to speciation inferred from times to secondary sympatry and ages of hybridizing species along a latitudinal gradient". In: *Am. Nat.* 177 (2011), pp. 462–469.

[368] J.B. Whittall and S.A. Hodges. "Pollinator shifts drive increasingly long nectar spurs in columbine flowers". In: *Nature* 447.7145 (2007), pp. 706–709.

[369] H. Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. ISBN: 978-0-387-98140-6. URL: `http://had.co.nz/ggplot2/book`.

[370] R.T. Wiedmann, T.P.L. Smith, and D.J. Nonneman. "SNP discovery in swine by reduced representation and high throughput pyrosequencing". In: *BMC Genetics* 9 (2008), p. 81.

[371] C. Wiley et al. "Post-zygotic Isolation Over Multiple Generations of Hybrid Descendants in a Natural Hybrid Zone: How Well Do Single-Generation Estimates Reflect Reproductive Isolation?" In: *Evol.* 63 (2009), pp. 1731–1739.

[372] S.E. Williams, R.G. Pearson, and J. Walsh. "Distributions and biodiversity of the terrestrial vertebrates of Australia's Wet Tropics: a review of current knowledge". In: *Pac. Cons. Bio.* 2 (1996), pp. 327–362.

[373] S.E. Williams et al. "Distributions, life history characteristics, ecological specialization and phylogeny of the rainforest vertebrates in the Australian Wet Tropics bioregion". In: *Ecology* 91 (2010), p. 2493.

[374] S.H. Williamson et al. "Simultaneous inference of selection and population growth from patterns of variation in the human genome". In: *PNAS* 102 (2005), pp. 7882–7.

[375] C.C. Wilson and L. Bernatchez. "The ghost of hybrids past: fixation of artic charr (*Salvelinus alpinus*) mitochondrial DNA in an introgressed population of lake trout (*S. namaycush*)". In: *Mol. Ecol.* 7 (1998), pp. 127–132.

[376] S. Wilson. *Field Guide to Reptiles of Queensland*. CSIRO, 2005.

[377] J.D.S. Witt, D.L. Threloff, and P.D.N. Hebert. "DNA barcoding reveals extraordinary cryptic diversity in an amphipod genus: implications for desert spring conservation". In: *Molecular Ecology* 15.10 (2006), pp. 3073–3082.

[378] K.M. Wright et al. "Indirect evolution of hybrid lethality due to linkage with selected locus in Mimulus guttatus". In: *PLoS biology* 11.2 (2013), e1001497.

[379] S. Wright. "Isolation by distance". In: *Genetics* 28.2 (1943), p. 114.

[380] S. Wright. "The genetical structure of populations". In: *Ann. Eugenics* 15 (1951), pp. 323–54.

[381] C. Wu. "The genic view of the process of speciation". In: *J. of Evol. Bio.* 14 (2001), pp. 851–865.

[382] A. Yanchukov et al. "Hybridization of *Bombina bombina* and *B. variegata* (Anura,Discoglossidae) at a sharp ecotone in western Ukraine: comparisons across transects and over time". In: *Evol.* 60 (2006), pp. 583–600.

[383] S.S. Yang et al. "Using RNA-Seq for gene identification, polymorphism detection and transcript profiling in two alfalfa genotypes with divergent cell wall composition in stems". In: *BMC Genomics* 12 (2011), p. 199.

[384] Z. Yang. "A likelihood ratio test of speciation with gene flow using genomic sequence data". In: *Genome biology and evolution* 2 (2010), p. 200.

[385] Z. Yang. "PAML 4: Phylogenetic analysis by maximum likelihood". In: *Mol. Bio. Evol.* 24 (2007), pp. 1586–1591.

[386] C.K.L. Yeung et al. "Testing Founder Effect Speciation: Divergence Population Genetics of the Spoonbills *Platalea regia* and *Pl. minor* (Threskiornithidae, Aves)". In: *Mol. Biol. Evol.* 28 (2011), pp. 473–482.

[387] X. Yi et al. "Sequencing of 50 human exomes reveals adaptation to high altitude". In: *Science* 329 (2010), pp. 75–78.

[388] T. Yuri et al. "The effect of marker choice on estimated levels of introgression across an avian (Pipridae: *Manacus*) hybrid zone". In: *Mol. Ecol.* 18 (2009), pp. 4888–4903.

[389] D.R. Zerbino and E. Birney. "Velvet: algorithms for *de novo* short read assembly using *de Bruijn* graphs". In: *Genome Research* 18 (2008), pp. 821–829.

[390] C. Zhang et al. "Evaluation of a Bayesian coalescent method of species delimitation". In: *Systematic biology* 60.6 (2011), pp. 747–761.

[391] Y. Zhu et al. "Empirical Validation of Pooled Whole Genome Population Re-Sequencing in *Drosophila melanogaster*". In: *PLoS ONE* 7 (2012), e41901.

Appendix A

Supplementary Information for Chapter 2

A.1 Phylogeny

To infer the phylogeny shown in Figure 1 (main text), we concatenated and aligned sequences from previously published loci for the lineages in this group – the mitochondrial locus *ND4* and the nuclear loci β-globin intron and C-mos exon [35, 245, 328, 85] using MUSCLE [95]. We used RAxML to infer a maximum-likelihood tree for the concatenated alignment [340]. The approximate root age for the tree was estimated based on data from [331].

A.2 Supplementary Figures

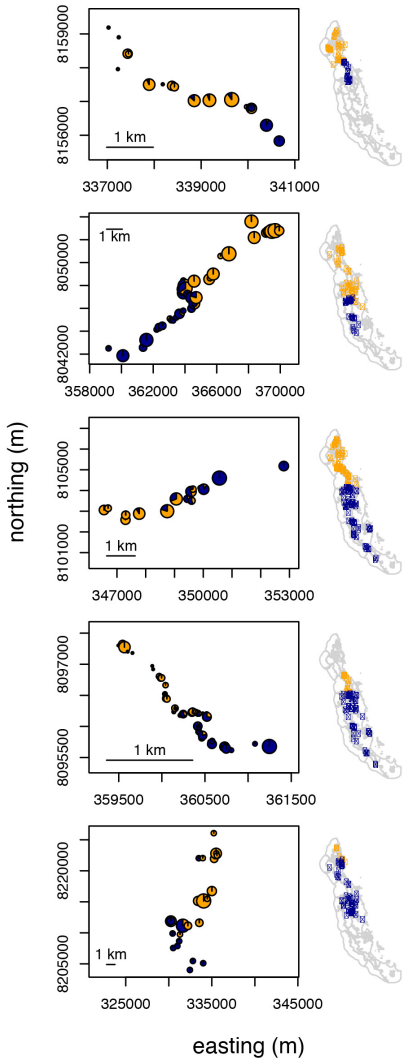

Figure A.1: On left, transect for each contact zone, showing mitochondrial composition of unique localities with localities scaled according to sample size; on right, map of the Australian Wet Tropics showing the range of the phylogeographic lineages. From top to bottom, *Lampropholis coggeri* N/C, *Saproscincus basiliscus* N/C, *Carlia rubrigularis* N/S, *L. coggeri* C/S, and *S. lewisi/S. basiliscus* N.

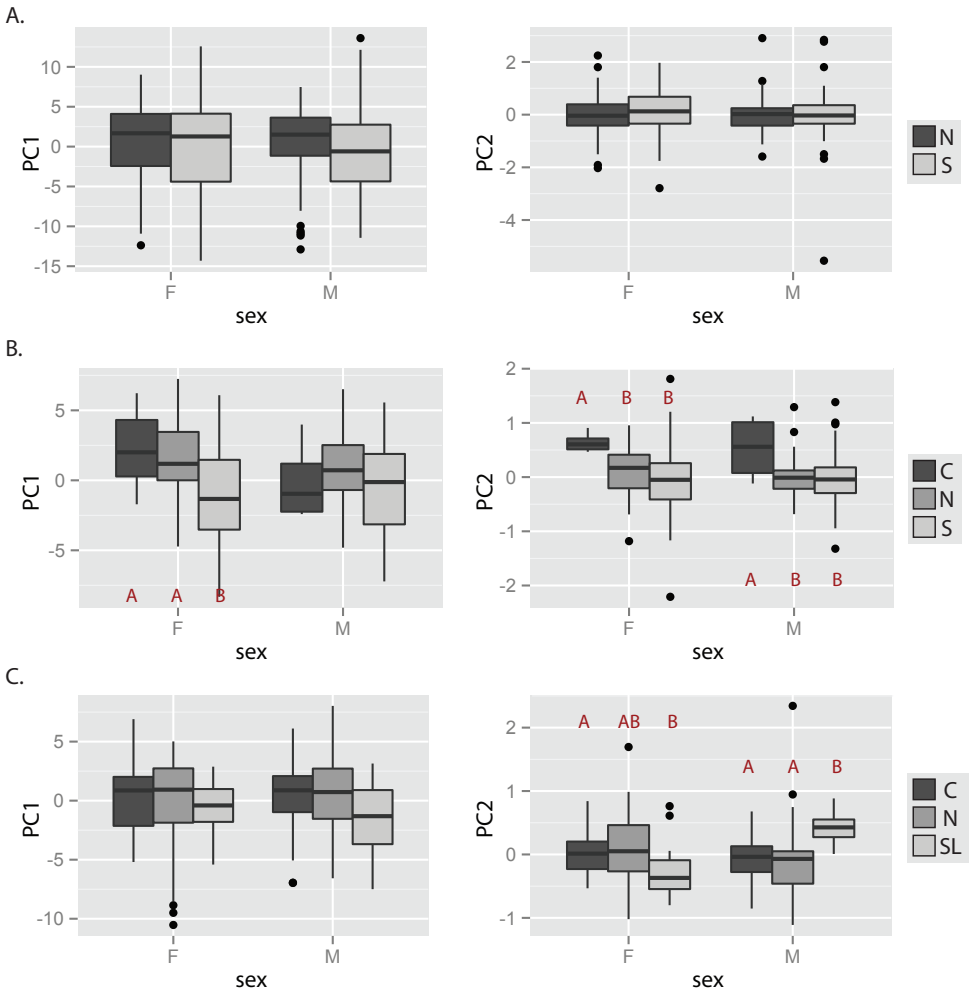

Figure A.2: Morpological data summarized across sexes and across phylogeographic lineages within the four morphologically defined species in this study: A. *Carlia rubrigularis* ($N_{♀} = 223$, $N_{♂} = 156$), B. *Lampropholis coggeri* ($N_{♀} = 174$, $N_{♂} = 143$), and C. *Saproscincus basiliscus* and *S. lewisi* ($N_{♀} = 119$, $N_{♂} = 119$). For each species, we present the first two axes of variation, as summarized by a principal components analysis. Significant differences are labeled in red.

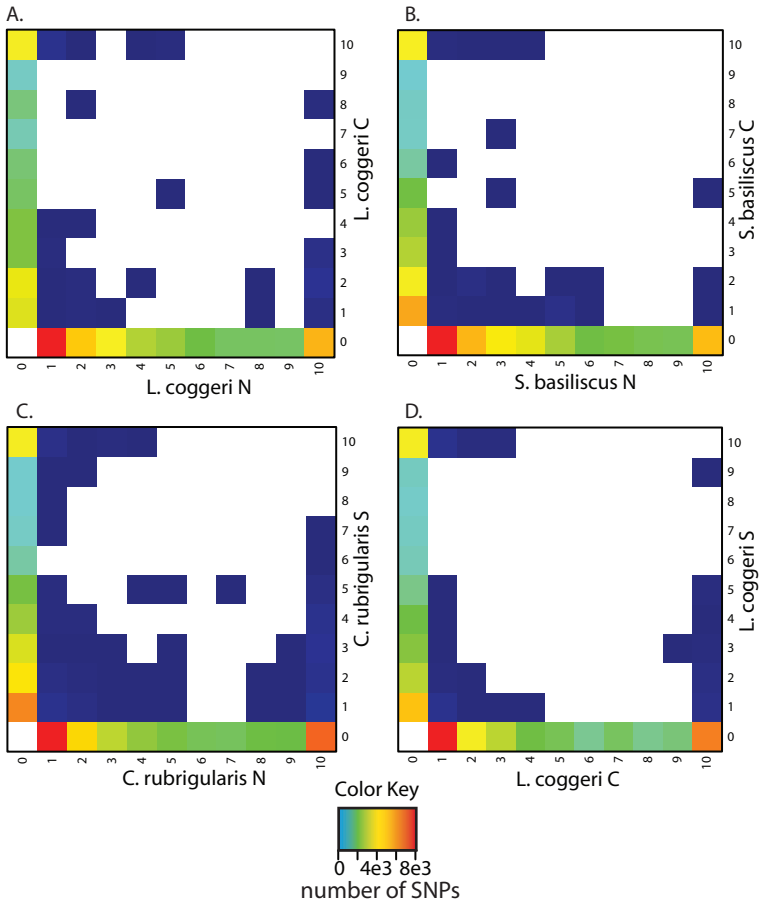

Figure A.3: Two-dimensional site-frequency spectra (2D-SFS), as inferred by ANGSD, for A. *Lampropholis coggeri* N/C, B. *Saproscincus basiliscus* N/C, C. *Carlia rubrigularis* N/S, and D. *L. coggeri* C/S. For each lineage-pair, we used a total of ten individuals, or twenty chromosomes, evenly split between the two lineages. Details on single nucleotide polymorphisms used to construct the 2D-SFS can be found in Table S2.

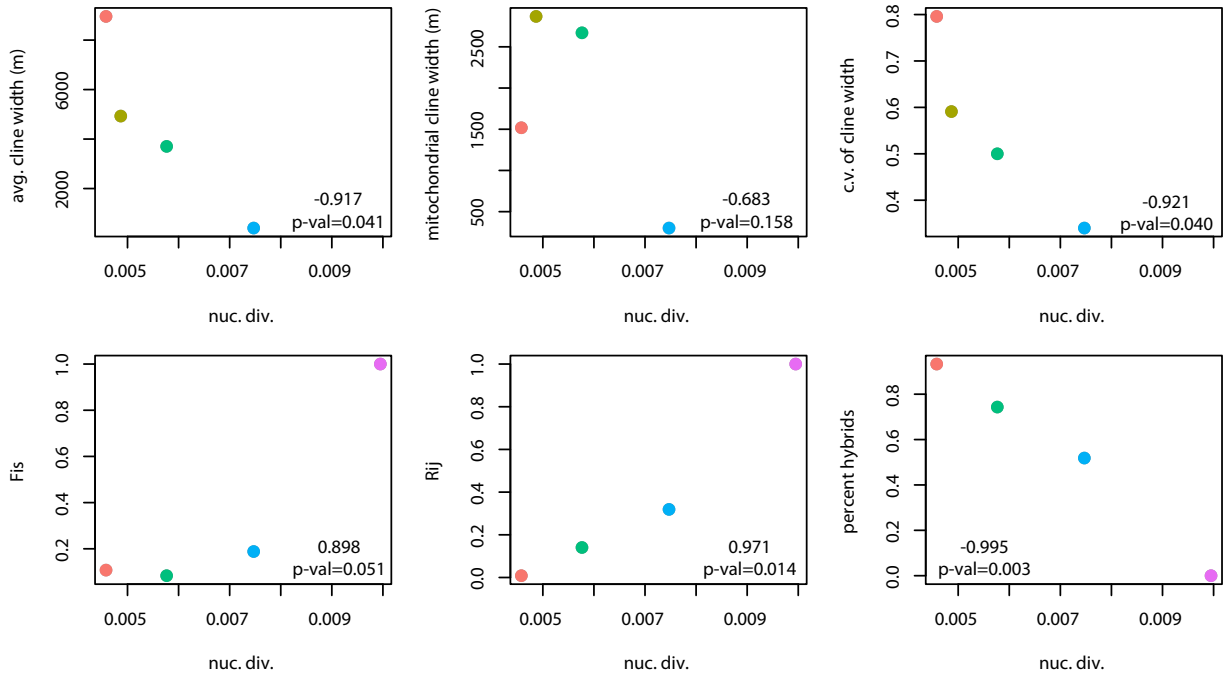

Figure A.4: Comparative results showing the correlation between nuclear divergence and different indices of reproductive isolation: average nuclear cline width, mitochondrial cline width, coefficient of variance in nuclear cline width, Hardy-Weinberg disequilibrium (F_{IS}), linkage disequilibrium (R_{ij}), and percent of hybrids in the contact zone. Graphs are labeled with correlation coefficients.

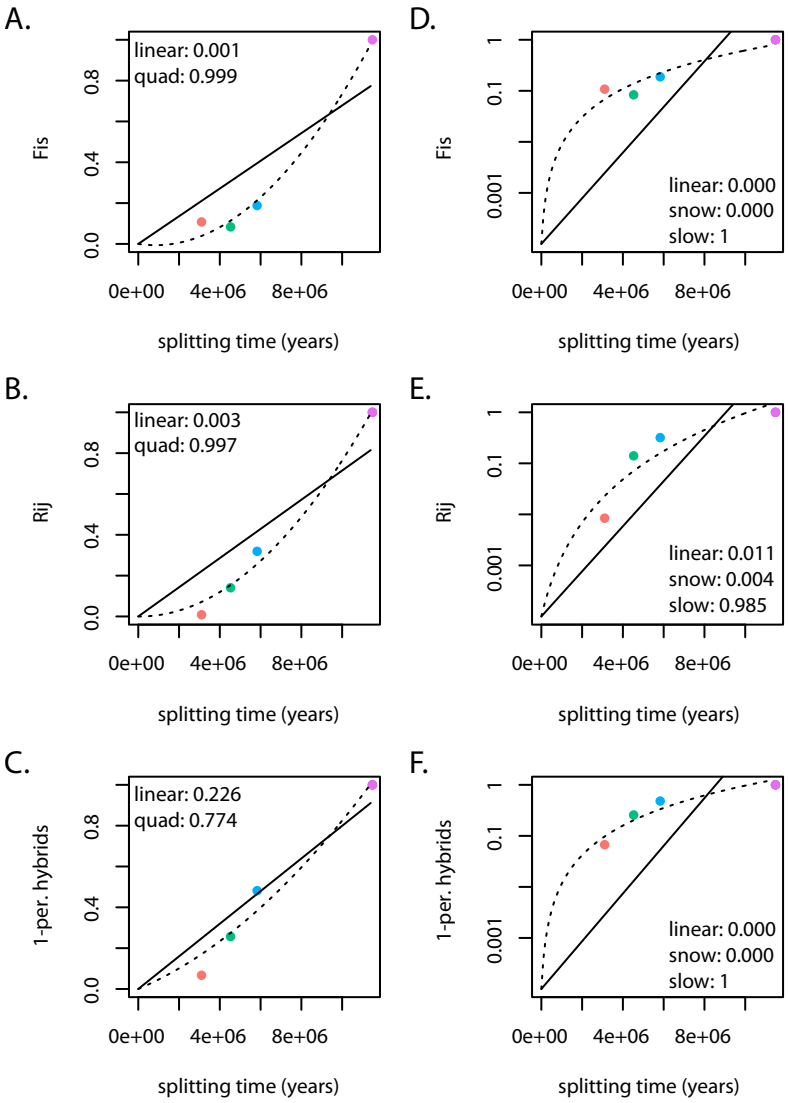

Figure A.5: Model fitting for three indices of reproductive isolation. On the left, we fit linear (solid) and quadratic models (dotted) to the increase of reproductive isolation through time. On the right, we fit linear (solid), quadratic/snowball (solid), and slowdown models (dotted) to the log-linear increase of reproductive isolation through time [130]. Note that only one solid line is visible; model-fitting under the linear and snowball models gave the same result. Relative weights for the different models (as calculated via AIC scores) are shown for each model for each index. Colors follow Figure 2.

A.3 Supplementary Tables

contact zone	number of samples	number of transect populations	transect length
L. coggeri N/C	202	11	16 km
S. basiliscus N/C	209	10	12 km
C. rubrigularis N/S	308	10	7 km
L. coggeri C/S	406	17	2 km
S. basiliscus N & *S. lewisi*	55	NA	15 km

Table A.1: Sampling details for each contact zone. Transect populations are those used in estimation of clines.

LocusID	Forward Primer	Reverse Primer	PCR Temp.	Location/Type	R. Enz.	Cutting Pattern	Reference	contact
GLB1L2	GGGTTGGAGGCTCCCTGCCT	CCGTCACAGCTTCACTAAGGTCCGT	65	3'UTR, non-coding	XbaI	410 | (190+220)	this paper	*C. rubrigularis* N/S
GRN	TGCCCTGTCCCATGAGGGCT	TCGGGAGCTGAACCTCCACCC	65	3'UTR, non-coding	BclI	300 | (50+250)	this paper	*C. rubrigularis* N/S
IDE	TGGGACATCCCAAAGCAGACACT	TGGACACCTGGGTTCTCTGTGC	65	3'UTR, non-coding	MspI	270 | (100+170)	this paper	*C. rubrigularis* N/S
IRS4	CCAGCACAGGCACCGACAGG	CAGGGCATCGGCTTCCAGGC	65	3'UTR, non-coding	KpnI	300 | (250+50)	this paper	*C. rubrigularis* N/S
KIAA2013	TCAGCGCCAGCACTTGCCTC	AGAGGCCACGGGAGGAGCAG	65	3'UTR, non-coding	XbaI	300 | (200+100)	this paper	*C. rubrigularis* N/S
LMBR1	TGCTGACTTGCTCACCACTGCC	GCCTAGCCCAGGAACAAGAGCG	65	3'UTR, non-coding	Eco53KI	370 | (50+320)	this paper	*C. rubrigularis* N/S
MST4	TAGGCGGTTGGCAGCAGCAC	CAAAAGCTGCCCCCTCCCCG	65	3'UTR, non-coding	BstUI	260 | (90+170)	this paper	*C. rubrigularis* N/S
ND4	CACCTATGACTACCAAAAGCTCATGTAGAAGC	CATTACTTTTACTTGGATTTGCACCA	50	CDS, syn.	HhaI	900 | (500+400)	Arevalo et al, 1994	*C. rubrigularis* N/S
NT5C2	CGGTTCTGCTGCAGGCCCAA	ACAAATGCGCCACATGCCAAGG	65	3'UTR, non-coding	SspI	290 | (50+240)	this paper	*C. rubrigularis* N/S
SF1	TCAGGACGCTAGCGCCGAGT	CACACCGGCCCCACACAGAC	65	3'UTR, non-coding	XmnI	290 | (50+240)	this paper	*C. rubrigularis* N/S
SF3A1	CCTGGGAAGCAACAGCCGGG	CTGCGCTGGGCAAGACACCA	65	3'UTR, non-coding	AlwNI	200 | (40+60+100)	this paper	*C. rubrigularis* N/S
ABHD5	ACCCCACTTGTTCTTCTCCA	TGAGTAAGCAGCTGCCCAAAA	60	CDS, syn.	BstBI	230 | (160+70)	Singhal and Moritz 2012	*L. coggeri* C/S
AUTO	TGAGCAGGAAAGGCAAATCT	GTGCCAGTGTGTCCTTGATG	62	CDS, syn.	BanI	190(170+20)	Singhal and Moritz 2012	*L. coggeri* C/S
BGLO (intron 2)	GCGAACTGCACTGYGACAAG	GCTGCCAAGCGGGTGGTGA	63	intron	ApeKI	670 | (640+30)	Dolman and Phillips 2004	*L. coggeri* C/S
LEMD2	GTGCATTCAAGCAGACCAGA	GGCTAGCACTCTCCACCAAG	60	3'UTR, non-coding	HindIII	240 | (140+100)	Singhal and Moritz 2012	*L. coggeri* C/S
ND4	CACCTATGACTACCAAAAGCTCATGTAGAAGC	CATTACTTTTACTTGGATTTGCACCA	50	CDS, syn.	HhaI	900 | (530+370)	Arevalo et al, 1994	*L. coggeri* C/S
NDST2	TCTTGGGGTTGTTTCCAGAC	CACTTGGCATTGTCAGCAGT	60	3'UTR, non-coding	NcoI	440 | (230+210)	Singhal and Moritz 2012	*L. coggeri* C/S
PCBD1	TCTCTCTTGGCTGTGTGGAA	TAAATCATGTGCCCCCAAAT	60	3'UTR, non-coding	HinfI	250 | (140+110)	Singhal and Moritz 2012	*L. coggeri* C/S
RPS8 (intron 3)	CTCTTGGGCGTAAGAAGGAG	CCGCTCATCGTATTTCTTCTG	53	intron	HhaI	630 | (400+230)	Bell et al., 2010	*L. coggeri* C/S
RTN3	AACCTGTTCCAACGCAATTC	TTGAGAAAGGGGAGTTGTGG	60	3'UTR, non-coding	Eco53KI	440 | (130+310)	Singhal and Moritz 2012	*L. coggeri* C/S
SAR1	TAATCACTTTGGCCCACCTC	TATCGCACAAATGCAAGAGC	56	3'UTR, non-coding	AccI	420 | (350+70)	Singhal and Moritz 2012	*L. coggeri* C/S
TPI (intron 5)	TTCTAGCCTATGAACCAGTTTGG	CCTCAACTTGTCATGAACTTCC	50	intron	--	--	Bell et al., 2010	*L. coggeri* C/S
AKT2	ATTCTTCCCCCACCCCGG	CCGCACCCCGGCAAACAGAA	53	3'UTR, non-coding	Hpy166II	380 | (120+180+80)	this paper	*L. coggeri* N/C
ELOVL2	GAGGCAGGGTGTGCAGCAGG	TGCAGCCTAAGAGCAAAAGCCAGT	65	3'UTR, non-coding	BstNI	240 | (60+180)	this paper	*L. coggeri* N/C
MAT1A	TGGTCAGGCTGCCAGATCCCA	GCCTCTCCATTCTCTTGGCCCC	65	3'UTR, non-coding	BglI	280 | (200+80)	this paper	*L. coggeri* N/C
PCYT1A	GCAGCCCCATGGCGGTATC	CTGGGCAGGCTCTGGGCAAC	65	3'UTR, non-coding	RsaI	230 | (50+180)	this paper	*L. coggeri* N/C
PEX16	AGCCTTGGTTTGGTCATTTCAGCCA	GGACCCACCACCAGCTCTGGA	65	3'UTR, non-coding	SphI	300 | (270+30)	this paper	*L. coggeri* N/C
PNPLA2	TGCTGGAGCTGGACCTAGCGA	AGGAAGGGAGGCCCAGTTCACA	65	3'UTR, non-coding	KpnI	250 | (200+50)	this paper	*L. coggeri* N/C
PPP2R1A	GGGTGGAGGGATCCGGTCGA	GGGTTTGCGCACCTCCCAGG	65	3'UTR, non-coding	NcoI	380 | (300+80)	this paper	*L. coggeri* N/C
SDCBP	TCATGTGGGCATCCAGCTCT	ACTGTTGTGTAAAATGCAGGCAA	65	3'UTR, non-coding	BstNI	220 | (170+50)	this paper	*L. coggeri* N/C
SLC35F5	TGAGGCTCCTGCCTGAATGGAA	AGCGATGCAACAGTGCCAGC	65	3'UTR, non-coding	RsaI	300 | (200+100)	this paper	*L. coggeri* N/C
SLC37A4	TCCCCTGCCCCAGCACTGTGG	TGGAAGGGCGCTGAGTGGGTCC	65	3'UTR, non-coding	DraI	400 | (250+150)	this paper	*L. coggeri* N/C
ND4	CACCTATGACTACCAAAAGCTCATGTAGAAGC	CATTACTTTTACTTGGATTTGCACCA	57 & 50	CDS, syn.	HhaI	900 | (300+600)	Arevalo et al, 1994	*L. coggeri* N/C
ACY1	CCCAAGGAGGCAAAGGGCCC	TCACCCCCATGCCTTCCCCT	65	3'UTR, non-coding	MslI	190 | (110+80)	this paper	*S. basiliscus* N/C
DPP4	TGCAGTTTCTTTGTCCCATGTTGGGA	GCACACACCCCACTGGCACA	65	3'UTR, non-coding	HhaI	280 | (60+220)	this paper	*S. basiliscus* N/C
ECE2	ACTTCTGGGCGGGGTGTCGT	GTGGGAGCTCTGGGGTGGGT	65	3'UTR, non-coding	Hpy166II	180 | (45+135)	this paper	*S. basiliscus* N/C
GOLGA2	AGTGGGCCTATGTGCTGAGCA	AGCACCGTAAGGGGTGTGCA	65	3'UTR, non-coding	MspI	250 | (200+50)	this paper	*S. basiliscus* N/C
LCAT	ACCCTGCCGGACAACTGGAGA	AGTTGCGCACAAGCCCTGCA	65	3'UTR, non-coding	BclI	180 | (80+100)	this paper	*S. basiliscus* N/C
PNPLA2	AGAGTGCAACTCCCCACACC	GGTCCCTTCAAGACAGCAGCA	65	3'UTR, non-coding	BstUI	230 | (60+170)	this paper	*S. basiliscus* N/C
SLC9A7	TGATGAACAGAGGCACTCGGCT	AGTCCACCTGCCCTTGTGCA	65	3'UTR, non-coding	BglI	250 | (200+50)	this paper	*S. basiliscus* N/C
TOMM70A	TGACTCATGTTTTGAGGGTTGTTGTGA	TGCAGCAGAACAGCAGGCTG	65	3'UTR, non-coding	BsmAI	280 | (90+190)	this paper	*S. basiliscus* N/C
TXNRD3	TGGGGAAACCTATTCGTCCAGTCA	AGGCCACTGCTTTCTTAGCAGC	65	3'UTR, non-coding	BsmAI	330 | (130+200)	this paper	*S. basiliscus* N/C
UGT1A1	GCAAGTGCTCTGCCAGCATGC	AGCACCGATTCAGCGCAGCT	65	3'UTR, non-coding	DraIII	220 | (50+170)	this paper	*S. basiliscus* N/C
ND4	CACCTATGACTACCAAAAGCTCATGTAGAAGC	CATTACTTTTACTTGGATTTGCACCA	57 & 50	CDS, syn.	RsaI	900 | (200+700)	Arevalo et al, 1994	*S. basiliscus* N/C
B-globin	GCGAACTGCACTGYGACAAG	GCTGCCAAGCGGGTGGTGA	61	intron	sequenced		Dolman and Phillips 2004	*S. basiliscus* N/*S. lewisi*
cmos	GCGGTAAAGCAGGTGAAGAAA	TGAGCATCCAAAGTCTCCAAT	57	CDS, syn	sequenced		Saint 1998	*S. basiliscus* N/*S. lewisi*
TPI (intron 5)	TTCTAGCCTATGAACCAGTTTGG	CCTCAACTTGTCATGAACTTCC	57	intron	sequenced		Bell *et al.*, 2010	*S. basiliscus* N/*S. lewisi*
CRISP	TGCTGTAGCCTACTGTCCTCAA	TGCTTATCATGCTCCGCTAAGTT	57	intron	sequenced		this paper	*S. basiliscus* N/*S. lewisi*
RPS8 (intron 3)	CTCTTGGCGTAAGAAGGAG	CCGCTCATCGTATTTCTTCTG	57	intron	sequenced		Bell *et al.*, 2010	*S. basiliscus* N/*S. lewisi*
rho	CCTTGCCTGGACACCCTATGCTG	CAGGAGAGACCCTCACATTG	61	intron	sequenced		Dolman and Phillips 2004	*S. basiliscus* N/*S. lewisi*
ND4	CACCTATGACTACCAAAAGCTCATGTAGAAGC	CATTACTTTTACTTGGATTTGCACCA	57	CDS, syn.	sequenced		Arevalo *et al*, 1994	*S. basiliscus* N/*S. lewisi*

Table A.2: The loci used in this study and their associated details.

contact zone	total number of SNPs	fixed SNPs	polymorphic SNPs	shared SNPs
L. coggeri N/C	19884	3510 (17.7%)	16220 (81.6%)	154 (0.8%)
S. basiliscus N/C	29664	4712 (15.9%)	24798 (83.6%)	206 (0.7%)
C. rubrigularis N/S	32264	6365 (19.7%)	25693 (79.6%)	369 (1.1%)
L. coggeri C/S	41618	9260 (22.2%)	31989 (76.9%)	330 (0.8%)

Table A.3: Details on the number of single nucleotide polymorphisms (SNPs), and their proportions, used in the two-dimensional site frequency spectrum (2D-SFS) for the contact zones analyzed with genomic data.

contact zone	nuc. div.	mt. div.	theta (θ)	div. time	M_{12}	M_{21}	N_1	N_2	N_A
L. coggeri N/C	0.0046	0.028	3090	3.1 my	0.0268	0.0117	408881	574006	1352453
S. basiliscus N/C	0.0049	0.056	3644	3.4 my	0.0123	0.0112	239822	919316	1352327
C. rubrigularis N/S	0.0058	0.141	3775	4.5 my	0.0112	0.0359	464585	1200178	1362782
L. coggeri C/S	0.0075	0.132	4608	5.8 my	0.0097	0.0204	628695	1176557	2227376
S. lewisi/*S. basiliscus* N	0.0100	0.185	NA	11.4 my	0.0186	0.0040	278501	740017	NA

Table A.4: Parameter estimates for the isolation-with-migration model, as fit to the lineage-pairs. Populations labelled '1' are the northern lineage in each contact; populations labelled '2' the southern lineage.

Appendix B

Supplementary Information for Chapter 3

B.1 Simulations of Anonymous Pooling

To determine the pooling strategy – both with respect to number of individuals to include per pool and desired coverage – we conducted simple simulations in R [288]. These simulations were designed to determine how sampling drift affected our inference of allele frequency, considering both the bias we would see when sampling just a portion of the population and when we would sequence anonymous pools of that portion. Here, for "known" allele frequencies (0.01, 0.1, 0.25, 0.5), we used the binomial sampling distribution to simulate the effects of this sampling drift. Importantly, we assumed that the DNA had been pooled in exactly equimolar amounts, such that each chromosome in the population was equally likely to be sampled. As summarized in Fig. S11, we had two major findings:

1. increasing the number of individuals pooled had a much bigger effect on reducing error than increasing coverage
2. increasing coverage beyond $50\times$ had negligible effects on the error in inferring allele frequency.

B.2 Evaluating Success of Exome Capture

Applications of exome capture to non-model organisms are still in their infancy [39, 194], and thus, it was crucial to evaluate the efficacy of our exome capture method to validate our results. To do so, we used several metrics and statistics, which we outline below.

1. **Sequencing**: The first step in a next-generation experiment is to filter raw data for quality. Here, through a rigorous filtration, we lost 45% to 55% of our data, largely

because ≈70% of our paired-end reads could be merged into a single read. The resulting quality of the data was high; for all but one capture experiment, the average Phred quality score was the maximum possible (36). Despite this aggressive filtration, we retained enough data to get high coverage of both nuclear exon regions (>100×) and the mitochondrial DNA (>1000×).

2. **Assembly**: In exome capture experiments with non-model organisms, researchers typically do not know the sequence flanking targets. Because of edge effects, inclusion of just the target sequence will lead to reduced mapping efficiency [39]. Luckily, in exome capture experiments, a portion of the flanking region is also captured and can be reconstructed using *de novo* assemblers. Using this approach, we recovered an average of 60% additional sequence (Table 2), which largely represents non-coding sequence surrounding our target exons.

3. **Annotation**: We annotated the *de novo* assemblies of the cleaned sequence reads to identify the targets to which they matched. We successfully assembled 100% of targeted exons, and the majority of assembled contigs were longer than the target exons. Many of the assembled contigs did not match any of our targeted exons; annotating them with the non-redundant ('nr') NCBI database showed that many of them were unique genes or genomic sequence (Fig. S10). Although some of these contigs are thus likely "biologically real", we opted for a conservative approach and excluded them from downstream analyses. These contigs could be analyzed in future work.

4. **Sensitivity**: Sensitivity is a measure of what portion of in-target assemblies are represented by sequence data. Here, every single exon was covered by at least 1× coverage.

5. **Specificity**: Specificity is measured as the percentage of cleaned reads that map onto targeted regions. Depending on the technology, specificity can range from 10% to 90% across experiments [343], but crucially, variance in specificity should be low within an experiment. Low variance suggests that the procedure worked uniformly across captures. In this study, we found that specificity ranged from 58.2% to 75.2% across captures and that variance within an array was low (Table 3; Fig. S4).

6. **Coverage metrics**: Coverage metrics can show how uniform results are across libraries on the same capture and can indicate if coverage responds to other characteristics of the data set as expected. Here, we measured several metrics:

   - *Correlation of coverage across libraries in a capture*: high correlation suggests consistency of capture. As seen in Fig. S5, coverage was highly correlated across libraries on the same capture ($r^2 > 0.97$).
   - *Correlation of loci across captures*: high correlation suggests that differences in capture efficiency across loci is due more to locus-specific effects and less due

to stochastic effects of a given capture experiment. As seen in Fig. S6, coverage of orthologous loci across different capture experiments was significant and high ($r^2 = 0.53 - 0.70$).

- *Density plots of coverage*: ideally, coverage across loci should be tightly distributed, indicating no strong bias in capture efficiency at a given locus. As seen in Fig. S7, most loci had about $200\times$ coverage, although there is some spread in the distribution of coverage across loci.
- *Correlation of divergence with coverage*: in our experiments, we were capturing orthologous loci across two sister lineages in a lineage-pair. To ensure no bias in capture efficiency if the two orthologs were divergent, we included both orthologs on our array. As such, we would expect to see little correlation of coverage with sequence divergence. As seen in Fig. S8, we see significant but low correlation between divergence and coverage ($r^2 = 0.16 - 0.19$).
- *Correlation of coverage with GC-content*: coverage is expected to have a hump-shaped relationship with GC-content, such that coverage is low at low- and high-GC content. We recover this pattern in our data (Fig. S9).

In sum, our results indicate that our exome capture experiments were successful and that our downstream inference should not be affected by the technical vagaries of the experiments themselves.

B.3 Supplemental Figures

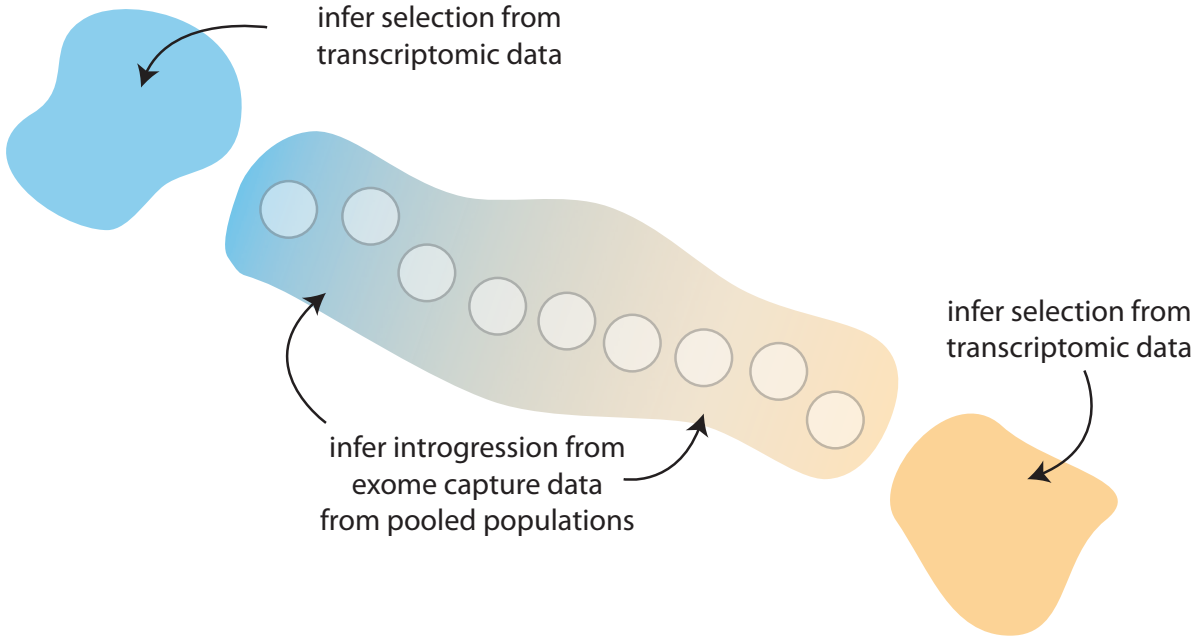

Figure B.1: Basic sampling scheme used in this study. Transcriptomic data from geographically isolated populations were used to infer selection history of loci and to design arrays; anonymously pooled exome capture data from populations in the hybrid zones were used to infer introgression extent.

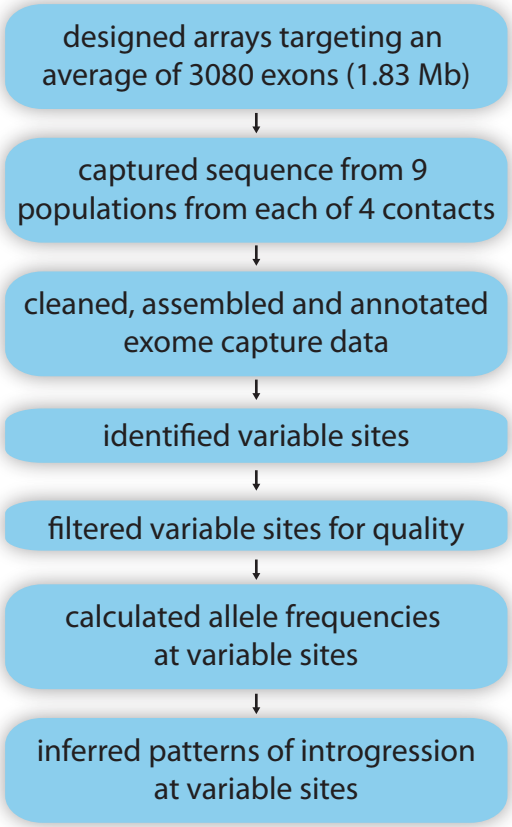

Figure B.2: Summary of bioinformatics and inference pipeline used in this study.

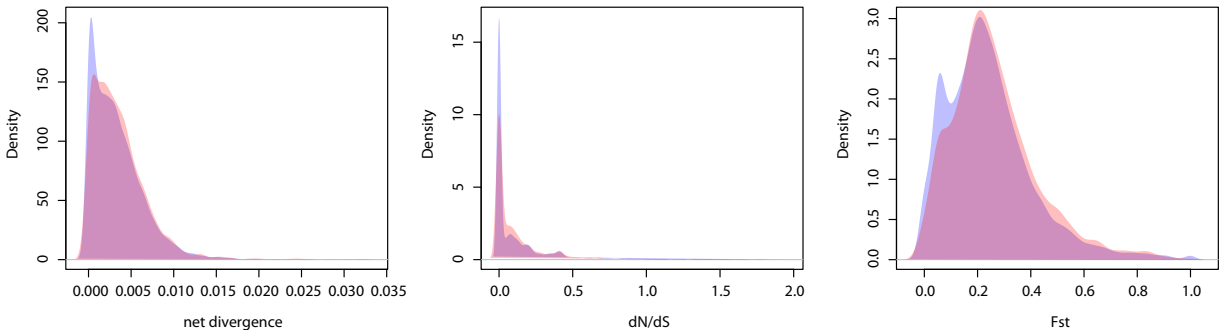

Figure B.3: Density histograms comparing distributions of summary statistics for all transcripts sequenced for focal lineages (in red) and for the subset of transcripts used on exome capture arrays (in blue).

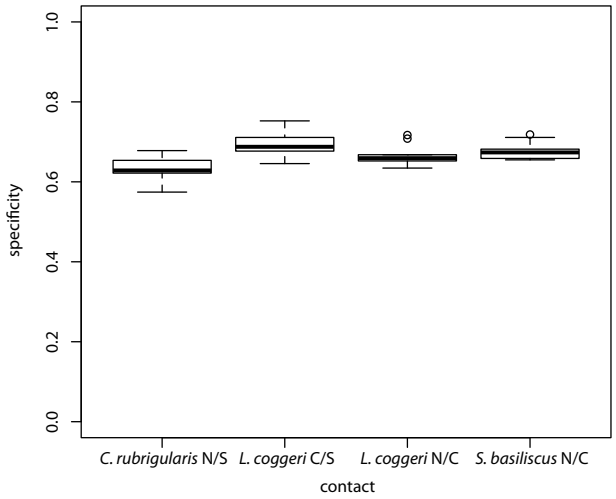

Figure B.4: Specificity, or proportion of cleaned reads mapping onto target, summarized across all libraries for each contact.

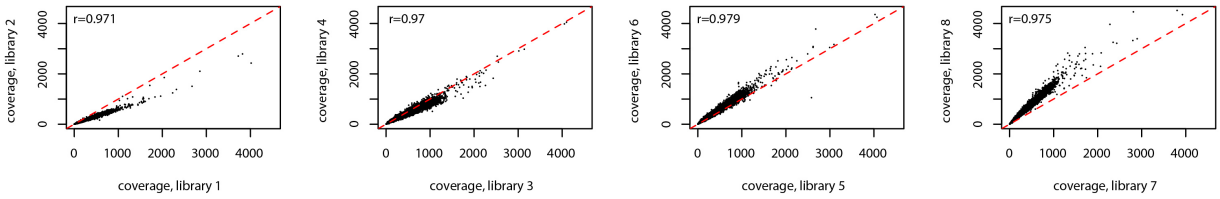

Figure B.5: For a randomly selected contact zone (*Carlia rubrigularis* N/S), correlation in coverage between different libraries from the same transect. The red dotted line is at unity.

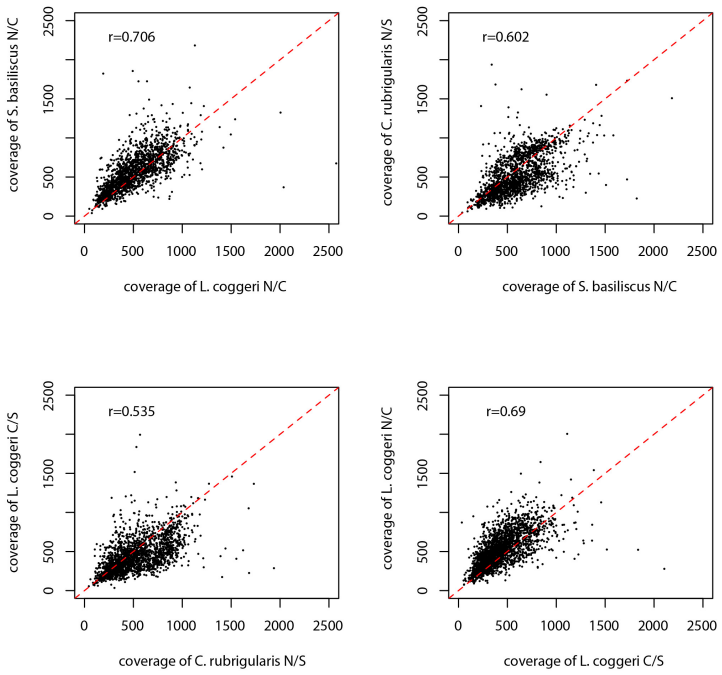

Figure B.6: For orthologs across multiple arrays, correlation in coverage between different contact zones. The red dotted line is at unity.

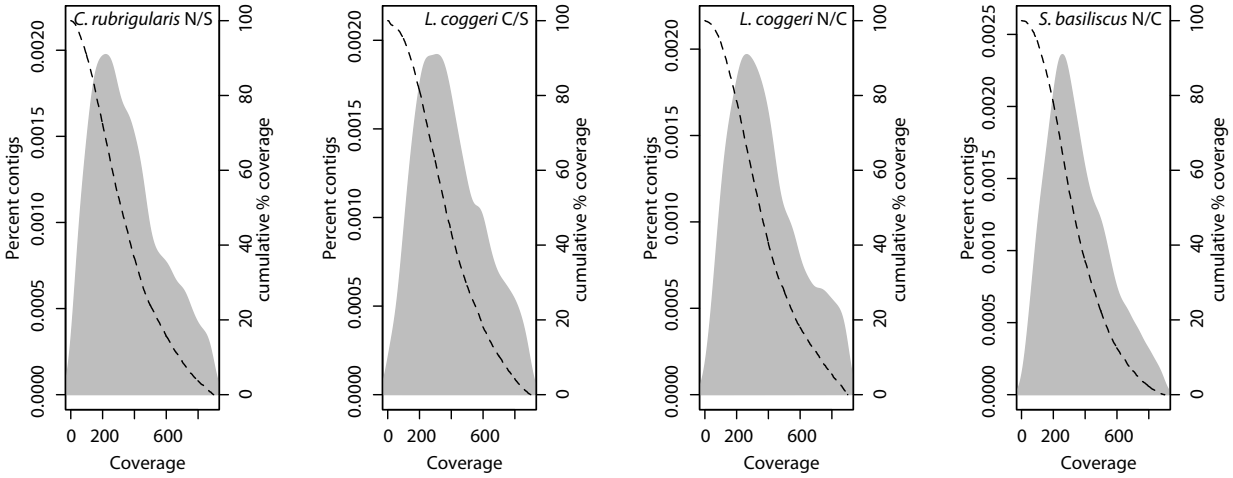

Figure B.7: Density plots of locus-wide coverage, with frequencies shown on the left *y*-axis. The dotted line shows accumulation of coverage across increasing coverage levels, with percentages shown on the right *y*-axis.

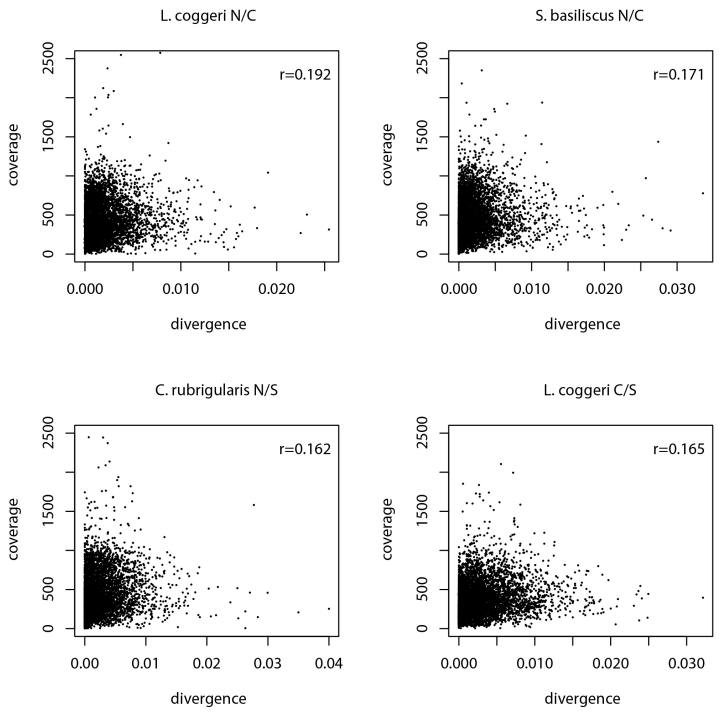

Figure B.8: Correlation between coverage and net divergence at loci.

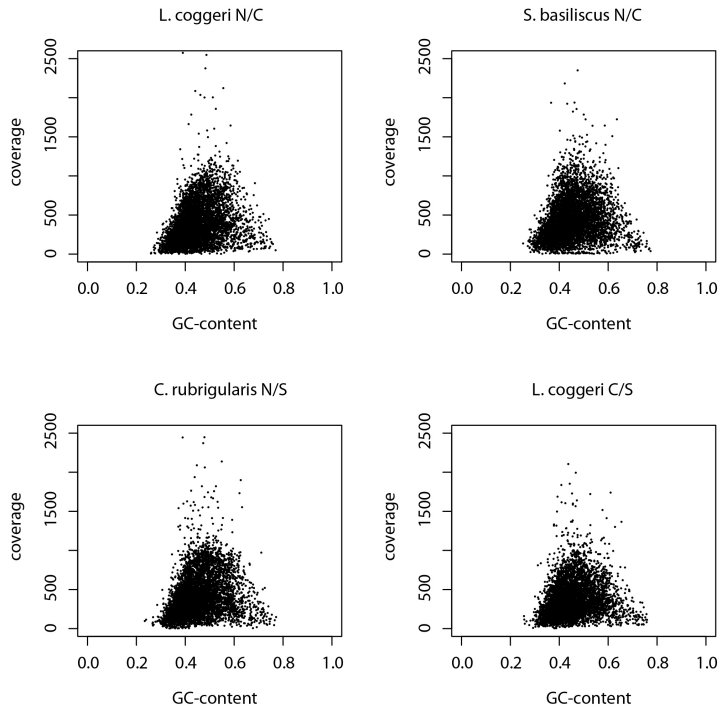

Figure B.9: Correlation between coverage and GC-content at loci.

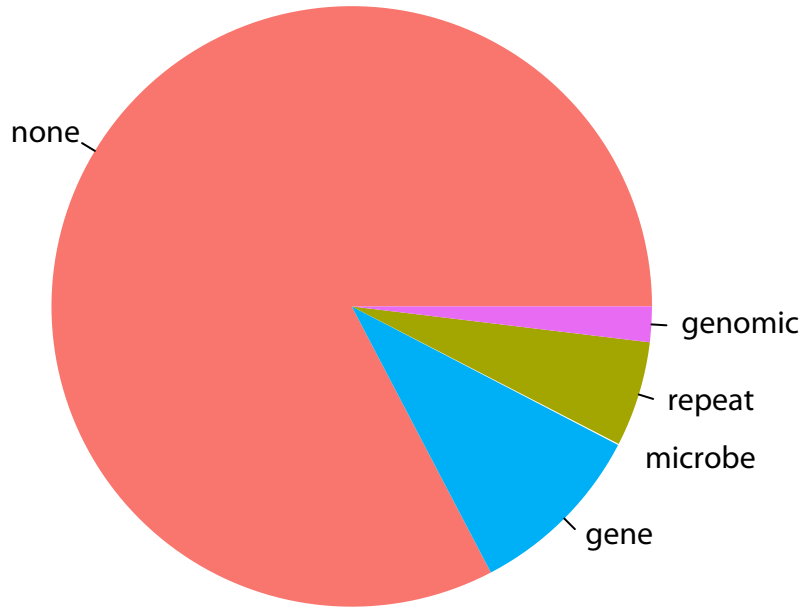

Figure B.10: Analysis of where unannotated contigs from the exome capture assemblies matched, based on BLAST searches against the NCBI 'nr' database.

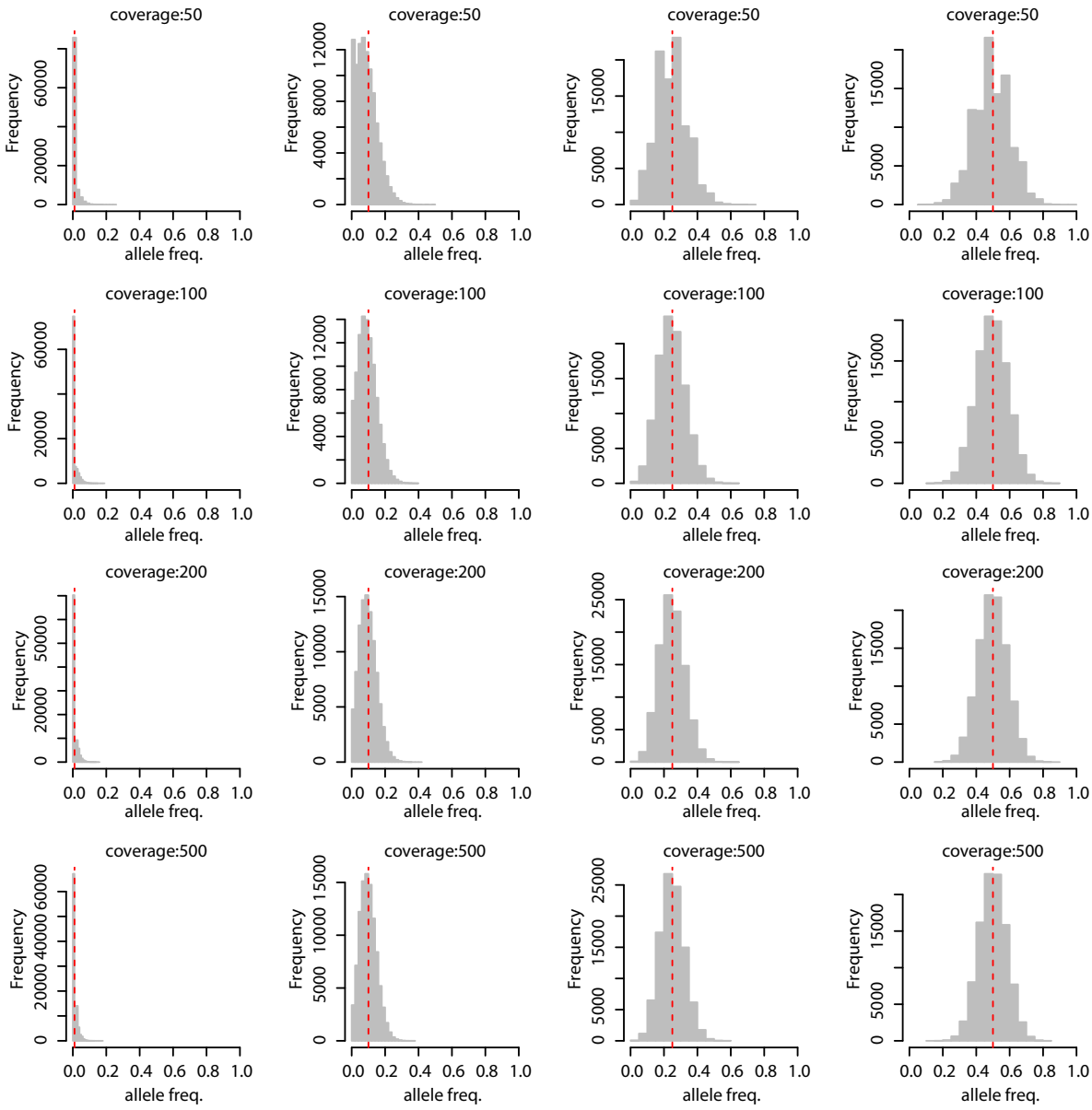

Figure B.11: Results from toy simulations exploring role of sampling drift in inferring allele frequencies from pooled populations. The dotted red line indicates the true allele frequency.

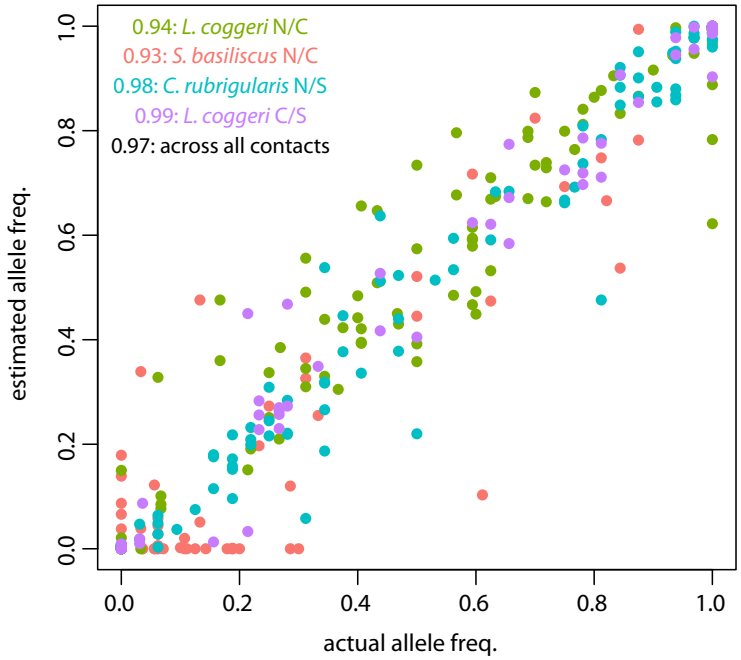

Figure B.12: Correlation between known allele frequencies (as measured in [326]) and estimated allele frequencies, as calculated by sequencing anonymously pooled populations.

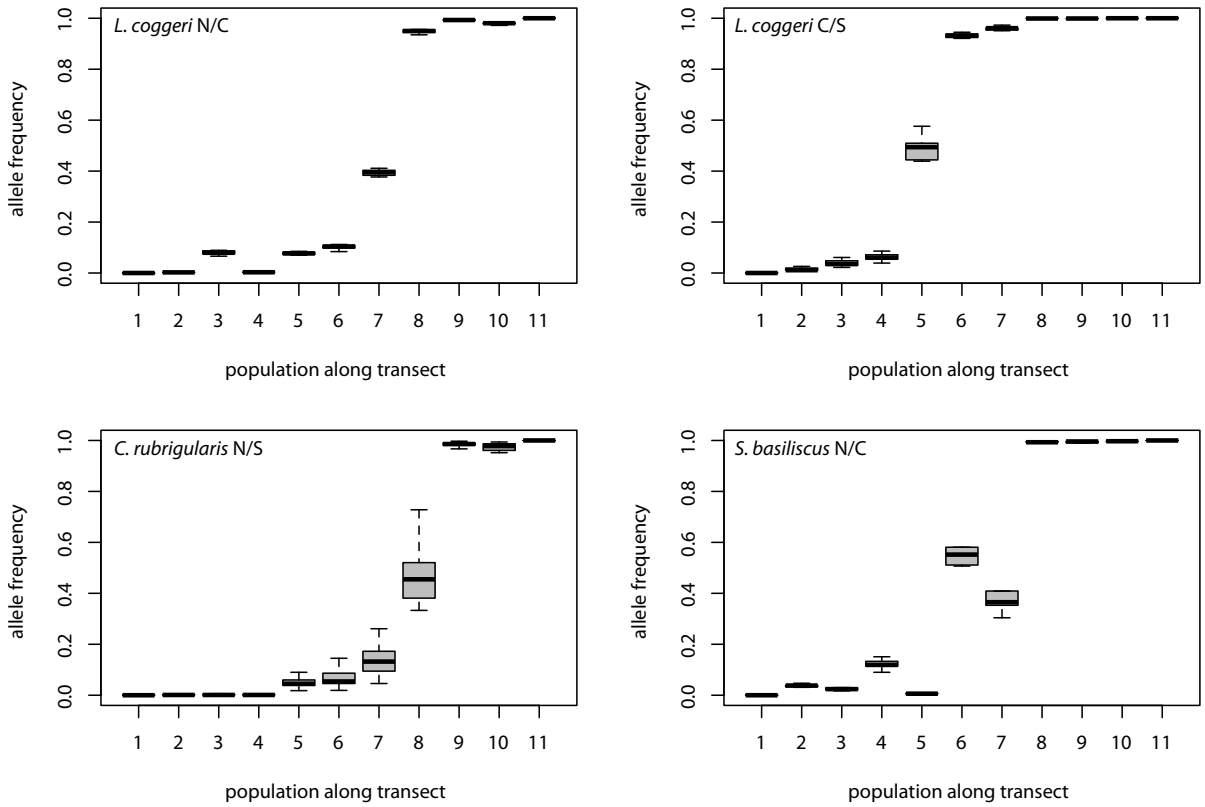

Figure B.13: Variance in allele frequency estimates across fixed single nucleotide polymorphism (SNPs) in the mitochondrial genome (mtDNA). As mtDNA does not recombine, all SNPs should have the same allele frequency.

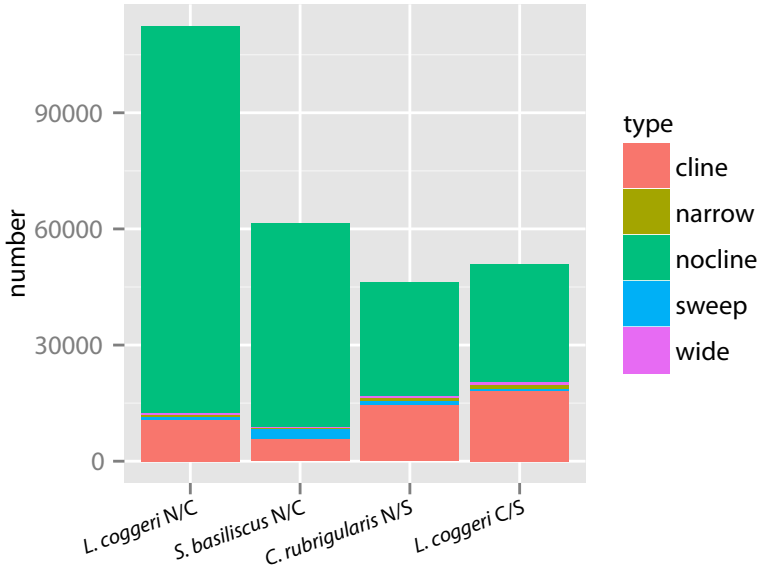

Figure B.14: Type of clines inferred at those SNPs that passed through filtering.

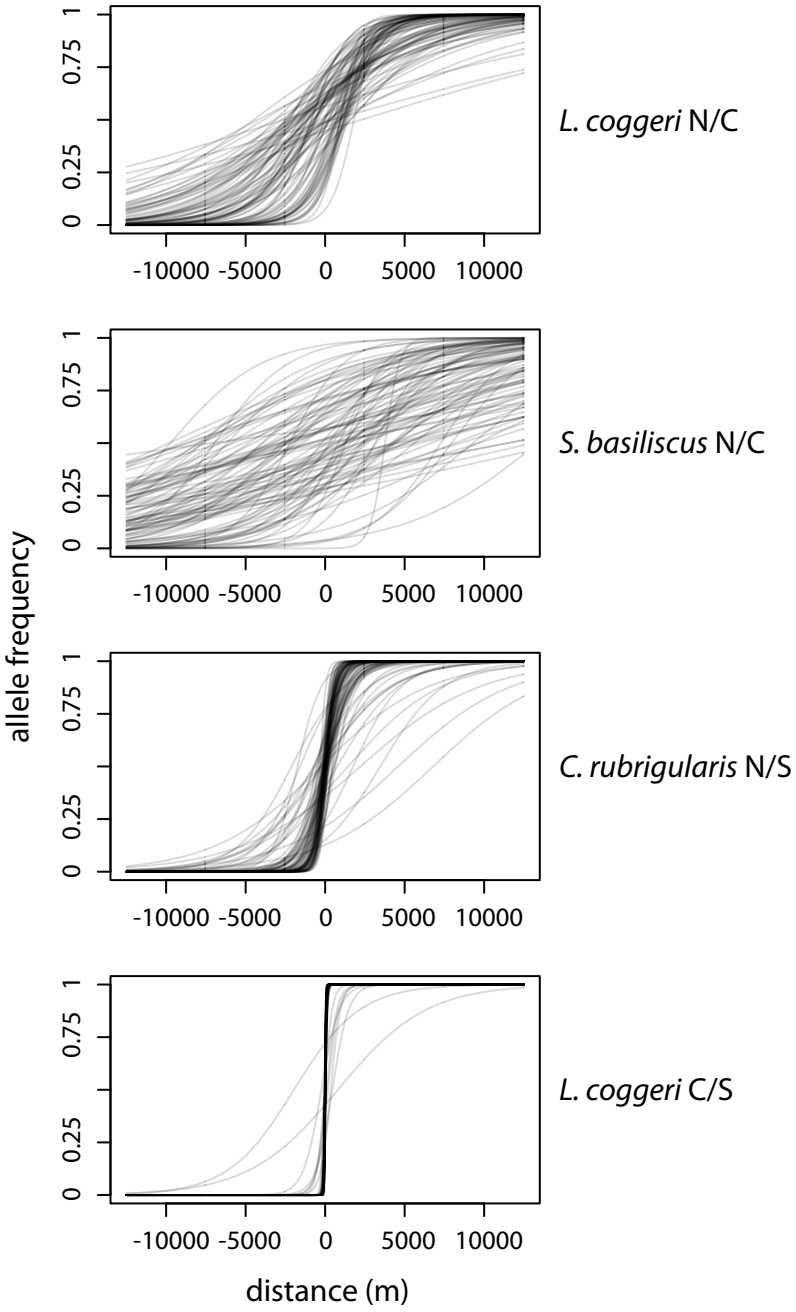

Figure B.15: A random sampling of clines ($N = 100$) for the four contact zones in this study, with lineage-pairs arranged from least divergent to most divergent from top to bottom.

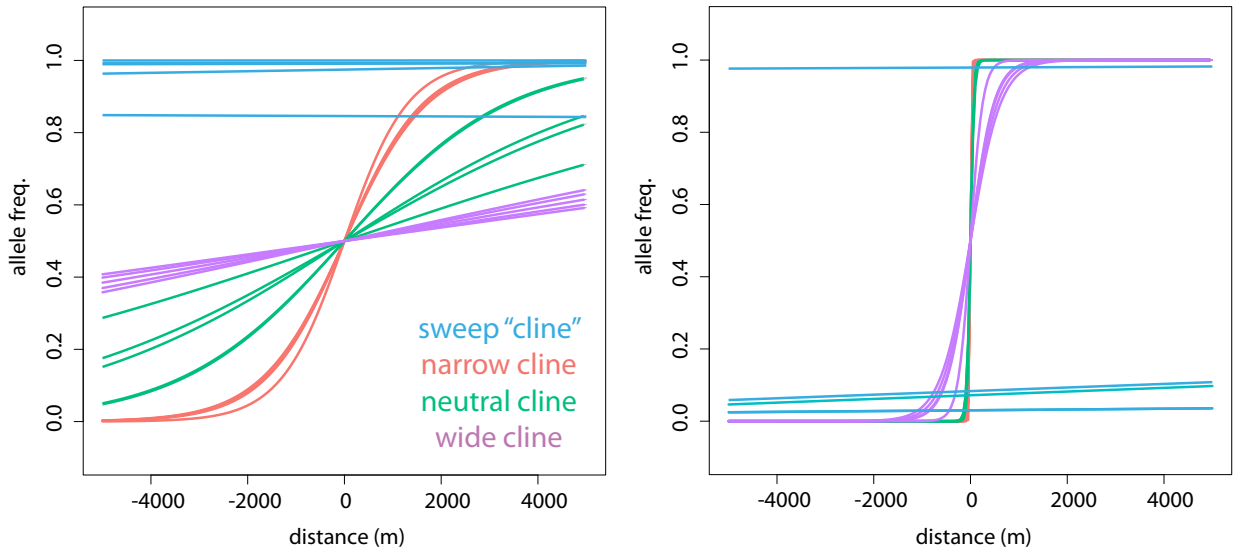

Figure B.16: Outlier types, illustrated here by five randomly chosen examples from *Lampropholis coggeri* N/C (the least divergent lineage-pair) and *L. coggeri* C/S (the most divergent lineage-pair).

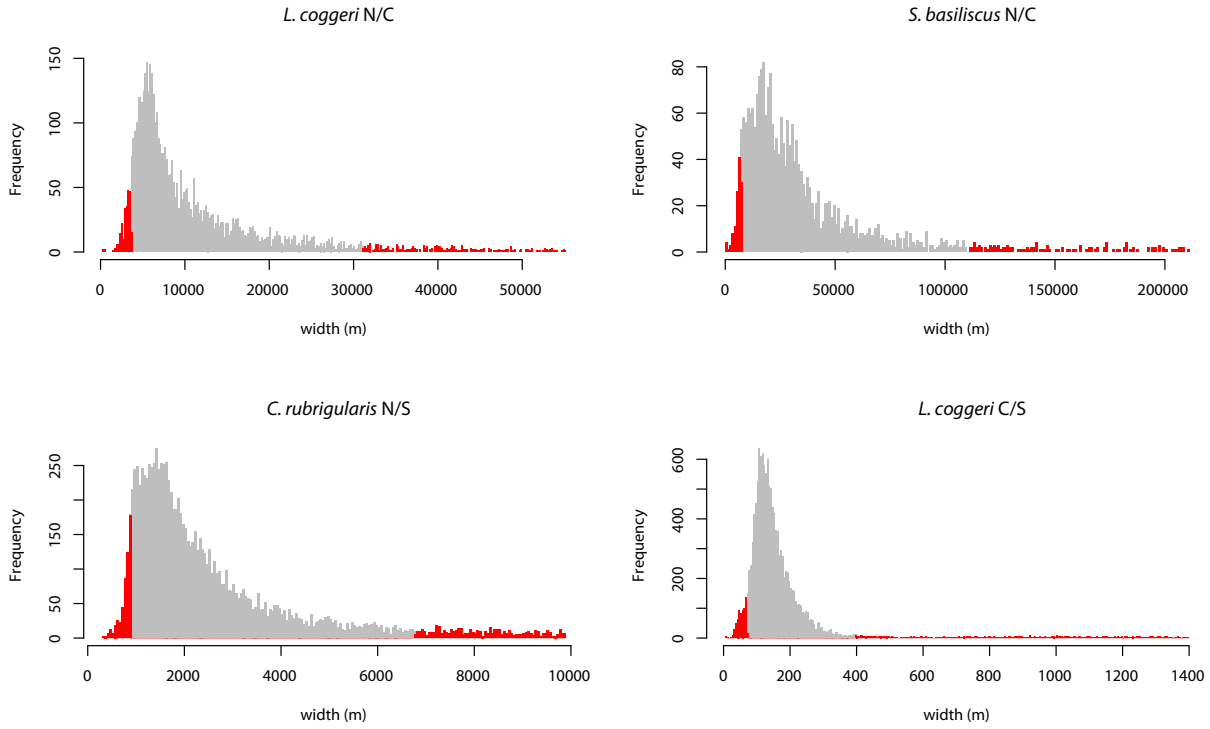

Figure B.17: Frequency histograms for width of clines in each hybrid zone; clines identified as ″narrow″ (left side of distribution) and ″wide″ outliers (right side of distribution) highlighted in red.

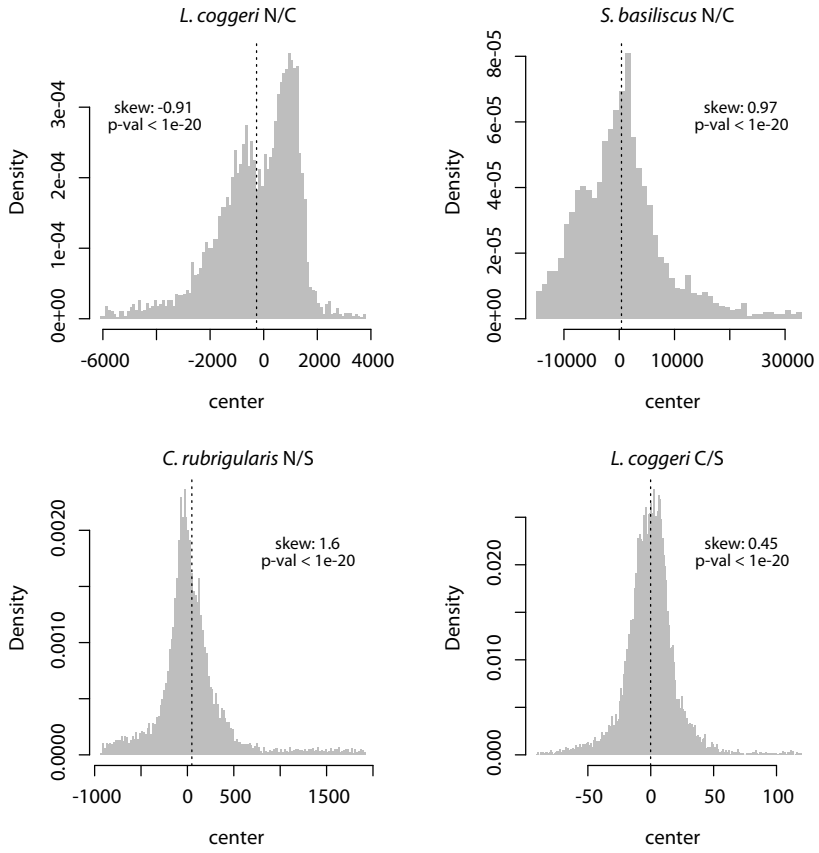

Figure B.18: Distributions of cline center values for each contact zone, shifted so that the median of the distribution is centered at zero. The dotted black line marks the mean of the distribution.

B.4 Supplemental Tables

contact	population name	sample size	latitude	longitude	transect location (m)
C. rubrigularis N/S	pop1	16	-17.156	145.564	764
C. rubrigularis N/S	pop2	16	-17.154	145.569	1278
C. rubrigularis N/S	pop3	16	-17.153	145.578	2334
C. rubrigularis N/S	pop4	16	-17.147	145.581	2676
C. rubrigularis N/S	pop5	16	-17.148	145.586	3175
C. rubrigularis N/S	pop6	16	-17.143	145.590	3689
C. rubrigularis N/S	pop7	16	-17.133	145.616	6732
C. rubrigularis N/S	parental N	5	-16.611	145.452	NA
C. rubrigularis N/S	10 km N	16	-17.075	145.596	NA
C. rubrigularis N/S	10 km S	16	-17.205	145.679	NA
C. rubrigularis N/S	parental S	5	-17.694	145.695	NA
L. coggeri C/S	pop1	14	-17.172	145.687	0
L. coggeri C/S	pop2	16	-17.205	145.679	3730
L. coggeri C/S	pop3	16	-17.215	145.687	4764
L. coggeri C/S	pop4	16	-17.215	145.686	4803
L. coggeri C/S	pop5	16	-17.215	145.688	4829
L. coggeri C/S	pop6	16	-17.220	145.695	5319
L. coggeri C/S	pop7	16	-17.273	145.663	11295
L. coggeri C/S	parental N	5	-16.976	145.777	NA
L. coggeri C/S	10 km N	16	-17.142	145.629	NA
L. coggeri C/S	10 km S	16	-17.295	145.712	NA
L. coggeri C/S	parental S	5	-17.676	145.713	NA
L. coggeri N/C	pop1	15	-16.659	145.480	3497
L. coggeri N/C	pop2	16	-16.660	145.485	4075
L. coggeri N/C	pop3	16	-16.664	145.492	4988
L. coggeri N/C	pop4	16	-16.664	145.496	5546
L. coggeri N/C	pop5	15	-16.666	145.500	6029
L. coggeri N/C	pop6	16	-16.671	145.503	6419
L. coggeri N/C	pop7	16	-16.675	145.506	6740
L. coggeri N/C	parental N	5	-16.579	145.315	NA
L. coggeri N/C	10 km N	15	-16.617	145.458	NA
L. coggeri N/C	10 km S	16	-16.753	145.593	NA
L. coggeri N/C	parental S	5	-16.976	145.777	NA
S. basiliscus N/C	pop1	16	-17.608	145.772	0
S. basiliscus N/C	pop2	10	-17.608	145.768	679
S. basiliscus N/C	pop3	7	-17.626	145.744	4494
S. basiliscus N/C	pop4	16	-17.665	145.723	7881
S. basiliscus N/C	pop5	14	-17.655	145.717	8893
S. basiliscus N/C	pop6	8	-17.673	145.715	9414
S. basiliscus N/C	pop7	8	-17.694	145.695	12707
S. basiliscus N/C	parental N	15	-17.292	145.634	NA
S. basiliscus N/C	10 km N	5	-17.579	145.697	NA
S. basiliscus N/C	10 km S	5	-17.699	145.523	NA
S. basiliscus N/C	parental S	16	-18.199	145.849	NA

Table B.1: Summary of geographic locations and sample sizes of populations in the transect zone.

contact	number of targets	total length as designed (bp)	number of probes	total target length as assembled (bp)
C. rubrigularis N/S	3224	1.86e6	9.70e5	3.02e6
L. coggeri C/S	3333	1.83e6	9.57e5	3.08e6
L. coggeri N/C	2889	1.81e6	9.69e5	3.62e6
S. basiliscus N/C	2870	1.82e6	9.68e5	2.95e6

Table B.2: Summary of exome capture array designs and resulting assemblies.

contact	population name	raw data (bp)	cleaned data (bp)	% of raw data kept	avg. quality score (post-cleanup)	specificity	avg. exonic cov.	avg. mtDNA cov.
C. rubrigularis N/S	10 km N	4.8e9	2.2e9	46.7%	35	62.9%	351	1026
C. rubrigularis N/S	10 km S	3.5e9	1.6e9	46.4%	35	65.4%	267	260
C. rubrigularis N/S	pop1	2.9e9	1.3e9	45.4%	35	67.8%	224	943
C. rubrigularis N/S	pop2	6.8e9	3.2e9	46.8%	35	62.8%	488	1057
C. rubrigularis N/S	pop3	5.9e9	2.8e9	47.4%	35	62.2%	433	1055
C. rubrigularis N/S	pop4	6.5e9	3.0e9	46.1%	35	62.6%	463	1005
C. rubrigularis N/S	pop5	8.0e9	3.8e9	47.9%	35	65.8%	545	1047
C. rubrigularis N/S	pop6	5.4e9	2.5e9	46.3%	35	58.2%	410	478
C. rubrigularis N/S	pop7	1.0e10	4.8e9	46.4%	35	57.4%	665	528
L. coggeri C/S	10 km N	2.2e9	1.2e9	55.3%	36	69.9%	381	7792
L. coggeri C/S	10 km S	2.5e9	1.4e9	55.3%	36	72.5%	325	7680
L. coggeri C/S	pop1	6.1e9	3.3e9	54.2%	36	65.8%	285	7721
L. coggeri C/S	pop2	5.4e9	3.0e9	55.7%	36	64.6%	265	7682
L. coggeri C/S	pop3	5.5e9	3.1e9	55.4%	36	68.3%	273	7781
L. coggeri C/S	pop4	4.3e9	2.4e9	55.9%	36	68.8%	505	7778
L. coggeri C/S	pop5	5.3e9	2.9e9	55.2%	36	71.1%	743	7788
L. coggeri C/S	pop6	4.0e9	2.2e9	55.0%	36	67.7%	198	7691
L. coggeri C/S	pop7	1.4e9	8.1e8	58.6%	36	75.2%	514	7681
L. coggeri N/C	10 km N	4.9e9	2.9e9	58.6%	36	65.9%	216	1011
L. coggeri N/C	10 km S	4.3e9	2.5e9	57.8%	36	65.2%	248	7296
L. coggeri N/C	pop1	3.3e9	2.0e9	59.3%	36	70.8%	534	2557
L. coggeri N/C	pop2	3.6e9	2.0e9	57.6%	36	63.6%	478	6370
L. coggeri N/C	pop3	3.5e9	2.0e9	58.2%	36	66.8%	506	2539
L. coggeri N/C	pop4	6.8e9	3.8e9	56.7%	36	66.3%	408	7253
L. coggeri N/C	pop5	1.0e10	5.8e9	56.6%	36	71.7%	485	7336
L. coggeri N/C	pop6	2.3e9	1.4e9	59.0%	36	65.6%	381	7383
L. coggeri N/C	pop7	7.1e9	4.1e9	57.8%	36	63.4%	149	7266
S. basiliscus N/C	10 km N	4.8e9	2.7e9	55.3%	36	65.9%	424	7459
S. basiliscus N/C	10 km S	5.5e9	2.9e9	53.9%	36	67.4%	478	7476
S. basiliscus N/C	pop1	3.0e9	1.7e9	56.3%	36	67.4%	272	5272
S. basiliscus N/C	pop2	2.0e9	1.1e9	55.1%	36	66.2%	197	6392
S. basiliscus N/C	pop3	5.8e9	3.2e9	55.5%	36	65.5%	504	7562
S. basiliscus N/C	pop4	4.9e9	2.7e9	54.9%	36	65.6%	437	7595
S. basiliscus N/C	pop5	2.5e9	1.4e9	55.3%	36	71.1%	247	6331
S. basiliscus N/C	pop6	6.7e9	3.8e9	56.8%	36	71.8%	609	7480
S. basiliscus N/C	pop7	8.5e9	4.7e9	55.4%	36	68.2%	769	7529

Table B.3: Summary of data collected, coverage, and specificity across sequenced populations.

contact	SNPs pre-filtering	SNPs post-filtering	non-coding SNPs	non-synonymous SNPs	synonymous SNPs
C. rubrigularis N/S	112098	44505	17290	9062	18153
L. coggeri C/S	129153	49354	16397	10291	22666
L. coggeri N/C	242578	79192	29388	15364	34440
S. basiliscus N/C	180056	57007	15098	13192	28717

Table B.4: Summary of single nucleotide polymorphisms (SNPs) identified. Non-coding SNPs are those in intronic and untranslated regions.

contact	clines fit	sweep outliers	narrow outliers	wide outliers
C. rubrigularis N/S	14691	826	735	735
L. coggeri C/S	18081	638	904	906
L. coggeri N/C	10623	927	531	530
S. basiliscus N/C	5836	2621	292	290

Table B.5: Summary of clines fit.

Appendix C

Supplementary Information for Chapter 4

C.1 Supplementary Figures

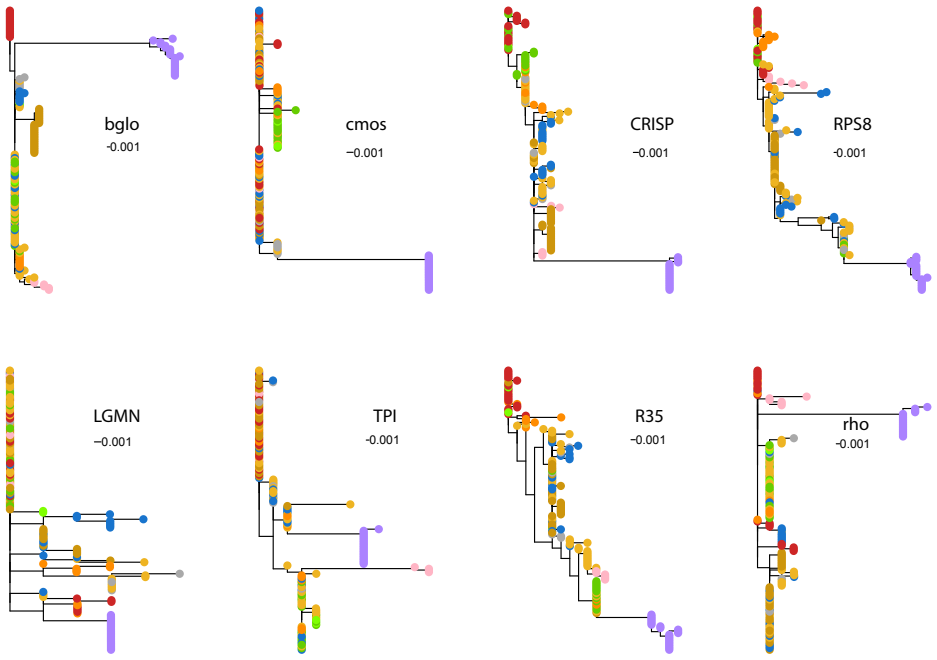

Figure C.1: Gene trees for eight nuclear genes for *Saproscincus basiliscus* and *S. lewisi* based on individual haplotypes, as inferred by maximum-likelihood in RAxML. Color scheme follows that used in Figure 1.

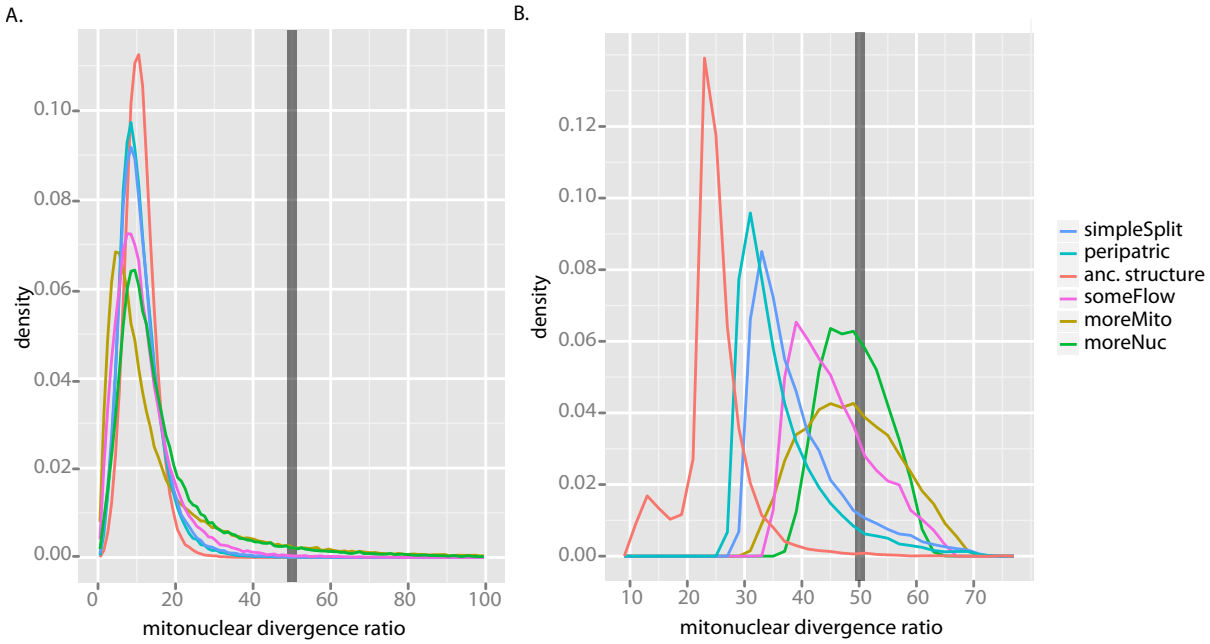

Figure C.2: Expected distribution of mito-nuclear divergence ratios for all of the modelled scenarios across the complete parameter space, A. before fitting and B. after fitting. The mito-nuclear divergence ratio found in this study is outlined in darker grey.

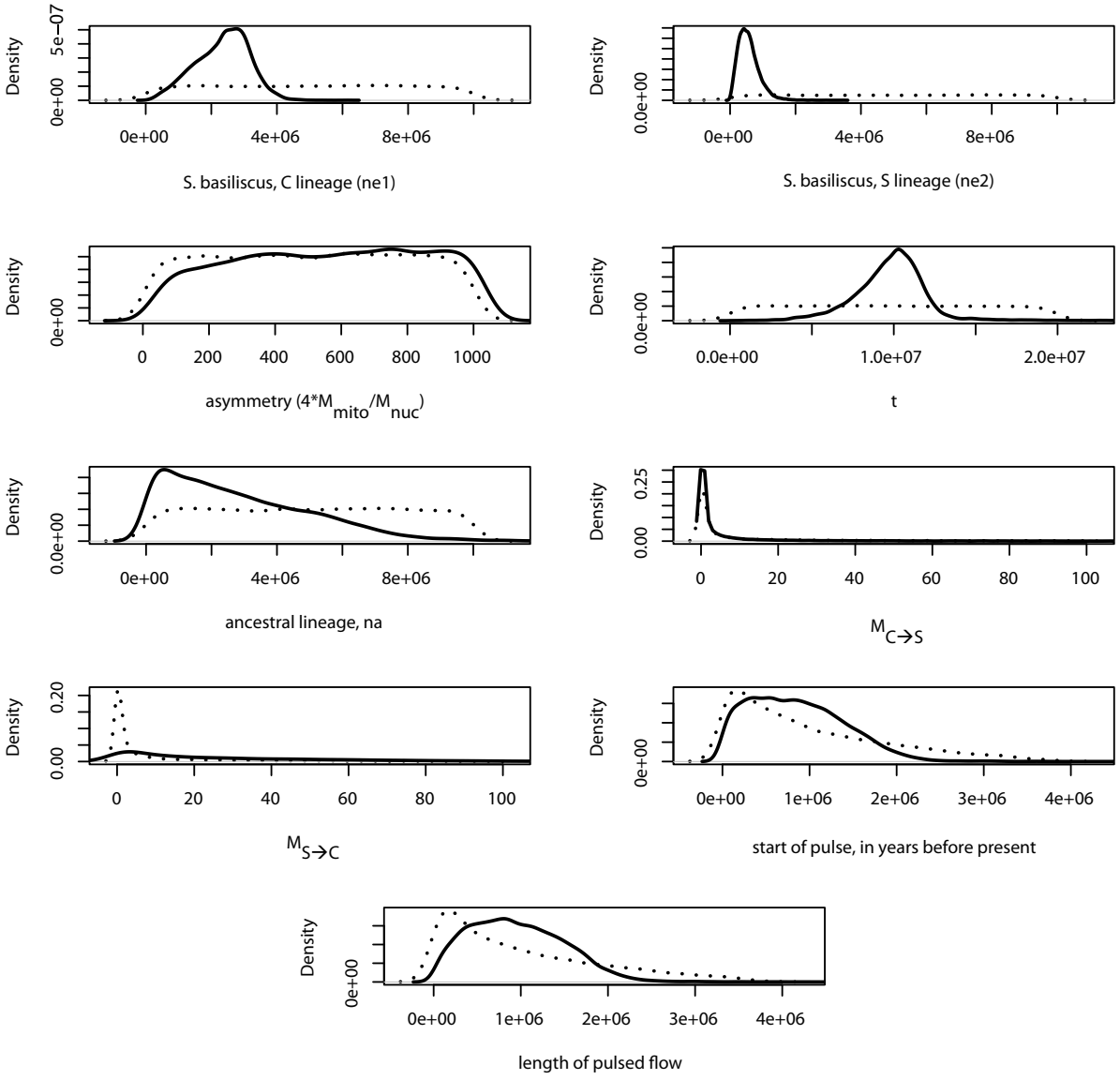

Figure C.3: Prior (shown by dotted line) and posterior (shown by bold black line) probability distributions for parameters of the most likely inferred model (model 5), in which there is pulsed introgression with more mitochondrial than nuclear gene flow.

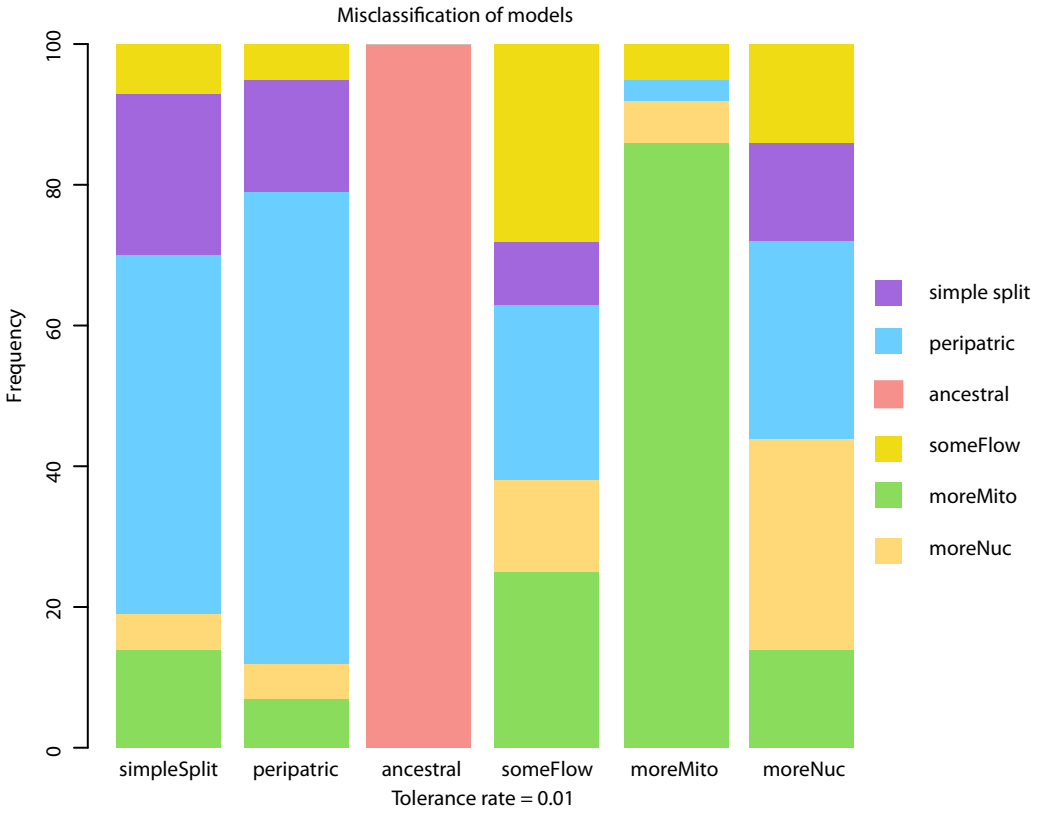

Figure C.4: Results from 100 pseudo-observed data sets, showing the frequency of misclassification among the models simulated for the ABC analysis.

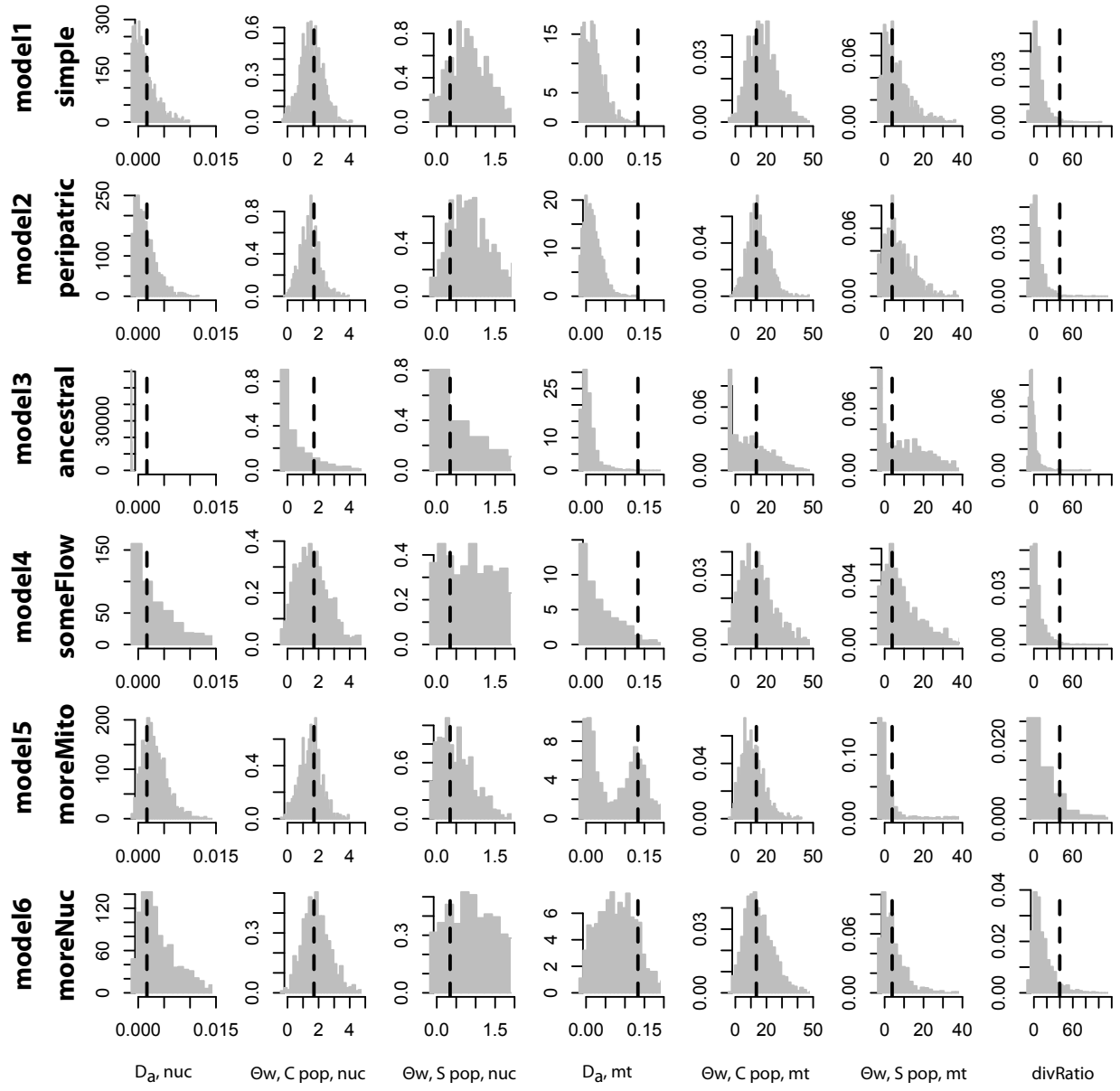

Figure C.5: Posterior predictive results for all six models across all seven summary statistics. Dashed black lines reflect true value of summary statistic for the empirical data. Some graphs cropped for ease of visibility.

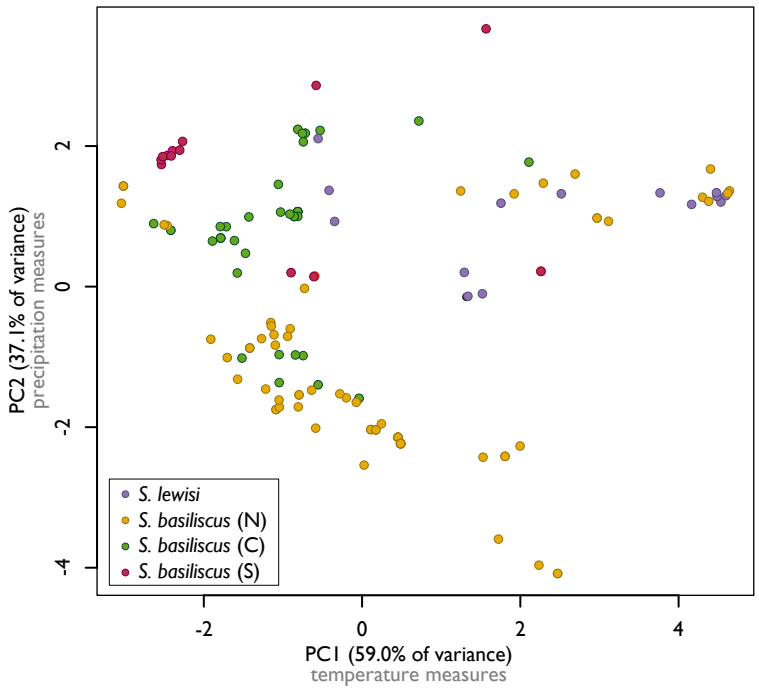

Figure C.6: PCA of climatic variables (Bioclim) grouped by mitochondrial lineage; PCA performed using prcomp in R. Color scheme follows that used in Figure 1.

C.2 Supplementary Tables

SampleID	Mito Seq	Nuc Seq	Mito Type	Bioregion	Latitude	Longitude
AA1054	yes	no	C	LE	-18.63355	145.873
AA1078	yes	no	C	LE	-18.63355	145.873
AA1091	yes	no	C	CC	-17.7767	145.555517
AA1098	yes	no	C	LE	-18.626866	145.876
BP185	yes	no	N	AU-EE	-17.217	145.717
BP935	yes	no	N	LU	-17.15200016	145.55732647
CJS1105	yes	no	N	MT-S	-17.0916661	145.8783332
CJS1106	yes	no	N	MT-S	-17.0916661	145.8783332
CJS1107	yes	no	N	MT-S	-17.0916661	145.8783332
CM13	yes	no	N	AU-WR	-17.4393191	145.85802863
CM15	yes	no	N	CC	-17.4393191	145.85802863
CONX1093	yes	yes	S	EU	-19.47726321	146.9863644
CONX1508	yes	no	N	AU-HR	-17.419093	145.837212
CONX1511	yes	no	N	BM	-16.79257934	145.648749
CONX1560	yes	no	N	AU-HR	-17.419093	145.837212
CONX1570	yes	no	N	BM	-16.84452	145.64165
CONX1571	yes	no	N	AU-HR	-17.419093	145.837212
CONX1923	yes	no	N	BM	-16.82672377	145.6474281
CONX1923	yes	yes	N	BM	-16.82672377	145.6474281
CONX582	yes	no	C	CC	-17.7767	145.555517
Elliot1	yes	yes	S	EU	-19.47726321	146.9863644
Elliot1	yes	yes	S	EU	-19.47726321	146.9863644
NSF107	yes	no	N	AU-KO	-17.609193	145.772248

SampleID	Mito Seq	Nuc Seq	Mito Type	Bioregion	Latitude	Longitude
NSF111	yes	no	N	AU-KO	-17.609193	145.772248
NSF156	yes	no	N	CC	-17.86911586	146.0671699
NSF190	yes	no	N	AU-KO	-17.609193	145.772248
NSF54	yes	no	N	BM	-16.73142476	145.56873628
NSF95b	yes	no	SL	FUS	-15.80077343	145.30805244
S3324	yes	no	S	EU	-19.455117	146.955233
S3831	yes	no	C	AU-WR	-17.607714	145.771707
SEW00002	yes	no	C	KU	-18.18893517	145.7430523
SEW00004	yes	no	C	KU	-18.20429394	145.759158
SEW00064	yes	no	C	AU-KO	-17.7026933	145.52683015
SEW00090	yes	no	C	AU-KO	-17.7467055	145.53227959
SEW00094	yes	no	C	AU-KO	-17.74641775	145.5336776
SEW00103	yes	no	C	AU-KO	-17.7467285	145.5293183
SEW00134	yes	no	C	AU-KO	-17.73954783	145.5663896
SEW00146	yes	no	C	AU-KO	-17.70066965	145.5244897
SEW00153	yes	no	C	AU-KO	-17.69954396	145.5238012
SEW00207	yes	no	C	KU	-18.19161507	145.7493582
SEW00246	yes	no	C	KU	-18.20755576	145.7618865
SEW00286	yes	no	C	KU	-18.20735781	145.7606209
SEW00299	yes	no	C	KU	-18.22828306	145.8113783
SEW00308	yes	no	C	KU	-18.22828306	145.8113783
SEW00380	yes	no	C	AU-KO	-17.70193123	145.52522419
SEW00409	yes	no	C	AU-KO	-17.70615356	145.5267359
SEW00507	yes	no	S	SU	-19.00315739	146.2025699
SEW00530	yes	no	S	SU	-19.01432954	146.2094526
SEW00537	yes	no	C	AU-KO	-17.74849494	145.5238814
SEW00544	yes	no	C	AU-KO	-17.74739204	145.5274268
SEW00602	yes	yes	S	SU	-19.0113899	146.1737139
SEW00604	yes	no	S	SU	-19.01432344	146.2080274
SEW00618	yes	no	S	SU	-19.00315739	146.2025699
SEW00671	yes	no	S	SU	-18.93975617	146.1482446
SEW00672	yes	no	S	SU	-18.93975617	146.1482446
SEW00686	yes	no	S	SU	-18.93280954	146.1429618
SEW00687	yes	no	S	SU	-18.93280954	146.1429618
SEW00763	yes	no	N	AU-EE	-17.37413903	145.7179621
SEW00764	yes	no	N	AU-EE	-17.37413903	145.7179621
SEW00865	yes	no	N	AU-EE	-17.37939275	145.7628676
SEW00867	yes	no	N	AU-EE	-17.38038636	145.7613267
SEW00891	yes	no	C	KU	-18.16899315	145.7255172
SEW00896	yes	no	C	KU	-18.16898374	145.7240993
SEW00914	yes	no	C	KU	-18.21463992	145.7972869
SEW00941	yes	no	C	KU	-18.17053157	145.6337713
SEW00944	yes	yes	S	SU	-18.94557696	146.192919
SEW00970	yes	no	N	TL	-16.26076434	145.44197545
SEW00971	yes	no	N	TL	-16.18733066	145.41198963
SEW00976	yes	no	SL	TL	-15.96592289	145.3565003
SEW01199	yes	yes	C	KU	-18.16639336	145.7286651
SEW01200	yes	yes	C	KU	-18.16639336	145.7286651
SEW01201	yes	no	C	KU	-18.16639336	145.7286651
SEW01202	yes	no	C	KU	-18.16639336	145.7286651
SEW01253	yes	no	C	KU	-18.20718751	145.7636384
SEW01260	yes	no	S	SU	-18.94566319	146.1919688
SEW01263	yes	no	S	SU	-18.94488885	146.190548
SEW01276	yes	yes	S	SU	-18.93573731	146.1647133
SEW01315	yes	yes	S	HIU	-18.41273076	146.2808844
SEW01338	yes	no	S	HIU	-18.41419035	146.2820808
SEW01354	yes	yes	S	HIU	-18.41419035	146.2820808
SEW01365	yes	no	S	HIU	-18.41527554	146.2822562
SEW01413	yes	no	S	HIU	-18.36104827	146.2468732

SampleID	Mito Seq	Nuc Seq	Mito Type	Bioregion	Latitude	Longitude
SEW01454	yes	no	S	HIU	-18.36059899	146.2452472
SEW01524	yes	no	SL	FUN	-15.71031105	145.2631983
SEW01532	yes	no	SL	FUN	-15.70887087	145.2605143
SEW01533	yes	no	SL	FUN	-15.70887087	145.2605143
SEW01591	yes	no	N	AU-HR	-17.43279395	145.4865371
SEW01596	yes	no	N	AU-HR	-17.43279395	145.4865371
SEW01601	yes	no	N	AU-HR	-17.30021379	145.4226987
SEW01602	yes	no	N	AU-HR	-17.30021379	145.4226987
SEW01603	yes	no	N	AU-HR	-17.30141511	145.422632
SEW01613	yes	no	N	AU-HR	-17.30141511	145.422632
SEW01614	yes	no	N	AU-HR	-17.30141511	145.422632
SEW02010	yes	no	C	AU-KO	-17.70193123	145.5252242
SEW02041	yes	no	C	AU-KO	-17.70435718	145.5269863
SEW02073	yes	yes	S	SU	-19.01432954	146.2094526
SEW02097	yes	no	C	KU	-18.19300284	145.7459919
SEW02107	yes	yes	C	KU	-18.19010301	145.7447552
SEW02110	yes	yes	C	KU	-18.19010301	145.7447552
SEW02149	yes	no	C	KU	-18.2060096	145.7645544
SEW02156	yes	no	C	KU	-18.19300284	145.7459919
SEW02209	yes	no	N	AU-EE	-17.38496653	145.7436762
SEW02220	yes	no	N	AU-BF	-17.37222402	145.772798
SEW02223	yes	yes	S	SU	-18.94566319	146.1919688
SEW02337	yes	no	N	TU	-16.10239947	145.33158153
SEW02358	yes	no	N	TU	-16.10239947	145.33158153
SEW02390	yes	yes	N	TU	-16.10239947	145.3315815
SEW02903	yes	yes	SL	TL	-16.14070433	145.5418915
SEW02923	yes	yes	SL	TL	-16.07100738	145.4609778
SEW02925	yes	no	SL	TL	-16.03966553	145.4597888
SEW03075	yes	yes	SL	FUN	-15.70887087	145.2605143
SEW03079	yes	yes	SL	FUN	-15.70887087	145.2605143
SEW03123	yes	no	SL	FUN	-15.71428072	145.2771873
SEW03125	yes	yes	SL	FUN	-15.71428072	145.2771873
SEW03173	yes	no	N	AU-CE	-17.35788151	145.5853862
SEW03195	yes	no	N	AU-HR	-17.42582472	145.48507898
SEW03567	yes	no	N	AU-WR	-17.60389203	145.6316353
SEW03568	yes	no	N	AU-WR	-17.60369044	145.6299312
SEW03569	yes	yes	N	AU-WR	-17.60369044	145.6299312
SEW03571	yes	no	N	AU-WR	-17.60369044	145.6299312
SEW03572	yes	no	N	AU-WR	-17.6029758	145.6324338
SEW03627	yes	no	C	AU-KO	-17.70435718	145.52698628
SEW03646	yes	no	C	AU-WR	-17.67066916	145.7167358
SEW03662	yes	no	N	AU-WR	-17.65453809	145.71675588
SEW03673	yes	no	N	AU-WR	-17.646815	145.716994
SEW03808	yes	no	N	AU-WR	-17.60822594	145.76978113
SEW03810	yes	no	N	AU-WR	-17.607714	145.771707
SEW03824	yes	no	N	AU-WR	-17.60771377	145.77170705
SEW03831	yes	no	N	AU-WR	-17.60771377	145.77170705
SEW03833	yes	no	N	AU-WR	-17.60771377	145.77170705
SEW03835	yes	no	N	AU-WR	-17.60771377	145.77170705
SEW03842	yes	no	N	AU-WR	-17.60782846	145.77416589
SEW04059	yes	no	N	AU-WR	-17.60389203	145.6316353
SEW04074	yes	no	C	AU-WR	-17.65453809	145.7167559
SEW04085	yes	yes	C	AU-WR	-17.67300424	145.7145039
SEW04273	yes	no	N	TU	-16.17365844	145.36560323
SEW04280	yes	yes	N	AU-CE	-17.29164954	145.633634
SEW04309	yes	no	N	AU-CE	-17.29164954	145.633634
SEW04314	yes	no	N	AU-CE	-17.29164954	145.633634
SEW04316	yes	no	N	AU-CE	-17.29164954	145.633634
SEW04317	yes	no	N	AU-CE	-17.29164954	145.633634

SampleID	Mito Seq	Nuc Seq	Mito Type	Bioregion	Latitude	Longitude
SEW04320	yes	no	N	AU-CE	-17.29164954	145.633634
SEW04412	yes	yes	N	TL	-16.22568908	145.4356663
SEW04420	yes	yes	N	TL	-16.23876843	145.4323075
SEW04425	yes	no	N	CC	-17.27028595	145.9000977
SEW04426	yes	no	N	CC	-17.27028595	145.9000977
SEW04446	yes	yes	N	CC	-17.27028595	145.9000977
SEW04495	yes	yes	C	LE	-18.56287121	145.7781074
SEW04498	yes	no	C	LE	-18.56287121	145.7781074
SEW04499	yes	no	C	LE	-18.56287121	145.7781074
SEW04500	yes	yes	C	LE	-18.56287121	145.7781074
SEW04502	yes	no	C	LE	-18.56287121	145.7781074
SEW04503	yes	no	C	LE	-18.56287121	145.7781074
SEW04509	yes	yes	C	LE	-18.64804384	145.8743075
SEW04510	yes	yes	C	LE	-18.64804384	145.8743075
SEW04511	yes	no	C	LE	-18.64804384	145.87430752
SEW04521	yes	no	C	LE	-18.60066521	145.7997491
SEW04522	yes	yes	C	LE	-18.60066521	145.7997491
SEW04523	yes	yes	C	LE	-18.60066521	145.7997491
SEW04526	yes	no	C	LE	-18.60066521	145.79974907
SEW04527	yes	yes	C	LE	-18.60066521	145.7997491
SEW04528	yes	yes	C	LE	-18.60066521	145.7997491
SEW04540	yes	yes	C	IL	-18.41659035	145.9443368
SEW04549	yes	no	N	AU-WR	-17.60041464	145.75773438
SEW04550	yes	no	N	AU-WR	-17.60041464	145.75773438
SEW04552	yes	no	N	AU-WR	-17.60041464	145.75773438
SEW04553	yes	yes	N	AU-WR	-17.60041464	145.7577344
SEW04560	yes	no	N	AU-WR	-17.60041464	145.75773438
SEW04663	yes	no	N	AU-CE	-17.270286	145.900098
SEW04709	yes	no	N	AU-EE	-17.37596268	145.72898
SEW04733	yes	no	N	CC	-17.27028595	145.9000977
SEW06069	yes	yes	C	AU-KO	-17.70066965	145.5244897
SEW06140	yes	yes	N	CC	-17.71884301	145.8582907
SEW06142	yes	no	N	CC	-17.71884301	145.85829071
SEW06149	yes	yes	N	CC	-17.71884301	145.8582907
SEW06164	yes	yes	C	AU-KO	-17.70066965	145.5244897
SEW06165	yes	yes	C	AU-KO	-17.70066965	145.5244897
SEW06166	yes	yes	C	AU-KO	-17.70066965	145.5244897
SEW06185	yes	yes	N	CC	-17.71884301	145.8582907
SEW06217	yes	no	N	AU-WR	-17.46244433	145.47372287
SEW06226	yes	yes	C	AU-KO	-17.70066965	145.5244897
SEW06227	yes	yes	C	AU-KO	-17.70066965	145.5244897
SEW06239	yes	no	S	SU	-19.00970563	146.2359729
SEW06240	yes	no	S	SU	-19.00970563	146.2359729
SEW06241	yes	yes	S	SU	-19.00970563	146.2359729
SEW06554	yes	yes	N	CU	-16.58640134	145.2976211
SEW06585	yes	no	N	AU-WR	-17.60822594	145.76978113
SEW06587	yes	no	N	AU-WR	-17.60822594	145.76978113
SEW06590	yes	no	N	AU-WR	-17.60822594	145.76978113
SEW06593	yes	no	N	AU-WR	-17.60834065	145.76788616
SEW06676	yes	yes	N	WU	-16.2933914	145.0550757
SEW06677	yes	yes	N	WU	-16.2933914	145.0550757
SEW06678	yes	yes	N	WU	-16.2933914	145.0550757
SEW06681	yes	no	SL	FUS	-15.798097	145.292914
SEW06682	yes	no	SL	FUS	-15.798097	145.292914
SEW06684	yes	no	SL	TL	-16.06932544	145.462103
SEW06685	yes	no	SL	TL	-16.06932544	145.462103
SEW06686	yes	no	SL	TL	-16.06932544	145.462103
SEW06687	yes	no	SL	TL	-16.06932544	145.462103
SEW06688	yes	no	SL	TL	-16.07247659	145.4629481

SampleID	Mito Seq	Nuc Seq	Mito Type	Bioregion	Latitude	Longitude
SEW06689	yes	no	SL	TL	-16.07247659	145.4629481
SEW06708	yes	yes	N	ML	-16.39489254	145.3263035
SEW06709	yes	yes	N	ML	-16.39489254	145.3263035
SEW06710	yes	yes	N	BM	-16.59619409	145.3386099
SEW06711	yes	yes	N	BM	-16.59619409	145.3386099
SEW06748	yes	no	SL	TL	-16.12400242	145.4570634
SEW06749	yes	yes	SL	TL	-16.12400242	145.4570634
SEW06750	yes	no	SL	TL	-16.12400242	145.4570634
SEW06751	yes	no	N	TL	-16.13770459	145.44109023
SEW06752	yes	no	SL	TL	-16.13530768	145.451983
SEW06753	yes	no	SL	TL	-16.06932544	145.462103
SEW06842	yes	no	N	AU-WR	-17.60771377	145.77170705
SEW06847	yes	no	N	AU-WR	-17.60822594	145.76978113
SEW06848	yes	no	N	AU-WR	-17.60929842	145.76640954
SEW06849	yes	no	N	AU-WR	-17.60834065	145.76788616
SEW06851	yes	no	N	AU-WR	-17.60929842	145.76640954
SEW07192	yes	yes	N	CC	-17.34212279	145.8715333
SEW07463	yes	no	N	AU-WR	-17.60822594	145.76978113
SEW07464	yes	yes	N	AU-WR	-17.60782846	145.7741659
SEW07467	yes	no	N	CC	-17.71700607	145.85944334
SEW07469	yes	no	C	AU-WR	-17.67300424	145.71450386
SEW07476	yes	yes	N	AU-WR	-17.65453809	145.7167559
SEW07484	yes	yes	N	BM	-16.59615922	145.3387696
SEW07510	yes	no	N	AU-WR	-17.59933189	145.63606075
SEW07518	yes	no	C	AU-WR	-17.67300424	145.71450386
SEW07519	yes	no	C	AU-WR	-17.67300424	145.71450386
SEW07523	yes	no	C	AU-WR	-17.67300424	145.71450386
SEW07561	yes	no	C	AU-WR	-17.67465474	145.71400192
SEW07673	yes	yes	S	SU	-19.01401504	146.1179481
SEW07780	yes	no	N	AU-WR	-17.60834065	145.76788616
SEW07781	yes	no	N	AU-WR	-17.60929842	145.76640954
SEW07784	yes	no	N	AU-WR	-17.57892097	145.69414024
SEW07785	yes	yes	N	AU-WR	-17.65745561	145.7193086
SEW07786	yes	no	C	AU-WR	-17.65745561	145.71930856
SEW07787	yes	no	C	AU-WR	-17.65633657	145.71677141
SEW07788	yes	yes	N	AU-WR	-17.65837891	145.7209328
SEW07789	yes	yes	N	AU-WR	-17.65645828	145.717421
SEW07790	yes	no	N	AU-WR	-17.599331	145.63606
SEW07791	yes	yes	N	AU-WR	-17.57942142	145.69739669
SEW07812	yes	yes	S	SU	-19.01490811	146.166295
SEW07814	yes	yes	S	SU	-19.00228035	146.208778
SEW07817	yes	yes	S	SU	-19.01118481	146.2244222
SEW07820	yes	yes	S	SU	-19.00300099	146.2449817
SEW07853	yes	yes	N	AU-EE	-17.215078	145.68531418
SEW07857	yes	yes	N	AU-CE	-17.25901629	145.6523658
SEW07858	yes	no	N	AU-CE	-17.25901629	145.6523658
SEW08016	yes	yes	N	AU-CE	-17.24900958	145.6306001
SEW08037	yes	no	N	AU-EE	-17.37479509	145.7429832
SEW08042	yes	yes	N	AU-EE	-17.21500597	145.6881263
SEW08047	yes	yes	N	AU-EE	-17.17218046	145.6569432
SEW08049	yes	no	N	AU-EE	-17.17470235	145.65837287
SEW08071	yes	yes	N	AU-CE	-17.25942694	145.6091822
SEW08400	yes	no	N	AU-WR	-17.645835	145.732304
SEW08401	yes	no	N	AU-WR	-17.645835	145.732304
SEW08402	yes	no	N	AU-WR	-17.645835	145.732304
SEW08403	yes	no	N	AU-WR	-17.645835	145.732304
SEW08404	yes	yes	SL	TL	-16.069117	145.462404
SEW08405	yes	yes	SL	TL	-16.174416	145.430684
SEW08406	yes	no	SL	TL	-16.174416	145.430684

SampleID	Mito Seq	Nuc Seq	Mito Type	Bioregion	Latitude	Longitude
SEW08437	yes	yes	S	HIL	-18.401477	146.32485
SEW08439	yes	yes	S	HIL	-18.401477	146.32485
SEW08440	yes	yes	S	HIL	-18.401477	146.32485
SEW08558	yes	no	C	AU-WR	-17.63499	145.630958
SEW08559	yes	no	C	AU-WR	-17.63499	145.630958
SEW08560	yes	no	C	AU-WR	-17.63499	145.630958
SEW08561	yes	no	C	AU-WR	-17.652706	145.639373
SEW08562	yes	no	C	AU-WR	-17.652706	145.639373
SEW08563	yes	yes	C	AU-WR	-17.652706	145.639373
SEW08564	yes	no	C	AU-WR	-17.652706	145.639373
SEW08565	yes	no	C	AU-WR	-17.652706	145.639373
SEW08566	yes	no	C	AU-WR	-17.652706	145.639373
SEW08567	yes	no	C	AU-WR	-17.652706	145.639373
SEW08568	yes	no	C	AU-WR	-17.668272	145.64922
SEW08569	yes	no	C	AU-KO	-17.70004	145.672186
SEW08583	yes	no	SL	TL	-16.138241	145.441993
SEW08587	yes	no	SL	TL	-16.138241	145.441993
SEW08588	yes	no	SL	TL	-16.138241	145.441993
SEW08589	yes	yes	SL	TL	-16.138349	145.44807
SEW08590	yes	no	SL	TL	-16.138349	145.44807
SEW08591	yes	no	SL	TL	-16.138349	145.44807
SEW08593	yes	no	SL	TL	-16.138349	145.44807
SEW08594	yes	no	SL	TL	-16.138349	145.44807
SEW08595	yes	no	SL	TL	-16.138349	145.44807
SEW08596	yes	no	SL	TL	-16.138349	145.44807
SEW08597	yes	yes	SL	TU	-16.076997	145.459913
SEW08598	yes	no	SL	TU	-16.076997	145.459913
SEW08599	yes	no	SL	TU	-16.075896	145.447611
SEW08609	yes	yes	N	TU	-16.076054	145.443226
SEW08611	yes	yes	N	TL	-16.22903	145.446856
SEW08612	yes	no	SL	TU	-16.170097	145.443099
SEW08613	yes	yes	SL	TU	-16.170097	145.443099
SEW08626	yes	no	N	AU-WR	-17.613102	145.757893
SWS14	yes	no	C	LE	-18.49666612	145.7649959

Table C.1: Data on sampled individuals

locus	primer sequence	annealing temperature	length	reference
B-globin	Bglo1CR 5′GCG AAC TGC ACT GYG ACA AG 3' Bglo2CR 5′GCT GCC AAG CGG GTG GTG A 3′	61°C	660 bp	Dolman and Phillips 2004
cmos	G73 5′GCG GTA AAG CAG GTG AAG AAA 3′ G74 5′ TGA GCA TCC AAA GTC TCC AAT 3′	57°C	380 bp	Saint 1998
TPI triosephosphate isomerase (intron 5)	LC5 5′TTC TAG CCT ATG AAC CAGTTT GG 3′ LC6 5′CCT CAA CTT GTC ATG AAC TTC C 3′	57°C	230 bp	Bell *et al*., 2010
CRISP cysteine-rich secretory protein	LC13 5′TGCTGTAGCCTACTGTCCTCAA 3' LC14 5′TGCTTATCATGCTCGCTAAGTT 3′	57°C	730 bp	this paper
RPS8 40S ribosomal protein S8 (intron 3)	LC17 5′CTC TTG GGC GTA AGA AAG GAG 3′ LC18 5′CCG CTC ATC GTA TTT CTT CTG 3′	57°C	670 bp	Bell *et al*., 2010
LGMN legumain precursor asparaginyl endopeptidase	LC29 5′CATTGCCTATATGTATCGTCACAA 3′ LC30 5′AtCCAGATTCACATGCTTCAAT 3′	57°C	300 bp	this paper
r35	r35F 5′ GAC TGT GGA YGA YCT GAT CAG TGT GGT GCC 3′ r35R 5′ GCC AAA ATG AGS GAG AAR CGC TTC TGA GC 3′	65°C	640 bp	Leache 2009
rho	Rho3CR 5′CCTTGCCTGGACACCCTATGCTG 3′ Rho4CR 5′ CAGGAGAGACCCTCACATTG 3′	61°C	370 bp	Dolman and Phillips 2004
ND4	ND4 5′ CACCTATGACTACCAAAAGCTCATGTAGAAGC 3′ LEU 5′ CATTACTTTTACTTGGATTTGCACCA 3′	57°C	850 bp	Arevalo *et al*, 1994

Table C.2: Loci used in this study, including their associated information.

model	parameter	prior distribution
all	generation time	1 $\frac{gen}{year}$
all	nuclear mutation rate	5e-10 $\frac{mutation}{site \cdot year}$
all	standard deviation in nuclear mutation rate	0.2
all	mitochondrial mutation rate	7.3e-9 $\frac{mutation}{site \cdot year}$
all	standard deviation in mitochondrial mutation rate	0.2
all	nuclear recombination rate	1e-8 $\frac{recombinations}{site \cdot year}$
all	length of mitochondrial locus	850 bp
all	lengths of nuclear loci	200, 200, 280, 350, 500, 550, 560, 600
all	population size of *S. basiliscus* C lineage, N_e	U$\sim$[1e4,1e7]
all	population size of *S. basiliscus* S lineage, N_e	U$\sim$[1e4,1e7]
all	population size of ancestral lineage, N_a	U$\sim$[1e4,1e7]
all	split time, τ	U$\sim$[1e3,2e7]
model 2	relative size of peripatric lineage	U$\sim$[0.001,0.5]
model 2	growth rate of peripatric lineage	$10^{U\sim[0.3,2.7]}$
model 3	migration rate between ancestral populations, M	U$\sim$[0,1.0]
model 3	length for ancestral population structure in years	U$\sim$[1e3,2e7]
models 4 - 6	length of pulsed gene flow in years	$\tau \cdot$ U$\sim$[0.0001,0.2]
models 4 - 6	start of pulsed gene flow, years before present	$\tau \cdot$ U$\sim$[0.0001,0.2]
models 4 - 6	gene flow from *S. basiliscus* C lineage to S lineage, M	$10^{U\sim[-2,2]}$
models 4 - 6	gene flow from *S. basiliscus* S lineage to C lineage, M	$10^{U\sim[-2,2]}$
model 5	asymmetry in gene flow, $\frac{M_{mito} \cdot 4}{M_{nuc}}$	U$\sim$[10,1000]
model 6	asymmetry in gene flow, $\frac{M_{nuc}}{M_{mito} \cdot 4}$	U$\sim$[10,1000]

Table C.3: Prior distributions for parameters used in simulating data sets for the Approximate Bayesian Computation (ABC) analysis.

model	Type I error	Type II error
model 1, simple split	0.48	0.638
model 2, peripatric	0.36	0.632
model 3, ancestral structure	0	0
model 4, some gene flow	0.77	0.558
model 5, more mitochondrial	0.16	0.475
model 6, more nuclear	0.77	0.597

Table C.4: Type I and Type II errors for model mis-classification based on pseudo-observed data set analysis.

Appendix D

Supplementary Information for Chapter 5

D.1 Additional Information on Simulations

To investigate the behavior of clines and disequilibrium statistics under a number of scenarios, we simulated secondary contact between two isolated lineages using the forward-time program simuPOP [273]. Here, we report additional rationale for how we designed the simulations and why we chose the parameter space we did (Fig. S5, Table S4). First, we simulated a hybrid zone as a one-dimensional chain of demes, even though species exist in two-dimensional space. This simplification was appropriate because (1) cline theory is applied to linear transects [100] and (2) one-dimensional systems generally approximate the behavior of two-dimensional systems. Second, we simulated sixty populations, as we found (1) it minimized edge effects, (2) unlike if we used fewer demes, it allowed us to look at changes in clines over longer periods of time, and (3) any greater number of demes significantly slowed down simulation speed. Third, dispersal, as used here, does not translate directly to any physical measure as it is given in terms of deme number. Thus, these simulations should apply to a wide range of organisms with varied dispersal capabilities. Fourth, fitness was characterized as a multiplicative selection model, which is most appropriate for a complex trait like fitness that likely has a polygenic model of inheritance. Further, although this model's assumption that all loci contribute evenly to fitness is biologically unrealistic, it is a necessary simplification. We simulated under the full possible range of fitness values, from no selection to nearly complete selection against hybrids. Finally, to model assortative mating, we used a "group-based model", in which individuals mate preferentially with those who are of similar ancestry [105]. Here, this effectively becomes a multilocus model of assortative mating. Most other hybrid zone models use a one-locus or two-locus assortative mating model [320]; although this model has support from biological systems [183, 323], it is likely that both the phenotypes of female preference and male traits generally have a polygenic basis. In this case, an

ancestry-based model is appropriate. As with selection, we looked at the outcomes of the model under the full range of assortative mating, from random mating to nearly complete assortative mating.

D.2 Supplementary Figures

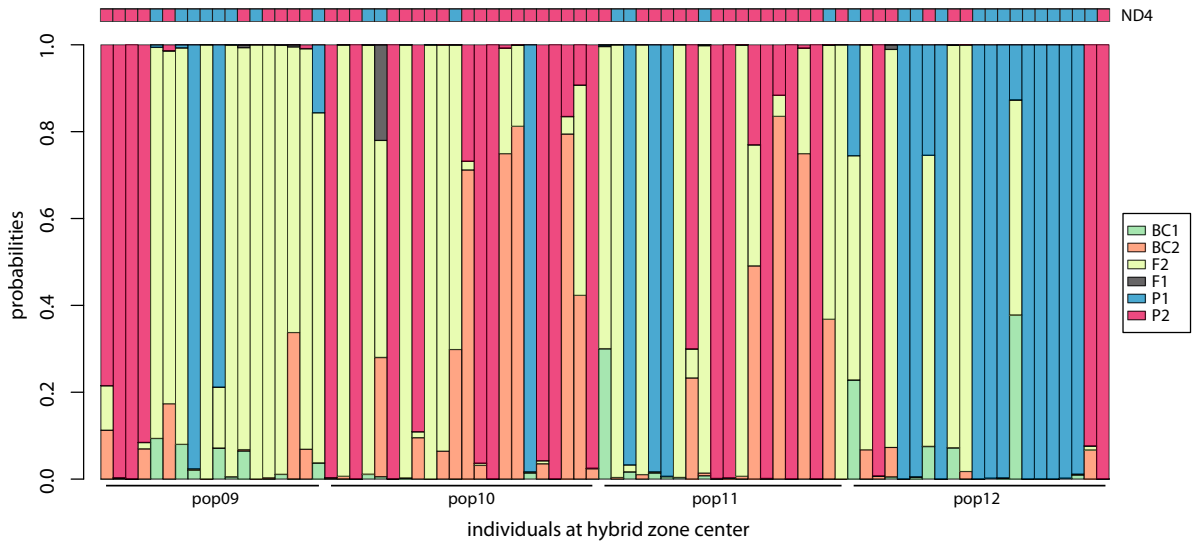

Figure D.1: NewHybrids classification of hybrid class of individuals located in the hybrid zone center.

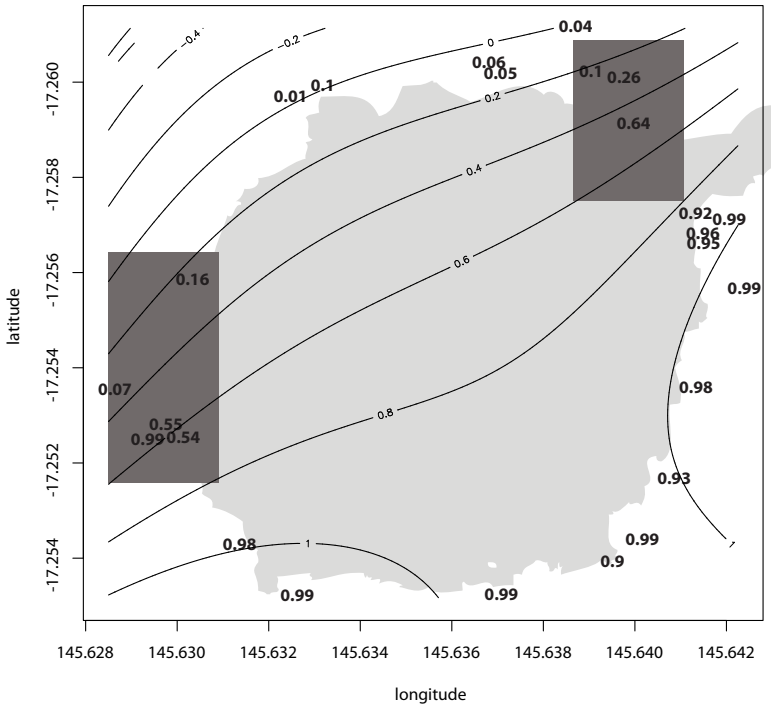

Figure D.2: A map of Lake Barrine (shown in light gray), with individual hybrid indices shown in bold. Regions where hybrids are found are shown in dark grey. Isoclines represent projected hybrid index values along the lake, as based on a generalized least squares model (implemented in R package nmle [279]).

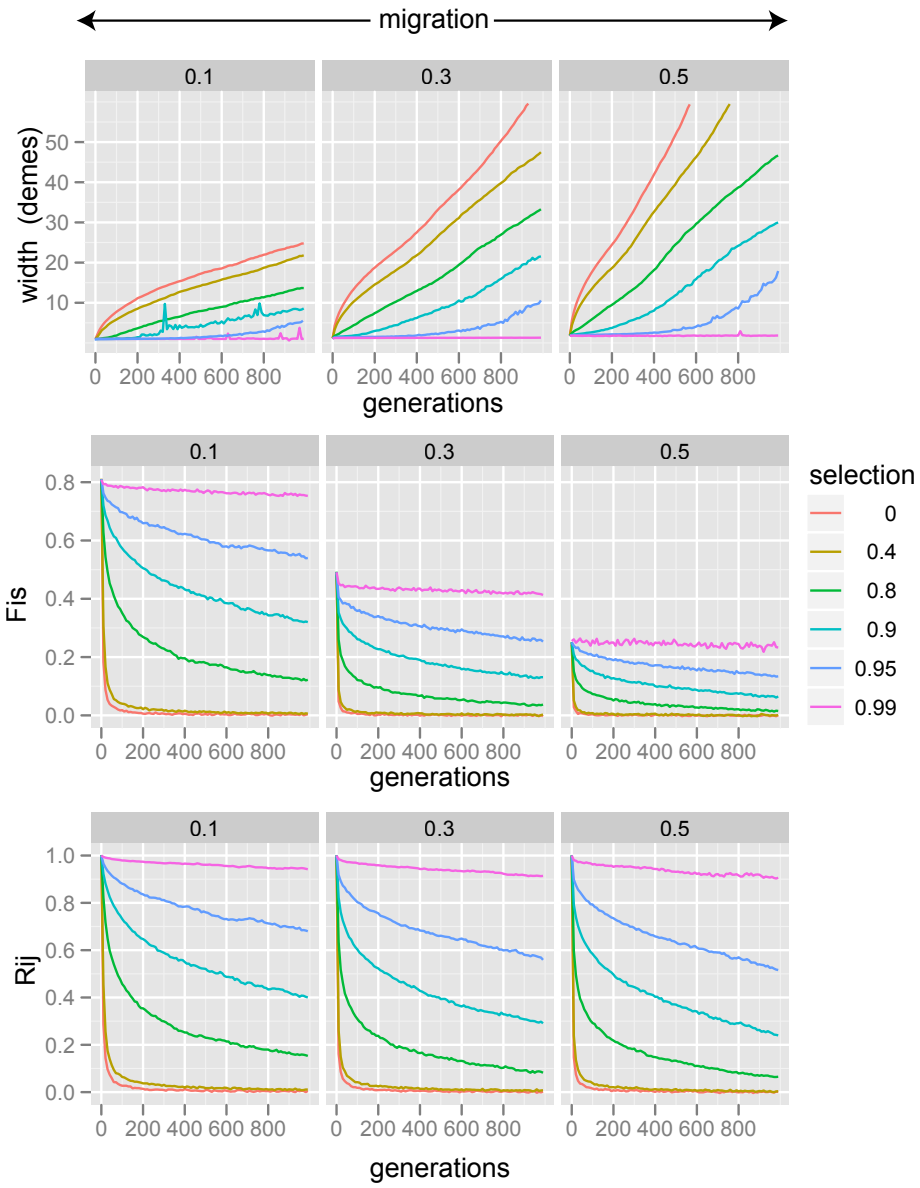

Figure D.3: Cline width (in demes) at a neutral locus, F_{IS} at a neutral locus, and R_{ij} at a neutral locus for a range of values for selection against hybrids and migration rates. Shown for no assortative mating and with 10 loci under selection.

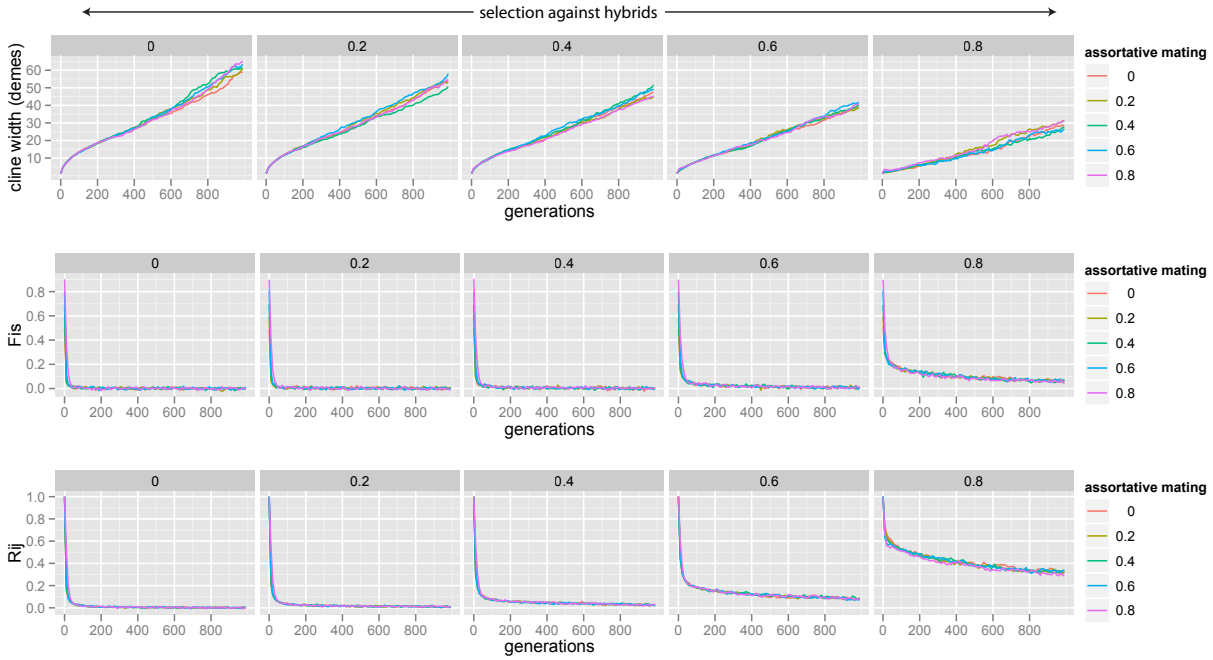

Figure D.4: Cline width (in demes) at a neutral locus, F_{IS} at a neutral locus, and R_{ij} at a neutral locus for a range of values for selection against hybrids and strength of assortative mating. Shown for migration rate of 0.3 and with 10 loci under selection.

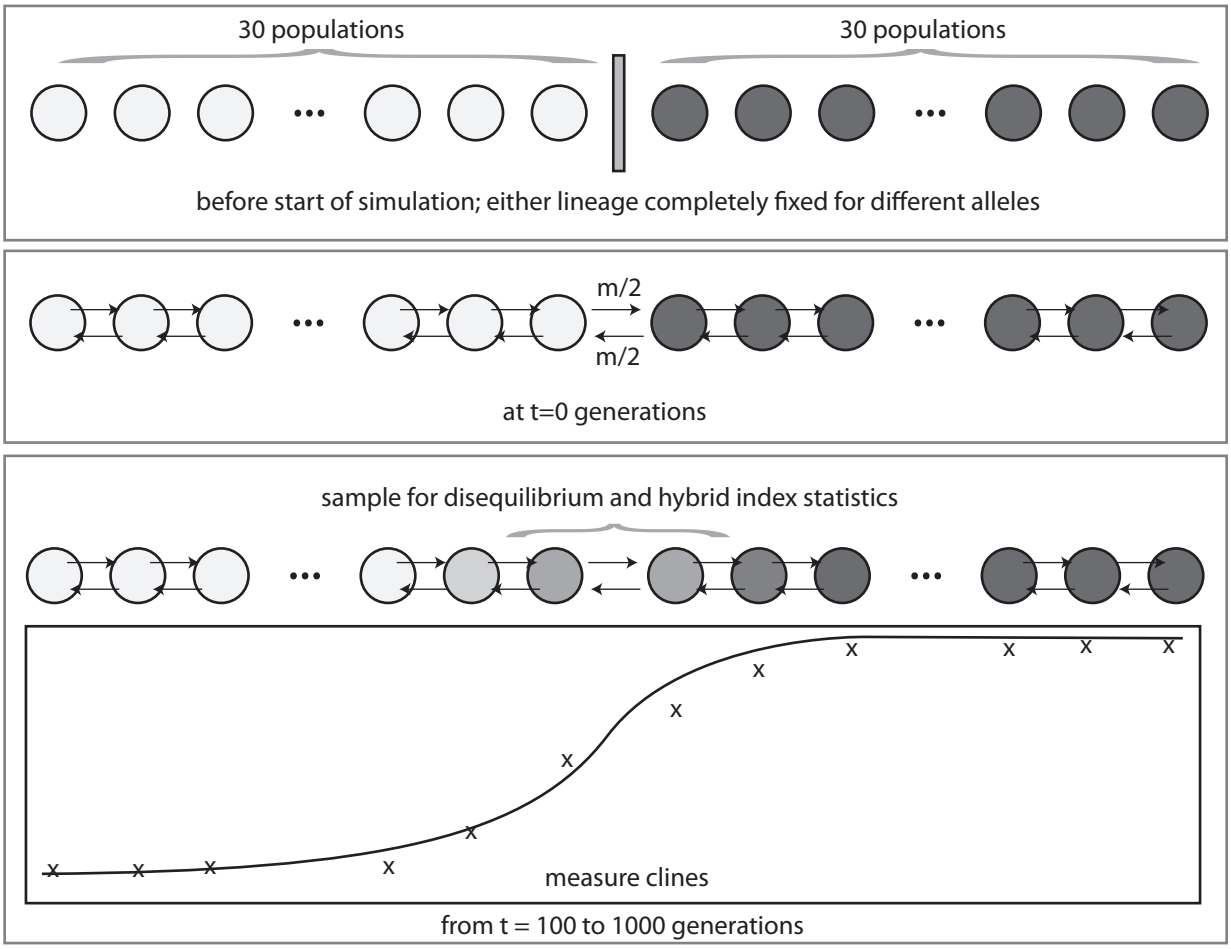

Figure D.5: A cartoon schematic of how simulations were conducted.

Name	Latitude	Longitude	N	location on transect
Lake Morris, QLD	-16.824109	145.640146	12	38 km N
Gillies Lookout, QLD	-17.171538	145.68705	17	3.8 km N
Gadgarra Forest, QLD	-17.273448	145.663125	18	2.8 km S
S Johnstone River, QLD	-17.623924	145.594004	12	45 km S
GilliesTransect_1	-17.204725	145.679049	11	0 m
GilliesTransect_2	-17.205197	145.679384	20	52 m
GilliesTransect_3	-17.209406	145.683135	19	515 m
GilliesTransect_4	-17.210666	145.683765	4	654 m
GilliesTransect_5	-17.212456	145.683809	22	852 m
GilliesTransect_6	-17.214125	145.684747	12	1036 m
GilliesTransect_7	-17.214935	145.685569	18	1041 m
GilliesTransect_8	-17.214589	145.686652	22	1086 m
GilliesTransect_9	-17.214737	145.687169	20	1102 m
GilliesTransect_10	-17.215187	145.688181	21	1151 m
GilliesTransect_11	-17.216708	145.687258	13	1320 m
GilliesTransect_12	-17.21745	145.687347	11	1402 m
GilliesTransect_13	-17.21813	145.687718	15	1477 m
GilliesTransect_14	-17.219121	145.688736	18	1586 m
GilliesTransect_15	-17.219673	145.690143	24	1646 m
GilliesTransect_16	-17.220009	145.690423	14	1683 m
GilliesTransect_17	-17.219659	145.695023	28	1701 m
LakeBarrine_1	-17.252015	145.630107	1	
LakeBarrine_2	-17.254058	145.631531	1	
LakeBarrine_3	-17.245581	145.63276	1	
LakeBarrine_4	-17.255035	145.636904	1	
LakeBarrine_5	-17.254427	145.639231	1	
LakeBarrine_6	-17.245229	145.639496	1	
LakeBarrine_7	-17.246107	145.639687	1	
LakeBarrine_8	-17.252836	145.64056	1	
LakeBarrine_9	-17.25114	145.640929	1	
LakeBarrine_10	-17.248131	145.641093	1	
LakeBarrine_11	-17.248276	145.641091	1	
LakeBarrine_12	-17.248403	145.641109	1	
LakeBarrine_13	-17.247808	145.641368	1	
LakeBarrine_14	-17.247954	145.641639	1	
LakeBarrine_15	-17.249258	145.64194	1	
LakeBarrine_16	-17.252039	145.629656	2	
LakeBarrine_17	-17.251924	145.630033	2	
LakeBarrine_18	-17.24535	145.633345	2	
LakeBarrine_19	-17.245101	145.636723	2	
LakeBarrine_20	-17.245139	145.636958	2	
LakeBarrine_21	-17.254006	145.639808	2	
LakeBarrine_22	-17.248529	145.64109	2	
LakeBarrine_23	-17.248131	145.641093	2	
LakeBarrine_24	-17.249009	145.6306	3	
LakeBarrine_25	-17.255033	145.632709	4	
LakeBarrine_26	-17.251095	145.628995	5	
LakeBarrine_27	-17.245125	145.638857	5	
LakeBarrine_28	-17.244263	145.638346	10	

Table D.1: Sampling points for this study.

D.3 Supplementary Tables

LocusID	Primer Sequence	PCR Temp.	SNP	Location/Type	R. Enz.	Cutting Pattern	Reference
ND4	ND4 5′ CACCTATGACTACCAAAAGCTCATGTAGAAGC 3′ LEU 5′ CATTACTTTTACTTGGATTTGCACCA 3′	50	G(C|A)GC	CDS, syn.	HhaI	900|(530+370)	Arevalo et al, 1994
beta-globin (intron 2)	Bglo1CR 5′GCG AAC TGC ACT GYG ACA AG 3′ Bglo2CR 5′GCT GCC AAG CGG GTG GTG A 3′	63	GCAG(C|T)	intron	ApeKI	670|(640+30)	Dolman and Phillips 2004
LC5/LC6 triosephospohote isomerase (intron 5)	LC5 5′TTC TAG CCT ATG AAC CAGTTT GG 3′ LC6 5′CCT CAA CTT GTC ATG AAC TTC C 3′	50	15 bp indel	intron	--	--	Bell et al., 2010
LC17/LC18 40S Ribosomal protein S8 (intron 3)	LC17 5′CTC TTG GGC GTA AGA AAG GAG 3′ LC18 5′CCG CTC ATC GTA TTT CTT CTG 3′	53	(A|G)C(A|G)C	intron	HhaI	630|(400+230)	Bell et al., 2010
ABHD5 abhydrolase domain containing protein 5	F 5′ ACCCCACTTGTTCTTCTCCA 3′ R 5′ TGAGTAAGCAGCTGCCAAAA 3′	60	TTC(G|A)AA	CDS, syn.	BstBI	230|(160+70)	this paper
AUTO autophagy-related protein 101	F 5′ TGAGCAGGAAAGGCAAATCT 3′ R 5′ GTGCCAGTGTGTCCTTGATG 3′	62	G(G|A)CGCC	CDS, syn.	BanI	190(170+20)	this paper
NDST2 N-deacetylase/N-sulfotransferase 2	F 5′ TCTTGGGGTTGTTTCCAGAC 3′ R 5′ CACTTGGCATTGTGAGCAGT 3′	60	CC(A|T)TGG	3′UTR, non-coding	NcoI	440|(230+210)	this paper
LEMD2 LEM domain-containing protein 2	F 5′ GTGCATTCAAGCAGACCAGA 3′ R 5′ GGCTAGCACTCTCCACCAAG 3′	60	AAGCT(T|G)	3′UTR, non-coding	HindIII	240|(140+100)	this paper
PCBD1 Pterin-4-alpha-carbinolamine dehydratase	F 5′ TCTCTCTTGGCTGTGTGGAA 3′ R 5′ TAAATCATGTGCCCCCAAAT 3′	60	GAA(T|A)C	3′UTR, non-coding	HinfI	250|(140+110)	this paper
RTN3 reticulon-3	F 5′ AACCTGTTCCAACGCAATTC 3′ R 5′ TTGAGAAAGGGGAGTTGTGG 3′	60	GAGCT(C|A)	3′UTR, non-coding	Eco53KI	440|(130+310)	this paper
SAR1 GTP-binding protein SAR1a	F 5′ TAATCACTTTGGCCCACCTC 3′ R 5′ TATCGCACAAATGCAAGAGC 3′	56	GTCTA(C|T)	3′UTR, non-coding	AccI	420|(350+70)	this paper

Table D.2: Loci used in this study, including their diagnostic SNPs and cutting patterns with listed restriction enzyme.

Parameter	Values
Number of loci under selection	2,5, 10
Assortative mating	0, 0.2, 0.4, 0.6, 0.8
Selection against hybrids	0, 0.2, 0.4, 0.6, 0.8, 0.9, 0.95, 0.99
Migration rates	0.1, 0.3, 0.5
Number of populations	60
Population size	1000
Recombination rates	0.5
Number of neutral loci	10

Table D.3: Parameters for simulation for this study.

Appendix E

Supplementary Information for Chapter 6

E.1 Supplementary Figures

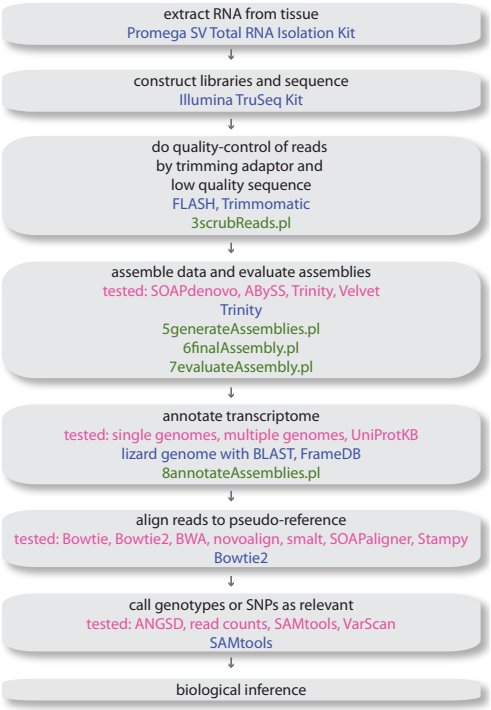

Figure E.1: Pipeline used in this work, annotated to show (1) different approaches tested [pink], (2) the approach used for the final analysis [blue], and (3) scripts used, as named in the DataDryad package [green].

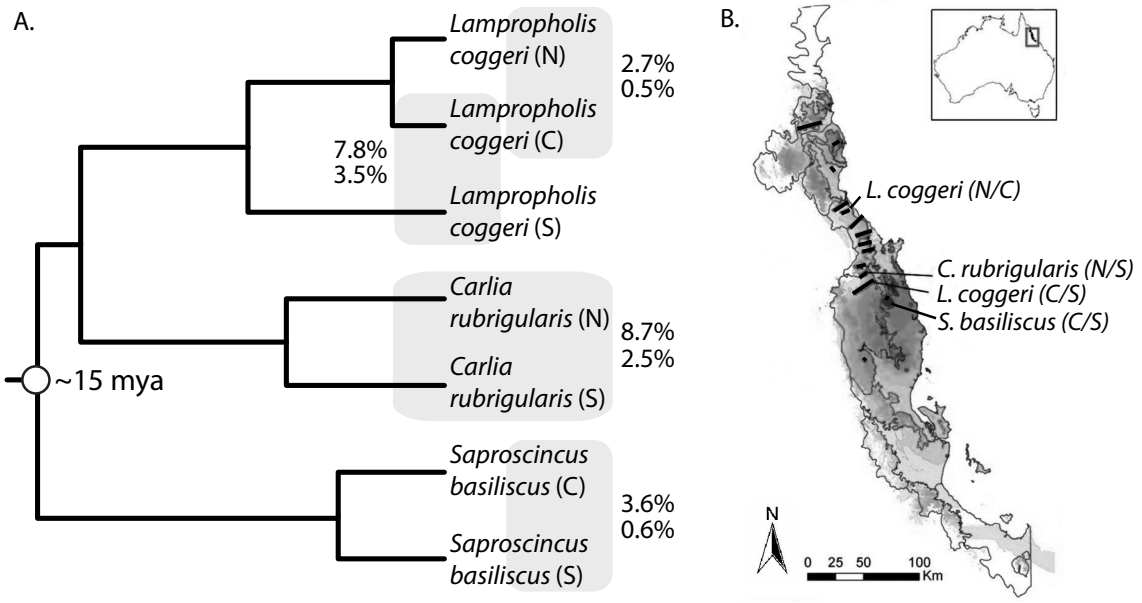

Figure E.2: A. Phylogeny of the lineages studied in this work. Boxes indicate contacts studied; the top percentage reflects the mitochondrial divergence between lineages and the bottom is nuclear. B. A map of the Australian Wet Tropics, with all identified contact zones represented by black lines. Contacts of interest in this study are labelled.

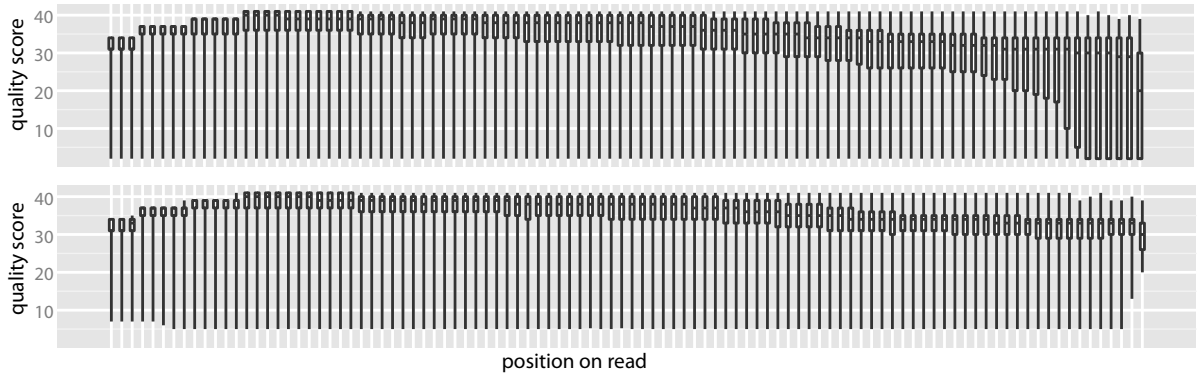

Figure E.3: Quality scores in Phred along a read; top graph shows quality prior to cleaning and filtering, bottom shows quality after cleaning.

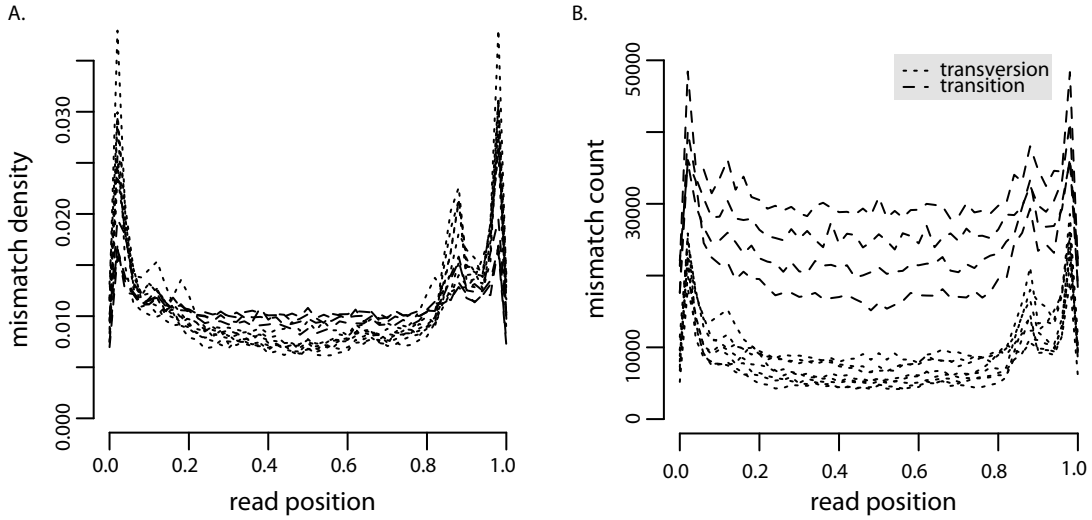

Figure E.4: Identified mismatches between reads from a randomly-selected individual and the reference sequence, A. expressed in raw numbers and B. as a density distribution.

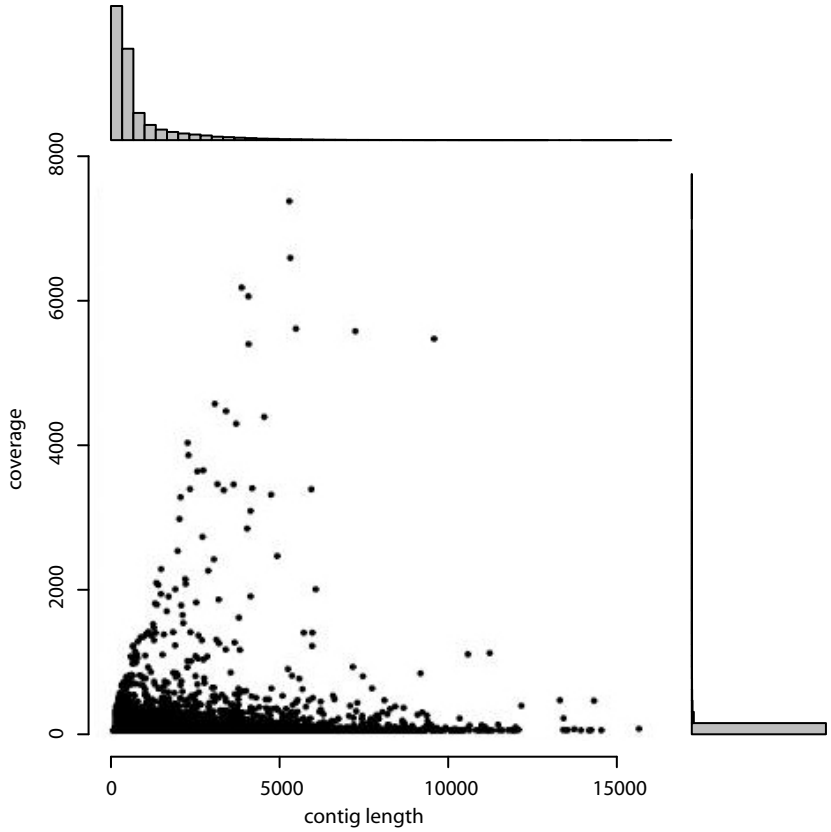

Figure E.5: Correlation between contig length and coverage for a randomly-selected final assembly.

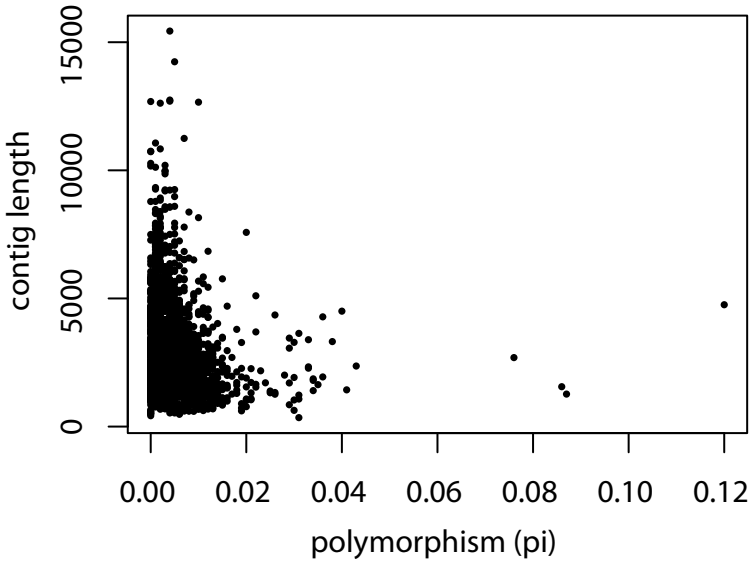

Figure E.6: Correlation between contig length and polymorphism for a randomly-selected final assembly.

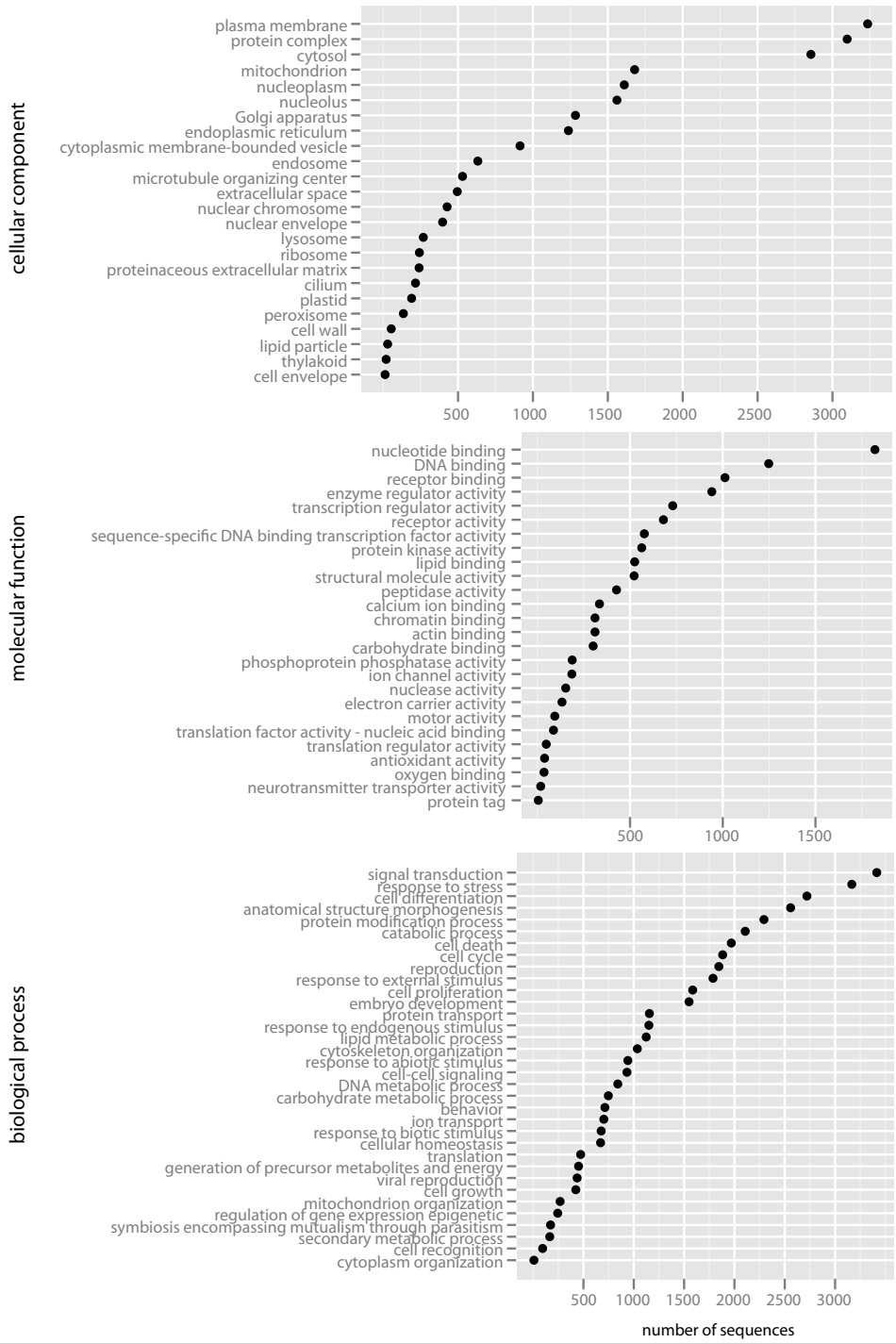

Figure E.7: Gene ontology for annotated contigs for a randomly-selected lineage, with respect to cellular component, biological process, and molecular function.

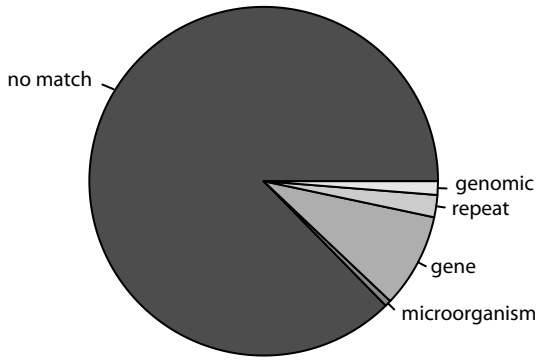

Figure E.8: Identify of unannotated contigs from a randomly selected assembly, as identified from a BLAST search to the NCBI 'nr' nucleotide database.

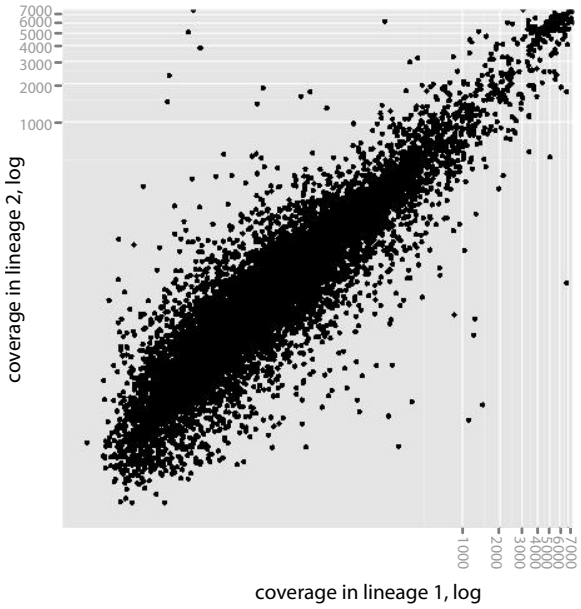

Figure E.9: Correlation in coverage between homologous, annotated contigs for a randomly-selected lineage-pair.

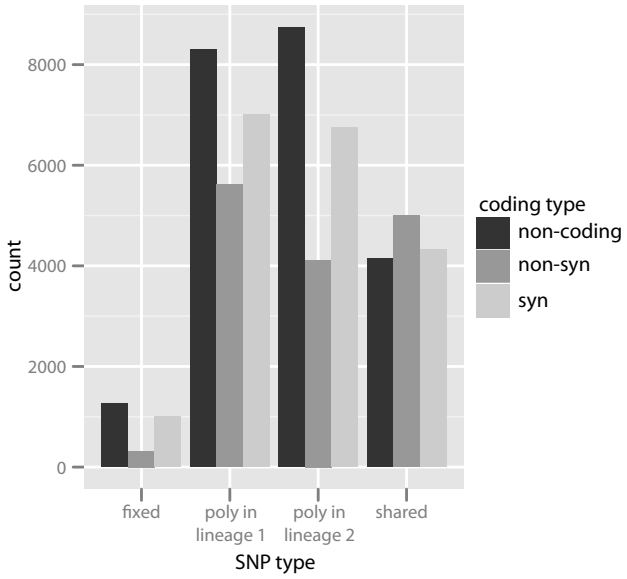

Figure E.10: Summary of SNPs found in a randomly-selected lineage-pair, annotated with respect to SNP and coding type.

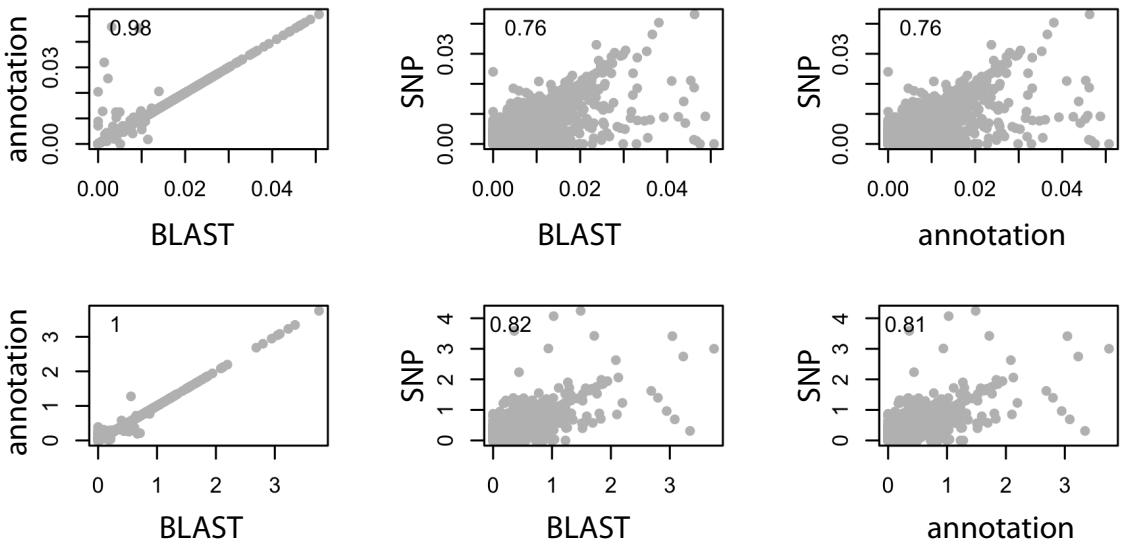

Figure E.11: Top row shows correlation in sequence divergence and bottom row shows correlation in inferred $\frac{dN}{dS}$ ratios for homologs for a randomly-selected lineage-pair for three methods of homolog discovery: annotation, in which contigs which share the same annotation are inferred to be homologous, BLAST, in which reciprocal best-hit BLAST is used to identify homologs, and SNP methods, in which variant information is used to reconstruct one homolog with respect to another.

E.2 Supplementary Tables

individual	lineage	latitude	longitude	Locality
SS34	*C. rubrigularis* N	-16.617	145.458	Mount Harris
SS35	*C. rubrigularis* N	-16.617	145.458	Mount Harris
SS37	*C. rubrigularis* N	-16.611	145.452	Mount Harris
SS40	*C. rubrigularis* N	-16.611	145.452	Mount Harris
SS41	*C. rubrigularis* N	-16.611	145.452	Mount Harris
SS48	*C. rubrigularis* S	-17.694	145.694	S. Johnstone River, Sutties Gap Rd
SS50	*C. rubrigularis* S	-17.694	145.694	S. Johnstone River, Sutties Gap Rd
SS52	*C. rubrigularis* S	-17.660	145.722	S. Johnstone River, Sutties Gap Rd
SS56	*C. rubrigularis* S	-17.678	145.710	S. Johnstone River, Sutties Gap Rd
SS57	*C. rubrigularis* S	-17.678	145.710	S. Johnstone River, Sutties Gap Rd
SEW08448	*L. coggeri* C	-16.976	145.777	Lake Morris Rd
SEW08452	*L. coggeri* C	-16.976	145.777	Lake Morris Rd
SS135	*L. coggeri* C	-16.976	145.777	Lake Morris Rd
SS136	*L. coggeri* C	-16.976	145.777	Lake Morris Rd
SS138	*L. coggeri* C	-16.976	145.777	Lake Morris Rd
SS64	*L. coggeri* N	-16.579	145.315	Mount Lewis
SS65	*L. coggeri* N	-16.572	145.322	Mount Lewis
SS67	*L. coggeri* N	-16.578	145.308	Mount Lewis
SS72	*L. coggeri* N	-16.585	145.289	Mount Lewis
SS74	*L. coggeri* N	-16.584	145.302	Mount Lewis
SS54	*L. coggeri* S	-17.660	145.722	S. Johnstone River, Sutties Gap Rd
SS59	*L. coggeri* S	-17.700	145.693	S. Johnstone River, Sutties Gap Rd
SS60	*L. coggeri* S	-17.700	145.693	S. Johnstone River, Sutties Gap Rd
SS62	*L. coggeri* S	-17.676	145.713	S. Johnstone River, Sutties Gap Rd
SS63	*L. coggeri* S	-17.628	145.740	S. Johnstone River, Sutties Gap Rd
SS25	*S. basiliscus* C	-17.295	145.712	Butchers Creek
SS28	*S. basiliscus* C	-17.299	145.701	Butchers Creek
SS29	*S. basiliscus* C	-17.299	145.701	Butchers Creek
SS30	*S. basiliscus* C	-17.299	145.701	Butchers Creek
SS32	*S. basiliscus* C	-17.299	145.701	Butchers Creek
SS127	*S. basiliscus* S	-18.199	145.849	Kirrama Range Rd
SS128	*S. basiliscus* S	-18.199	145.849	Kirrama Range Rd
SS129	*S. basiliscus* S	-18.199	145.849	Kirrama Range Rd
SS130	*S. basiliscus* S	-18.199	145.849	Kirrama Range Rd
SS131	*S. basiliscus* S	-18.199	145.849	Kirrama Range Rd

Table E.1: Individuals included in this study and their associated locality data; individuals are accessioned at the Museum of Vertebrate Zoology at University of California, Berkeley.

filtering type	rate
duplication	$1.4 \pm 0.2\%$
contamination	$0.4 \pm 1.1\%$
low-complexity reads	$0.004 \pm 0.003\%$
merging reads	$68.7 \pm 4.7\%$

Table E.2: Quality control filtering and their rates for raw data, summarized across seven lineages.

database	annotated contigs	unique, annotated contigs
A. carolinensis	23804	12218
G. gallus	22324	11146
UniProt90 database	26089	12324
Ensembl 9-species database	25838	NA
Ensembl 54-species database	26601	NA

Table E.3: Number of contigs annotated according to different reference databases for a randomly selected assembly.

assembly	initial chimerism	final chimerism	initial stop codons	final stop codons
C. rubrigularis, N	4.6%	0.0%	2.6%	0.6%
C. rubrigularis, S	3.7%	0.0%	2.8%	0.8%
L. coggeri, N	10.3%	0.0%	3.3%	1.1%
L. coggeri, C	5.5%	0.0%	3.1%	1.0%
L. coggeri, S	3.9%	0.0%	3.3%	1.0%
S. basiliscus, C	4.4%	0.0%	2.6%	0.6%
S. basiliscus, S	4.0%	0.0%	2.8%	0.7%

Table E.4: Prevalence of chimerism, or percentage of contigs that appeared to consist of multiple genes misassembled together, and stop codons, or percentage of contigs that had nonsense mutations, in assemblies, summarized across seven lineages both before and after the data were run in the annotation pipeline.

coverage	number of contigs within lineage	number of contigs between lineages
10x	3326 ± 494	2606 ± 399
20x	1888 ± 316	1439 ± 245
30x	1311 ± 245	981 ± 178
40x	994 ± 190	741 ± 133
50x	808 ± 157	602 ± 108

Table E.5: Number of annotated contigs which have given coverage for each individual; shown for one randomly selected lineage-pair.