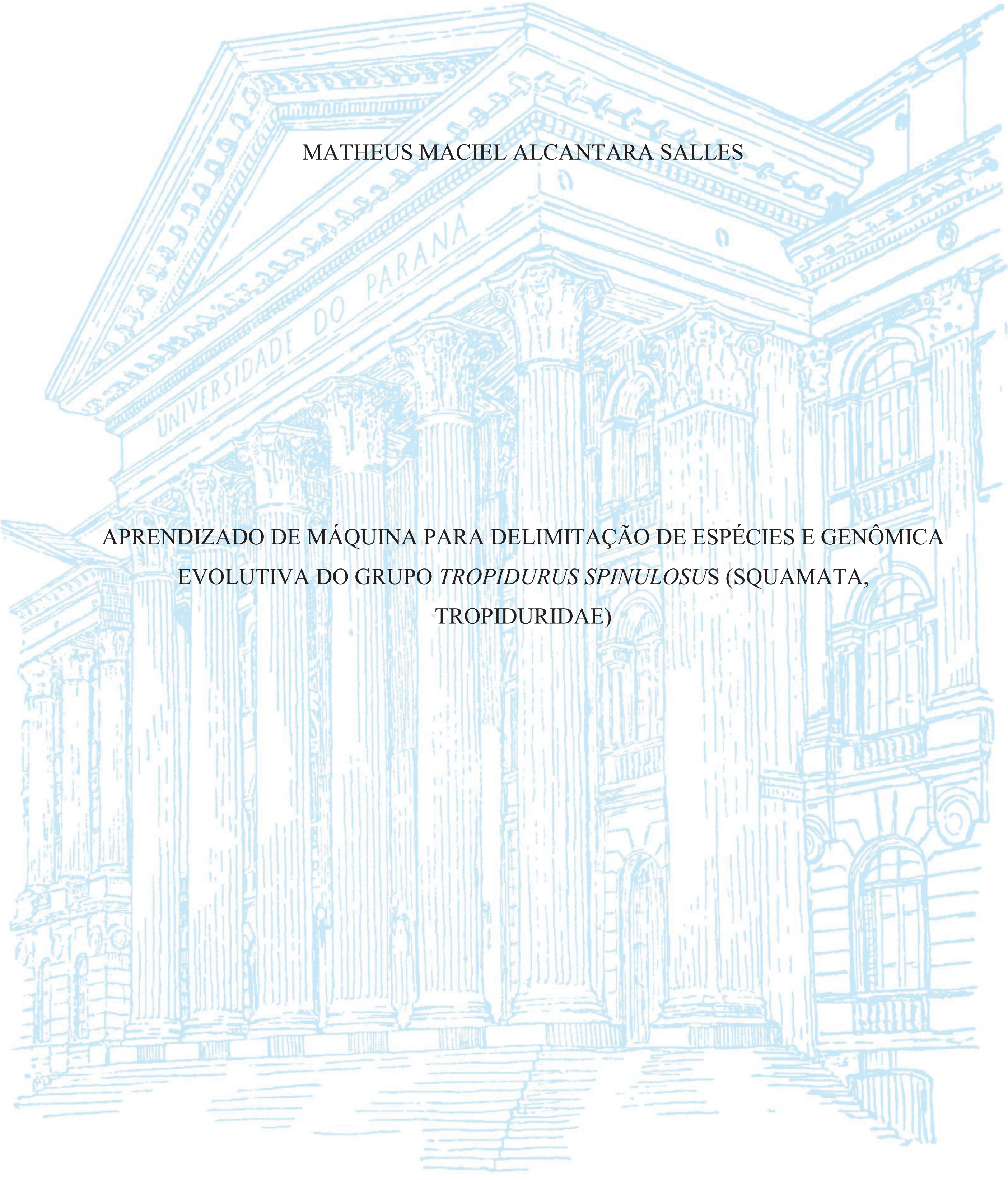
UNIVERSIDADE FEDERAL DO PARANÁ



MATHEUS MACIEL ALCANTARA SALLES

APRENDIZADO DE MÁQUINA PARA DELIMITAÇÃO DE ESPÉCIES E GENÔMICA EVOLUTIVA DO GRUPO *TROPIDURUS SPINULOSU*S (SQUAMATA, TROPIDURIDAE)

CURITIBA

2025

MATHEUS MACIEL ALCANTARA SALLES

APRENDIZADO DE MÁQUINA PARA DELIMITAÇÃO DE ESPÉCIES E GENÔMICA EVOLUTIVA DO GRUPO *TROPIDURUS SPINULOSU*S (SQUAMATA, TROPIDURIDAE)

Tese apresentada ao Programa de Pós-Graduação em Zoologia, no Departamento de Zoologia, Setor de Ciências Biológicas, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Doutor em Zoologia.

Orientador: Dr. Fabricius M. C. B. Domingos
Co-orientador: Dr. André L. G. Carvalho

CURITIBA
2025

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP) UNIVERSIDADE FEDERAL DO PARANÁ
SISTEMA DE BIBLIOTECAS – BIBLIOTECA DE CIÊNCIAS BIOLÓGICAS

Salles, Matheus Maciel Alcantara, 1996-.
Aprendizado de máquina para delimitação de espécies e genômica evolutiva do grupo *Tropidurus Spinulosus* (Squamata, Tropiduridae). / Matheus Maciel Alcantara Salles. – Curitiba, 2025.
1 recurso on-line : PDF.

Tese (Doutorado) – Universidade Federal do Paraná, Setor de Ciências Biológicas. Programa de Pós-Graduação em Zoologia.
Orientador: Dr. Fabricius M. C. B. Domingos.
Coorientador: Dr. André L. G. Carvalho.

1. Zoologia. 2. Bioinformática. 3. Evolução. 4. Inteligência artificial. 5. Lagarto. I. Domingos, Fabricius Maia Chaves Bicalho. II. Carvalho, André Luiz Gomes de. III. Universidade Federal do Paraná. Setor de Ciências Biológicas. Programa de Pós-Graduação em Zoologia. IV. Título.

Bibliotecária: Rosilei Vilas Boas CRB-9/939

MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIÊNCIAS BIOLÓGICAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO ZOOLOGIA - 40001016008P4

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação ZOOLOGIA da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **MATHEUS MACIEL ALCANTARA SALLES**, intitulada: **APRENDIZADO DE MÁQUINA PARA DELIMITAÇÃO DE ESPÉCIES E GENÔMICA EVOLUTIVA DO GRUPO *TROPIDURUS SPINULOSUS* (SQUAMATA, TROPIDURIDAE)**, sob orientação do Prof. Dr. FABRICIUS MAIA CHAVES BICALHO DOMINGOS, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de doutor está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 29 de Outubro de 2025.

Assinatura Eletrônica
31/10/2025 14:37:04.0
FABRICIUS MAIA CHAVES BICALHO DOMINGOS
Presidente da Banca Examinadora

Assinatura Eletrônica
03/11/2025 00:20:40.0
RENATO JOSE PIRES MACHADO
Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
03/11/2025 15:31:16.0
FERNANDA DE PINHO WERNECK
Avaliador Externo (INSTITUTO NACIONAL DE PESQUISAS DA AMAZÔNIA)

Assinatura Eletrônica
31/10/2025 15:21:08.0
MARCIO ROBERTO PIE
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica 04/11/2025
01:24:04.0
ANDRÉ LUIZ GOMES DE CARVALHO
Coorientador(a) (UNIVERSITY OF WASHINGTON)

Dedico este trabalho à minha família, que sempre esteve ao meu lado, me apoiando incondicionalmente, e que foi essencial para que eu realizasse este sonho.

AGRADECIMENTOS

Ao terminar o Mestrado, em plena (e maldita) pandemia de COVID-19, fazer um Doutorado não estava exatamente entre as minhas prioridades. Ter concluído o Mestrado foi uma experiência fantástica, claro, mas a vida (não apenas a acadêmica) naquele período foi muito complicada, para todo mundo. Eu estava meio desacreditado quanto a seguir na universidade e não levava muito a sério a ideia de engatar no doutorado. Pelo menos até conhecer o Fabricius. Acho que nossas primeiras conversas, especialmente na preparação para (e durante) a disciplina de Filogeografia Estatística, em 2021, reacenderam em mim a alegria de conversar, trocar ideias sobre ciência e, sobretudo, *fazer* ciência. Portanto, é difícil colocar em palavras a gratidão que sinto por ele. Além de um orientador fantástico, o Fabricius se tornou um amigo que quero levar para a vida toda. E, de quebra, ele ainda me deu a oportunidade de conhecer melhor sua linda família. Penso que pessoas incríveis costumam atrair pessoas incríveis, e esse certamente é o caso de Marina e Fabricius. A Marina, inclusive, foi responsável por uma das disciplinas mais legais (senão a mais legal) de todo o meu doutorado. E claro, tive o privilégio de conhecer desde pequenas as duas estrelas da família: Flora (que ameaça seriamente o meu posto de maior fã do Studio Ghibli) e Aurora (que certamente está entre as crianças mais fofas do universo). Espero continuar acompanhando o crescimento das pequenas e a felicidade dessa família maravilhosa, que tanto me apoiou ao longo dos últimos quatro anos.

Ao André Carvalho, que foi o melhor coorientador que eu poderia ter. Sinto que os capítulos da minha tese sobre *Tropidurus* são tão meus quanto dele. Sem a experiência e o conhecimento do André sobre esses animais, dificilmente eu teria conseguido investigar com tanta profundidade a história evolutiva desses lagartinhos que aprendi a gostar tanto. Aos poucos, conversa a conversa, reunião a reunião, o André se tornou uma das minhas maiores referências profissionais, uma verdadeira inspiração. Mais que isso, é um ser humano extremamente solidário e acolhedor, que tive o privilégio de conhecer melhor, mesmo à distância. Em meio a uma academia tantas vezes poluída por egos e competições vazias, o André nos faz lembrar que a ciência também pode ser feita de boas pessoas, gentileza e companheirismo.

Aos membros da banca, Fernanda Werneck, Marcio Pie e Renato Machado. Também ao Alexandre Palaoro, que topou entrar na enrascada de ficar como suplente. Agradeço pela gentileza de aceitarem o convite e dedicarem parte de seu tempo à avaliação deste trabalho. Em especial à Fernanda e ao Renato, que me acompanharam desde o início do Doutorado como membros do comitê de acompanhamento. Tenho certeza de que suas contribuições foram fundamentais para o amadurecimento desta tese. Como fiz ao final do Mestrado, não posso deixar de agradecer também a todos os professores e professoras que contribuíram para minha formação, desde a infância. Muito do que sou hoje (biólogo e agora doutor em Zoologia) devo às professoras e aos professores que tive.

À Universidade Federal do Paraná, por estar presente na minha vida desde a graduação. É um orgulho poder dizer que praticamente toda a minha formação acadêmica aconteceu aqui. Em especial, ao Programa de Pós-Graduação em Zoologia, e a todos os professores, técnicos, estudantes e amigos com quem tive o privilégio de conviver nesses últimos anos. Um abraço ainda mais especial à minha amiga Marina de Souza, companheira de representação discente, que esteve comigo praticamente desde o início do doutorado (em poucas e boas). E claro, ao meu grande amigo Henrique “El Mago” Schipanski, que me atura desde o mestrado e certamente é a pessoa de toda a universidade que já me viu nas minhas piores (ou melhores?) condições. Valeu por tudo, Schipa. Te devo (mais uma) cerveja no Seu Nilson.

A todos os colegas do Laboratório de Evolução e Diversidade Zoológica (FAZ O LEDZ 🤘), obrigado pelos bons momentos compartilhados ao longo destes quatro anos. Foram muitas pessoas queridas que passaram pelo lab nesse tempo, e sou grato a todas elas. Fazer parte do LEDZ sempre foi motivo de alegria para mim e tive apenas boas companhias nesse percurso. Em especial, um abraço carinhoso em Junior, Marcos e Naiane, que foram aqueles com quem vivenciei mais momentos para além dos corredores sombrios do Biológicas.

À minha família, especialmente à minha mãe, Marledes, e ao meu padrasto, Ricardo. Sem vocês eu não teria chegado até aqui. Minha mãe me apoia incondicionalmente desde o dia em que decidi prestar vestibular para Biologia, nunca deixando de estar ao meu lado. Ao meu irmão, Lucas, que tanto me inspira a ser uma pessoa melhor, a cada dia (apesar de ele ter tido a audácia de ter ficado mais alto do que eu justamente em 2025). Dizem que sou muito tranquilo, e certamente isso só é possível porque vocês fazem de mim ser quem eu sou. Amo vocês.

À Laura. Assim como no mestrado, só você sabe o quanto foi importante para mim durante estes quatro anos de doutorado (período em que também juntamos nossos farrapos), e o quanto é parte essencial da minha vida. Como diz aquele filme que nós dois gostamos tanto: *meu rosto é meu, minhas mãos são minhas, minha boca é minha, mas eu não; eu sou seu*.

Por fim, a todos os lagartos que já estudei, ouvi, observei ou simplesmente tive a sorte de contemplar. Foi, no fim das contas, a linda história evolutiva desses animais que sustentou a maior parte do desenvolvimento desta tese. Minha admiração e gratidão por eles vai muito além do que qualquer palavra pode expressar.

Meu processo é pensar... pensar... e pensar. [...]

Se você tiver uma maneira melhor, por favor, me avise.

Hayao Miyazaki

RESUMO

Compreender os mecanismos que influenciam a história evolutiva dos organismos (incluindo processos de especiação, evolução molecular e biogeográficos) exige a adoção de estruturas analíticas robustas. Neste contexto, conjuntos de dados genômicos fornecem aos cientistas ferramentas poderosas para abordar questões evolutivas complexas e intrincadas. A presente tese explora esses temas por meio de investigações multifacetadas, tanto com ênfase em um grupo de organismos—répteis escamados (Squamata), em particular os lagartos do grupo de espécies *Tropidurus spinulosus*—como em um domínio específico da Biologia Evolutiva: a delimitação de espécies. Utilizando conjuntos diversos de dados genômicos (como mitogenomas, elementos ultraconservados nucleares e polimorfismos de nucleotídeo único) aliados a análises filogenômicas, modelagem demográfica e métodos de *machine learning* (ML), esta tese proporciona avanços tanto em aspectos metodológicos quanto empíricos dentro da Biologia Evolutiva, organizados em cinco eixos interligados: (1) os desafios e oportunidades do uso de ML na delimitação de espécies; (2) a comparação empírica de métodos de delimitação de espécies baseados em ML sob diferentes cenários de evolução molecular; (3) dinâmicas das taxas de substituição mitocondrial em Squamata e suas implicações para a calibração de tempos de divergência; (4) os processos evolutivos por trás de padrões de discordância mitonuclear em lagartos do grupo *Tropidurus spinulosus*; e (5) padrões filogenéticos e biogeográficos em *T. spinulosus*. Especificamente, o primeiro capítulo revisa criticamente o uso de ML na delimitação de espécies, destacando sua flexibilidade para lidar com dados complexos, mas também limitações inerentes ao seu funcionamento. Propõem-se boas práticas para tornar sua utilização na delimitação de espécies mais produtiva, posicionando o ML como ferramenta complementar a métodos tradicionalmente aplicados na área. No segundo capítulo comparam-se métodos de delimitação de espécies baseados em ML em cenários diversos de evolução molecular, avaliando seu desempenho em cenários demográficos variados. Os resultados indicam que classificadores supervisionados são robustos e computacionalmente eficientes para inferir limites entre espécies, mesmo sob diferentes níveis de variação nos modelos de substituição molecular. No terceiro capítulo estimam-se taxas de substituição mitocondrial em Squamata utilizando mitogenomas, revelando valores entre 0,006 e 0,02 substituições por sítio por milhão de anos. Essas taxas variam significativamente entre regiões codificantes e não codificantes, bem como entre posições de códons, reforçando a necessidade de calibrações rigorosas e modelos particionados para inferências temporais. O quarto capítulo investiga a discordância mitonuclear no grupo *T. spinulosus* por meio de genomas mitocondriais, UCEs nucleares e SNPs. Análises filogenéticas e de genética de populações identificam eventos ancestrais de captura mitocondrial, associados a uma história que combina demografia, contato secundário e fluxo gênico, como as principais causas da discordância entre os genomas nuclear e mitocondrial. Por fim, o quinto capítulo reconstrói processos biogeográficos e de especiação em *T. spinulosus*, integrando evidências geoclimáticas, técnicas de modelagem de distribuição populacional ancestral e dados genômicos para testar hipóteses sobre a diversificação desses animais. Os resultados revelam uma história complexa de especiação influenciada por mudanças paleoambientais e eventos de fluxo gênico ancestral, com padrões filogenéticos consistentes com a ação de multiplos fatores geoclimáticos na divergência entre as linhagens. Coletivamente, esta tese auxilia no progresso científico em Biologia Evolutiva através da compreensão da evolução de organismos não modelo, como a maioria das espécies de Squamata, ao sintetizar dados genômicos, ferramentas computacionais e estudos de caso empíricos. Oferece também abordagens metodológicas robustas para resolver discordâncias filogenéticas, refinar calibrações de taxas de substituição e aprimorar práticas de delimitação de espécies. Ao integrar revisões teórico-conceituais e investigações empíricas, a presente tese contribui para debates em diversas áreas da Biologia Evolutiva, destacando a importância de abordagens integrativas e multidisciplinares para desvendar histórias evolutivas complexas.

Palavras-chave: Bioinformática. Evolução. Inteligência Artificial. Lagartos. Zoologia.

ABSTRACT

Understanding the mechanisms that shape the evolutionary history of organisms (including processes of speciation, molecular evolution, and biogeography) requires the use of robust analytical frameworks. In this context, genomic datasets provide scientists with powerful tools to address complex and intricate evolutionary questions. This PhD thesis explores these themes through multifaceted investigations, with an emphasis both on a group of organisms—squamate reptiles (Squamata), particularly lizards of the *Tropidurus spinulosus* species group—and on a specific domain of Evolutionary Biology: species delimitation. By combining diverse genomic datasets (such as mitogenomes, nuclear ultraconserved elements, and single nucleotide polymorphisms) with phylogenomic analyses, demographic modeling, and machine learning (ML) approaches, this dissertation advances both methodological and empirical aspects of Evolutionary Biology, organized into five interconnected axes: (1) the challenges and opportunities of using ML in species delimitation; (2) the empirical comparison of ML-based species delimitation methods under different scenarios of molecular evolution; (3) the dynamics of mitochondrial substitution rates in Squamata and their implications for divergence-time calibration; (4) the evolutionary processes underlying mitonuclear discordance in lizards of the *Tropidurus spinulosus* group; and (5) phylogenetic and biogeographic patterns in *T. spinulosus*. Specifically, Chapter 1 critically reviews the use of ML in species delimitation, highlighting its flexibility for handling complex data while also pointing out inherent limitations. Best practices are proposed to make its application more effective, positioning ML as a complementary tool to methods traditionally employed in the field. Chapter 2 compares ML-based species delimitation methods across diverse molecular evolution scenarios, assessing their performance under different diversification contexts. The results indicate that supervised classifiers are robust and computationally efficient for inferring species limits, even under different molecular substitution models. Chapter 3 estimates mitochondrial substitution rates in Squamata using mitogenomes, revealing values ranging from 0.006 to 0.02 substitutions per site per million years. These rates vary substantially among coding and noncoding regions as well as among codon positions, underscoring the need for rigorous calibrations and partitioned models for reliable temporal inferences. Chapter 4 investigates mitonuclear discordance in the *T. spinulosus* group using mitochondrial genomes, nuclear UCEs, and SNPs. Phylogenetic and population genetic analyses identify ancestral mitochondrial capture events—linked to a history shaped by demography, secondary contact, and gene flow—as the main drivers of discordance between nuclear and mitochondrial genomes. Finally, Chapter 5 reconstructs biogeographic and speciation processes in *T. spinulosus* by integrating geoclimatic evidence, ancestral population distribution modeling, and genomic data to test hypotheses about the diversification of these lizards. The results reveal a complex history of speciation influenced by paleoenvironmental changes and ancestral gene flow events, with phylogenetic patterns consistent with the action of multiple geoclimatic factors in the divergence between lineages. Altogether, this dissertation contributes to the advancement of Evolutionary Biology by improving our understanding of the evolution of non-model organisms, such as most Squamata species, through the integration of genomic data, computational tools, and empirical case studies. It also provides robust methodological approaches for resolving phylogenetic discordances, refining substitution-rate calibrations, and improving species delimitation practices. By combining theoretical and conceptual reviews with empirical investigations, this dissertation contributes to debates across multiple areas of Evolutionary Biology and underscores the importance of integrative and multidisciplinary approaches to unravel complex evolutionary histories.

Key words: Artificial Inteligence. Bioinformatics. Evolution. Lizards. Zoology.

SUMÁRIO

INTRODUÇÃO GERAL

A incorporação de dados moleculares revolucionou o estudo de processos ecológicos e evolutivos, permitindo análises em escalas temporais e geográficas sem precedentes (Hudson, 2008; Bleidorn, 2016). Essa transformação, impulsionada por avanços em genômica e bioinformática, refinou nossa capacidade de testar hipóteses na grande área da Biologia Evolutiva, particularmente em regiões megadiversas como o Neotrópico—uma região que reconhecidamente apresenta altos níveis de endemismo (Rull & Carnaval, 2020) e que foi moldado por eventos geoclimáticos complexos (Hoorn et al., 2010; Turchetto-Zolet et al., 2013; Antonelli et al., 2018). Consequentemente, a integração de dados genômicos, morfológicos e ecológicos tem sido essencial para reconstruir histórias evolutivas intrincadas.

No entanto, a integração dessas abordagens é dificultada pela escassez de dados genômicos para muitos grupos, particularmente os compostos majoritariamente por espécies não-modelo, como Squamata. Esse grupo, que abrange mais de 11.000 espécies (Uetz et al., 2025), possui uma história evolutiva ainda pouco compreendida devido a essa lacuna de dados. Isso se torna um obstáculo significativo para desvendar a história de linhagens como Pleurodonta, cujas radiações adaptativas moldaram as comunidades de lagartos atuais, inclusive na região Neotropical (Blankers et al., 2013; Alencar et al., 2024). A carência de estimativas confiáveis para diferentes parâmetros evolutivos, por exemplo, limita a resolução de reconstruções filogenéticas e biogeográficas neste grupo, afetando nossa compreensão mais ampla dos seus processos de diversificação. Diante desse cenário, estudos filogeográficos, particularmente, destacam-se como uma das principais ferramentas capazes de vincular padrões genéticos a processos históricos como vicariância, dispersão e adaptação (Antonelli et al., 2010; Turchetto-Zolet et al., 2013; Leal et al., 2016). Ao mesmo tempo, abordagens de filogeografia também tem revelado diversidade críptica em táxons amplamente distribuídos, desafiando noções tradicionais de limites de espécies (Domingos et al., 2014; Werneck et al., 2015; Melo et al., 2016).

Nesse âmbito, a delimitação de espécies emerge como um desafio central na biologia evolutiva, pois define a unidade básica para qualquer inferência biológica (Carstens et al., 2013; Rannala, 2015). Por outro lado, há de se destacar que a tensão entre conceitos teóricos e critérios operacionais para delimitar uma espécie (de Queiroz, 2007; Sukumaran & Knowles, 2021) exige abordagens inovadoras, especialmente diante da explosão de dados genômicos gerados por tecnologias de sequenciamento de última geração (NGS). Embora métodos estatísticos baseados em coalescência tenham avançado significativamente nossa capacidade de compreender e inferir limites entre espécies (Rannala & Yang, 2010; 2020), limitações persistem em cenários evolutivos mais complexos, como aqueles que envolvem fluxo gênico contínuo e especiação incompleta (Smith & Carstens, 2020). O *machine learning* (ML) tem emergido como paradigma promissor, combinando eficiência computacional e flexibilidade analítica para resolver problemas multidimensionais (Tang et al., 2019; Greenner et al., 2021). Ainda assim, embora algoritmos de ML sejam analiticamente poderosos e aplicados em áreas diversas da Biologia Evolutiva, incluindo delimitação de espécies (e.g., Derkarabetian et al., 2019; Pyron, 2023), sua utilização sugere que eles também apresentam limitações específicas e não devem ser considerados como alternativas definitivas aos métodos tradicionalmente aplicados em Biologia Evolutiva.

Dado este contexto, a presente tese utiliza o grupo de lagartos *Tropidurus spinulosus* (Squamata: Tropiduridae) como modelo para explorar tanto a área de delimitação de espécies sob uma perspectiva conceitual, como as interações entre processos geoclimáticos e sua diversificação. Distribuídos em paisagens dinâmicas do Cerrado e Pantanal, esses organismos testemunharam eventos neotectônicos críticos da América do Sul, como o soerguimento dos Andes e a subsidência da Bacia do Pantanal (Prates et al., 2016). Apesar de avanços filogenéticos recentes (Carvalho, 2013; Carvalho et al., 2016), lacunas persistem na resolução das suas relações evolutivas e na identificação de linhagens crípticas, agravadas pela escassez de dados populacionais de alta resolução (Carvalho, 2013; Carvalho et al., 2013; Carvalho et al., 2016).

Para preencher essas lacunas, integramos dados genômicos—sobretudo mitogenomas completos e elementos nucleares ultra-conservados (do inglês, *Ultraconserved Elements*, UCEs)—com análises de Biologia Comparada e modelagem de nicho ecológico, testando hipóteses sobre diversificação em resposta a mudanças paleoambientais. Paralelamente, avaliamos o potencial de algoritmos de ML na delimitação de espécies, analisando seu desempenho em termos de eficiência biológica e computacional.

Portanto, além de elucidar padrões biogeográficos em *Tropidurus spinulosus*, esta tese propõe diretrizes para o uso responsável de aprendizado de máquina em Biologia Evolutiva, combinando rigor estatístico e relevância biológica. A síntese entre dados empíricos e avanços metodológicos reforça a necessidade de abordagens multidisciplinares para desvendar histórias evolutivas complexas, especialmente em organismos não modelo como o grupo *Tropidurus*. Cabe destacar, ainda, que todo o processo de pesquisa foi conduzido sob os princípios da ciência aberta. Dados brutos, scripts analíticos, fluxos de trabalho e versões preliminares dos manuscritos foram disponibilizados em repositórios públicos, seguindo padrões internacionais de transparência e reprodutibilidade. Essa prática não só permite o escrutínio independente dos resultados aqui apresentados, mas também fomenta a reutilização e adaptação dessas informações, contribuindo para uma ciência mais colaborativa, acessível e responsiva às demandas contemporâneas da Biologia Evolutiva.

OBJETIVO GERAL DA TESE

Investigar a história evolutiva de *Tropidurus spinulosus* mediante métodos filogenéticos, filogeográficos e de delimitação de espécies, bem como avaliar o potencial de algoritmos de ML para superar limitações de métodos tradicionais em cenários de alta complexidade evolutiva.

OBJETIVOS ESPECÍFICOS

1. Através de um estudo de revisão, discutir os desafios e as oportunidades do uso de aprendizado de máquina na delimitação de espécies.
2. Comparar empiricamente métodos de delimitação de espécies baseados em aprendizado de máquina sob diferentes cenários de evolução molecular.
3. Explorar as dinâmicas das taxas de substituição mitocondrial em um grupo de Squamata e suas implicações para a calibração de tempos de divergência.
4. Investigar os processos evolutivos por trás de padrões de discordância mitonuclear em lagartos do grupo *Tropidurus spinulosus*.
5. Analisar padrões filogenéticos e biogeográficos em *T. spinulosus*.

REFERÊNCIAS

ANTONELLI, A; et al. Molecular studies and phylogeography of Amazonian tetrapods and their relation to geological and climatic models. **Amazonia, landscape and species evolution: a look into the past**, v. 386, p. 404, 2010.

ANTONELLI, A.; ZIZKA, A.; CARVALHO, F. A. SCHARN, R., BACON, C. D., SILVESTRO, D., & CONDAMINE, F. L. Amazonia is the primary source of Neotropical biodiversity. **Proc Natl Acad Sci USA**, v. 115, p. 6034–6039, 2018.

BLEIDORN, C. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. **Systematics and biodiversity**, v. 14, n. 1, p. 1-8, 2016.

CARSTENS, B.C., PELLETIR, T.A., REID, N.M. & SATLER, J.D. 2013. How to fail at species delimitation. **Molecular Ecology**, v. 22, p. 4369–83.

CARVALHO, A. L. G. On the distribution and conservation of the South American lizard genus Tropidurus Wied-Neuwied, 1825 (Squamata: Tropiduridae). **Zootaxa**, v. 3640, n. 1, p. 42-56, 2013.

CARVALHO, A. L. G. 2016. Three New Species of the *Tropidurus spinulosus* Group (Squamata: Tropiduridae) from Eastern Paraguay. **American Museum Novitates**, 3853(3853): 1–44, doi:10.1206/3853.1.

CARVALHO, A. L. G.; DE BRITTO, M. R.; FERNANDES, D. S. Biogeography of the Lizard Genus Tropidurus Wied-Neuwied, 1825 (Squamata: Tropiduridae): Distribution, Endemism, and Area Relationships in South America. **PLoS One**, v. 8, 2013.

CARVALHO, A. L.G.; SENA M. A.; PELOSO, P. L.V.; MACHADO, F. A.; MONTESINOS, R.; SILVA H. R.; CAMPBELL G.; RODRIGUES M. T. A. New *Tropidurus* (Tropiduridae) from the Semiarid Brazilian Caatinga: Evidence for Conflicting Signal between Mitochondrial and Nuclear Loci Affecting the Phylogenetic Reconstruction of South American Collared Lizards. **American Museum Novitates**, v. 3852, p. 1–68, 2016.

DERKARABETIAN, S.; CASTILLO, S.; KOO, P. K.; OVCHINNIKOV, S.; HEDIN, M. A demonstration of unsupervised machine learning in species delimitation. **Molecular Phylogenetics and Evolution**, v. 139, 106562, 2019.

DOMINGOS, F.M.C.B.; BOSQUE, R.J.; CASSIMIRO, J.; COLLI, G.R; RODRIGUES, M.T.; SANTOS, M.G.; BEHEGARAY, L.B. Out of the deep: Cryptic speciation in a Neotropical gecko (Squamata, Phyllodactylidae) revealed by species delimitation methods. **Molecular Phylogenetics and Evolution**, v. 80, p. 113–24, 2014.

GRENNER, J. G.; KANDATHIL, S. M.; MOFFAT, L.; JONES, D. T. A guide to machine learning for biologists. **Molecular Cell Biology**, v. 23, n. 1, p. 40–55, 2021. https://doi.org/10.1038/s41580-021-00407-0.

HOORN, C.; et al. Amazonia Through Time: Andean Uplift, Climate Change, Landscape Evolution, and Biodiversity. **Science,** v. 330, p. 927–931, 2010.

HUDSON, M. E. Sequencing breakthroughs for genomic ecology and evolutionary biology. **Molecular ecology resources**, v. 8, n. 1, p. 3-17, 2008.

LEAL, B. S. S.; PALMA DA SILVA, C.; PINHEIRO, F. Phylogeographic studies depict the role of space and time scales of plant speciation in a highly diverse Neotropical region. **Critical Reviews in Plant Sciences**, v. 35, n. 4, p. 215-230, 2016.

MELO, B. F.; et al. Cryptic species in the Neotropical fish genus Curimatopsis (Teleostei, Characiformes). **Zoologica Scripta**, v. 45, n. 6, p. 650-658, 2016.

PRATES, I.; XUE, A.T.; BROWN, J.L.; ALVARADO-SERRANO, D.F.; RODRIGUES, M.T.; HICKERSON, M.J.; CARNAVAL, A.C. Inferring responses to climate dynamics from historical demography in neotropical forest lizards. **Proceedings of the National Academy of Sciences**, v. 113, n. 7978-7985, 2016.

PYRON, R. A. Unsupervised machine learning for species delimitation, integrative taxonomy, and biodiversity conservation. **Molecular Phylogenetics and Evolution**, v. 189, p. 107939, 2023.

DE QUEIROZ, K. Species concepts and species delimitation. **Syst. Biol.**, v. 56, p. 879–886, 2007.

RANNALA, B. The art and science of species delimitation. **Current Zoology**, v. 61, n. 5, p. 846-853, 2015.

RANNALA, B; YANG, Z. Bayesian species delimitation using multilocus sequence data. **Proc. Natl. Acad. Sci. USA,** v. 107, p. 9264–9269, 2010.

RANNALA, B; YANG, Z. Species delimitation. In: Scornavacca, C and Delsuc, F and Galtier, N, (eds.) **Phylogenetics in the Genomic Era.** (5.5:1-5.5:18). Self published, 2020.

RULL, V.; CARNAVAL, A. C. (Ed.). **Neotropical diversification: patterns and processes**. Berlin: Springer, 2020.

SMITH, M. L.; CARSTENS, B. C. Process-based species delimitation leads to identification of more biologically relevant species. **Evolution**, v. 74, n. 2, p. 216-229, 2020.

SUKUMARAN, J.; HOLDER, M. T.; KNOWLES, L. L. Incorporating the speciation process into species delimitation. **PloS computational biology**, v. 17, n. 5 (e1008924), 2021.

TANG, B.; PAN, Z.; YIN, K.; KHATEEB, A. Recent advances of deep learning in bioinformatics and computational biology. **Frontiers in genetics,** v. 10, n. 214, 2019.

TURCHETTO-ZOLET, A. C.; PINHEIRO, F.; SALGUEIRO, F.; PALMA-SILVA, C. Phylogeographical patterns shed light on evolutionary process in South America. **Mol Ecol,** v. 22, p. 1193–1213, 2013.

WERNECK, F. P.; et al. Biogeographic history and cryptic diversity of saxicolous Tropiduridae lizards endemic to the semiarid Caatinga. **BMC Evolutionary Biology**, v. 15, p. 1-24, 2015.

1. CAPÍTULO I: TOWARDS THE NEXT GENERATION OF SPECIES DELIMITATION METHODS: AN OVERVIEW OF MACHINE LEARNING APPLICATIONS

Esta seção apresenta um dos artigos desenvolvidos ao longo do meu doutorado, já publicado. A formatação segue as normas da revista em que o artigo foi publicado.

REFERÊNCIA: Salles, M. M. A., & Domingos, F. M. C. B. (2025). Towards the next generation of species delimitation methods: An overview of machine learning applications. *Molecular Phylogenetics and Evolution*, 108368. https://doi.org/10.1016/j.ympev.2025.108368.

ABSTRACT

Species delimitation is the process of distinguishing between populations of the same species and distinct species of a particular group of organisms. Various methods exist for inferring species limits, whether based on morphological, molecular, or other types of data. In the case of methods based on DNA sequences, most of them are rooted in the coalescent theory. However, coalescence-based models have limitations, for instance regarding complex evolutionary scenarios and large datasets. In this context, machine learning (ML) can be considered as a promising analytical tool, and provides an effective way to explore dataset structures when species-level divergences are hypothesized. In this review, we examine the use of ML in species delimitation and provide an overview and critical appraisal of existing workflows. We also provide simple explanations on how the main types of ML approaches operate, which should help uninitiated researchers and students interested in the field. Our review suggests that while current ML methods designed to infer species limits are analytically powerful, they also present specific limitations and should not be considered as definitive alternatives to coalescent methods for species delimitation. Future ML enterprises to delimit species should consider the constraints related to the use of simulated data, as in other model-based methods relying on simulations. Conversely, the flexibility of ML algorithms offers a significant advantage by enabling the analysis of diverse data types (e.g., genetic and phenotypic) and handling large datasets effectively. We also propose best practices for the use of ML methods in species delimitation, offering insights into potential future applications. We expect that the proposed guidelines will be useful for enhancing the accessibility, effectiveness, and objectivity of ML in species delimitation.

Key words: bioinformatics, molecular data, speciation, phylogenetics, artificial intelligence, deep learning.

1. Introduction

1.1. Inferring species limits

Species represent fundamental entities across all biological disciplines. Consequently, the review, categorization, and characterization of taxa within this level constitute a pivotal aspect of biodiversity research (Bortolus, 2008; Vink et al., 2012; Ely et al., 2017). The process of identifying, characterizing, and defining a species is data-intensive and entails various practical dimensions. This complexity arises from managing extensive biological data and dealing with a range of theoretical elements, from the establishment of homologies, to taxon-specific traits, and the very philosophical notion of species. Furthermore, conceptual issues surrounding the definition of species concepts still attract debates among taxonomists and evolutionary biologists (Pante et al., 2015; Zachos, 2016). These discussions reach the realms of philosophy, because a multitude of data and methodologies will probably not fully solve many fundamental questions surrounding the nature of species (Zachos, 2016; Wilkins et al., 2022), or the 'species ontology' (what a species really is or represents). A complete resolution on this subject remains elusive, as it intertwines the empirical evidence biologists are able to extract from nature with philosophical definitions surrounding species concepts (Pigliucci, 2003).

One of the most popular modern definitions is the 'Biological Species Concept' (de Queiroz, 2005a; Zachos, 2016), which defines species as interbreeding populations reproductively isolated from others (Mayr, 1969; 1996; 2000). Yet, many challenges to this concept emerged throughout the years as empirical data clearly shows that the history of life on Earth does not fit into a bifurcating process (Edwards et al., 2016; Mallet et al., 2016), and a clear delineation of reproductive barriers is hindered by instances of asexual reproduction, natural hybridization and gene flow (Arnold, 1992; Shurtliff, 2013; Gompert et al., 2017). Hence, taxonomists and evolutionary biologists must recognize that multiple species definitions will coexist in the practice of species delimitation, and these are usually chosen based on the biological context of the organisms under study.

An important concept in this context is the **General Lineage Concept** (**GLC**, terms in bold are defined in the Glossary, available in Appendix A), which unifies diverse contemporary views on the nature of species, prioritizing the recognition of independently evolving lineages over specific biological criteria such as reproduction or morphology (de Queiroz, 1998; 1999; 2007). According to the GLC, a species is defined as an independently evolving metapopulation lineage, emphasizing each species' unique evolutionary identity across time and space (de Queiroz, 2007). While unique morphological, ecological, or any other biological trait might be considered relevant in supporting the investigation of the speciation process, they are not mandatory criteria for species definition under the GLC perspective, but rather additional evidence supporting lineage separation (de Queiroz, 2007). Thus, this concept accounts for the contingent nature of the speciation process, where different biological properties may support species limits in varying degrees. It also emphasizes the need for multiple lines of evidence to corroborate hypotheses of species divergence, aligning with **Integrative Taxonomy** approaches (Wiens and Penkrot, 2002; Dayrat, 2005; Padial et al., 2010; Fujita et al., 2012; Karbstein et al., 2024).

The GLC also provides a theoretical distinction between the ‘species ontology problem’ (what a species is) and the ‘delimitation problem’ (how to operationally distinguish among putative species) (de Queiroz, 2007). Interestingly, while a clear relationship exists between these components, namely the species concept and species delimitation, historically, a significant part of the scientific efforts has focused on the former (see Sites Jr and Marshall, 2004; Wiens, 2007; de Queiroz, 2011; Hausdorf, 2011). The development of theoretical considerations related to species delimitation, in particular that based on molecular data, occurred mainly in the last two decades, accompanied by the introduction of new criteria and statistical methods (Lukhtanov, 2019; Rannala and Yang, 2020). Historically, identifying species limits, and describing new species, have primarily relied on morphological data (Wiens, 2007; Rannala, 2015; Rannala and Yang, 2020). However, morphological traits can be influenced by environmental factors, leading to convergence or divergence without necessarily reflecting genetic or evolutionary relationships between lineages (Price et al., 2003; Wake et al., 2011; Jarvis et al., 2014). Thus, genomic data has emerged as a crucial tool for inferring species limits, offering a more objective approach for species delimitation (Fujita et al., 2012), while complementing traditional morphological methods (Jörger and Schrödl, 2013).

Modern species delimitation methods (SDMs) aiming at identifying evolutionary units (Tautz et al., 2003; Vogler and Monaghan, 2007) have grown due to advancements in statistical frameworks for phylogenetic inference (Edwards, 2009; O’Meara, 2012), along with Molecular Biology tools (e.g., next-generation sequencing (NGS); Slatko et al., 2018) and Bioinformatics (Searls, 2010). They mostly operate with molecular data under the principles of Coalescent Theory, notably, the multispecies coalescent (MSC;

Rannala and Yang, 2003; Degnan and Rosenberg, 2009; Rannala et al., 2020). The MSC analytical framework has many evolutionary assumptions, such as the absence of recombination and hybridization, independence of gene trees and their coalescent processes, random mating within species, among others (a review on the subject can be found in Mirarab et al., 2021). However, these conditions are typically only met in tree-like speciation scenarios involving diploid, sexually reproducing organisms. In any case, MSC methods are capable of managing common problems in phylogenetic inference, such as conflicts among different gene trees due **incomplete lineage sorting** (**ILS**; Knowles and Carstens, 2007; Carstens et al., 2013; Jacobs et al., 2018).

Therefore, while they are valuable for inferring evolutionary relationships, coalescence-based SDMs may fail to distinguish population structure from species-level divergence (Sukumaran and Knowles, 2017), and may also be affected by the above-mentioned assumptions of the MSC model (Rannala and Yang, 2003; Degnan and Rosenberg, 2009; Edwards, 2009; Fujita et al., 2012). Some methods have their functionality and performance compromised in scenarios when there is introgression between putative species (Rannala and Yang, 2010; Leaché et al., 2014; Jackson et al., 2017), and are more reliable in situations where gene flow ceases immediately after population divergence (Fujita et al., 2012). Besides, simulations have shown that ignoring gene flow leads the MSC to overestimate **population sizes** and underestimate divergence times (e.g., Leaché et al., 2014). Hence, the effectiveness of the MSC framework is limited, to some extent, when additional processes influence divergence during speciation (Smith and Carstens, 2020). Naturally, different coalescence-based SDMs have varying capabilities to address particular evolutionary scenarios, and while such methods may be biased under certain evolutionary and analytical conditions, they are certainly an important part of the evolutionary biologist toolkit. For a more detailed discussion on SDMs based on the MSC for genomic data, see Rannala and Yang (2020).

1.2. Machine learning, evolutionary biology, and species delimitation

Machine learning (ML), a branch of artificial intelligence (AI) known for its computational efficiency and predictive accuracy, has recently gained popularity in Evolutionary Biology mainly due to its ability to analyze and process large, complex, and high-dimensional datasets (Chicco, 2017; Fountain-Jones et al., 2021; Greener et al., 2021; Morimoto et al., 2021; Borowiec et al., 2022). In general terms, ML can be defined as a group of computational programs that can learn through experience I with respect to a class of tasks (T), and an evaluation measure (P), if its performance on the tasks of T, evaluated by P, increases with E (Mitchell, 1997). Many ML algorithms are known to be useful in various aspects of biology. This includes photo-based species identification (Wäldchen and Mäder 2018), morphology-based species delimitation and description (Domingos et al., 2014; Breitman et al., 2018), biodiversity monitoring (McClure et al., 2020), behavioural studies (Valletta et al., 2017; Wang, 2019), DNA sequencing (Libbrecht and Noble, 2015; Liu, 2019), population genetics (Sheehan and Song 2016; Schrider and Kern, 2018; Fonseca and Carstens, 2024), ecology (Christin et al., 2019; Scalon et al., 2020; Pichler et al., 2020; Lürig et al., 2021; Silva et al., 2024), medicine (Sidey-Gibbons and Sidey-Gibbons, 2019), microbiology (Qu et al., 2019), and more (see Fountain-Jones et al., 2021; Morimoto et al., 2021; Borowiec et al., 2022).

Therefore, ML's potential in evolutionary biology, and particularly in species delimitation, is evident (Karbstein et al., 2024). Specific examples can also be found in studies involving model selection in demography and phylogeography (Pudlo et al., 2016; Fonseca et al., 2021), speciation (Blischak et al., 2021), phylogenetics (Suvorov et al., 2020; Solis-Lemus et al., 2022 preprint; Smith and Hahn, 2023; Zaharias et al., 2022; Mo et al., 2024), and species delimitation (Pei et al., 2018; Derkarabetian et al., 2019; Smith and Carstens, 2020; Pyron et al., 2023), with the last one forming the primary focus of this review.

In the following sections, we provide an overview of ML applications in the context of species delimitation, with an emphasis on those that operate using molecular data.

2. Current ML applications for species delimitation

In the same way that there are two primary categories of ML, namely supervised and unsupervised learning (SML and UML, respectively), species delimitation methods can also be broadly categorized into two main groups: discovery and validation (see Carstens et al., 2013; Rannala, 2015). Discovery approaches involve grouping samples without prior information (Pons et al., 2006; O'Meara, 2010; Huelsenbeck et al., 2011), while validation approaches require researchers to first assign the samples to potential lineages (species hypotheses) before testing them (Flouri et al., 2018; Sukumaran et al., 2021). This draws a conceptual parallel between traditional discovery approaches and UML methods, and between validation methods and supervised algorithms (Fig. 1). Also, it is important to note that ML methods are likelihood-free species delimitation approaches, offering several advantages over **likelihood-based approaches**. For example, by avoiding the need for explicit likelihood calculations, these methods might be computationally advantageous, particularly when combined with approaches optimized for high-throughput data processing, making them particularly suitable for analyzing large datasets with many taxa.

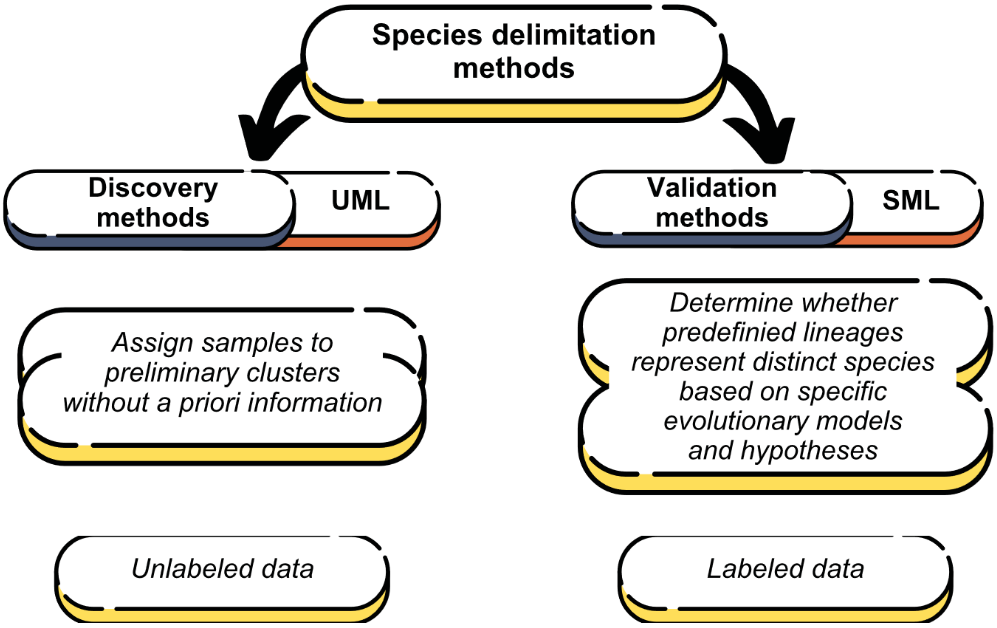

Species delimitation methods
Discovery methods
UML
Validation methods
SML
Assign samples to preliminary clusters without a priori information
Determine whether predefinied lineages represent distinct species based on specific evolutionary models and hypotheses
Unlabeled data
Labeled data

Fig. 1. Comparative diagram categorizing species delimitation methods and machine learning algorithms, along with some of their key characteristics. Species delimitation methods can be broadly categorized as discovery and validation methods, akin to unsupervised and supervised machine learning algorithms, respectively.

Below, we present a comprehensive overview of recently applied ML methods in the domain of molecular species delimitation, emphasizing their computational attributes and underlying assumptions. Our selection process involved a thorough search across scientific literature repositories, databases, and online journals, with a specific emphasis on studies featuring ML methods and workflows explicitly designed for species limits inference. We prioritized studies that either introduced novel methodologies (see Table 1) or enhanced and tested existing techniques in this context (Table A.1 in Appendix B). In our selection process, we focused exclusively on projects directly dedicated to species delimitation, despite the abundant literature on ML within related fields such as demography, population genetics, and phylogeography. Additionally, our emphasis is on methods designed for analyzing DNA sequence data. The categorized methods include SML, UML, and **deep learning**. While the backend processes may differ among such ML categories, their main goal when it comes to species delimitation usually remains the same: to analyze a given set of test data and classify it into distinct outcomes that define the species represented within the data.

Some studies applied ML techniques using other types of data rather than molecular information, such as morphology or ecology, for species delimitation and integrative taxonomy. A brief exploratory section regarding these particular studies can be found in Appendix B.

Table 1. List of proposed ML applications specifically designed to work on inferences about species limits.

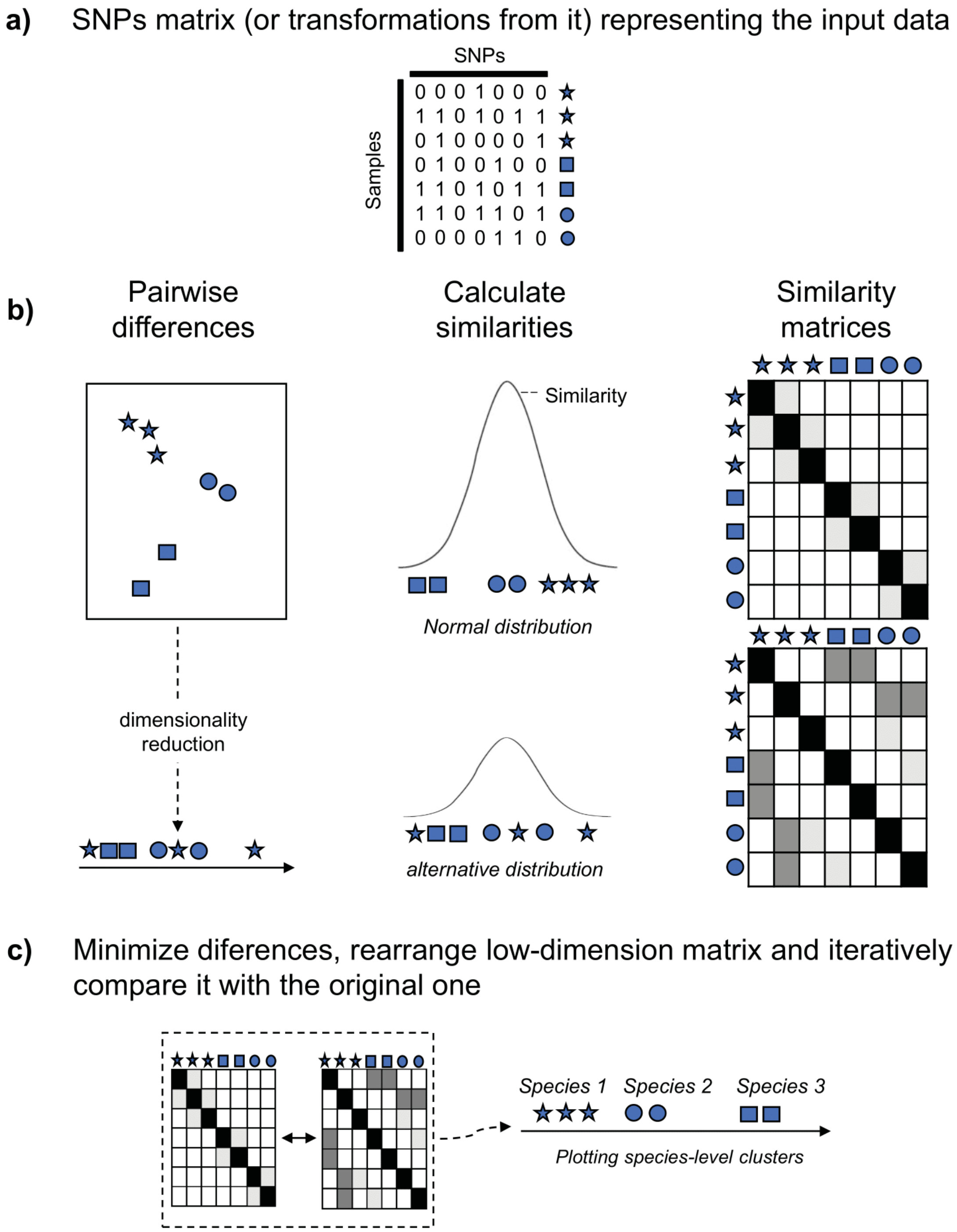
Reference	Languages	Category	Algorithms	Simulator	Input	Data representation
CLADES: A Classification-based Machine Learning Method for Species Delimitation from Population Genetic Data (Pei et al., 2018)[1]	python	SML	Support vector machines	Mccoal	Multiple sequence alignment (MSA) or SNP matrix	Population genetics summary statistics
A demonstration of unsupervised machine learning in species delimitation (Derkarabetian et al., 2019)[2]	R/python	SML & UML	t-Distributed Stochastic Neighbor Embedding, Random Forest, Variational autoencoders	NA	SNP data matrix	One-hot-encoding of the SNP data matrix (VAE), *axis* from a discriminant analysis of principal components (t-SNE), scaled data from DAPC + cMDS and isoMDS ouput (Random forest)
Process-based species delimitation leads to identification of more biologically relevant species (Smith and Carstens, 2020)[3]	python	SML	Random forest	fastsimcoal	SNP data matrix	Folded multi-dimensional SFS
Coalescent-based species delimitation meets deep learning: Insights from a highly fragmented cactus system (Perez et al., 2021)[4]	python	Deep learning	Convolutional neural networks	ms	SNP data matrix	Matrices (as images), with genotypes encoded as higher or lower frequency states
Speciation Hypotheses from Phylogeographic Delimitation Yield an Integrative Taxonomy for Seal Salamanders (*Desmognathus monticola*) (Pyron et al., 2023)[5]	R	UML	Self-organizing maps (SOMs)	NA	SNP data matrix	SNP matrix, in which the rows are individual specimens, the columns are the 2-4 possible states at each SNP locus, and the entries are the frequency of that state

Online repositories where it is possible to find more information about the currently existing platforms. [1] https://github.com/pjweggy/CLADES; [2]https://www.sciencedirect.com/science/article/abs/pii/S1055790319301721; [3]https://github.com/meganlsmith/delimitR; [4]https://github.com/manolofperez/CNN_spDelimitation_Piloso; [5]https://github.com/kyleaoconnell22/Pyron_et_al_UML_sp_delim/tree/main

2.1 Discovery and unsupervised methods

Unsupervised machine learning (UML) relies on the inherent data structure to find patterns within the data, whether by clustering similar data points together, reducing the dimensionality of the data while retaining essential information, a combination of both, or by identifying unusual patterns or outliers, which may indicate errors or novel phenomena (Hastie et al., 2009; Libbrecht and Noble, 2015; Dike et al., 2018). UML algorithms are often regarded as methods lacking strong predefined assumptions about the underlying structure of the dataset (such as population parameters, species numbers, or sample categorization, in the case of species delimitation). Nevertheless, it is possible to incorporate heuristic or pragmatic assumptions in an UML framework to facilitate their operation. Either way, UML will be particularly useful in cases where prior hypotheses are limited or unavailable, provided that the assumptions of the chosen method are evaluated.

UML clustering methods group input data into subsets, where samples with high similarities are placed in the same cluster and exhibit less similarity with samples in other clusters. Meanwhile, UML dimensionality reduction techniques compress data to identify a smaller distinct set of variables that retain essential features of the original data, while minimizing information loss. We highlight this as, when it comes to species delimitation, UML approaches often operate through clustering and/or dimensionality reduction algorithms (Fig. 2), extracting and condensing the necessary information to identify limits between biological groups (Derkarabetian et al., 2019; Pyron, 2023; Pyron et al., 2023), while also enabling the simultaneous use of different types of data (e.g., genetical, phenotypical and ecological).


a) SNPs matrix (or transformations from it) representing the input data
SNPs
Samples
b)
Pairwise differences
Calculate similarities
Similarity matrices
Similarity
Normal distribution
dimensionality reduction
alternative distribution
c) Minimize diferences, rearrange low-dimension matrix and iteratively compare it with the original one
Species 1
Species 2
Species 3
Plotting species-level clusters

0	0	0	1	0	0	0	★
1	1	0	1	0	1	1	★
0	1	0	0	0	0	1	★
0	1	0	0	1	0	0	■
1	1	0	1	0	1	1	■
1	1	0	1	1	0	1	●
0	0	0	0	1	1	0	●

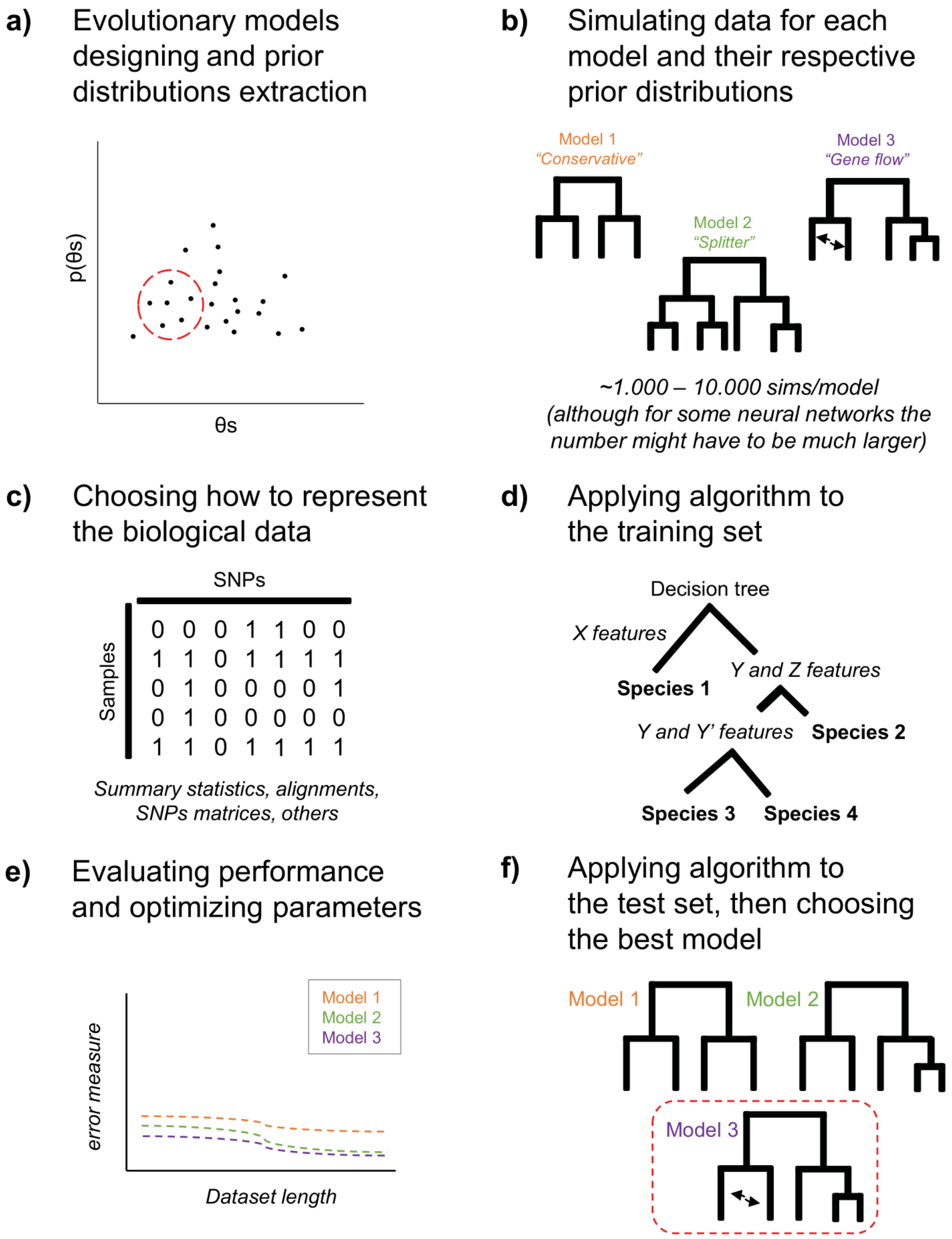
Fig. 2. Diagram outlining a potential UML workflow for species delimitation, utilizing the t-SNE algorithm as an example (inspired by Derkarabetian et al., 2019). A) Data representation is the initial step, and it varies depending on the chosen ML tool, which may work with sequence data, SNP matrices, or population genetics metrics extracted from them; in this case, samples from different populations are represented by distinct symbols. B) t-SNE, as a dimensionality reduction technique, iteratively finds a lower-dimensional representation of the original data. It identifies local similarity spaces between sample pairs by analyzing Gaussian and lower-dimensional distributions, such as the Cauchy or t-student with one degree of freedom. C) The algorithm's goal is to align the new similarity matrix with the original data by iteratively moving data points closer to their nearest neighbors in the higher-dimensional space and away from more distant ones. This process continues until the maximum number of iterations is reached or no further improvements can be made, resulting in the proper grouping of samples based on their similarities (e.g., individuals or populations assigned to a species based on the chosen data representation).

Derkarabetian et al. (2019) evaluated the performance of different ML methods for species delimitation, including both SML and UML algorithms. Specifically, they evaluated the capacity of three approaches: **Random Forest** (RF, including a supervised and a non-supervised alternative), and two unsupervised models, **variational autoencoders (VAE)** and **t-Distributed Stochastic Neighbor Embedding (t-SNE).** In the t-SNE approach, data derived from a **principal component analysis (PCA)** were used as input variables, followed by clustering techniques using the output from the UML algorithms. In the VAE approach, **single-nucleotide polymorphism (SNP)** matrices were converted via **one-hot encoding,** where nucleotides were transformed into binary variables. In this case, the encoder takes the transformed SNP data and infers the distribution of latent variables, given as a normal distribution with a mean (μ) and standard deviation (σ). The decoder maps the latent distribution to a reconstruction of the one-hot encoded SNP data, offering a two-dimensional depiction. Finally, the RF approaches were performed with scaled data derived from a **Discriminant Analysis of Principal Components (DAPC)**, and the resulting proximity matrix was then used for **classical multidimensional scaling (cMDS)** and **isotonic multidimensional scaling (isoMDS)**. In sum, all approaches yielded, through different clustering strategies (depending on the algorithm being investigated), more readily interpretable outcomes compared to other traditional delimitation methods (or population structure detection methods) assessed by the authors, revealing distinct species groupings (Derkarabetian et al., 2019). Notably, the identified groups also corresponded to those of an integrative taxonomy approach, suggesting that the limits identified by UML algorithms probably correspond to species-level divergence rather than population structure (Derkarabetian et al., 2019).

Pyron et al. (2023) introduced a novel UML approach designed for delineating species limits from extensive genomic datasets, primarily based on **self-organizing maps (SOMs)**. This approach produces discrete outcomes rather than continuous ones, grouping genotypes based on similarity. Additionally, the authors propose determining the number of species by analyzing the degree of grid occupancy in the SOM output. This quantification establishes how many units, representing distinct clusters of genotypes, have been effectively mapped from the original SNP matrix. Subsequently, the method estimates the cumulative distances from each sample to its immediate neighbors. To effectively separate candidate species, Pyron et al. (2023) recommend performing cluster analyses, such as k-means. The determination of the optimal number of **classes,** or species, is achieved by selecting the value that maximizes the sequential reduction in the weighted sum of squares from k to $k + 1$. An extension of this method has been proposed in the form of a SuperSOM approach, incorporating the possibility of using several trait classes simultaneously, such as genetic, morphological and ecological variables (Pyron, 2023).

2.2. Validation and supervised methods

While UML approaches are powerful and widely applicable, SML offers distinct analytical advantages in certain scenarios, contributing to its widespread use in population genetics and evolutionary biology (Schrider and Kern, 2018). A primary requirement for SML is the availability of **labeled training data**, which is used to teach the algorithm to recognize patterns and make predictions. In the context of population genetic analyses, such labeled datasets are often unavailable or insufficient in size. To overcome this limitation, simulated genetic data based on known evolutionary models are usually generated to represent evolutionary scenarios. This simulated data is then encoded, along with observed genetic data, into feature vectors used to train the algorithm, which is used to recognize specific patterns in new observed data points (Fig. 3).



a) Evolutionary models designing and prior distributions extraction

b) Simulating data for each model and their respective prior distributions

c) Choosing how to represent the biological data

d) Applying algorithm to the training set

e) Evaluating performance and optimizing parameters

f) Applying algorithm to the test set, then choosing the best model

0	0	0	1	1	0	0
1	1	0	1	1	1	1
0	1	0	0	0	0	1
0	1	0	0	0	0	0
1	1	0	1	1	1	1

Fig. 3. Diagram illustrating a potential SML workflow for species delimitation, here using a decision-tree based algorithm as an example (inspired by the work of Smith and Carstens, 2020). A) The initial step involves, from a wider set of priors, extracting relevant subsets of priors for the evolutionary models of interest (clusters of dots circled in red). B) Simulated data is generated for each model, typically ranging from 1,000 to 10,000 simulations per model, using relevant simulation software; the specified models may involve variations in topology, such as scenarios with differing numbers of potential species ('*conservative*' versus *'splitter'*), and can also account for the possibility of gene flow. C) The data is represented according to the requirements of the chosen ML tool. D) Following data simulation and representation, ML model training begins, involving various preliminary steps like data pre-processing, dataset division, feature selection, and algorithm choice. E) Model performance (both in terms of biological accuracy and computationally) is assessed using statistical metrics, allowing for retraining and adjustment based on the results. F) After the algorithm is properly trained and evaluated, it can be used to predict species limits for new datasets (whether they are newly simulated or empirical data), using the model identified as best representing the species limits in the biological system (indicated here by the dashed red line).

The reliability of SML methods rely on the resemblance between the training data (typically simulated) and the actual biological data. Thus, the process of applying ML algorithms in species delimitation is influenced by the assumptions of the underlying evolutionary processes, such as population size, selection strength, and gene flow. Anyhow, SML algorithms generally demand a significantly smaller amount of simulated data compared to other methods based on simulations, such as **Approximate Bayesian Computation (ABC)**, resulting in reduced computational effort (e.g., a few thousand simulated datasets *versus* hundreds of thousands of simulations per scenario in most ABC approaches; Csilléry et al., 2010; Pudlo et al., 2016; Raynal et al., 2019).

CLADES (Pei et al., 2018) is an SML-based approach for species delimitation that employs **classification models** trained and tested on *multilocus* sequence data. Within this framework, **support vector machines (SVM)** are used to classify population pairs as either belonging to the same species or distinct species. Regarding model training, datasets at the population level are simulated, with and without gene flow. Then, each training sample is represented as a list of **summary statistics**, and a SVM **regression** is estimated, through iterative training, to minimize the misclassification cost. Subsequently, the SVM classifier computes the probability of the training samples belonging to each potential grouping.

Notably, the training dataset in this study was simulated based on a two-species model (A and B) where both species diverged at time τ with identical population size parameters (θA = θB = θ). Each species further consisted of two populations that recently split at time τp. **Migration** between A and B was allowed at a rate of M = N*m* migrants per generation, with *m* representing the migration rate per generation. Additionally, symmetrical migration between A and B was accounted for prior to their divergence into two distinct populations each (A1 and A2, and B1 and B2). Multilocus sequence data of length L were simulated under diverse parameter combinations using the Mccoal software (Rannala and Yang, 2003). For each possible parameter combination (θ, τ, M), sequences were simulated for 100 loci with a length of L = 100Kbp for all populations. For each locus, 40 sequences were sampled, with 10 sequences per population. All training samples were combined to train a global classifier, enabling it to adapt to various values of θ and M instead of assuming fixed parameters. With regard to CLADES' performance, longer loci improved its efficiency, and this approach was robust to different modeling structures, accommodating various demographic events and evolutionary parameters.

Smith and Carstens (2020) introduced delimitR, a SML approach designed to conduct species delimitation in a model selection task; delimitR employs the multidimensional **site**

frequency spectrum (mSFS) with a **binning** strategy as a predictor variable for a RF classifier. In essence, this framework aims to discriminate between various divergence models compatible with virtually any species concept, as asserted by the authors. Besides, working with data summarized through the mSFS, delimitR facilitates the evaluation of models that vary in terms of lineage numbers. Either way, given its supervised nature, delimitR demands researchers to define reasonable priors, such as divergence times or migration rates, and decide which models will be assessed. For each model evaluated in their study, Smith and Carstens (2020) simulated 10,000 mSFS. A RF classifier was constructed using 1,000 **decision trees** to accommodate the extensive number of models. delimitR's performance improved with larger SNP matrices and increasing divergence times. Compared to ABC methods, delimitR showed lower error rates, even though the detection of migration was challenging in cases of recent divergence between lineages (Smith and Carstens, 2020). The authors acknowledge that further research is needed to elucidate the association between the model space, number of parameters, and delimitation accuracy.

2.3. Deep learning

Deep learning is a subset of ML that focuses on training **artificial neural networks (ANNs)** with multiple layers (hence "deep") to perform complex tasks (Sheehan and Song, 2016). In terms of data labeling, deep learning algorithms can be both supervised or unsupervised. Deep learning techniques have found success in various fields in the Biological Sciences (Angermueller et al., 2016; Sheehan and Song, 2016; Schrider and Kern, 2018). However, its adoption in Evolutionary Biology is relatively recent (see Angermueller et al., 2016; Sheehan and Song, 2016; Fonseca et al., 2021; Blischak et al., 2021; Yelmen and Jay, 2023). The popularity of deep learning can be attributed to their highly flexible data input and output structure, allowing networks trained for one task to be repurposed for another by modifying their final **layers**, for instance, through **transfer learning** approaches. This versatility enables the resolution of intricate tasks that might prove challenging for **shallow learning** algorithms. Conversely, deep learning often demands meticulous and more specific fine-tuning compared to shallow learning methods. For a detailed description of how neural networks work, and their general structure, see Sheehan and Song (2016), Borowiec et al. (2022), and Korfmann et al. (2023).

The fundamental stages involved in creating a deep learning framework for species delimitation, especially a supervised one, closely parallel those of a shallow SML workflow, as both typically involve formulating evolutionary models and simulating data. Broadly, these steps include data simulation and representation, **model** training and optimization, and ultimately, predicting relevant categories from empirical data (Fig. 3).

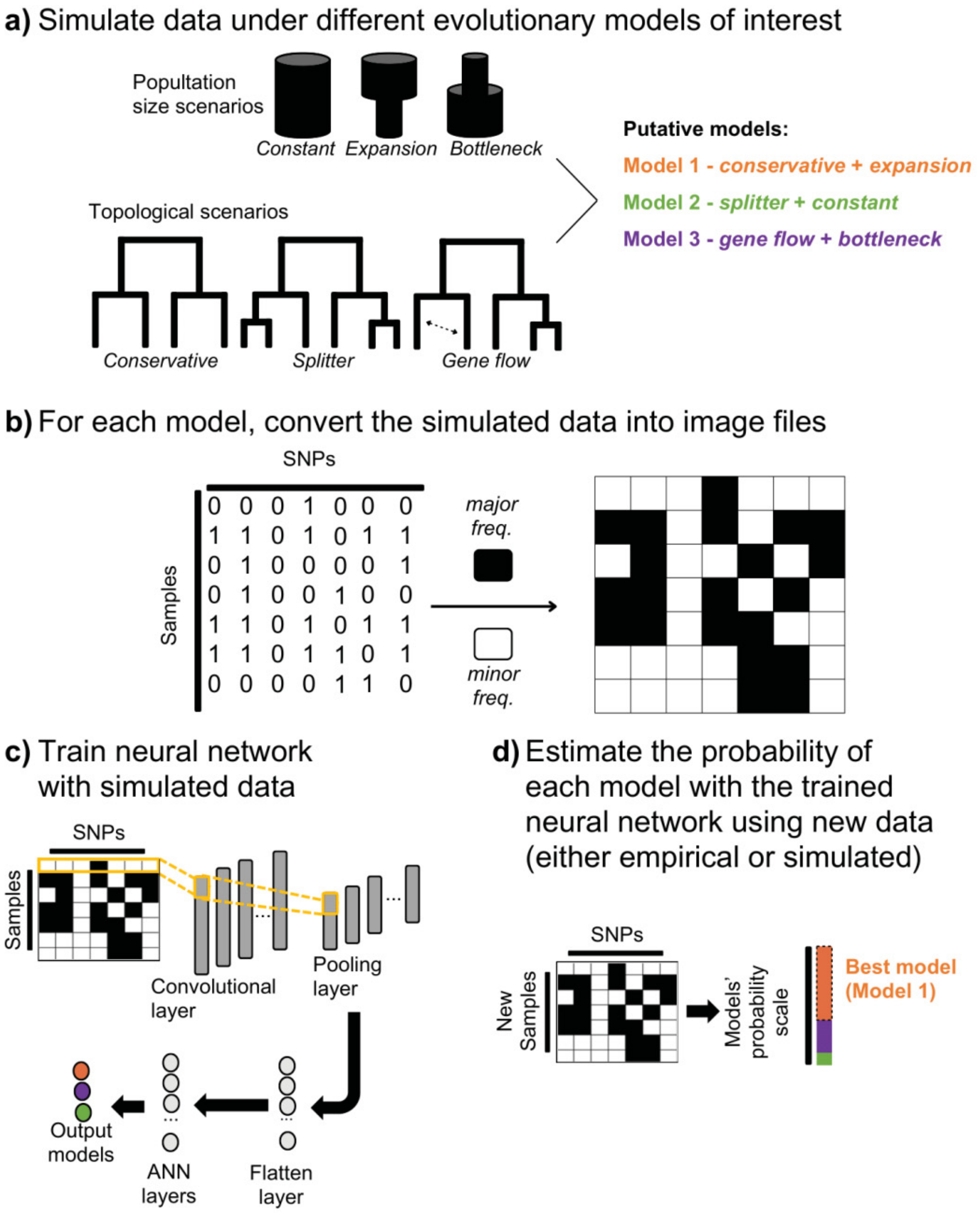

a) Simulate data under different evolutionary models of interest
Popultation size scenarios
Constant
Expansion
Bottleneck
Topological scenarios
Conservative
Splitter
Gene flow
Putative models:
Model 1 - conservative + expansion
Model 2 - splitter + constant
Model 3 - gene flow + bottleneck
b) For each model, convert the simulated data into image files
SNPs
Samples
0 0 0 1 0 0 0
1 1 0 1 0 1 1
0 1 0 0 0 0 1
0 1 0 0 1 0 0
1 1 0 1 0 1 1
1 1 0 1 1 0 1
0 0 0 0 1 1 0
major freq.
minor freq.
c) Train neural network with simulated data
SNPs
Samples
Convolutional layer
Pooling layer
Output models
ANN layers
Flatten layer
d) Estimate the probability of each model with the trained neural network using new data (either empirical or simulated)
SNPs
New Samples
Models' probability scale
Best model (Model 1)

Fig. 4. Diagram illustrating a potential deep learning workflow applied in the context of species delimitation, using CNNs as an example of algorithm that can be used in this context (inspired by Perez et al., 2021). A) The process typically begins with the simulation of biological data under various evolutionary models, considering factors like topology (e.g., scenarios with differing numbers of potential species, namely 'conservative' and 'splitter'), population size (considering potential demographic variations over time, whether of population contraction or expansion), gene flow, and many more, similar to a SML pipeline. B) Next, data representation is crucial. For CNNs, SNP matrices are often converted into arrays or image files, where pixel contrast reflects differences in minor and major frequencies between samples. C) With the simulated and properly represented data, the network training phase can commence. The parameter configuration and network architecture may vary, depending on the specific study's requirements. D) Once each model is trained and its performance is rigorously evaluated, the final stage of the workflow involves predicting categories for new data. This can include using new simulated data with slight parametric modifications, still within the trained model's limits, as well as empirical data whose evolutionary history aligns with the proposed model. In both cases, the goal is to determine which delimitation model best applies to the biological system being investigated.

Perez et al. (2021) proposed a delimitation approach that accommodates the integration of coalescence-based methods with model selection using **convolutional neural networks (CNNs)**. In short, this approach can integrate models from coalescent analyses, such as BPP (Flouri et al., 2018; 2020), to compare different evolutionary scenarios while incorporating information from multiple sources. Specifically, it allows users to combine insights from genetic analyses (e.g., coalescent-based methods) with hypotheses derived from other data types (e.g., phenotypic traits) that reflect different taxonomic arrangements. This flexibility enables the formulation of models informed by multiple lines of evidence. The initial steps in this approach involve simulating genetic data for each delimitation hypothesis, with the study encompassing 10,000 simulations per model. These simulations are then converted into images, which serve as input for training a neural network. It is worth noting that while CNNs used 10,000 simulations per model in this study, ABC required 100,000 simulations per model. Finally, each species hypothesis probability can be predicted through the trained CNNs using a **test set**. Perez et al. (2021) also compared their model selection approach with ABC using empirical data. The CNNs consistently demonstrated superior performance in distinguishing between the simulated evolutionary scenarios, outperforming ABC in all cases, with fewer simulations and faster execution times (Perez et al., 2021).

2.4. How has machine learning changed our approach to delimit species so far?

To date, relatively few studies (<20, see Appendix B) have specifically explored ML techniques for species delimitation, in particular focusing on molecular data. Among these, only five introduced novel ML approaches for species delimitation, providing comprehensive details from initial simulations to statistical performance evaluations (Pei et al., 2018; Derkarabetian et al., 2019; Smith and Carstens, 2020; Perez et al., 2021; Pyron et al., 2023).

These approaches, and also other ML frameworks applied in demographic inferences, phylogeography and population genetics, are often advocated by the researchers and developers themselves on the following arguments: i) challenges and limitations associated with the assumptions of coalescent methods (Derkarabetian et al., 2019; Smith and Carstens, 2020; Blischak et al., 2021; Martin et al., 2021; Derkarabetian et al., 2022); ii) ML computational efficiency and the capacity of handling complex evolutionary models (Pei et al., 2018; Martin et al., 2021; Perez et al., 2021; Derkarabetian et al., 2022; Pyron et al., 2023); and iii) ML acting as a likelihood-free approach, enabling the consideration of models where likelihood computation would be intractable (Smith and Carstens, 2020; Martin et al., 2021; Perez et al., 2021; Sanchez et al., 2020). Also, while ML algorithms are often used similarly to simulation-based approaches like ABC, additional steps are generally incorporated, such as: i) selecting a more comprehensive subset of summary statistics based on specific criteria (Smith and Carstens, 2020; Martin et al., 2021); ii) handling larger or more complex genetic datasets more efficiently compared to model selection tools such as ABC (Smith and Carstens, 2020; Collin et al., 2021; Ghirotto et al., 2021). These advantages stem from the fact that ML approaches usually require less specificity in summary statistic selection and can manage high-dimensional data with fewer concerns about the **curse of dimensionality**.

2.5. What types of species ML methods might be detecting?

A significant part of the studies we analyzed were philosophically based on species concepts grounded on evolutionary or genealogical independence criteria. This might stem from our focus on workflows using molecular data, which generally aims at identifying lineages and genetic clusters characterized by significant levels of genetic divergence and restricted amounts of gene flow. Also, some studies specifically model parameters like migration, which make them in line with concepts focused on reproductive criteria. While evolutionary and genealogical independence evidence (or reproductive criteria) may have their limitations in investigating species limits, results generated by ML methods in this context can still serve as hypotheses for further investigations (e.g., Fujita et al., 2012), aligning with the GLC perspective (de Queiroz, 1998; 1999; 2005b).

So far, there are no definitive coalescent-based solutions to differentiate between population structure and species (Sukumaran and Knowles, 2017; Leaché et al., 2019). In this context, it is reasonable to assert that ML-based delimitation methods, just as coalescence-based methods, might not always be identifying species *per se*, but rather: i) incompletely separated (or incipient) species, which may eventually be classified as distinct (Burbrink et al., 2021), or even as ‘subspecies’ (de Queiroz, 2020); or ii) population or phylogeographic variation (Rosenblum et al., 2012; Sukumaran et al., 2021). Consequently, while genetic structure (either at population or species level) detected through ML can be biologically relevant for species delimitation, additional data and an evolutionary process-based perspective remain crucial to discern the nature of the inferred biological entities (Smith and Carstens, 2020; Sukumaran et al., 2021). Just as phenotypic, ecological, or other biological attributes are not mandatory criteria for designating an evolutionary lineage as a species (de Queiroz, 2007; Pyron et al., 2023), genetic or genealogical groupings identified using ML-based delimitation methods can be similarly interpreted.

Within this context, while the primary criterion for recognizing a species can still be evolutionary independence, other characteristics could serve as secondary evidence of divergence and may be also analyzed using ML frameworks. Given ML's versatility in handling diverse data types, future applications in species delimitation should prioritize the explicit incorporation and evaluation of diverse biological properties—such as genomic divergence, ecological adaptation, and phenotypic differentiation—to enhance species hypothesis testing (e.g., Karbstein et al., 2024; Pyron et al., 2024). Several strategies can achieve this, including integrating different feature categories as distinct layers within a deep learning architecture. Besides, investigating how the contribution of various traits impacts species delimitation across different biological systems also represents a key avenue for future research. Only a few detailed ML pipelines have been proposed in this context, aiming to explore the relationship between evolutionary models and divergence scenarios in terms of distinct characteristics, whether genetic, phenotypic, geographic or ecological. Pyron (2023), for instance, implemented a UML method using SOMs for learning high-dimensional associations between observations (e.g., specimens) across a wide set of input features (e.g., genetics, geography, environment, and phenotype). Yang et al. (2022) is another great example, which introduced a CNN method that successfully integrates morphological and molecular data for species identification.

In sum, integrating genetic, ecological, and phenotypic data may be essential for achieving robust and reliable species limits assessments, particularly in systems with complex evolutionary histories—such as cryptic species complexes or hybridizing lineages. In this context, ML-based species delimitation offers a powerful framework to combine domain expertise with quantitative hypothesis testing, optimizing the reconciliation of conflicting evidence. This approach also paves the way for establishing comprehensive frameworks rooted in modern Integrative Taxonomy, potentially enabling the automated synthesis of diverse data to accurately define taxonomic units (Karbstein et al., 2024).

3. Advantages, limitations and future perspectives

3.1. Strengths and benefits of using ML to delimit species

In general, ML methods applied to infer species limits based on genetic data offer some advantages over coalescent or traditional simulation-based methods. Despite particular constraints, ML algorithms can be as accurate (in biological terms) as traditional model selection tools and likelihood-based species delimitation methods (Pei et al., 2018; Smith and Carstens, 2020; Perez et al., 2021; Derkarabetian et al., 2022). Because of their likelihood-free nature, they are computationally more efficient and generally can be trained on models that are at times too intricate for formal statistical estimators (Pei et al., 2018; Kuzenkov et al., 2020; Smith and Carstens, 2020; Suvorov et al., 2020; Martin et al., 2021; Perez et al., 2021). Some of these algorithms have proven to be highly efficient in complex evolutionary scenarios, including situations involving gene flow or population size fluctuations (Pei et al., 2018; Perez et al., 2021). This efficiency does not compromise the ability to distinguish between different models (Smith et al., 2017), and even simple SML methods provide high selection accuracy when comparing multiple models in a single analysis (Gehara et al., 2020 preprint).

A major advantage of deep learning, in particular, is the capacity to automatically extract information from alignments (commonly treated as images), as opposed to relying on summary statistics typically required by other methods. This facilitates accurate and efficient classification or regression tasks, as observed in studies by Sanchez et al. (2020), Fonseca et al. (2021), Perez et al. (2021), and Borowiec et al. (2022), holding promise in future species delimitation studies. Besides, especially in supervised approaches, which often use explicit evolutionary models to validate species (e.g., Smith and Carstens, 2020), ML enables a more in-depth exploration of the speciation and phylogeographic processes that underlie the formation of independent evolutionary lineages. Thus, given that properly sampled genomic datasets can offer sufficient data for analyzing complex evolutionary models, ML might serve a dual role: providing primary evidence for examining species limits patterns, and assisting in the investigation and reconstruction of the evolutionary processes responsible for these patterns.

3.2. Factors requiring careful consideration in ML-based species delimitation

Certain algorithms, especially supervised ones trained on simulated data, may become overly specialized. Modern ML methods are proficient at interpolating within the observed range of values in the training data, even in cases where specific values have not been encountered before, being adaptive and not solely reliant on memorizing specific training instances. Even so, as models are typically trained on simulated data with specific values of evolutionary parameters, such as θ and M, their performance might be compromised when applied far outside the training parameter space (Schrider and Kern, 2018; Borowiec et al., 2022). Besides, ML algorithms have some degree of **inductive bias** (Hüllermeier et al., 2013). Therefore, exploring in further details the association between training capacity and predictive power should be a priority for future studies.

Methods relying on a substantial volume of simulated data across diverse evolutionary scenarios must carefully design prior distributions to ensure that the data generated under these models closely reflect the actual biological system being studied. This is a challenge for non-model organisms, where data availability may limit the quality of parameter estimates (Tagu et al., 2014; Fonseca et al., 2016; Cerca et al., 2021; Jorna et al., 2021). Importantly, simulation problems are not exclusive to ML-based workflows, as model selection frameworks such as ABC also employ simulated data (Beaumont et al., 2002; Bertorelle et al., 2010). Furthermore, it may be unfeasible to simulate data or train an ML algorithm across an entire parameter space, especially in complex evolutionary models (Rannala and Yang, 2020), and important phenomena may be entirely missing from the simulations (e.g., background selection, Mo and Siepel (2023), or missing data Arnab et al. (2023)). Also, limited information is available regarding the asymptotic statistical performance of most ML methods applied for species delimitation. Thus, such models may never be comprehensive enough, have limitations in representing real data, and demand substantial computational resources (Arenas, 2012; Mangul et al., 2019; Zaharias et al., 2022). This leads to an inherent challenge in avoiding some degree of misspecification in the training data, even considering the variety of powerful genetic data simulators currently available.

Moreover, all model-based methods depend on the chosen models and their parameters, whether they are used for simulations or for direct likelihood estimation. As a result, even methods that do not rely on simulations can still be sensitive to model misspecification. For example, coalescence-based approaches depend on MSC assumptions, which may not always accurately represent specific biological systems. Likelihood-based approaches offer advantages in exploring parameter space within a defined model—due to their optimality and iterative nature—though they can be computationally intensive (e.g., Flouri et al., 2018; Sukumaran et al., 2021). Thus, these methods remain important alternatives especially when there is no clear reference for simulations. ML approaches, on the other hand, have the potential to explore a broader model space. That said, no approach can account for all possible evolutionary processes, leaving both traditional SDMs and ML approaches limited in their ability to comprehensively explore the broad space of evolutionary models. Only UML approaches might be relatively immune to some of these constraints, as they do not rely on predefined models. Either way, regardless of the analytical framework, misrepresenting evolutionary relationships can lead to misleading outcomes. This underscores the need for biologically informed feature processing and modeling.

Another important perspective to consider is related to data representation. While ML can uncover patterns in high-dimensional datasets, its performance heavily relies on the quality and relevance of the input data and how it is representated (Guyon and Elisseeff, 2003; Domingos, 2012; LeCun et al., 2015). In the context of the present study, there are ML pipelines that utilize data derived from SNPs matrices (Derkarabetian et al., 2019; Sanchez et al., 2020; Smith and Carstens, 2020; Blischak et al., 2021; Fonseca et al., 2021; Martin et al., 2021), and only a few are extensible to different genetic markers (e.g., Collin et al., 2021). Also, numerous studies using ML frameworks, whether focusing on species delimitation, demography, or population genetics, use genetic summary statistics as the main input data (e.g., Pei et al., 2018; Collin et al., 2021; Ghirotto et al., 2021).

While summary statistics can be valuable for distinguishing between models, some may not be suitable for making inferences about species limits. Besides, the practical implementation of such statistics on the detection of specific evolutionary processes often encounters confounding factors that can mimic similar effects on gene histories (Flagel et al., 2019). For example, Tajima's D is a statistic sensitive to both positive selection and changes in population size (Simonsen et al., 1995). Therefore, unless we have a clear understanding of which type of data is truly sufficient to capture the target biological phenomena, relying solely on a specific set of statistics can lead to inevitable information loss (Rannala and Yang, 2020). An alternative to this is to consider the sequence alignment itself as input, as demonstrated in the deep learning approach introduced by Perez et al. (2021), which implicitly enables dimensionality reduction while capturing structures within the input data. Notably, deep learning techniques are valuable tools in this context, offering the capability to analyze both raw sequence data and summary statistics (Korfmann et al., 2023).

Even considering that data representation is a critical component of any analytical framework, its role in species delimitation demands particular attention, where complex evolutionary processes such as gene flow and incomplete lineage sorting (ILS) can leave distinct signatures in the data. For instance, gene flow may produce localized discordance in allele frequencies and introgressed genomic segments, whereas ILS typically results in widespread gene tree incongruence due to ancestral polymorphisms. Consequently, how data is represented (e.g., via full sequence alignments, SNP matrices, or gene tree reconstructions) can impact the ability to detect and distinguish these phenomena, and consequently the robustness and interpretability of delimitation results. Therefore, priority should be given to representations that retain detailed information for detecting key processes in species delimitation while preserving the inherent variability and structure of the data. Future research should focus on understanding the extent to which the flexibility of ML to handle various input data types provides a true advantage for species delimitation. Moreover, despite the challenges associated with comparing ML approaches due to differing assumptions, employing diverse training data representations—from genomic sequences to summary statistics—could help illuminate the strengths and limitations of each method for detecting species limits.

3.3. Possible avenues and prospects for future studies

Mitigating the effects of misspecification during simulation might involve designing or using a simulator that enforces high compatibility between simulated and actual data. Generative adversarial networks (GANs), a type of deep learning algorithm commonly used for creating synthetic images and voices (Chadha et al., 2021), have shown promise in this regard (see Wang et al., 2021; Callier, 2022). GANs operate with two networks, the generator and the discriminator, trained together (Goodfellow et al., 2014). While the generator simulates data, the discriminator distinguishes between real and synthetic data. During training, the generator network becomes more powerful at producing realistic **examples**, and the discriminator network becomes more skilled at distinguishing between real and synthetic data. When training is complete, the generator network can generate new examples that are indistinguishable from real data, providing a reliable way to work with labeled data. Researchers have already assessed the utility of GANs in various fields, including genomics, phylogenetics, and population genetics (Nesterenko et al., 2022 preprint; Booker et al., 2023; Yelmen and Jay, 2023). Smith and Hahn (2023) introduced phyloGAN, a workflow that takes a concatenated alignment (or a set of alignments) as input and infers a phylogenetic tree, potentially accounting for gene tree heterogeneity.

The application of GANs is still incipient in Evolutionary Biology. Although the above-mentioned approaches perform well in relatively simple scenarios, methodological challenges arise as the complexity of the evolutionary model space increases. This can result from additional variables in evolutionary models or larger phylogenetic trees and sequence alignments, potentially affecting both accuracy and computational efficiency (Nesterenko et al., 2022 preprint; Smith and Hahn, 2023). Therefore, future advancements in the use of GANs in should focus on enhancing the efficiency of exploring parameter spaces, reducing computational training times, and accommodating more complex evolutionary models (Smith and Hahn, 2023). To fully harness the potential of this tool in species delimitation, further efforts are required to refine the population genetics parameters estimates (e.g., Wang et al., 2021), and to improve the accuracy of species limits inferences based on these parameters. Future GAN applications in this context should also focus on generating synthetic datasets to model realistic scenarios of species divergence under complex evolutionary processes.

Potential errors in data simulation can be linked to a "domain adaptation" problem as well, where a model trained on one data distribution is applied to a dataset originating from a different distribution (Farahani et al., 2021; Mo and Siepel, 2023). A classic illustration of domain adaptation is found in image classification: consider a situation in which a recognition model needs to identify different dog breeds from photographs ("target domain"), but the only labeled training data available are cartoon drawings of dogs ("source domain"). In such cases, a ML model must be trained on one dataset with the expectation of performing well on another, even in the presence of systematic differences between the two distributions. Recent solutions involve learning a "domain-invariant" data representation through a feature extractor neural network. This is accomplished by minimizing domain disparities (Rozantsev et al., 2018), using adversarial networks (Ganin and Lempitsky, 2015; Liu and Tuzel, 2016; Bousmalis et al., 2017), or employing auxiliary reconstruction tasks (Ghifary et al., 2016).

Domain adaptation techniques have found applications in fields such as genomics (Cochran et al., 2022) and population genetics (Mo and Siepel, 2023), particularly as an unsupervised domain adaptation problem. Through extensive simulation studies, Mo and Siepel (2023) convincingly demonstrated that their domain-adapted models significantly outperformed standard networks across various simulation misspecification scenarios. This outcome underscores the potential of domain adaptation techniques as a promising avenue for developing robust deep learning models in evolutionary biology inference (Mo and Siepel, 2023), potentially including species delimitation. In this area, future efforts should focus on mitigating problems introduced by sampling bias or model misspecifications across diverse evolutionary scenarios, particularly in supervised frameworks that rely on simulated data. Employing domain adaptation strategies to facilitate the integration of naturally heterogeneous datasets—such as genomic and morphological, environmental and geographical data—by extracting of domain-invariant features will also enhance the resolution and reliability of delimitation outcomes.

Future ML applications to infer species limits should also focus on the development of new transfer learning structures. For example, a deep learning **architecture** originally trained for inferring historical population sizes can be repurposed for classifying demographic scenarios (Pan and Yang, 2010), even though reusing trained models can be challenging due to differences in data dimensionality (Sanchez et al., 2020). Particularly regarding species delimitation, improving model generalizability could be achieved by transferring learning between well-studied taxonomic groups and those with limited data availability. In order to establish baseline models, a practical workflow could involve using ML algorithms to identify species limits in a broad training dataset, such as population-level genomic or morphological data from different species. Then, these baseline models could be fine-tuned with smaller, taxonomically specific datasets to validate or identify taxa in understudied groups. This iterative approach can also optimize computational efficiency by avoiding the need to train models from scratch. A methodology relatively aligned with this reasoning is exemplified in the study by Derkarabetian et al. (2022). Moreover, ML methods initially designed for other model selection purposes, such as phylogeography (Fonseca et al., 2021), could be reasonably adapted for species delimitation, provided that the simulated data and initial models adequately capture species limits nuances.

4. Optimizing the use of ML in the context of species delimitation

4.1. Enhancing Species Delimitation through accessible and purpose-built ML

The introduction of new ML approaches will increasingly enhance researchers' ability to make biologically precise decisions, especially when these methods are purpose-built, from conception to implementation, for the specific task of delimiting evolutionary lineages. In order to choose the appropriate species delimitation method, researchers must consider the available data and putative evolutionary scenarios, while considering the available statistical evaluation and performance optimization of each method (Greener et al., 2021; Morimoto et al., 2021). However, a comprehensive comparison of the recently proposed ML methods characteristics, advantages and disadvantages, and overall performance compared to other SDMs is still lacking.

Such evaluation should consider the inherent properties of the ML algorithms, such as how the workflows manipulate the data attributes, and the different types of data. In nearly all studies using ML methods to infer species limits, at least a minimal approach to estimating error or noise is employed (Pei et al., 2018; Smith and Carstens, 2020; Martin et al., 2021; Derkarabetian et al., 2022). For example, it is common for researchers to evaluate the ML model's performance using genetic datasets of varying sizes, or alignments of different dimensions. The quantity and quality of data clearly influences the effectiveness of ML applications, as analyses conducted on larger, well-filtered datasets consistently yield better delimitation results (Pei et al., 2018; Smith and Carstens, 2020; Martin et al., 2021; Derkarebetian, et al., 2022). This effect is pronounced in UML approaches, as they tend to be more susceptible to data-related issues (Martin et al., 2021).

From a practical perspective, as should be the case in any scientific field, evaluating the suitability of an ML tool for species delimitation also involves assessing its accessibility and reproducibility, particularly compared to traditional SDMs. For example, a thorough description of the ML method, but without a detailed reference to the dataset, can lead to significant issues within the workflow (Chicco, 2017; Greener et al., 2021). The same rationale extends to the availability of the trained models. A good example that circumvents these problems can be found on the study by Derkarabetian et al. (2022), where the authors assessed a ML approach capability to delimit cryptic species constructing a "customized" training dataset from a well-studied lineage with biological characteristics akin to their focal taxon, and clearly explained the rationale behind the customizations made to the datasets and pre-trained models. In cases like these, where a specific ML classifier has been designed and trained with a particular dataset based on a specific evolutionary model's parameters, it is important to ensure both the dataset and the classifier are meticulously described and made accessible to the public. Such precautions minimize the need to construct entirely new workflows for each study, involving tasks such as data simulation, model training, and the selection of evaluation metrics, enabling researchers to evaluate and enhance the method without needing to start from scratch (Greener et al., 2021; Heil et al., 2021).

4.2. Integrating analytical frameworks to investigate complex delimitation models

All models, while inherently limited in representing the underlying nature of species diversification and, hence, of the current species limits among the tested entities, will be more or less useful depending on their effectiveness in extracting relevant evolutionary information from the available data. In some systems, certain methods should be prioritized based on the processes driving divergence, and using multiple methods with similar biases might not enhance biological interpretability. Therefore, the choice of which species delimitation method to use should be done before or during the hypothesis-formulation process, considering the nature of the available data and, possibly, prior relevant biological information regarding the evolution of the organisms.

To effectively prioritize methods, researchers should consider key factors such as the evolutionary context (e.g., presence of gene flow, divergence times, demographic parameters) and the type and quality of the available data (e.g., genomic vs. phenotypic data). For instance, Smith and Carstens (2020) precisely argue that traditional methods like BPP can accurately infer the number of species but may overlook significant processes, such as secondary contact, something that ML workflows like delimitR could address more efficiently. Thus, while some ML-based methods may be particularly well-suited for systems where distinguishing between divergence with gene flow and strict isolation is critical, coalescent-based methods may perform better in detecting fine-scale population structure.

As previously discussed, incorporating multiple delimitation criteria is essential for capturing the diverse mechanisms driving speciation, as different evolutionary processes leave distinct signatures in genetic and non-genetic data. But even considering that inferring species limits from molecular data and integrating phenotypic data can be a solution in some cases, robust species delimitation still requires mechanistic hypotheses about the speciation process itself (Padial and De la Riva, 2021; Pyron et al., 2023; Pyron et al., 2024), as distinguishing between population structure and diverging or collapsing species require explicit hypotheses and quantifiable tests (Sukumaran and Knowles, 2017; Derkarabetian et al., 2019; Huang, 2020; Pyron et al., 2024). In this context, a promising approach that could shape the future of genetic-based species delimitation (and would greatly benefit from integrating different delimitation approaches) is the empirical validation of **speciation-based models**, which offers a more nuanced understanding of the speciation process (see Sukumaran et al., 2021; Hua and Moritz, 2025). For instance, divergence with gene flow can create a pattern of genomic heterogeneity where some loci reflect historical connectivity while others indicate reproductive isolation (see Harrison and Larson, 2016). Such cases may be better captured by models that incorporate allele frequency shifts across loci or explicit tests for introgression (e.g., network-based methods or ML classifiers trained on specific genomic features). Also, the temporal dynamics of divergence, such as recent speciation events with ILS, may be more appropriately addressed by coalescent-based models that account for stochasticity in gene tree variation. Therefore, such a process-based approach might be particular useful in distinguishing intraspecific genetic structure from interspecific divergence.

In the current state of affairs, while ML methods are still being developed and refined, explicitly considering the evolutionary mechanisms underlying species divergence, and strategically integrating this approach with different delimitation tools can enhance both the accuracy and biological realism of species limits. ML-based methods will probably not replace coalescent or tree-based approaches in the near future, but rather complement them by leveraging their particular strengths. Even so, ML is sure to become an integral part of the toolkit used by scientists not only in the field of species delimitation, but for various Evolutionary Biology applications.

Beyond the technical advances of ML workflows and process-based approaches in species delimitation, it is also crucial to ensure that such tools are accessible and properly interpreted by the broader scientific community. Recognizing the methodological complexity and the potential for misapplication, we developed a science communication piece aimed at demystifying the use of machine learning for species delimitation (Appendix C). This material, designed for both specialists and early-career researchers, provides a structured overview of the strengths, limitations, and appropriate contexts for ML-based methods, highlighting how they complement (rather than replace) classical approaches in Evolutionary Biology. By bridging the gap between technical development and scientific dissemination, such efforts can foster more responsible and informed use of these methods, ultimately enhancing the reproducibility and transparency of species delimitation research.

5. Conclusions

- Relatively few studies have yet applied ML techniques to species delimitation using molecular data. Nonetheless, these approaches have already proved to be computationally efficient, and capable of being readily integrated into diverse analytical frameworks, providing a robust way to explore dataset structure when species-level divergences are hypothesized.
- Existing ML-based genetic species delimitation frameworks use various data representation as input (e.g., sequences treated as images, summary statistics, SNPs). Although this flexibility can be advantageous, the reliance on particular representations may bias the accurate delineation of species limits. Assessing the impact of data transformation on delimitation outcomes remains a challenge, and is a key avenue for future research.
- Overly specialized ML algorithms might perform well within the specific ranges of evolutionary parameters present in their training data, but struggle when applied beyond that parameter space. This is particularly critical given the heavy reliance on simulated data in evolutionary biology, where overspecialization can compromise generalizability and transferability—especially in supervised pipelines. Emerging approaches offer promising solutions, including but not limited to the use of transfer learning approaches to exchange knowledge across datasets, GANs to produce more realistic simulated data, and domain adaptation techniques to address the challenges of working with inherently heterogeneous datasets.
- Given the flexibility of ML workflows in handling different types of data—and the multifactorial nature of divergence processes among evolutionary lineages—future research should focus on quantifying how different biological traits contribute to and influence species delimitation results across distinct biological systems.

- A key priority is the development of robust ML-based species delimitation frameworks within the context of Integrative Taxonomy. This will enable the automated integration of multiple lines of evidence to accurately define taxonomic units, while facilitating the reconstruction of the evolutionary processes underlying species limits patterns.
- Although no universally superior species delimitation method currently exists, ML algorithms present promising prospects for their integration into systematic protocols tailored for species delimitation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank André Luiz Gomes de Carvalho, Fernanda de Pinho Werneck and Renato José Pires Machado for their helpful suggestions in earlier versions of the text. We extend our gratitude to Daniel R. Schrider for providing feedback on the first draft of this manuscript. Finally, we wish to thank Emanuel Masiero da Fonseca for his suggestions, which helped improve the final version of the manuscript. This work was supported by the Brazilian federal institution "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior" (CAPES) through a PhD scholarship to MMAS.

REFERENCES

References identified with an asterisk (*) are cited only within the Appendices.

ANGERMUELLER, C., et al. Deep learning for computational biology. **Molecular Systems Biology**, 12, 2016.

ARENAS, M. Simulation of molecular data under diverse evolutionary scenarios. **PloS Computational Biology**, 8, 2012.

ARNAB, S. P., et al. Uncovering footprints of natural selection through spectral analysis of genomic summary statistics. **Molecular Biology and Evolution**, 40, 2023.

ARNOLD, M. L. Natural hybridization as an evolutionary process. **Annual Review of Ecology and Systematics**, 23, 237–261, 1992.

BEAUMONT, M. A., et al. Approximate Bayesian computation in population genetics. **Genetics**, 162, 2025–2035, 2002.

BERTORELLE, G., et al. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. **Molecular Ecology**, 19, 2609–2625, 2010. Doi:10.1111/j.1365-294X.2010.04690.x

BLISCHAK, P. D., et al. Chromosome-scale inference of hybrid speciation and admixture with convolutional neural networks. **Molecular Ecology Resources**, 21, 2676–2688, 2021. https://doi.org/10.1111/1755-0998.13355

BOOKER, W. W., et al. This population does not exist: learning the distribution of evolutionary histories with generative adversarial networks. **Genetics**, 224(2), iyad063, 2023.

BOROWIEC, M. L., et al. Deep learning as a tool for ecology and evolution. **Methods in Ecology and Evolution**, 13, 1640–1660, 2022.

BORTOLUS, A. Error cascades in the biological sciences: the unwanted consequences of using bad taxonomy in ecology. **AMBIO: A Journal of the Human Environment**, 37, 114–118, 2008.

BOUSMALIS, K., et al. Unsupervised pixel-level domain adaptation with generative adversarial networks. **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**, 3722–3731, 2017.

BREITMAN, M. F., et al. A new species of Enyalius (Squamata, Leiosauridae) endemic to the Brazilian Cerrado. **Herpetologica**, 74, 355–369, 2018.

BURBRINK, F. T., & RUANE, S. Contemporary philosophy and methods for studying speciation and delimiting species. **Ichthyology & Herpetology**, 109, 874–894, 2021.

CALLIER, V. Machine learning in evolutionary studies comes of age. **Proceedings of the National Academy of Sciences**, 119, 2022.

CARSTENS, B. C., et al. How to fail at species delimitation. **Molecular Ecology**, 22, 4369–4383, 2013.

CERCA, J., et al. Removing the bad apples: A simple bioinformatic method to improve loci-recovery in de novo RADseq data for non-model organisms. **Methods in Ecology and Evolution**, 12, 805–817, 2021.

CHADHA, A., et al. Deepfake: An overview. **Proceedings of Second International Conference on Computing, Communications, and Cyber-Security**, 557–566, Springer, Singapore, 2021.

CHICCO, D. Ten quick tips for machine learning in computational biology. **BioData Mining**, 10, 1–17, 2017. https://doi.org/10.1186/s13040-017-0155-3

CHRISTIN, S., et al. Applications for deep learning in ecology. **Methods in Ecology and Evolution**, 10, 1632–1644, 2019.

COCHRAN, K., et al. Domain adaptive neural networks improve cross-species prediction of transcription factor binding. **Genome Research**, 32, 512–523, 2022.

COLLIN, F. D., et al. Extending approximate Bayesian computation with supervised machine learning to infer demographic history from genetic polymorphisms using DIYABC Random Forest. **Molecular Ecology Resources**, 21, 2598–2613, 2021. https://doi.org/10.1111/1755-0998.13413.

CSILLÉRY, K., et al. Approximate Bayesian computation (ABC) in practice. **Trends in Ecology & Evolution**, 25, 410–418, 2010.

DAYRAT, B. Towards integrative taxonomy. **Biological Journal of the Linnean Society**, 85, 407–417, 2005.

DEGNAN, J. H., & ROSENBERG, N. A. Gene tree discordance, phylogenetic inference and the multispecies coalescent. **Trends in Ecology & Evolution**, 24, 332–340, 2009.

DERKARABETIAN, S., et al. A demonstration of unsupervised machine learning in species delimitation. **Molecular Phylogenetics and Evolution**, 139, 2019. https://doi.org/10.1016/j.ympev.2019.106562

DERKARABETIAN, S., et al. Using natural history to guide supervised machine learning for cryptic species delimitation with genetic data. **Frontiers in Zoology**, 19, 1–15, 2022.

DIKE, H. U., et al. Unsupervised learning based on artificial neural network: A review. **IEEE International Conference on Cyborg and Bionic Systems (CBS)**, 322–327, 2018.
DOMINGOS, P. A few useful things to know about machine learning. **Communications of the ACM**, 55, 78–87, 2012.

DOMINGOS, F. M., et al. Out of the deep: cryptic speciation in a Neotropical gecko (Squamata, Phyllodactylidae) revealed by species delimitation methods. **Molecular Phylogenetics and Evolution**, 80, 113–124, 2014.

*DUAN, L., et al. Species delimitation of the liquorice tribe (Leguminosae: Glycyrrhizeae) based on phylogenomic and machine learning analyses. **Journal of Systematics and Evolution**, 61, 22–41, 2023. https://doi.org/10.1111/jse.12902.

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

EDWARDS, S. V., et al. Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. **Molecular Phylogenetics and Evolution**, 94, 447–462, 2016.

EDWARDS, S. V., et al. Reticulation, divergence, and the phylogeography–phylogenetics continuum. **Proceedings of the National Academy of Sciences**, 113, 8025–8032, 2016.

ELY, C. V., et al. Implications of poor taxonomy in conservation. **Journal for Nature Conservation**, 36, 10–13, 2017.

*FAN, X. K., WU, et al. Phylogenomic, morphological, and niche differentiation analyses unveil species delimitation and evolutionary history of endangered maples in Acer series Campestria (Sapindaceae). **Journal of Systematics and Evolution**, 61, 284–298, 2023. https://doi.org/10.1111/jse.12919.

FARAHANI, A., et al. A brief review of domain adaptation. **Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020**, 877–894, 2021.

FLAGEL, L., et al. The unreasonable effectiveness of convolutional neural networks in population genetic inference. **Molecular Biology and Evolution**, 36, 220–238, 2019.

FLOURI, T., et al. Species Tree Inference with BPP using Genomic Sequences and the Multispecies Coalescent. **Molecular Biology and Evolution**, 35, 2585–2593, 2018. Doi:10.1093/molbev/msy147.

FLOURI, T., et al. A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. **Molecular Biology and Evolution**, 37, 1211–1223, 2020.

FONSECA, R. R., et al. Next-generation biology: sequencing and data analysis approaches for non-model organisms. **Marine Genomics**, 30, 3–13, 2016.

FONSECA, E. M., et al. Phylogeographic model selection using convolutional neural networks. **Molecular Ecology Resources**, 21, 2661–2675, 2021. https://doi.org/10.1111/1755-0998.13427.

FONSECA, E. M., & CARSTENS, B. C. Artificial intelligence enables unified analysis of historical and landscape influences on genetic diversity. **Molecular Phylogenetics and Evolution**, 108116, 2024.

FOUNTAIN-JONES, N. M., et al. Machine learning in molecular ecology. **Molecular Ecology Resources**, 21, 2589–2597, 2021. https://doi.org/10.1111/1755-0998.13532.

FUJITA, M. K., et al. Coalescent-based species delimitation in an integrative taxonomy. **Trends in Ecology & Evolution**, 27, 480–488, 2012.

GANIN, Y., & LEMPITSKY, V. Unsupervised domain adaptation by backpropagation. **International Conference on Machine Learning**, 1180–1189, 2015.

GEHARA, M., et al. PipeMaster: inferring population divergence and demographic history with approximate Bayesian computation and supervised machine-learning in R. **bioRxiv**, 2020-12, 2020. https://doi.org/10.1101/2020.12.04.410670

GHIFARY, M., et al. Deep Reconstruction Classification Networks for Unsupervised Domain Adaptation. In: LEIBE, B., MATAS, J., SEBE, N., WELLING, M., eds. **Computer Vision ECCV 2016. Lecture Notes in Computer Science.** Cham: Springer International Publishing, p. 597, 2016.

GHIROTTO, S., et al. Distinguishing among complex evolutionary models using unphased whole-genome data through random forest approximate Bayesian computation. **Molecular Ecology Resources**, 21, 2614–2628, 2021. https://doi.org/10.1111/1755-0998.13263.

GOMPERT, Z., et al. Analysis of population genomic data from hybrid zones. **Annual Review of Ecology, Evolution, and Systematics**, 48, 207–229, 2017.

GOODFELLOW, I., et al. Generative adversarial nets. **Advances in Neural Information Processing Systems**, 2672–2680, 2014.

GREENER, J. G., et al. A guide to machine learning for biologists. **Molecular Cell Biology**, 23, 40–55, 2021. https://doi.org/10.1038/s41580-021-00407-0.

GUYON, I., & ELISSEEFF, A. An introduction to variable and feature selection. **Journal of Machine Learning Research**, 3, 1157–1182, 2003.

HARRISON, R. G., & LARSON, E. L. Heterogeneous genome divergence, differential introgression, and the origin and structure of hybrid zones. **Molecular Ecology**, 25, 2454–2466, 2016.

HASTIE, T., et al. Unsupervised learning. In: **The Elements of Statistical Learning**, 485–585, Springer, New York, NY, 2009.

HAUSDORF, B. Progress toward a general species concept. **Evolution**, 65, 923–931, 2011.
HEIL, B. J., et al. Reproducibility standards for machine learning in the life sciences. **Nature Methods**, 18, 1132–1135, 2021.

*HODEL, R. G., et al. A phylogenomic approach, combined with morphological characters gleaned via machine learning, uncovers the hybrid origin and biogeographic diversification of the plum genus. **bioRxiv**, 2023-09, 2023. https://doi.org/10.1101/2023.09.13.557598.

HÜLLERMEIER, E., et al. Inductive bias. **Encyclopedia of Systems Biology**, 1018–1019, 2013.

HUA, X., & MORITZ, C. A phylogenetic approach to delimitate species in a probabilistic way. **Systematic Biology**, syaf004, 2025.

HUANG, J. P. Is population subdivision different from speciation? From phylogeography to species delimitation. **Ecology and Evolution**, 10, 6890–6896, 2020.

HUELSENBECK, J. P., et al. Structurama: Bayesian inference of population structure. **Evolutionary Bioinformatics**, 7, 55–59, 2011.

JACKSON, N. D., et al. Species delimitation with gene flow. **Systematic Biology**, 66, 799–812, 2017.

JACOBS, S. J., et al. Incongruence in molecular species delimitation schemes: What to do when adding more data is difficult. **Molecular Ecology**, 27, 2397–2413, 2018.

*JAMDADE, R., et al. Multilocus marker-based delimitation of Salicornia persica and its population discrimination assisted by supervised machine learning approach. **PloS ONE**, 17, 2022. https://doi.org/10.1371/journal.pone.0270463.

JARVIS, E. D., et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. **Science**, 346, 1320–1331, 2014.

JÖRGER, K. M., & SCHRÖDL, M. How to describe a cryptic species? Practical challenges of molecular taxonomy. **Frontiers in Zoology**, 10, 1–27, 2013.

JORNA, J., et al. Species boundaries in the messy middle—A genome-scale validation of species delimitation in a recently diverged lineage of coastal fog desert lichen fungi. **Ecology and Evolution**, 11, 18615–18632, 2021.

KARBSTEIN, K., et al. Species delimitation 4.0: integrative taxonomy meets artificial intelligence. **Trends in Ecology & Evolution**, 2023.

*KHALIGHIFAR, A., et al. Deep learning improves acoustic biodiversity monitoring and new candidate forest frog species identification (genus Platymantis) in the Philippines. **Biodiversity and Conservation**, 30, 643–657, 2021.

KNOWLES, L. L., & CARSTENS, B. C. Delimiting species without monophyletic gene trees. **Systematic Biology**, 56, 887–895, 2007.

KORFMANN, K., et al. Deep learning in population genetics. **Genome Biology and Evolution**, 2023. https://doi.org/10.1093/gbe/evad008.

KUZENKOV, O., et al. Exploring evolutionary fitness in biological systems using machine learning methods. **Entropy**, 23, 1–35, 2020.

LEACHÉ, A. D., et al. The influence of gene flow on species tree estimation: a simulation study. **Systematic Biology**, 63, 17–30, 2014.

LEACHÉ, A. D., et al. The spectre of too many species. **Systematic Biology**, 68, 168–181, 2019.

LECUN, Y., et al. Deep learning. **Nature**, 521, 436–444, 2015.LIBBRECHT, M.W. & NOBLE, W.S. 2015. Machine learning applications in genetics and genomics. **Nature Reviews Genetics** 16, 32–332.

*LIMA, A.P. et al. Not as widespread as thought: Integrative taxonomy reveals cryptic diversity in the Amazonian nurse frog Allobates tinae Melo-Sampaio, Oliveira and Prates, 2018 and description of a new species. **Journal of Zoological Systematics and Evolutionary Research**, 58(4), 1173–1194. 2020a.

*LIMA, L. R. et al. Below the waterline: cryptic diversity of aquatic pipid frogs (Pipa carvalhoi) unveiled through an integrative taxonomy approach. **Systematics and Biodiversity**, 18(8), 771–783. 2020b.

LIU, B. BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. **Briefings in bioinformatics** 20, 1280–1294. 2019.

LIU, M.Y. & TUZEL, O. Coupled Generative Adversarial Networks. In: **Advances in Neural Information Processing Systems 29**. Curran Associates, Inc. https://papers.nips.cc/paper/2016/hash/502e4a16930e414107ee22b6198c578f-Abstract.html. 2016.

LUKHTANOV, V. A. Species Delimitation and Analysis of Cryptic Species Diversity in the XXI Century. **Entmol. Rev.** 99, 463–472. https://doi.org/10.1134/S0013873819040055. 2019.

LÜRIG, M.D., et al. Computer vision, machine learning, and the promise of phenomics in ecology and evolutionary biology. **Frontiers in Ecology and Evolution** 9. 2021.

*MAGALHÃES, F. D. M., et al. Taxonomic review of South American Butter Frogs: Phylogeny, geographic patterns, and species delimitation in the Leptodactylus latrans species group (Anura: Leptodactylidae). **Herpetological Monographs**, 34(1), 131–177. 2020.

MALLET, J., et al. How reticulated are species? **BioEssays**, 38, 140–149. 2016.

MANGUL, S. et al. Systematic benchmarking of omics computational tools. **Nature communications** 10. 2019.

MARTIN, B. T., et al. The choices we make and the impacts they have: Machine learning and species delimitation in North American box turtles (Terrapene spp.). **Molecular Ecology Resources** 21, 2801–2817. 2021.

MAYR, E. M. The biological meaning of species. **Biological Journal of the Linnean society**, 1, 311–320. 1969.

MAYR, E. M. What is a species, and what is not? **Philosophy of science**, 63, 262–277. 1996.

MAYR, E. M. The biological species concept. **Species concepts and phylogenetic theory: a debate**, 17–29. 2000.

MCCLURE, E. C., et al. Artificial intelligence meets citizen science to supercharge ecological monitoring. **Patterns** 1. 2020.

MITCHELL, T. M. **Machine Learning**. McGraw-Hill, New York. 1997.

MIRARAB, S., et al. Multispecies coalescent: theory and applications in phylogenetics. **Annual Review of Ecology, Evolution, and Systematics**, 52, 247-268. 2021.

MO, Z., & SIEPEL, A. Domain-adaptive neural networks improve supervised machine learning based on simulated population genetic data. **PLOS Genetics**, 19. 2023.

MO, Y. K., et al. Applications of machine learning in phylogenetics. **Molecular Phylogenetics and Evolution**, 196, 108066. 2024.

MORIMOTO, J., et al. Editorial: Applications of Machine Learning to Evolutionary Ecology Data. **Frontiers in Ecology and Evolution**. 2021.

NESTERENKO, L., et al. Phyloformer: towards fast and accurate phylogeny estimation with self-attention networks. **bioRxiv**, 2022-06. https://doi.org/10.1101/2022.06.24.496975. 2022.

*NEWTON, L. G., et al. Integrative species delimitation reveals cryptic diversity in the southern Appalachian Antrodiaetus unicolor (Araneae: Antrodiaetidae) species complex. **Molecular Ecology** 29, 2269–2287. 2020.

O'MEARA B. C. New heuristic methods for joint species delimitation and species tree inference. **Systematic Biology** 59, 59–73. 2010.

O'MEARA B. C. Evolutionary inferences from phylogenies: a review of methods. **Annual Review of Ecology, Evolution, and Systematics** 43, 267–285. 2012.

PADIAL, J. M., et al. The integrative future of taxonomy. **Frontiers in zoology**, 7, 1–14. 2010.

PAN, S. J. & YANG, Q. A survey on transfer learning. **IEEE Transactions on Knowledge and Data Engineering** 22, 1345–1359. 2010.

PANTE, E., et al. Species are hypotheses: avoid connectivity assessments based on pillars of sand. **Molecular Ecology** 24, 525–544. 2015.

PEI, J., et al. CLADES: A classification-based machine learning method for species delimitation from population genetic data. **Molecular Ecology Resources** 18, 1144–1156. https://doi.org/10.1111/1755-0998.12887. 2018.

PEREZ, M. F., et al. Coalescent-based species delimitation meets deep learning: Insights from a highly fragmented cactus system. **Molecular Ecology Resources**. 2021.

PICHLER, M., et al. Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. **Methods in Ecology and Evolution** 11, 281–293. 2020.

PIGLIUCCI, M. Species as family resemblance concepts: the (dis-)solution of the species problem? **BioEssays**, 25, 596–602. 2003.

PONS, J., et al. Sequence-based species delimitation for the DNA taxonomy of unde-scribed insects. **Systematic Biology** 55, 595–609. 2006.

PRICE, T. D., et al. The role of phenotypic plasticity in driving genetic evolution. **Proceedings of the Royal Society of London. Series B: Biological Sciences** 270, 1433–1440. 2003.

*PRITCHARD, J. K., et al. Inference of population structure using multilocus genotype data. **Genetics** 155, 945–959. 2000.

PUDLO, P., et al. Reliable ABC model choice via random forests. **Bioinformatics** 32, 859–866. https://doi.org/10.1093/bioinformatics/btv684. 2016.

PYRON, R. A. Unsupervised Machine Learning for Species Delimitation, Integrative Taxonomy, and Biodiversity Conservation. **Molecular Phylogenetics and Evolution**, 189. 2023.

PYRON, R. A., et al. Speciation hypotheses from phylogeographic delimitation yield an integrative taxonomy for Seal Salamanders (Desmognathus monticola). **Systematic Biology**, 72, 179–197. 2023.

PYRON, R. A., et al. Discerning structure versus speciation in phylogeographic analysis of Seepage Salamanders (Desmognathus aeneus) using demography, environment, geography, and phenotype. **Molecular Ecology**, 33, e17219. 2024.

DE QUEIROZ, K. The general lineage concept of species, species criteria, and the process of speciation. **Endless forms: species and speciation**. 1998.

DE QUEIROZ, K. The General Lineage Concept of Species and the Defining Properties of the Species Category. In book: **Species: New Interdisciplinary Essays**, Chapter: 3, Publisher: MIT Press, Editors: Robert A. Wilson. 1999.

DE QUEIROZ, K. Ernst Mayr and the modern concept of species. **Proceedings of the National Academy of Sciences**, 102, 6600–6607. 2005a.

DE QUEIROZ, K. Different species problems and their resolution. **BioEssays** 27, 1263–1269. 2005b.

DE QUEIROZ, K. Species concepts and species delimitation. **Syst. Biol.** 56, 879–886. 2007.

DE QUEIROZ, K. Branches in the lines of descent: Charles Darwin and the evolution of the species concept. **Biol. J. Linn. Soc.** 103, 19–35. 2011.

DE QUEIROZ, K. An updated concept of subspecies resolves a dispute about the taxonomy of incompletely separated lineages. **Herpetological Review**. 2020.

QU, K., et al. Application of machine learning in microbiology. **Frontiers in Microbiology** 10. 2019.

RANNALA, B. The art and science of species delimitation. **Current Zoology** 61, 846–853. 2015.

RANNALA, B., & YANG, Z. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. **Genetics** 164, 1645–1656. 2003.

RANNALA, B., & YANG, Z. Bayesian species delimitation using multilocus sequence data. **Proceedings of the National Academy of Sciences** 107, 9264–9269. 2010.

RANNALA, B., & YANG, Z. Species Delimitation. In: **Phylogenetics in the genomic era**. 2020.

RANNALA, B., et al. The Multi-species Coalescent Model and Species Tree Inference. SCORNAVACCA, CELINE; DELSUC, FRÉDÉRIC; GALTIER, NICOLAS. **Phylogenetics in the Genomic Era**, No commercial publisher | Authors open access book. 2020.

RAYNAL, L., et al. ABC random forests for Bayesian parameter inference. **Bioinformatics** 35, 1720–1728. 2019.

ROZANTSEV, A., SALZMANN, M. & FUA, P. Beyond sharing weights for deep domain adaptation. **IEEE transactions on pattern analysis and machine intelligence** 41, 801–814. 2018.

*SARYAN, P., et al. Species complex delimitations in the genus Hedychium: A machine learning approach for cluster discovery. **Applications in Plant Sciences** 8. https://doi.org/10.1002/aps3.11377. 2020.

SANCHEZ, T., et al. Deep learning for population size history inference: Design, comparison and combination with approximate Bayesian computation. **Molecular Ecology Resources** 21, 2645–2660. 2020.

SCALON, M.C., et al. Diversity of functional trade-offs enhances survival after fire in Neotropical savanna species. **Journal of Vegetation Science**, 31, 139-150. 2020.

SCHRIDER, D. R. & KERN, A. D. Supervised Machine Learning for Population Genetics: A New Paradigm. **Trends in Genetics** 34, 301–312. https://doi.org/10.1016/j.tig.2017.12.005. 2018.

SEARLS, D. B. The Roots of Bioinformatics. **PloS Comput Biol** 6. https://doi.org/10.1371/journal.pcbi.1000809. 2010.

SHEEHAN, S., & SONG, Y.S. Deep learning for population genetic inference. **PloS computational biology** 12. 2016.

SHURTLIFF, Q. R. Mammalian hybrid zones: a review. **Mammal Review**, 43, 1–21. 2013.

SIDEY-GIBBONS, J. A., & SIDEY-GIBBONS, C. J. Machine learning in medicine: a practical introduction. **BMC medical research methodology** 19, 1–18. 2019.

SILVA, D. C., et al. Cerrado bat community assembly is determined by both present-day and historical factors. **Journal of Biogeography**. 2024.

SIMONSEN, K. L., et al. Properties of statistical tests of neutrality for DNA polymorphism data. **Genetics** 1411, 413–429. 1995.

SITES, JR J. W. & MARSHALL, J. C. Operational criteria for delimiting species. **Annual Review of Ecology, Evolution, and Systematics**, 199-227. 2004.

SLATKO, B. E., et al. Overview of next-generation sequencing technologies. **Current protocols in molecular biology** 122. 2018.

SMITH, M. L., et al. Demographic Model Selection using Random Forests and the Site Frequency Spectrum. **Molecular Ecology**. 2017.

SMITH, M. L. & CARSTENS B. C. Process-based species delimitation leads to identification of more biologically relevant species. **Evolution** 74, 216–229. https://doi.org/10.1111/evo.13878. 2020.

SMITH, M. L., & HAHN, M. W. Phylogenetic inference using generative adversarial networks. **Bioinformatics**, 39. 2023.

SOLIS-LEMUS, C., et al. Accurate phylogenetic inference with a symmetry-preserving neural network model. **arXiv preprint arXiv:2201.04663**. 2022.

SUKUMARAN, J. & KNOWLES, L.L. Multispecies coalescent delimits structure, not species. **Proceedings of the National Academy of Sciences** 114, 1607–1612. 2017.

SUKUMARAN, J., et al. Incorporating the speciation process into species delimitation. **PloS Computational Biology** 17. 2021.

SUVOROV, A., et al. Accurate inference of tree topologies from multiple sequence alignments using deep learning. **Systematic biology** 69, 221–233. 2020.

TAGU, D., et al. Genomic data integration for ecological and evolutionary traits in non-model organisms. **BMC genomics** 15, 1–16. 2014.

TAUTZ, D., et al. A plea for DNA taxonomy. **Trends Ecol. Evol.** 18, 70–74. 2003.

VALLETTA, J. J., et al. Applications of machine learning in animal 58otaling studies. **Animal Behaviour** 124, 203–220. 2017.

VINK, C. J., et al. Taxonomy and irreproducible biological science. **BioScience** 62, 451–452. 2012.

VOGLER, A. P., MONAGHAN, M. T. Recent advances in DNA taxonomy. **J. Zool. Syst. Evol. Res.** 45, 1–10. 2007.

WAKE, D. B., et al. Homoplasy: from detecting patterns to determining process and mechanism of evolution. **Science** 331, 1032–1035. 2011.

WÄLDCHEN, J. & MÄDER, P. Machine learning for image-based species identification. **Methods in Ecology and Evolution** 9, 2216–2225. 2018.

WANG, G. Machine learning for inferring animal behavior from location and movement data. **Ecological informatics** 49, 69–76. 2019.

WANG, Z., et al. Automatic inference of demographic parameters using generative adversarial networks. **Molecular ecology resources** 21, 2689–2705. 2021.

WIENS, J. J., & PENKROT, T. A. Delimiting species using DNA and morphological variation and discordant species limits in spiny lizards (Sceloporus). **Syst. Biol.**, 51, 69–91. 2002.

WIENS, J. J. Species delimitation: new approaches for discovering diversity. **Syst. Biol.** 56, 875–8. 2007.

WILKINS, J. S., ZACHOS, F. E., & PAVLINOV, I. Y. (Eds.). **Species Problems and Beyond: Contemporary Issues in Philosophy and Practice**. CRC Press. 2022.

YANG, B., et al. Identification of species by combining molecular and morphological data using convolutional neural networks. **Systematic Biology**, 71, 690–705. 2022.

YELMEN, B. & JAY, F. An Overview of Deep Generative Models in Functional and Evolutionary Genomics. **Annual Reviews of Biomedical Data Science**. https://doi.org/10.1146/annurev-biodatasci-020722. 2023.

ZACHOS, F. E. **Species concepts in biology** (Vol. 801). Cham: Springer. 2016.

ZACHOS, F. E. (New) Species concepts, species delimitation and the inherent limitations of taxonomy. **Journal of genetics**, 97, 811–815. 2018.

ZAHARIAS, P., et al. Re-evaluating Deep Neural Networks for Phylogeny Estimation: The Issue of Taxon Sampling. **Journal of Computational Biology** 29, 74–89. https://doi.org/10.1089/cmb.2021.0383. 2022.

Appendices

(A) Glossary

Architecture: the configuration of neurons, layers, and connections among them in a neural network.

Artificial neural network/ANN: a network of layers of one or more ‘neurons’ which receive inputs from each neuron in the previous layer, and perform a linear combination on these inputs, which is then passed through an activation function. The first layer is the input layer (i.e., the feature vector) and the last layer is the output layer yielding the predicted responses. Intervening layers are referred to as ‘hidden’ layers.

Binning: generally, this is the process of grouping reads, contigs or similar genetic markers and assigning them to individual genomes.

Class: a category of data in a discrete category classification problem.

Classical Multidimensional Scaling (cMDS): method generally used to visualize the similarity or dissimilarity of data in a lower-dimensional space. It works by taking a distance or dissimilarity matrix and finding a configuration of points in a Euclidean space that best preserves those distances; the primary aim of cMDS is to project high-dimensional data into two or three dimensions for easier interpretation, while retaining as much of the original variance as possible.

Classification: an ML task where the value to be predicted for each example is a categorical label. This kind of algorithm classifies input variables to discrete categories, or target variables. Contrast with regression models.

Convolutional Neural Network/CNN: a particular type of ANN commonly used in visual recognition, in which connections between different layers allow performing convolutions.

Curse of dimensionality: briefly, this refers to challenges that arise when analyzing high-dimensional data, where the number of features approaches or exceeds the number of observations. As dimensionality increases, data becomes increasingly sparse in the feature space, leading to computational inefficiencies, overfitting of models, and reduced statistical power. Key issues include inflated distances between points (making similarity metrics less meaningful), exponential growth in the volume of the space (requiring disproportionately larger sample sizes for reliable inference), and difficulty in visualizing or interpreting patterns.

Decision tree: a hierarchical structure that predicts the response variable of an example by examining a feature, and branching to the right subtree if the value of that feature is greater than some threshold, and branching to the left otherwise. At the next level of the tree another feature is examined. The predicted value is determined by which leaf of the tree is reached at the end of this process.

Deep learning: type of learning using ANNs or similarly networked algorithmic models that contain multiple 'hidden' layers between the input and output layers.

Discriminant Analysis of Principal Components (DAPC): It combines two steps: (1) principal component analysis (PCA) to reduce the dimensionality of the dataset while retaining most of the genetic variation, and (2) linear discriminant analysis (LDA) to maximize the separation between predefined groups or clusters. Unlike traditional methods, DAPC does not make assumptions about Hardy-Weinberg equilibrium or linkage disequilibrium, making it well-suited for analyzing complex genetic data. It is often applied to identify population structure, infer species boundaries, and investigate patterns of genetic variation.

Example: a dataset is a collection of numbers or values that relate to a particular subject. Each one of these numbers or values can be referred to as an *example*.

Feature vector: a multidimensional representation of a datapoint made up of measurements (or features) taken from it (e.g., a vector of population genetic summary statistics measured in a genomic region).

General Lineage Concept (GLC): a major framework in Evolutionary Biology that defines a species as a population or group of populations that represents a distinct, evolving lineage. Under this concept, species are considered separately evolving metapopulation lineages, where metapopulations consist of connected populations of individuals that interbreed or interact across time and space. The GLC emphasizes that the central unifying feature of all species is their evolutionary independence, derived from processes like restricted gene flow and divergence. Unlike other species concepts (e.g., biological, morphological, or phylogenetic), which often focus on specific criteria like reproductive isolation, diagnosability, or shared derived traits, the GLC posits that these criteria are secondary properties (or lines of evidence) that emerge as a result of lineage divergence. These secondary properties can be used as operational criteria to recognize species, but they are not what fundamentally defines a species. In essence, the GLC unifies different species concepts under the idea that species are different evolutionary lineages, and previous concepts should be viewed as tools or methods for identifying and describing those lineages.

Incomplete lineage sorting: persistence of ancestral genetic polymorphisms, especially during rapid speciation events, inducing incongruences between gene trees and species trees.

Inductive bias: in the context of machine learning, this term refers to a set of assumptions (implicit or explicit) made by a learning algorithm in order to perform induction, that is, to generalize a finite set of observations (generally referred as training data) into a general model of a particular domain. Treating all these possibilities equally, or without any bias in the sense of a preference for specific types of generalization, predictions for new situations could not be made.

Integrative Taxonomy: branch of systematics and taxonomy focused on classifying and naming organisms by employing multiple data sources (taxonomical evidence; such as genetic and morphological) to achieve a more comprehensive and holistic understanding of the group's diversity.

Isotonic Multidimensional Scaling (isoMDS): a non-metric multidimensional scaling (NMDS) method designed to represent pairwise dissimilarities in a low-dimensional Euclidean space. Unlike cMDS, which preserves actual distances, isoMDS focuses on preserving the rank order of dissimilarities (i.e., ensuring that objects with higher dissimilarities remain further apart in the low-dimensional space). This method is particularly useful when the dissimilarity data are non-Euclidean, noisy, or derived from complex models.

Layers: a logical collection of nodes/neurons layer in a deep learning model, taking information from the previous layers and then passes it to the next layer; usually, there are at least three types of layers in every ANN: input, hidden and output.

Labeled data: data examples for which the true response value (or label) is known.

Likelihood-based approaches: when it comes to species delimitation, such approaches refer to methods that use statistical likelihood functions to assess the fit between observed data and various delimitation models. These models typically describe how genetic or other data are generated under different scenarios of species divergence and gene flow. By comparing the likelihood of the data under different species delimitation models, researchers can infer the most probable number and limits of species within a study system.

Machine learning/ML: subset of artificial intelligence that allows a system to learn and improve autonomously using mathematical algorithms, without being explicitly programmed, by feeding it large amounts of data.

*Migration parameter/*M: generally referred to as the parameter describing migration from one population to another; can also be referred to as m_{ij}, where *I* and *j* represent two different populations.

Model: in the context of deep learning, the word "model" is often used to refer to a trained neural network. This can be confusing in studies that use neural network classifiers to distinguish among data generated under different evolutionary models or models of demographic evolution.

One-hot encoding: a method for representing categorical variables as binary vectors, where each category is converted into a binary (0 or 1) feature. For a variable with *n* categories, one-hot encoding generates *n* binary columns, each corresponding to a unique category, with a value of 1 indicating the presence of that category and 0 otherwise. This technique ensures categorical data (e.g., nucleotide bases, amino acids, or population labels) can be processed by algorithms requiring numerical input while avoiding implicit ordinal relationships between categories. In molecular phylogenetics and evolution, one-hot encoding is commonly applied to DNA or protein sequence data (e.g., encoding nucleotides A, T, C, G as [1,0,0,0], [0,1,0,0], etc.), enabling compatibility with models for sequence classification, phylogenetic inference, or trait prediction. It is also used to represent discrete morphological traits or population identifiers in multivariate analyses.

*Population size/*θ: the population size parameter reflects the mutation–drift balance occurring within a population with an effective size of 'Ne' individuals, and a mutation rate of 'μ' per site per generation. The most common equation is: $\theta = 4Ne\mu$ (with $2Ne\mu$ for haploid organisms).

Principal Component Analysis (PCA): a multivariate statistical technique used to reduce the dimensionality of a dataset while preserving as much variability as possible. It does this by transforming the original variables into a new set of uncorrelated variables, called principal components, which are ordered by the amount of variance they explain. The first principal component captures the most variation in the data, the second captures the next highest variation (orthogonal to the first), and so on. PCA is commonly used for data visualization, noise reduction, and identifying patterns in high-dimensional datasets.

Random forest: an ensemble of semi-randomly generated decision trees. An example is run through each tree in the forest, and these trees then vote to determine the predicted value. Random forests can perform both classification and regression.

Regression: an ML task where the value to be predicted for each example is a continuous number. Example: a neural network estimating demographic parameter values of a population from genetic data. Contrasts with classification.

Self-organizing maps/SOMs: artificial neural networks trained using competitive learning (rather than error correction via back-propagation with gradient descent), assessed by relative distance to the closest unit across clusters. They can be used to produce two-dimensional representations of complex datasets preserving the higher-dimensional topological structure of the input data.

Shallow learning: contrasts with deep learning, where we have several nonlinear data processing layers; in shallow learning, we have only two data processing layers, with the second one generally being linear. Consequently, shallow learning models are comparatively less complex than deep learning architectures, often consisting of only a few layers. Shallow learning methods use simple algorithms and perform relatively simple tasks, usually working with lower-dimensional data and solving straightforward problems. While these models might be less powerful for capturing intricate patterns, they are faster to train and require less data. They are particularly effective for tasks with straightforward patterns or where complex feature hierarchies are not necessary.

Single nucleotide polymorphisms (SNPs): single-base-pair differences in DNA sequences among individuals within a population.

Site frequency spectrum/SFS: the distribution of allele frequencies in a population sample.

Speciation-based models: theoretical or computational frameworks in evolutionary biology that explicitly incorporate the processes and mechanisms driving the formation of new species. These models account for factors such as genetic divergence, reproductive isolation, natural selection, and ecological differentiation to delineate species boundaries. By directly modeling speciation events, they provide insights into how evolutionary processes shape biodiversity and help predict or infer the emergence and maintenance of distinct species from genetic and phenotypic data.

Splitter: a classification approach that prioritizes dividing organisms into multiple distinct lineages rather than grouping them into fewer, broader species. This perspective often emphasizes subtle morphological, genetic, or ecological differences to justify species recognition, leading to a more fine-grained classification of biodiversity. A splitter approach contrast with those which tend to group similar populations under fewer species, emphasizing continuity and intraspecific variation over differentiation.

Support vector machine/SVM: an ML approach that seeks to find the hyperplane that optimally separates two classes of training data. These data are often mapped to high-dimensional space using a kernel function. Variations of this approach can be performed to accomplish multiclass classification or regression.

t-Distributed Stochastic Neighbor Embedding (t-SNE): a nonlinear dimensionality reduction technique designed to visualize high-dimensional data in a low-dimensional space (typically 2D or 3D) while preserving the local structure of the data. It employs a probabilistic approach to model pairwise similarities between data points in the original high-dimensional space and their counterparts in the reduced space. T-SNE emphasizes the retention of local relationships by converting distances between points into conditional probabilities, which are then optimized using gradient descent. A key feature of t-SNE is its use of the Student's t-distribution in the low-dimensional space. Note that t-SNE is primarily a visualization tool and does not preserve global geometric relationships or serve as a feature extraction method.

Test set: a set of labelled examples for use during testing that is independent of the training set.

Training: the process of generating from a training set a function that seeks to correctly predict the response variable of the datum by examining its feature vector. This process is generally used to measure error and improve performance, and complete training usually involves multiple iterations over the entire training set, or epochs. A t*raining set* is a set of labelled examples for use during training.

Variational autoencoders (VAE): a model that combines neural networks with variational inference to learn a compressed, probabilistic representation (latent space) of high-dimensional data. Unlike traditional autoencoders, VAEs encode inputs into a distribution over the latent space rather than discrete points, enabling the generation of new data samples by sampling from this distribution. The model consists of an encoder, which maps input data to parameters of a probability distribution (e.g., mean and variance), and a decoder, which reconstructs data from latent variables.

(B) Additional tests regarding ML and species delimitation problems

Martin et al. (2021) employed the same methods featured in Derkarabetian et al. (2019), supplemented by an assessment of algorithm performance under various data filtering strategies. In general, they observed that data filtering does influence the signal-to-noise ratio and the level of disagreement among resolved clusters. Nonetheless, both studies concur that ML represents a viable alternative for species delimitation. Newton et al. (2020) pursued a similar approach to Derkarabetian et al. (2019) by utilizing UML to explore cryptic diversity patterns within the *Antrodiaetus unicolor* species complex. They specifically employed the VAE algorithm, which operates as a neural network. Being a UML framework, this approach relies solely on the inherent data structure for sample grouping, reducing the necessity for a priori assumptions about the biological system. It’s worth noting that the clustering results derived from VAE differed from other analyses performed by the authors, such as STRUCTURE (Pritchard et al., 2000). Furthermore, they observed that the amount of missing data had an impact on the result quality obtained through VAE, likely due to a reduction in informative sites. Derkarabetian et al. (2022) employed the CLADES approach, tailoring their training data to be more biologically relevant to their specific study. They focused on genetic data from harvestmen of the *Theromaster* group, using two distinct sets of training data. First, they used the “All” training dataset from Pei et al. (2018), which was originally generated through simulations with varying values of population size, migration rate, and divergence time in a two-species model (as explained earlier). This dataset is broadly applicable to various taxa, as it covers a wide range of genetic metrics encompassing both plants and animals. Subsequently, they developed a “customized” training dataset, derived from a well-studied lineage with biological characteristics similar to those of *Theromaster*, namely *Metanonychus*. In analytical terms, the process of forming this “customized” training dataset involved sending UCE loci from *Metanonychus* to CLADES using the “general” training dataset. As the output files generated by this procedure contained pairwise comparisons between specified populations, such files were manually adjusted to represent the samples from the same species within populations as defined in Derkarabetian et al. (2019). Only after creating the “customized” training set the *T. brunneus* data was imputed to CLADES. The outcome indicated that analyses based on the training dataset provided by Pei et al. (2018) classified all populations as distinct species. In contrast, analyses relying on the “customized” training dataset developed by Derkarabetian et al. (2022) supported a different scenario, suggesting the presence of at least two cryptic species. Either way, the authors suggest that this approach allows for more informed, context-specific decisions regarding species limits, particularly when using genetic data from an evolutionary lineage with similar biological characteristics to the focal organism. They also propose that this approach can be considered a logical and analytical extension of the species delimitation process, applicable to a wide range of biological systems (Derkarabetian et al., 2022). Another integrative approach is the multi-layer Kohonen Self-Organizing Maps (“SuperSOMs”) proposed by Pyron (2023). In practical terms, this new R package expands the approach proposed in Pyron et al. (2023), incorporating the possibility of delimiting species based on allelic, spatial, climatic, and phenotypic data.

Table A.1. List of proposed ML applications specifically designed to work on inferences about species limits and supplementary studies further exploring their performance.

Reference	Additional tests	Languages	Category	Algorithms	Simulator	Input	Data representation
CLADES: A Classification-based Machine Learning Method for Species Delimitation from Population Genetic Data (Pei et al., 2018)[1]	Using natural history to guide supervised machine learning for cryptic species delimitation with genetic data (Derkarabetian et al., 2022).	Python	SML	Support vector machines	Mccoal	Multiple sequence alignment (MSA) or SNP matrix	Population genetics summary statistics
A demonstration of unsupervised machine learning in species delimitation (Derkarabetian *et al.*, 2019)[2]	The choices we make and the impacts they have: Machine learning and species delimitation in North American box turtles (*Terrapene* spp.) (Martin et al., 2021)[6] Integrative species delimitation reveals cryptic diversity in the southern Appalachian *Antrodiaetus unicolor* (Araneae: Antrodiaetidae) species complex (Newton et al., 2020)[7] Using natural history to guide supervised machine learning for cryptic species delimitation with genetic data (Derkarabetian et al. 2022).	R/python	SML & UML	t-Distributed Stochastic Neighbor Embedding, Random Forest, Variational autoencoders	NA	SNP data matrix	One-hot-encoding of the SNP data matrix (VAE), *axis* from a discriminant analysis of principal components (t-SNE), scaled data from DAPC + cMDS and isoMDS ouput (Random forest)
Process-based species delimitation leads to identification of more biologically relevant species (Smith & Carstens, 2020)[3]	-	python	SML	Random forest	fastsimcoal	SNP data matrix	Folded multi-dimensional SFS
Coalescent-based species delimitation meets deep learning: Insights from a highly fragmented cactus system (Perez et al., 2021)[4]	-	python	Deep learning	Convolutional neural networks	ms	SNP data matrix	NumPy matrices (as images), with genotypes encoded as higher or lower frequency states
Speciation Hypotheses from Phylogeographic Delimitation	Unsupervised machine learning for species delimitation,	R	UML	Self-organizing maps (SOMs)	NA	SNP data matrix	SNP matrix, in which the rows

Yield an Integrative Taxonomy for Seal Salamanders (*Desmognathus monticola*) (Pyron et al., 2023)[5]	integrative taxonomy, and biodiversity conservation (Pyron, 2023)[8]	are individual specimens, the columns are the 2–4 possible states at each SNP locus, and the entries are the frequency of that state

Online repositories where it is possible to find more information about the currently existing platforms. [1]https://github.com/pjweggy/CLADES; [2]https://github.com/shahanderkarabetian/uml_species_delim; [3]https://github.com/meganlsmith/delimitR; [4]https://github.com/manolofperez/CNN_spDelimitation_Piloso; [5]https://github.com/kyleaoconnell22/Pyron_et_al_UML_sp_delim/tree/main; [6]https://github.com/btmartin721/mecr_boxturtle; [7]https://github.com/lgnewton/A.unicolor_sp_delim.; [8]https://github.com/rpyron/delim-SOM

I Outreach material regarding machine learning applied to species delimitation

Dá para ensinar uma máquina a delimitar espécies?

Avaliações detalhadas sobre a performance das aplicações de aprendizado de máquina (AM) são essenciais para que essas ferramentas se consolidem como alternativas viáveis, objetivas e robustas dentro da Biologia Evolutiva.

Assim como em qualquer outra área científica, o uso de métodos de AM na delimitação de espécies deve ocorrer apenas após uma análise crítica de seus pontos fortes e limitações em relação ao problema biológico abordado.

Além disso, mesmo que determinados algoritmos de AM se destaquem como soluções promissoras para problemas complexos de delimitação, ainda é prudente reconhecer que os métodos coalescentes tradicionais permanecem como referências fundamentais e complementares na investigação de limites entre espécies.

Apresentamos aqui alguns pontos de reflexão sobre o desenvolvimento de aplicações de aprendizado de máquina voltadas à inferência de limites entre espécies.

Esperamos que gostem!

Autoria
Matheus Salles

Revisão do texto
Fabricius Domingos

Consultoria artística
Laura Laino

Contato:
matheusmaciel.salles@gmail.com

Quer saber mais sobre o tema?
Acesse: bit.ly/4owVV51

Guia do mochileiro do aprendizado de máquina *vol. 1*

Aprendizado de máquina e delimitação de espécies

Caracterização do método de AM

Analisar as vantagens e desvantagens em relação aos métodos já existentes · Verificar se o conjunto de dados está adequadamente descrito, seja em termos de sua estrutura, ou representação biológica · Esclarecer para a comunidade científica as etapas de transformação e padronização dos dados e pensar na forma de disponibilização online do fluxo de trabalho que está sendo desenvolvido · Detalhar e disponibilizar os modelos treinados

Método coalescente ou de AM?

Refletir se um novo método poderia desempenhar melhor em relação a outros já existentes · Analisar de que forma seria possível comparar um método de AM com outro(s) coalescente(s) · Verificar em quais cenários evolutivos um método de AM poderia ser mais eficiente · Avaliar a performance computacional do método de AM em questão · Avaliar se há a necessidade de simular dados e se os simuladores genéticos atualmente existentes são suficientes no contexto do fluxo de trabalho que está sendo desenvolvido

Avaliação estatística da performance do método de AM

Considerar as diferentes possibilidades de avaliação do método de AM em termos de performance preditiva · Verificar quais métricas estatísticas podem ser utilizadas · Refletir sobre a possibilidade de se analisar a performance do modelo em relação a diferentes configurações do conjunto de dados · Pensar em maneiras de reduzir o risco de *overfitting* durante o treinamento dos modelos

Forma de representação dos dados e parametrização dos modelos evolutivos

Determinar como os dados biológicos serão representados · Pensar em formas de criar um fluxo de trabalho extensível para diferentes tipos de marcadores genéticos · Explorar os limites de parametrização de cada modelo evolutivo treinado pelo algoritmo de aprendizado de máquina · Analisar em que medida os resultados obtidos com uma parametrização específica podem ser aplicados a outros contextos

Acessibilidade das ferramentas de AM

Investigar o novo método de AM em termos de sua reprodutibilidade · Fornecer aos usuários um fluxo de trabalho bem documentado, com modelos pré-treinados e parametrizações claramente descritas · Disponibilizar tutoriais e manuais de instrução para a instalação e uso da nova plataforma · Considerar as diferenças entre os ambientes de programação das ferramentas de AM, bem como sua distinção em relação aos métodos coalescentes tradicionais

AM como uma alternativa integrada ao conjunto de ferramentas da delimitação de espécies

Averiguar até que ponto recursos computacionais representariam um fator limitante para o uso da plataforma · Minimizar barreiras ao uso de ferramentas de AM por pesquisadores em contextos onde práticas de bioinformática ainda não estão plenamente consolidadas · Incentivar a promoção de oportunidades de treinamento em AM para cientistas que trabalhem com delimitação de espécies

2. CAPÍTULO II: CHOICES THAT MATTER: THE IMPACT OF SUBSTITUTION MODELS ON MACHINE LEARNING-BASED SPECIES DELIMITATION INFERENCE

RESUMO

A escolha do modelo de substituição de nucleotídeos é um elemento central na inferência filogenética, pois influencia diretamente a precisão das estimativas de parâmetros evolutivos e, em consequência, a seleção de modelos demográficos e de delimitação de espécies. Com o uso crescente de métodos de aprendizado de máquina baseados em dados simulados, ainda não está claro até que ponto o modelo de substituição empregado durante o treinamento pode afetar o desempenho e a robustez dos classificadores quando aplicados a dados empíricos, geralmente marcados por intensa heterogeneidade genômica. Para investigar essa questão, realizamos um estudo de simulação controlada que avaliou o impacto da escolha incorreta do modelo de substituição na inferência por aprendizado supervisionado. Treinamos classificadores do tipo *Random Forest* com dados simulados sob três modelos amplamente utilizados (JC69, HKY e GTR) e testamos seu desempenho na identificação do modelo de delimitação correto em conjuntos de dados com uma combinação de processos de substituição entre loci, um cenário que reproduz de forma realista a heterogeneidade observada em genomas. Os resultados mostraram que classificadores treinados com um único modelo simples foram capazes de generalizar bem para dados de teste mais complexos, identificando de forma consistente o modelo demográfico verdadeiro com alta probabilidade posterior (média superior a 0,86 mesmo com apenas 100 SNPs). O desempenho aumentou até cerca de 600–800 SNPs, ponto em que atingiu um platô. As diferenças de acurácia entre os classificadores treinados sob JC69, HKY ou GTR foram mínimas, sugerindo que o sinal demográfico capturado pelo espectro de frequência de sítios (*site frequency spectrum*) prevalece sobre possíveis distorções causadas pela escolha do modelo de substituição dentro do intervalo de parâmetros testado. Ainda assim, essa robustez depende do contexto evolutivo. Cenários mais extremos (como divergências profundas, forte variação de taxas entre sítios ou dados de regiões codificantes) provavelmente extrapolam as condições avaliadas aqui e podem reduzir significativamente o desempenho dos classificadores. Além disso, uma boa capacidade de seleção de modelos não garante estimativas precisas de parâmetros, já que comprimentos de ramos e taxas evolutivas continuam sensíveis à escolha do modelo de substituição. Concluímos que, em muitas aplicações práticas de delimitação de espécies, o treinamento sob modelos simples pode ser uma estratégia válida e eficiente, desde que acompanhado de testes rigorosos de validação, avaliação de adequação e uma compreensão clara das limitações impostas pela complexidade dos dados genômicos. Nossos resultados oferecem, assim, um caminho pragmático para integrar a seleção de modelos filogenéticos a fluxos de trabalho modernos baseados em aprendizado de máquina, equilibrando eficiência computacional e rigor biológico.

Palavras-chave: aprendizado supervisionado; *random forest*; espectro de frequência de sítios; filogenia; simulação.

ABSTRACT

The choice of nucleotide substitution models is a cornerstone of phylogenetic inference, influencing the accuracy of the estimated evolutionary parameters and, by extension, demographic and species delimitation model selection. With the growing adoption of machine learning methods trained on simulated data, it remains unclear how the substitution model used during simulation training influences classifier performance and robustness when applied to empirical data, usually characterized by pervasive genomic heterogeneity. To address this gap, we conducted a controlled simulation study to evaluate the impact of substitution-model misspecification on supervised machine learning inference. We trained Random Forest classifiers on data simulated under three common substitution models (JC69, HKY, and GTR) and evaluated their performance in selecting the correct delimitation model from test datasets featuring a mixture of substitution processes across loci, a realistic scenario mimicking genomic heterogeneity. Our results demonstrate that classifiers trained under a single, simplistic substitution model generalized effectively to mixed-model test data, consistently identifying the true demographic model with high posterior probability (mean probability > 0.86 even using 100 SNPs), with highest performance plateauing beyond 600–800 SNPs. Notably, the differences in accuracy among classifiers trained under JC69, HKY, or GTR were minimal, indicating that the demographic signal captured by the site frequency spectrum predominates over substitution-model artifacts within the tested parameter space. However, this robustness is context-dependent. We caution that some extreme, though realistic, evolutionary scenarios (such as deep divergence, strong among-site rate variation, or protein-coding data) likely exceeds the conditions tested here and may severely degrade classifier performance. Furthermore, robust model selection does not imply accurate parameter estimation, as branch lengths and evolutionary rates remain sensitive to model misspecification. We conclude that for many practical applications in species delimitation, faster and computationally efficient training under simple models can be sufficient, provided it is coupled with rigorous validation, model-adequacy assessment, and an awareness of the limitations imposed by complex genomic data. Our findings offer a pragmatic framework for integrating phylogenetic model selection with modern ML workflows, balancing computational efficiency with biological rigor.

Key words: supervised learning; random forest; site frequency spectrum; phylogenetics; simulation.

2.1 INTRODUCTION

The selection of nucleotide substitution models is a critical step in phylogenetic analysis, directly influencing the accuracy of evolutionary inferences, including tree topology, branch lengths, and divergence times (Yang, 1996; Naser-Khdour et al., 2019). These mathematical models approximate the complex process of sequence evolution by parameterizing factors such as transition-transversion bias, equilibrium base frequencies, rate heterogeneity across sites, and codon-position effects (see Reviews in Sullivan & Joyce, 2005; Yang & Rannala, 2012). Due to the super-exponential growth in possible tree topologies with increasing taxa and the computational cost of likelihood calculations, commonly used models can be interpreted as necessarily simplified abstractions of biological reality.

The simplest models (e.g., JC-69: Jukes & Cantor (1969)) assume uniform rates and state sets across sites, treating sequence evolution as an unconstrained random process. While tractable and parameter-sparse, such models are often biologically inadequate, as ignoring rate heterogeneity can produce biased estimates (Yang, 1996). Consequently, most models incorporate subsets of higher complexity, such as substitution biases and among-site rate variation modeled via a discrete gamma distribution (Sullivan & Joyce, 2005; Yang & Rannala, 2012; Arenas, 2015). Therefore, an inevitable trade-off exists between biological realism, statistical identifiability, and computational tractability.

Biological complexity further extends beyond rate variation to include site-specific constraints on permissible states. Functional pressures (such as those acting on protein active sites, codon positions, or RNA structures) restrict the repertoire of acceptable substitutions (Echave et al., 2016). This challenge is further compounded, for example, by pervasive heterogeneity in evolutionary processes across the genome. For instance, functionally constrained regions evolve under distinct patterns compared to more flexible regions, creating heterogeneity that a single model cannot capture (Xia, 2000). Furthermore, stationarity, reversibility, and homogeneity (SRH) violations frequently occur in specific genomic contexts (e.g., third codon positions, viral and mitochondrial genes), and these violations are genuine features of the data rather than artifacts of sampling noise (Naser-Khdour et al., 2019). This implies that different genomic partitions can yield statistically distinct phylogenetic patterns, underscoring the importance of partition-specific model assessment.

To address heterogeneity, partitioned analyses and codon models allow different substitution processes for distinct loci or codon classes (Lanfear et al., 2012). However, this approach introduces a new trade-off: too few partitions risk underfitting, while excessive partitioning can lead to overfitting and increased parameter error (Lanfear et al., 2012; Gupta & Vadde, 2023). Furthermore, although site-specific preferences can, in principle, be inferred (Meyer & Haeseler, 2003), accurate estimation of such specific substitution models requires extensive data to overcome phylogenetic correlations and the large number of involved parameters (Puller et al., 2020). Model selection criteria (e.g., AIC, BIC, among many others) are consequently used to identify an optimal balance between fit and complexity by grouping sites that evolve under similar processes (Posada & Crandall, 2001; Ripplinger et al., 2010).

Sensitivity to substitution model choice is greater for protein data than for nucleotide data, mainly due to the larger amino acid state space (20 vs. 4), which increases the influence of model assumptions on inference accuracy. This expectation is confirmed by empirical evidence for nucleotides, which indicates that while complex nucleotide models (e.g., GTR+I+G) can often function as reliable defaults (Abadi et al., 2019), the prevailing view in protein phylogenetics is that explicit model selection is essential. The consequences are especially acute for protein evolution and ancestral sequence reconstruction (ASR), where inferences under best-fitting models yield superior topologies and branch lengths (Del Amparo & Arenas, 2022; 2023). In contrast, inappropriate models, particularly those with divergent amino acid exchangeability matrices, introduce systematic biases that propagate across all internal nodes of a tree, jeopardizing downstream biological interpretations (Del Amparo & Arenas, 2022).

Naturally, these critical principles of model impact extend directly to fields like species delimitation and demographic inference. Failure to account for model heterogeneity can bias essential evolutionary parameters (such as divergence times, effective population sizes, and migration rates), thus potentially compromising the distinction of species limits (Momigliano et al., 2021; Tiley et al., 2023). Because substitution models shape the likelihood surface from which gene trees and branch lengths are estimated, the uncertainty in these trees directly influences coalescent-based estimates of population parameters and species boundaries. Consequently, misspecified models (for example, ignoring across-site rate variation, compositional heterogeneity, heterotachy, codon structure in coding loci, or differences among loci that merit separate partitions) can produce systematic errors in tree topology and branch-length scaling, increase gene-tree estimation error, and thereby alter inferred levels of genealogical discordance that delimitation methods interpret as evidence for or against lineage independence. Contemporary species delimitation methods, such as those available in BEAST (Baele et al., 2025) and BPP (Flouri et al., 2018), allow integration of substitution models within the multispecies coalescent, but they commonly require user choices about partitioning and model parametrization; if those choices are not tested or if model uncertainty is not propagated, downstream delimitation and demographic inferences may be imprecise.

Building upon this foundation, machine learning (ML) has emerged as a powerful tool for navigating complex model spaces in species delimitation and demographic inference (Salles & Domingos, 2025). These approaches, specially supervised ones, are typically trained on a vast amount of simulations (Schrider & Kern, 2018). Furthermore, ML applications used in species delimitation commonly learn to map patterns in multidimensional feature vectors, composed of population genetics summary statistics, or to perform model selection of demographic or delimitation models (Salles & Domingos, 2025). A critical, yet often underappreciated, assumption of such a framework is that the joint distribution of features in the training data (simulations) must match that of the empirical data to which the classifier is applied. This is because the learned decision boundaries are intrinsically tied to the data parameter space on which the model was trained. Consequently, the substitution model chosen for simulation becomes a fundamental pillar of the entire ML pipeline.

Genomic data is characterized by pervasive heterogeneity in evolutionary processes, meaning that a single, uniform substitution model is biologically implausible for most genome-scale datasets. If an ML classifier is trained on alignments simulated under a simplistic or uniform model (e.g., JC69, or GTR+G applied homogeneously), the distribution of its input features will likely reflect this artificial homogeneity. Consequently, if this classifier is then deployed on empirical data comprising distinct partitions (e.g., exons, introns) that evolved under divergent substitution processes, the feature vectors will likely be drawn from a different, distorted distribution. This distributional shift induces a covariation, causing the classifier to operate in a region of feature space it was not trained on, potentially leading to poorly calibrated probabilities, misclassification, and a lower predictive performance. Ultimately, if this error propagates forward, parameter estimation and model selection might become biased, thus potentially leading to significant mistakes in species delimitation. Therefore, the accuracy of ML-based inferences is not only a function of the evolutionary model being tested but is inextricably linked to the biological realism of the substitution models used in its training phase.

In this context, the present study evaluates the effect of substitution-model heterogeneity on ML-based inference of species limits and demographic models. Using controlled multi-locus simulations and supervised classifiers, we compare uniform *versus* partition-aware model selection, assess how partitioning schemes (that is, how finely the data are subdivided into partitions with distinct substitution models) trades off model fit and classifier performance across a range of SNP densities, and characterize the direction and magnitude of biases that arise when among-locus heterogeneity is ignored. Finally, we integrate these results into practical recommendations for combining phylogenetic model-selection and ML workflows in species-delimitation and demographic inference.

2.2 METHODS

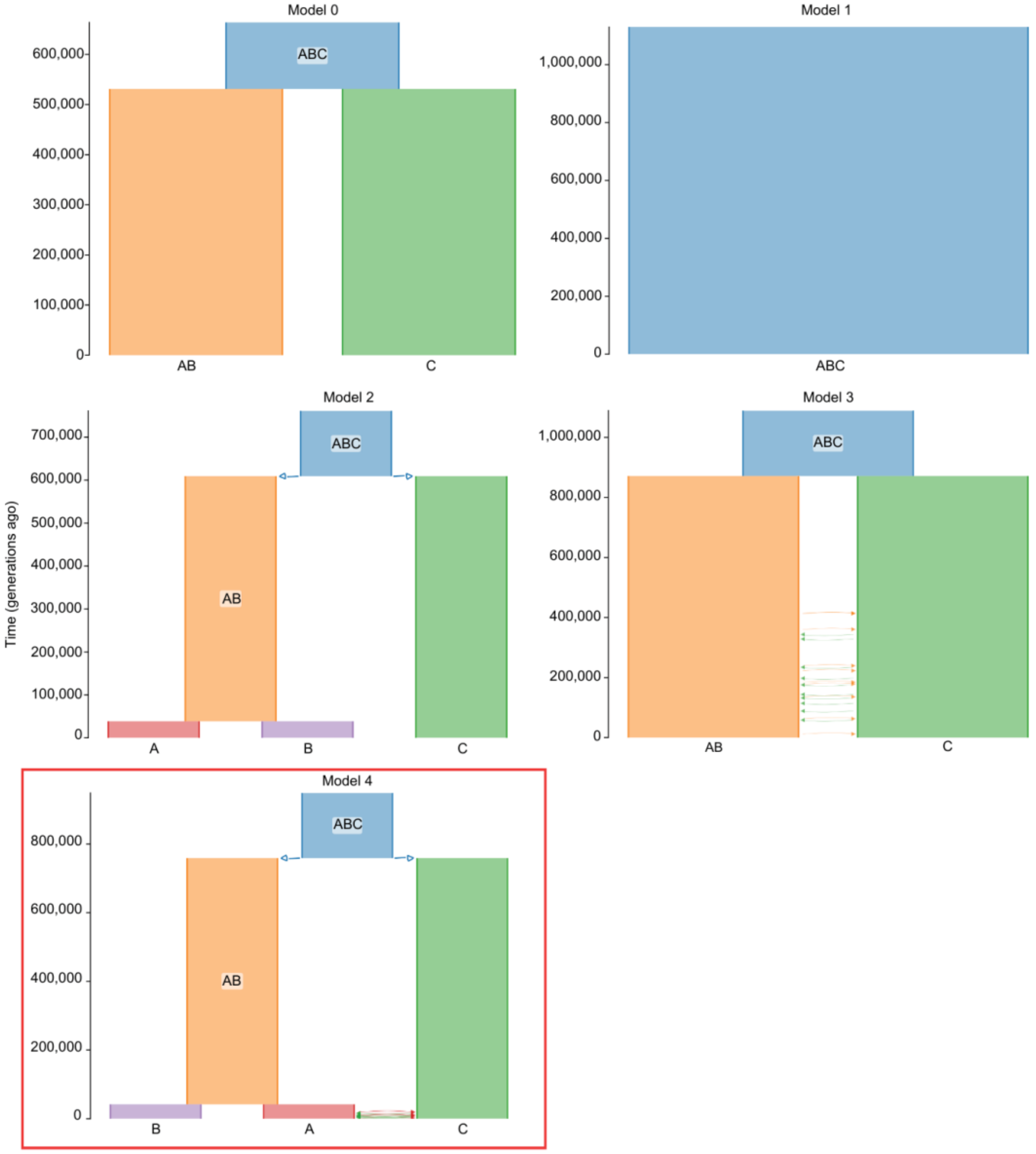
2.2.1 OVERVIEW AND SOFTWARE

All training and test datasets were generated with *popai v1.0* (https://github.com/SmithLabBio/popai), a Python-based framework for simulating population genomic data under user-specified demographic and mutational models, with an emphasis on species delimitation. *Popai* orchestrates simulation runs through configuration files that specify the phylogenetic background, migration matrix, population assignment, and parameter space. The software natively accommodates key evolutionary processes, including divergence, secondary contact, and divergence with ongoing migration. Global keys in the configuration file define elements such as random seed, number of replicates, substitution model, prior distribution for mutation rates, flags for symmetric versus asymmetric migration, and options to enable secondary contact or divergence-with-gene-flow models. *Popai* generates simulation commands for the underlying coalescent and sequence simulators, logs metadata for each replicate, and preserves the complete configuration, ensuring reproducibility and traceability of all experimental steps.

2.2.2 DEMOGRAPHIC MODELS

We evaluated five competing demographic scenarios (FIGURE 1), which included divergence-only, divergence with secondary contact, and divergence with gene flow between non-sister populations. The species tree topology, with three populations [(A, (B, C))], was fixed with 10 diploid individuals sampled per population and 20 loci per dataset. Effective population sizes were kept constant across all populations. Migration was modeled through a Boolean migration matrix that specified eligible population pairs. We restricted rates to symmetric values in the range 10^{-5}–10^{-4} per generation and limited each model to a single migration event. Secondary contact models were activated via a flag, introducing migration at the midpoint between the most recent divergence and the present, persisting until the present. In contrast, divergence-with-gene-flow models, in which migration begins immediately after divergence and ceases halfway to the next split, were not considered in this study. Mutation rates followed a uniform prior $U(5\times10^{-9}, 5\times10^{-7})$ substitutions/site/generation.

FIGURE 1. Five competing demographic models tested in the present study. Models include scenarios with divergence without gene flow, divergence with secondary contact between sister populations, and divergence with asymmetric gene flow between non-sister populations. The model simulated to generate the empirical test datasets (assumed as the true one) is highlighted in red. Time is given in generations before the 76otaling.



SOURCE: the author (2025).

2.2.3 SIMULATION DESIGN

Training data for the five demographic scenarios (Figure 1) were generated under three substitution models (JC69, HKY, and GTR). Sequence lengths ranged from 500 to 3,500 bp in 500 bp increments, corresponding approximately to 100, 300, 600, 800, 1,000, 1,200, and 1,400 segregating sites (SNPs). These SNP counts were used directly to build the site frequency spectrum (SFS). We here define a *regime* as a specific combination of substitution model and matrix size (i.e., one substitution model applied to one alignment length & SNP count). For each regime, we simulated the five fixed demographic scenarios, resulting in a total of 3 substitution models × 7 alignment sizes × 5 demographic scenarios = 105 training datasets. All simulations used the same tree topology (Figure 1), so the only sources of variation across regimes were the substitution model and alignment length. Importantly, we did not perform mixed-model training: each training dataset was generated under a single substitution model.

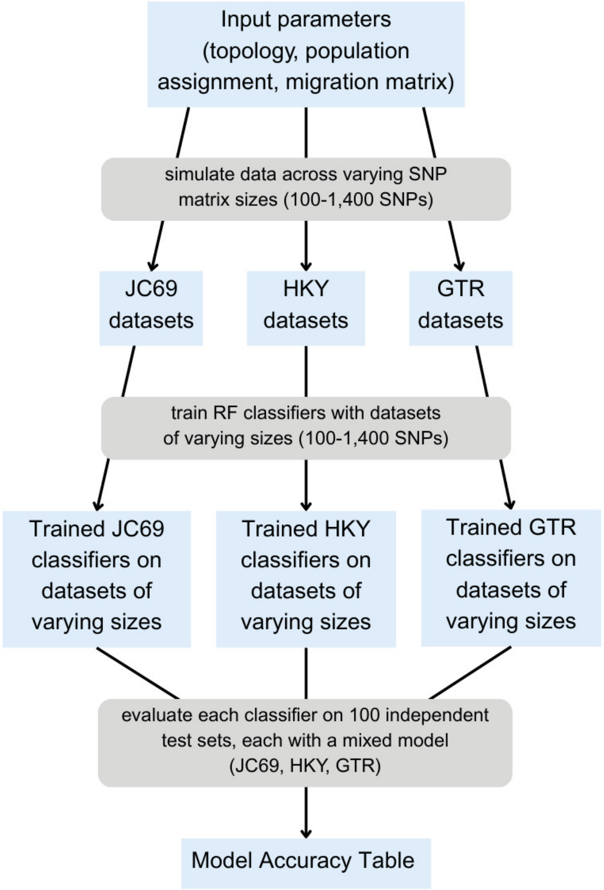
Test datasets were generated independently from the training sets and we assumed Model 4 (Figure 1) as the true demographic scenario. For each SNP matrix size (100–1400), we simulated 100 independent replicates per substitution model. Unlike the training sets, test datasets consisted of heterogeneous substitution regimes: each dataset included 20 loci drawn from a fixed mixture of models (e.g., 7 loci JC69, 7 loci HKY, 6 loci GTR). This design mimics realistic heterogeneity across loci. Because all data were fully simulated, no missing genotypes were present and down-projection was unnecessary.

2.2.4 CLASSIFIER TRAINING AND EVALUATION

The folded SFS (based on minor allele frequencies) was used as the main summary statistic. For each test replicate, 10 SFS were generated, preserving stochastic variance across replicates. These were used as input features for classification. For each substitution model regime, a separate Random Forest (RF) classifier was trained using the corresponding datasets. Features included SFS bins exported from *popai*. Each trained classifier was then evaluated against the 100 independent test datasets per SNP size. Evaluation mostly considered model-mismatched conditions (e.g., training under JC69, testing with mixed JC69/HKY/GTR). Probability outputs across demographic models were recorded for each replicate. Performance metrics included mean, median, standard deviation, amplitude, and 95% confidence intervals, which were estimated from the 100-test set resamples.

The full workflow of dataset generation, classifier training, and evaluation is summarized in FIGURE 2.

FIGURE 2. Workflow of the experimental design. Input files were processed with *popai* to generate multilocus datasets of varying SNP sizes. Then, training datasets simulated under different substitution regimes were used to train RF classifiers. Each classifier was evaluated on 100 independent test datasets per SNP size, simulated under the true demographic model and including a fixed mixture of substitution models.


Input parameters (topology, population assignment, migration matrix)
simulate data across varying SNP matrix sizes (100-1,400 SNPs)
JC69 datasets
HKY datasets
GTR datasets
train RF classifiers with datasets of varying sizes (100-1,400 SNPs)
Trained JC69 classifiers on datasets of varying sizes
Trained HKY classifiers on datasets of varying sizes
Trained GTR classifiers on datasets of varying sizes
evaluate each classifier on 100 independent test sets, each with a mixed model (JC69, HKY, GTR)
Model Accuracy Table

SOURCE: the author (2025).

2.2.5 REPRODUCIBILITY

All simulations were executed with fixed random seeds to ensure reproducibility. Complete configuration files, scripts, together with input and output data are archived in Zenodo [https://doi.org/10.5281/zenodo.17274456], including: 1. The species tree file (tree.nex); 2. The population assignment file (populations.txt), 3. The migration matrix file (migration.txt), and 4. Simulation configuration files for each substitution model × alignment length combination. Together with step-by-step execution instructions, these resources allow full replication of the workflow, from simulation to classifier training and evaluation.

2.3 RESULTS

Trained classifiers consistently identified the correct demographic model with the highest posterior probability across all simulated regimes. That is, for every dataset used for testing, the correct scenario (Model 4, as previously described) was always ranked as the most probable, demonstrating strong classification performance. Mean probabilities of correct model assignment, along with their 95% confidence intervals, are summarized in TABLE 1. Overall, accuracy increased with the number of SNPs included in the alignment, although performance stabilized beyond approximately 600–800 SNPs. For instance, under the JC69 training regime, the mean probability of correctly inferring the model increased from 0.8693 (95% CI: 0.7100–0.9780) with 100 SNPs to 0.9234 (95% CI: 0.8120–0.9800) with 800 SNPs, with comparable values at larger matrix sizes. A similar pattern was observed for HKY and GTR training regimes, both of which exhibited slightly higher accuracies for larger alignments, reaching mean probabilities above 0.93 with 1200 SNPs. Importantly, performance differences among training regimes (JC69, HKY, and GTR) were minimal, suggesting that the classifiers were robust to the specific substitution model used during training. For example, with 1200 SNPs, mean probabilities were 0.9264 (JC69), 0.9380 (HKY), and 0.9318 (GTR), with overlapping confidence intervals.

TABLE 1. Mean probability of correct model selection and corresponding 95% confidence intervals (in brackets) across different training regimes (JC69, HKY, and GTR) and matrix sizes (100–1400 SNPs). For every test set, the correct evolutionary scenario (Model 4) was always inferred with the highest posterior probability.

Substitution model used during training ↓ / matrix size →	100 SNPs	300 SNPs	600 SNPs	800 SNPs	1000 SNPs	1200 SNPs	1400 SNPs
JC69	0.8693, [0.7100, 0.9780]	0.9088, [0.8100, 0.9800]	0.9239, [0.8160, 0.9860]	0.9234, [0.8120, 0.9800]	0.9170, [0.8320, 0.9720]	0.9264, [0.8340, 0.9780]	0.9207, [0.8400, 0.9820]
HKY	0.8680, [0.7100, 0.9660]	0.9114, [0.8000, 0.9840]	0.9197, [0.8000, 0.9800]	0.9277, [0.8520, 0.9800]	0.9276, [0.8460, 0.9800]	0.9380, [0.8660, 0.9840]	0.9247, [0.8500, 0.9760]
GTR	0.8669, [0.6980, 0.9660]	0.9134, [0.8100, 0.9820]	0.9295, [0.8300, 0.9800]	0.9304, [0.8520, 0.9860]	0.9248, [0.8460, 0.9760]	0.9318, [0.8720, 0.9760]	0.9299, [0.8420, 0.9820]

SOURCE: the author (2025).

Across classifiers from all SNP sizes and training regimes, the demographic model most frequently assigned the second-highest probability (and thus the primary source of classification confusion) was consistently Model 2. The mean probability assigned to this alternative model decreased with increasing SNP counts, falling from approximately 0.089 with 100 SNPs to around 0.046 with 1200 SNPs, which aligns with the observed narrowing of confidence intervals and improved discriminatory power with larger datasets.

2.4 DISCUSSION

We used simulation-trained RF classifiers to evaluate how substitution-model assumptions influence the selection of species delimitation models. In our controlled framework (training separate classifiers under single substitution models and testing on independent datasets that deliberately mixed substitution models) classifiers were consistently able to recover the true model with high posterior probability. Even with only 100 SNPs the mean posterior for the correct model exceeded ~0.86, and performance rose with increasing SNP counts before reaching a plateau near ~600–800 SNPs. Confidence intervals narrowed with more SNPs, indicating greater precision as phylogenetic signals increased.

Two aspects of these results are particularly informative. First, classifier performance improved with additional SNPs but exhibited diminishing returns beyond a moderate data size; this suggests there is a practical “sweet spot” where additional data yield marginal gains. Second, differences among classifiers trained under JC69, HKY and GTR were small across the SNP range we explored. For example, at 1,200 SNPs mean correct-model posteriors were 0.9264 (JC69), 0.9380 (HKY) and 0.9318 (GTR) with overlapping confidence intervals. Within our simulation design these results indicate that delimitation signals (or demographic ones) encoded in the SFS and related summaries dominated the idiosyncratic differences induced by substitution models, so that even classifiers trained under simpler models often generalized well to heterogeneous test sets.

Furthermore, the pattern of model misassignment offers a critical biological interpretation. Across all SNP sizes and training regimes, Model 2 (divergence between three species, without gene flow) was consistently identified as the second-most probable scenario. This systematic confusion is phylogenetically meaningful. Although Model 2 lacks gene flow, it shares an identical topology with the true model (Model 4, divergence with ongoing migration), which likely results in similar site frequency spectrum signatures that require substantial data to disentangle. The consistent decrease in probability assigned to Model 2 as SNP number increased demonstrates that larger datasets progressively sharpen the discrimination between these topologically equivalent but demographically distinct histories. In contrast, the remaining models attracted negligible support, underscoring their clear distinguishability from the true scenario.

These findings are useful but must be interpreted within the limits of our experimental design. First, all results derive from simulations in which the evolutionary scenario, including species topology, sampling scheme, locus number, among many other factors, were fixed by design. Furthermore, the test sets were independent but drawn from the same family of delimitation scenarios. Real empirical data often violate simplifying assumptions in ways not captured by our simulations. For instance, Naser-Khdour et al. (2019) have shown that model violation in phylogenetic analysis is common and heterogeneous across genomic partitions, and that distinct partitions from a single dataset can produce statistically discordant trees. Such partition-level violations alter the distribution of summary statistics used by ML classifiers and therefore may degrade predictive performance if not identified and handled.

Second, several forms of model misspecification that we did not explore in this study can materially influence classifier performance and therefore limit the scope of our conclusions. In particular, molecular phenomena such as strong among-site rate heterogeneity, highly skewed equilibrium base frequencies, deep sequence divergence (saturation), and time-dependent shifts in site preferences, may all change the distribution of site patterns and hence the SFS and other summary statistics potentially used as inputs to the classifiers. These changes can move empirical feature vectors outside the parameter space sampled by our training simulations, producing miscalibration and misclassification. Work on site-specific models (Puller et al., 2020) shows that estimating per-site preferences can improve local fit when many moderately diverged sequences are available, but identifiability problems and residual branch-length errors persist, a reminder that even more realistic models do not always recover true divergence times or eliminate bias. Likewise, protein datasets (Del Amparo & Arenas, 2023), with a larger state space and more complex exchangeability structure, are substantially more sensitive to model choice than the SNP data we simulated. Taken together, these points imply that the robustness we observed for classifiers trained under simple nucleotide models should not be extrapolated to datasets exhibiting extreme heterogeneity or deep divergence. For such cases, strategies like model-adequacy screening, targeted simulations covering those regimes, and validation (e.g., mixed-model training or posterior-predictive checks) before applying classifiers to empirical data, may still be necessary.

It is also important to stress that robustness in model selection does not imply robustness in parameter estimation. Our experiments targeted model-choice accuracy (a classification task), but parameter inference (e.g., divergence times, migration rates, substitution rates) might be more sensitive to substitution-model misspecification because these estimates depend directly on branch-length scaling and substitution–time mapping. Potentially, even when model choice is correct, incorrectly specified substitution models may bias branch lengths, distort SFS summaries, and mislead parameter estimates. Thus, successful model choice should be treated as necessary but not sufficient, with follow-up sensitivity analyses under alternative models and posterior predictive calibration to evaluate estimator reliability.

Finally, the reproducibility and stability of classification accuracy across many independent replicates reinforce the robustness of our pipeline, built upon frameworks such as *popai*. Supervised learning approaches based on summary statistics like the SFS are now well established in evolutionary inference and species delimitation (Schrider & Kern, 2018; Smith & Carstens, 2020). Our results extend this literature by showing that, within the tested ranges, substitution-model differences alone do not undermine classifier accuracy. This finding supports a pragmatic strategy: training classifiers under simpler substitution models can be computationally more efficient without substantially compromising performance, provided that empirical safeguards (such as partition-level adequacy tests, calibration diagnostics, and validation on more realistic simulations) are employed.

This point becomes particularly relevant as we consider that the computational burden of data simulation escalates rapidly with increasing model complexity. Integrating realistic features such as heterogeneous substitution processes, heterotachy, or among-locus heterogeneity can impose prohibitive runtime and memory costs, especially for genome-wide datasets or diverse demographic scenarios. This trade-off is fundamental to supervised machine learning, where simulated data are paramount. When robust simulators exist (as in population genetics) they permit the generation of virtually unlimited training data, a capability bounded chiefly by computational resources (Korfmann et al., 2023). In line with this, we have documented comprehensive runtime benchmarks across our simulation regimes; these data, which clearly illustrate the associated computational trade-offs, will be presented in the final manuscript. Collectively, these factors underscore the critical need to balance biological realism with practical feasibility in the construction of training datasets for supervised inference.

2.4.1 FUTURE DIRECTIONS

Several avenues can deepen and generalize our findings. A first step is to implement mixed-model training, where substitution models are randomly assigned across training replicates. This would test whether robustness extends to heterogeneous training regimes and whether demographic signal truly overrides substitution-model artifacts. Expanding demographic heterogeneity in training (e.g., sampling a wider range of divergence times, migration rates, population sizes, bottlenecks, and expansions) would also test robustness under more realistic evolutionary scenarios. Likewise, testing across alternative tree topologies and sampling schemes would help confirm that robustness is not tied to a particular design.

Improving interpretability is another priority. Decision-tree algorithms like RF allow feature-importance analyses, which could identify which joint-SFS bins or summary statistics most influence classification decisions, and link them to population-genetic expectations (e.g., excess rare variants under growth or migration). Benchmarking against other approaches, such as DIYABC-RF (Collin et al., 2021), under the same simulations would provide context on accuracy, sensitivity to misspecification, and computational cost. Broader comparisons, including alternative machine learning algorithms, could reveal distinct strengths (higher accuracy, better interpretability, or improved scalability) relative to RF. Finally, constructing formal learning curves with larger SNP matrices (e.g., 2,000–5,000 SNPs) would quantify marginal gains of additional data and help define sequencing targets for empirical studies.

2.5 CONCLUSION

Within the controlled parameter space explored in this study, supervised classifiers trained under single substitution models generalized well to mixed-model test data and offered a computationally efficient route to demographic/delimitation model selection. The reproducibility of posterior probabilities across many replicates supports their practical reliability. At the same time, model violation in empirical datasets, deep divergence, and site-specific complexity impose real information and identifiability limits. Robust model choice does not guarantee accurate parameter estimation, underscoring the need for empirical adequacy checks, calibration diagnostics, and validation under more complex simulation regimes. For future applied work, we recommend pairing efficient training under simple substitution models with explicit adequacy testing, targeted validation under complex models, and interpretability analyses. This dual strategy may balance computational efficiency with empirical rigor, supporting stronger and more reliable biological inferences.

2.6 REFERENCES

ABADI, S.; AZOURI, D.; PUPKO, T.; MAYROSE, I. Model selection may not be a mandatory step for phylogeny reconstruction. **Nature Communications**, v. 10, n. 1, p. 934, 2019.

ARENAS, M. Trends in substitution models of molecular evolution. **Frontiers in Genetics**, v. 6, p. 319, 2015.

BAELE, G. et al. BEAST X for Bayesian phylogenetic, phylogeographic and phylodynamic inference. **Nature Methods**, p. 1-4, 2025.

COLLIN, F. D. et al. Extending Approximate Bayesian Computation with Supervised Machine Learning to infer demographic history from genetic polymorphisms using DIYABC Random Forest. **Molecular Ecology Resources**, v. 21, n. 8, p. 2598–2613, 2021.

DEL AMPARO, R.; ARENAS, M. Consequences of substitution model selection on protein ancestral sequence reconstruction. **Molecular Biology and Evolution**, v. 39, n. 7, msac144, 2022.

DEL AMPARO, R.; ARENAS, M. Influence of substitution model selection on protein phylogenetic tree reconstruction. **Gene**, v. 865, p. 147336, 2023.

ECHAVE, J., SPIELMAN, S. & WILKE, C. Causes of evolutionary rate variation among protein sites. **Nat Rev Genet**, v. 17, 109–121, 2016.

FLOURI, T.; JIAO, X.; RANNALA, B.; YANG, Z. Species tree inference with BPP using genomic sequences and the multispecies coalescent. **Molecular Biology and Evolution**, v. 35, n. 10, p. 2585-2593, 2018.

GUPTA, M. K.; VADDE, R. Next-generation development and application of codon model in evolution. **Frontiers in Genetics**, v. 14, p. 1091575, 2023.

JUKES, T. H.; CANTOR, C. R. Evolution of protein molecules. In: MUNRO, H. N. (ed.). **Mammalian Protein Metabolism**. New York: Academic Press, 1969. P. 21-132.

KORFMANN, K., GAGGIOTTI, O. E., & FUMAGALLI, M. Deep learning in population genetics. **Genome Biology and Evolution**, v. 15, n. 2, evad008, 2023.

LANFEAR, R.; CALCOTT, B.; HO, S. Y.; GUINDON, S. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. **Molecular Biology and Evolution**, v. 29, n. 6, p. 1695-1701, 2012.

MEYER, S.; VON HAESELER, A. Identifying site-specific substitution rates. **Molecular Biology and Evolution**, v. 20, n. 2, p. 182-189, 2003.

MOMIGLIANO, P.; FLORIN, A. B.; MERILÄ, J. Biases in demographic modeling affect our understanding of recent divergence. **Molecular Biology and Evolution**, v. 38, n. 7, p. 2967-2985, 2021.

NASER-KHDOUR, S. et al. The prevalence and impact of model violations in phylogenetic analysis. **Genome Biology and Evolution**, v. 11, n. 12, p. 3341-3352, 2019.

POSADA, D.; CRANDALL, K. A. Selecting the best-fit model of nucleotide substitution. **Systematic Biology**, v. 50, n. 4, p. 580-601, 2001.

PULLER, V.; SAGULENKO, P.; NEHER, R. A. Efficient inference, potential, and limitations of site-specific substitution models. **Virus Evolution**, v. 6, n. 2, veaa066, 2020.

RIPPLINGER, J.; SULLIVAN, J. Assessment of substitution model adequacy using frequentist and Bayesian methods. **Molecular Biology and Evolution**, v. 27, n. 12, p. 2790-2803, 2010.

SALLES, M. M. A. S.; DOMINGOS, F. M. C. B. Towards the next generation of species delimitation methods: An overview of machine learning applications. **Molecular Phylogenetics and Evolution**, v. 108368, 2025.

SCHRIDER, D. R.; KERN, A. D. Supervised machine learning for population genetics: a new paradigm. **Trends in Genetics**, v. 34, n. 4, p. 301-312, 2018.

SMITH, M. L.; CARSTENS, B. C. Process-based species delimitation leads to identification of more biologically relevant species. **Evolution**, v. 74, n. 2, p. 216-229, 2020.

SULLIVAN, J.; JOYCE, P. Model selection in phylogenetics. **Annual Review of Ecology, Evolution, and Systematics**, v. 36, n. 1, p. 445-466, 2005.

TILEY, G. P. et al. Estimation of species divergence times in presence of cross-species gene flow. **Systematic Biology**, v. 72, n. 4, p. 820-836, 2023.

XIA, X. Phylogenetic relationship among horseshoe crab species: effect of substitution models on phylogenetic analyses. **Systematic Biology**, v. 49, n. 1, p. 87-100, 2000.

YANG, Z. Among-site rate variation and its impact on phylogenetic analyses. **Trends in Ecology & Evolution**, v. 11, n. 9, p. 367-372, 1996.

YANG, Z.; RANNALA, B. Molecular phylogenetics: principles and practice. **Nature Reviews Genetics**, v. 13, n. 5, p. 303-314, 2012.

3. CAPÍTULO III: MITOCHONDRIAL GENOME EVOLUTION: THE INFLUENCE OF PARTITIONING, CALIBRATION, AND GENE HETEROGENEITY ON PLEURODONTAN SUBSTITUTION RATES

Esta seção apresenta um dos artigos desenvolvidos ao longo do doutorado, já submetido. A formatação do artigo segue as normas da revista em que o manuscrito foi divulgado na forma de pré-publicação.

REFERÊNCIA: Salles, M. M. A., & Domingos, F. M. C. B. (2025). Mitochondrial genome evolution: the influence of partitioning, calibration, and gene heterogeneity on pleurodontan substitution rates. *EcoEvoRxiv*. https://doi.org/10.32942/X2BW7B.

ABSTRACT

Substitution rate estimates are a key source of information in modern evolutionary biology, underpinning divergence time inference and other evolutionary analyses. Mitochondrial DNA nucleotide substitution rates, in particular, are commonly used for these purposes. However, these rates are typically derived from a small set of genes, closely related species, or from a limited number of model organisms. Such limitations become increasingly problematic at deeper phylogenetic levels, where errors in rate estimates and divergence times tend to accumulate with evolutionary distance. Here, we use nearly complete mitogenomes of 27 pleurodontan (Squamata: Pleurodonta) species to estimate substitution rates for the whole clade, paying special attention to the effect of data partitioning, calibrations and model choices on these estimations. The substitution rate estimates we obtained are consistent with previous findings for specific lineages within the group. Rates across individual genes ranged from approximately 0.004 to 0.02 substitutions/site/million years, with notable differences between coding and non-coding regions, and among codon positions. Calibrations had a less pronounced effect on the analyses than anticipated, although subtle differences were observed. These findings underscore the challenges of estimating targeted nucleotide substitution rates, especially for lineages with limited genomic data, as is the case for several Squamata lineages. Moreover, the results provide valuable insights into the evolutionary dynamics of Pleurodonta and emphasize the importance of incorporating robust data and models to improve accuracy in substitution rates and divergence time estimations.

Key words: divergence time, mitogenome, molecular evolution, phylogenomics, Squamata.

Introduction

Genomic datasets are essential for addressing complex questions in modern evolutionary biology. In this context, substitution rate estimates are a cornerstone, providing critical insights into molecular evolution and serving as a foundation for various applications. For instance, in the absence of fossils or other secondary calibration points, substitution rates often represent the main available data for estimating divergence times (e.g., Ho, 2007; Arcones et al., 2021). Mitochondrial DNA (mtDNA), in particular, has long been used for this purpose, mainly due to its relatively stable coding function, high mutation rates, small effective population size, matrilineal inheritance, and relatively fast coalescent times (Avise et al., 1987; Ballard & Rand, 2005). Besides, mitochondrial proteins play a critical role in the oxidative phosphorylation pathway and exhibit functional conservation across different metazoan lineages (Gray et al., 1999; Broughton & Reneau, 2006). Consequently, the accuracy of mitochondrial substitution rate estimates is fundamental to advancing evolutionary biology.

Substitution rates vary considerably across the mitochondrial genome and among different taxonomic groups. Empirical studies have revealed substantial variation among different mitochondrial genes (Williams & Hurst, 2002; Sloan et al., 2009; Pons et al., 2010; Duchêne et al., 2011; Zhu et al., 2014; Yang et al., 2018) as well as across lineages (Parkinson et al., 2005; Bininda-Emonds, 2007; Mower et al., 2007; Nabholz et al., 2008; Welch et al., 2008; Eo & DeWoody, 2010; Yan et al., 2021). Importantly, many studies have historically relied on a limited fraction of the mitogenome—primarily cytochrome b, cytochrome c oxidase I, II, and III, and the 12S and 16S ribosomal RNAs (Johns & Avise, 1998; Hebert et al., 2003; Roe & Sperling, 2007; Patwardhan et al., 2014)—and have been based on a few model organisms, typically at the intraspecific level or between closely related species (Avise et al., 1987; Ballard & Whitlock, 2004; Funk & Omland, 2005; Ballard & Rand, 2005; Rubinoff & Holland, 2005).

Consequently, despite their widespread use, molecular clock approaches based on mtDNA have important practical limitations. Overlooking those variations can introduce substantial biases in substitution rate estimates, posing challenges for accurate evolutionary inference. This is particularly concerning in deep-level phylogenies, where errors in phylogenetic inference tend to amplify with increasing branch length (Buckley, 2002; Lemmon et al., 2009). To mitigate this problem, some studies have attempted to calibrate molecular rates using complete (or nearly complete) mitogenomes, across different groups (Pons et al., 2010; Park et al., 2012; Plazzi et al., 2016; Mackiewicz et al., 2022). This specificity is crucial, as accurate divergence time estimates rely on the precision and accuracy of calibration points and the rates applied to each marker and lineage under investigation (Mello & Schrago, 2014; Zheng & Wiens, 2015; Ritchie et al., 2017; Smith et al., 2018). Also, effective calibrations help to counteract errors arising from clock model misspecification (Duchêne et al., 2014).

Squamates (lizards, snakes, and amphisbaenians; Order Squamata) form a globally distributed clade of reptiles comprising approximately 11,000 extant species (Simões & Pyron, 2021; Uetz et al., 2025), making them one of the most diverse vertebrate orders (Uetz et al., 2021). Despite recent advancements in next-generation sequencing, squamates remain underrepresented in genomic research compared to mammals and birds (Feng et al., 2020; Genereux et al., 2020; Gable et al., 2023). This limited genomic data availability hinders a comprehensive understanding of key evolutionary parameters within the group, including substitution rates. In particular, the Pleurodonta clade (the main focus of this study) encompasses a wide range of taxa predominantly distributed throughout the New World, with desert iguanas, horned, spiny, and collared lizards dominating many modern squamate faunas in North and South America (Pianka and Vitt, 2003; Losos, 2011; Avila et al., 2013; Carvalho et al., 2013). Although Pleurodontan evolutionary history is marked by multiple adaptive radiations in response to varied ecological pressures (Blankers et al., 2013; Alencar et al., 2024), mitochondrial evolutionary parameters remain scarce for the group. Commonly used substitution rate values broadly range from 0.005 to 0.02 substitutions per site per lineage per million years (subs/site/MY), depending on the gene (e.g., Zarza et al., 2008; Chan et al., 2012; Fontanella et al., 2012; Olave et al., 2015; Werneck et al., 2015; Román-Palacios et al., 2018; Bernardo et al., 2019; Camurugi et al., 2022; Carvalho et al., 2024; Rogers et al., 2024). However, as in most vertebrate groups, these estimates are often based on a limited number of species, typically at shallow evolutionary scales, and frequently rely on a small set of mitochondrial genes.

To address this issue, we integrated recently sequenced mitochondrial data with existing mitogenomic data to conduct comprehensive phylogenetic analyses, assessing evolutionary rate variation among Pleurodonta mitochondrial genes. Specifically, we analyzed their mitochondrial genomes to estimate its mitochondrial substitution rates. Using fossil-calibrated Bayesian phylogenetic analyses, we inferred molecular evolutionary rates across several families and characterized new nearly complete mitogenomes for seven *Tropidurus* species: *T. guarani*, *T. melanopleurus*, *T. sp. nov.* (species currently under formal description), *T. spinulosus*, *T. tarara*, *T. teyumirim*, and *T. xanthochilus*. We expect that these newly estimated rates will improve the precision of molecular clock dating and evolutionary inferences in squamates, offering deeper insights into the evolutionary processes influencing biodiversity patterns in this group.

Methods

We assembled a comprehensive dataset of Pleurodontan mitochondrial genomes available from GenBank by November 2024, including seven recently described sequences from different *Tropidurus* species (Salles et al., 2025). One Chamaleonidae species (*Calluma parsonii*) was included as an outgroup, resulting in a final dataset with 28 species (Table 1). Only coding regions (13 genes) and the two mitochondrially encoded ribosomal RNAs (12 and 16s) were used. We excluded additional mtDNA markers because they represent regions that are either non-coding and hyper-variable (D-loop) or ultra-conserved (tRNAs), and therefore inadequate for molecular clock calibrations. We separately aligned each mitochondrial gene with MAFFT v7.471 (Katoh & Standley, 2013) using specific customized settings (-globalpair, --maxiterate 1000, --adjustdirection). Alignments were broadly examined by eye, and AMAS (Borowiec, 2016) was used to concatenate alignments and compute final summary statistics.

Table 1. Species used in all analyses in the present study. New mitochondrial genomes are in bold.

Species	Family	GenBank accession number
Calluma parsonii	Chamaeleonidae	AB474915
Basiliscus vittatus	Corytophanidae	AB218883
Amblyrhynchus cristatus	Iguanidae	NC_028031
Conolophus subcristatus		NC_028030
Cyclura pinguis		NC_027089
Iguana delicatissima		NC_044899
Iguana iguana		NC_002793
Leiocephalus personatus	Leiocephalidae	AB266739
Liolaemus darwinii	Liolaemidae	NC_057242
Liolaemus millcayac		NC_057243
Liolaemus parthenos		NC_057244
Chalarodon madagascariensis	Opluridae	AB266748
Oplurus grandidieri		AB218720
Holbrookia lacerata	Phrynosomatidae	NC_041001
Phrynosoma blainvillii		NC_036492
Sceloporus occidentalis		AB079242
Urosaurus nigricaudus		NC_026308
Anolis punctatus	Polychrotidae*	NC_044125
Anolis cybotes		AB218960
Polychrus marmoratus		AB266749
Plica plica	Tropiduridae	AB218961
Tropidurus guarani		*will be submitted to genbank*
Tropidurus melanopleurus		*will be submitted to genbank*
Tropidurus sp. nov.		*will be submitted to genbank*
Tropidurus spinulosus		*will be submitted to genbank*
Tropidurus tarara		*will be submitted to genbank*
Tropidurus teyumirim		*will be submitted to genbank*
Tropidurus xanthochilus		*will be submitted to genbank*

* Traditionally, *Anolis* was classified within Polychrotidae. However, molecular phylogenetic studies have led to a major taxonomic reassessment. Recent evidence supports placing *Anolis* and related genera within Dactyloidae, rendering Polychrotidae paraphyletic or obsolete. While some taxonomic authorities now recognize Dactyloidae, references to Polychrotidae persist in the literature. Our option here was to consider *Anolis* and *Polychrus* to form a distinct phylogenetic group, despite of their taxonomical status. The group monophyly was not enforced and, hence, taxonomic arrangements had no influence in our analyses.

Effect of calibration points on substitution rate estimates

We implemented different calibration strategies to understand its possible effects on substitution rate estimates. Specifically, estimates were obtained separately through calibrated and non-calibrated analyses. Calibration points within the Pleurodonta clade were obtained consulting the specialized literature, prioritizing those that have been used in multiple evolutionary studies, and which are broadly supported by the fossil record (Table 2). Some possibly accurate calibrations, also commonly cited in the literature, but for groups whose monophyly is still under debate, were not used here, as monophyly was enforced for each calibrated node. We also note that estimating a fully resolved topology or divergence times for the entire group was not our primary objective, as the species included in this study represent only a limited sample of Pleurodontan diversity and exclude some of the group's most representative lineages.

Table 2. Values (million years, MY) of uniformly distributed calibration priors applied in dating analyses, based on both fossil and molecular data. Settings for calibration Bayesian prior mean, standard deviation and offset are provided. MRCA = most recent common ancestor.

Calibrated node (MRCA prior)	Species included	Lower value	Upper value	Offset	References
Pleurodonta	All except outgroup (*Calluma parsonii*)	65	85	0.5	Conrad & Norell (2007); Townsend et al. (2011); Prates et al. (2015); Scarpetta (2019)
Anolis	*Anolis cybotes*, *Anolis punctatus*	40	60	0.5	Sherratt et al. (2015); Zheng & Wiens (2016); Román-Palacios et al. (2018)
Phrynosomatidae	*Holbrookia lacerata*, *Phrysonoma blainvillii*, *Sceloporus occidentalis*, *Urosaurus nigricaudus*	35	55	0.5	Townsend et al. (2011); Leaché & Linkem (2015); Zheng & Wiens (2016)
Liolaemus 2	*Liolaemus darwinii*, *Liolaemus parthenos*, *Liolaemus millcayac*	30	45	0.5	Portelli et al. (2022)
Liolaemus 1	*Liolaemus darwinii*, *Liolaemus parthenos*	10	25	0.5	Fontanella et al. (2012); Portelli et al. (2022)

Bayesian estimation of mitochondrial nucleotide evolution rates

For each mitochondrial partition, mean nucleotide substitution rates were estimated using BEAST v2.7 (Drummond & Rambaut, 2007), applying a relaxed molecular clock with an uncorrelated log-normal distribution (*ucld*) and either a Yule or

Calibrated Yule speciation model, depending on the test. The relaxed *ucld*-model assumes independent substitution rates across branches, as there is no assumed correlation between the rate of a given lineage and that of its ancestor. This model requires a prior for the mean clock rate. For coding sequences, we set the mean clock rate to 0.01 substitutions/site/MY, and for rRNAs, to 0.0055, based on prior estimates for various Pleurodonta species (Supporting Table S1). A normal distribution was used for the *ucld* mean rate prior, with the above values as the mean, a standard deviation (Sigma) of 0.005 for coding sequences and 0.0015 for rRNAs, with these same values used as the Offset. These hyperprior values (Sigma and Offset) were determined based on preliminary analyses to ensure appropriate parameterization, and also computational and statistical demands.

For each calibration scheme (whether or not it included calibration points), we estimated substitution rates for each mitochondrial gene and for each codon position within protein-coding sequences. Tree topologies were linked across partitions, while clock models were unlinked both between genes and among codon positions within genes. Site models were linked across codon positions within individual genes but unlinked between genes, with model selection performed using BEAST Model Test (bModelTest; Bouckaert and Drummond, 2017) under the ‘namedExtended’ model set. Uncalibrated analyses consisted of two independent MCMC runs of 500 million generations each, with parameters sampled every 25,000 generations. Calibrated analyses followed the same sampling scheme, but each run was extended to 850 million generations. Convergence of all parameters was verified using Tracer v1.4 (Rambaut et al., 2007), ensuring effective sample sizes (ESS) ≥ 200 whenever possible. In summary, we performed four BEAST analyses (two calibrated and two uncalibrated) and reported the final results as the combination of two runs per analysis using LogCombiner (Rambaut and Drummond 2014).

Results

Alignments and evolutionary models

The alignment of protein-coding sequences alone comprised 11,426 bp, while the inclusion of non-coding sequences increased the total length to 14,059 bp. All coding genes exhibited multiple substitution models within the 95% highest posterior density (HPD) interval estimated through the bModelTest. Only the two rRNAs had a single best-fitting model to explain site substitution, specifically the GTR model (Supporting Table S2).

Substitution rates

Analysis of substitution rates across codon positions revealed heterogeneity in evolutionary rates across the mitochondrial genome of Pleurodontans; 95% posterior distributions of substitution rates for each gene can be observed in Fig. 1. When considering only protein-coding sequences, substitution rates appear relatively homogeneous across genes, with substantial overlap of the HPD intervals (Fig. 1A). Conversely, coding and non-coding regions exhibit markedly different substitution rates, with non-coding regions evolving approximately ten times slower (Fig. 1B). Calibrated analysis (for all genes) has consistently shown similar rates to non-calibrated ones, but with slightly smaller estimates. Median substitution rate estimates for individual genes, based exclusively on third codon positions, are presented in Table 3 (estimates for all codon positions are available in the original BEAST output files archived on Zenodo). In the case of non-calibrated analysis, the fastest mean rate was observed for ND4 and the slowest for *12s* and *16s*. Regarding calibrated analysis, rRNAs also exhibited the lowest estimates, but in this case ND2 presented the higher value.

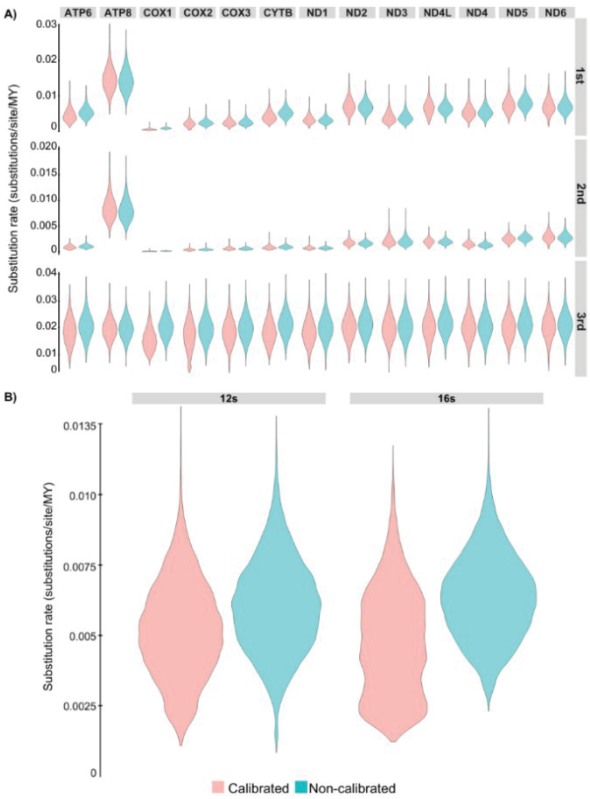


Fig. 1. Posterior distributions of mitochondrial substitution rates (substitutions/site/MY) from calibrated (pink) and non-calibrated (blue) analyses. **(A)** Violin plots for 13 protein-coding genes, illustrating the range and density of estimated rates for each codons position **(B)** Violin plots for two rRNA genes. In both cases, the width of each violin indicates the distribution density, and horizontal lines represent median values.

Table 3. Nucleotide substitution rates per site per million years estimated from 13 mitochondrial protein-coding genes (3rd codon position) and 2 rRNAs across 27 Pleurodontan species plus one outgroup. These rates were inferred using BEAST with a relaxed clock model assuming a lognormal distribution. The reported values represent the combined results from two independent runs.

Gene	Non-calibrated		Calibrated	
	ucld mean rate	Stdev	ucld mean rate	Stdev
12s	0.00598	0.0016623	0.00518	0.0017229
16s	0.00632	0.0015115	0.00468	0.0019285
ATP6	0.02059	0.0044696	0.01861	0.0049339
ATP8	0.01977	0.0038046	0.01958	0.0038725
COX1	0.02036	0.0039585	0.01578	0.0042038
COX2	0.01945	0.0055384	0.01706	0.0055097
COX3	0.01995	0.0043769	0.01836	0.0048264
CYTB	0.02136	0.0042930	0.01840	0.0045799
ND1	0.02011	0.0045590	0.01826	0.0047083
ND2	0.02114	0.0043318	0.02025	0.0044484
ND3	0.02049	0.0044802	0.01977	0.0046298
ND4	0.02151	0.0042928	0.01986	0.0048263
ND4L	0.02036	0.0045192	0.01996	0.0045306
ND5	0.02157	0.0042466	0.01991	0.0044204
ND6	0.02121	0.0044322	0.02006	0.0048847

Discussion

The influence of different partitioning schemes on the estimation of substitution rates

Our data support the widely accepted theory that nucleotides at the third codon position evolve at distinct rates and through different mechanisms compared to those at the first and second codon positions (Kimura, 1980). In practical terms, this also means that, within a gene, first codon positions are expected to evolve more similarly to other first codon than to second or third positions, with the same reasoning applying to each position (Bofkin & Goldman, 2007)—which is exactly what we detected here (Fig. 1). Thus, we emphasize that any studies drawing inferences from mitochondrial data must carefully account for the inherent heterogeneity in the composition of each gene, as ignoring this fact can introduce different phylogenetic artifacts (Hassanin, 2006).

In this context, codon-position models offer a more robust framework for capturing the evolution of coding sequences in most multiple sequence alignments. By accounting for site heterogeneity and other evolutionary parameters, these models should provide greater accuracy and biological relevance compared to simpler alternatives. Consequently, substitution model testing can be a fundamental approach on this regard, and our results once again reveal notable patterns. We observed substantial differences in the evolutionary models best suited for each mitochondrial gene, including complex models that consider heterogeneity both in rates and nucleotide frequencies (Table S2), reflecting the inherent heterogeneity in substitution rates across the mitogenome. While the limitations of using overly simplistic evolutionary models may vary depending on the dataset, simpler models might invariably misestimate different evolutionary parameters by failing to account for the occurrence of multiple substitutions at the same site (Yang & Nielsen, 2000; Anisimova & Kosiol, 2009; Duchêne et al., 2014)—which can also lead to errors in phylogenetic inferences (Buckley et al., 2001; Su et al., 2014).

Although the mitogenome evolves as a single non-recombining unit, and typically exhibits a largely consistent phylogenetic signal across genes, our results also align with established evidence that evolutionary pressures act differentially on individual mitochondrial genes (Saccone et al., 1999; Xu et al., 2006). While substitution rates showed broad similarity across the mitogenome, some degree of heterogeneity was observed among specific genes. In the calibrated analyses, genes such as ND2 and ND6 had substitution rates above 0.02 substitutions/site/MY, while others like COX1 and COX2 generally ranged between 0.015 and 0.017. These findings highlight the importance of using partitioned analyses that account for both site- and gene-specific rate variation, along with appropriately selected substitution models, given the observed heterogeneity across loci and codon positions (Table S2). Such approaches are critical for improving the precision of evolutionary inferences, including divergence time estimation and substitution rate calibration.

Substitution rates heterogeneity depending on the presence of calibration points

Our study also evaluated the influence of temporal calibrations on Pleurodontan mitochondrial substitution rate estimates. Analyses incorporating fossil calibrations yielded estimates slightly lower than those from uncalibrated analyses (Table 3), consistent with evidence that well-constrained calibrations reduce biases in molecular dating (Hipsley & Muller, 2014; Warnock et al., 2015). This underscores the importance of integrating multiple fossil calibrations, particularly at deep nodes, to improve the accuracy of divergence time inferences—a critical consideration for groups like Pleurodonta, that exhibit complex biogeographic histories and potential rate heterogeneity across subclades (Blankers et al., 2013; Alencar et al., 2024). Nonetheless, uncalibrated estimates did not depict large standard deviations, highlighting that, at least in the Bayesian framework we implemented, node calibrations were not as important as the used priors in the estimate's variation.

Difference between coding and non-coding regions

Mitogenomes are often established as superior to single genes-based approaches for divergence time estimation, as the latter typically overestimate node ages (e.g., Duchêne et al., 2011). In Pleurodontan squamates, our analyses revealed minimal substitution rate variation across mitochondrial coding regions, implying that, except for the rRNAs, practically any chosen gene subset may effectively capture their genome-wide evolutionary rate patterns. This finding offers practical advantages for research on this squamate group, where targeted sequencing of subsets could reduce costs and labor while preserving phylogenetic signal. Selecting loci with intermediate substitution rates and robust phylogenetic resolution might be an important strategy for future research design. For instance, the use of genes such as 12S and 16S rRNAs should be critically pondered, as they present considerably lower substitution rates (Table 3), likely driven by functional constraints on ribosome assembly and saturation in conserved domains (Mueller, 2006; Duchêne et al., 2011). Additionally, loci exhibiting reduced phylogenetic informativeness (whether due to limited variability, homoplasy, or alignment ambiguity) require rigorous evaluation to avoid compromising analytical resolution (Zardoya and Meyer, 1996; Non et al., 2007). Either way, delineating such gene subsets demand taxon-specific substitution models and rigorous calibration to minimize biases, underscoring the need for tailored analytical frameworks, which can now be achieved by using our provided estimates.

Pleurodontan evolutionary dynamics and future perspectives

The substitution rate estimates from this study (nearly 0.01–0.02 substitutions/site/MY) align with prior estimates reported for Pleurodontan lineages (Supporting Table S1). However, we note that many of these earlier values were extrapolated from studies of distantly related taxa rather than empirically derived from lineage-specific calibrations. This reinforces the reliability of our methodological framework, which incorporated different partitioning schemes, appropriate substitution models, and string prior calibration strategies. Furthermore, our chosen priors, which were informed by values for different taxa within the Pleurodontan clade already reported on the literature (Table 2), proved to be robust. The close agreement of our substitution rate estimates with those previously reported for Pleurodonta also highlights the relative stability of mitochondrial evolutionary rates within the group. Prior research has demonstrated that mitochondrial substitution rates tend to cluster within narrow ranges among closely related taxa, often reflecting shared evolutionary constraints (Päckert et al., 2007; Pons et al., 2010). On the other hand, while mitochondrial protein-coding genes show conserved rate variation patterns across vertebrates—a phenomenon stable for ~450 million years (Broughton & Reneau, 2006)—the drivers of this variation remain poorly understood. This gap highlights an opportunity to explore how structural, functional, and selective pressures differentially shape mitochondrial gene evolution.

Furthermore, by providing robust substitution-rate estimates for Pleurodontans as a whole, our study offers a valuable resource for future molecular dating analyses. In any case, lineage-specific rates for individual clades or species within the group might be warranted, depending on the study design. Such an approach could offer valuable insights into the evolutionary dynamics of particular species, particularly when ecological, physiological, or demographic factors influence mitochondrial evolution (e.g., Welch et al., 2008; Nabholz et al., 2016; Jing et al., 2024). Estimating lineage-specific rates could thus help identify these patterns and refine our understanding of the drivers of molecular evolution within Pleurodonta. However, the limited availability of complete mitochondrial genomes for several Pleurodontan lineages still hinders a full understanding of their evolutionary history from being achieved.

In this context, it is important to recognize that multiple methods exist for estimating substitution rates beyond the approach used here. For instance, germline-based estimates (e.g., Bergeron et al., 2023) are particularly relevant for assessing average nuclear genomic variation, a task that has only recently become feasible with advances in genomic sequencing and bioinformatics. However, obtaining such estimates is challenging, as it requires genomic data from multiple generations. Additionally, these methods remain taxonomically limited (Chintalapati & Moorjani, 2020; Bergeron et al., 2023), posing a major challenge in groups like Pleurodonta, where evolutionary parameters remain largely unknown for most species. Branch-specific substitution rate estimates, such as those generated using PAML (Yang, 2007), offer a robust alternative but are influenced by several factors that may affect their reliability for specific research objectives (e.g., sequence quality, alignment accuracy, and model assumptions) (Rasmussen & Kellis, 2007; Yan et al., 2023).

Lineage-specific rate estimation approaches can be computationally demanding, as it requires constructing tailored substitution models that account for codon position variation, partition-specific evolutionary dynamics, and rate heterogeneity across the mitochondrial genome, as demonstrated here. Also, the accuracy of such specific estimates depends on the availability of high-quality sequence data and well-supported calibration points, both of which remain limited for many species, including those within Pleurodonta. This gains further importance as incomplete or biased sampling and poorly chosen calibrations can introduce substantial uncertainty into divergence time estimates (Zheng & Wiens, 2015; Schenk, 2016). Therefore, while lineage-specific rate estimation has the potential to refine our understanding of evolutionary rates, it must be applied cautiously, weighing the benefits of increased resolution against computational and methodological challenges.

Conclusion

In this study we examined the phylogenetic utility of nearly complete mitogenomes regarding the estimation of substitution rates, offering critical insights into the application of mitochondrial data in evolutionary studies. Despite the study’s focus on a specific taxonomic scope (the Pleurodonta clade), the framework applied here may be broadly applicable across different taxa and divergence times.

Our findings reveal relatively homogeneity in substitution rates across Pleurodontan mitochondrial protein-coding genes, but heterogeneity between these and non-coding regions. Also, there is a considerable amount of difference in substitution rates when accounting for codon positions. Although this heterogeneity is relatively localized, employing rate estimates specific to the genes or genomic regions under study clearly enhance the accuracy of evolutionary inferences. Future research will be essential to determine whether this heterogeneity arises primarily from conserved replication mechanisms that drive variation in mutation rates across genomic regions, the effects of natural selection on individual genes, a combination of these factors, or other evolutionary processes.

Furthermore, evaluating the best modelling and partitioning schemes when conducting evolutionary analyses constitute a key factor and must not be overlooked when using mitochondrial markers. While subsets of informative genes might approximate these results, their effectiveness depends on robust methodological frameworks and careful taxon-specific selection. On this regard, our results provide a valuable reference for future investigations into evolutionary dynamics specifically within the Pleurodonta clade and its closely related lineages, offering a foundation for comparative studies across Squamata. We then hope that our findings establish a foundation for optimizing mitochondrial phylogenetics in squamates, facilitating more accurate evolutionary reconstructions across diverse taxa and timescales.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data underlying this article, including phylogenetic datasets, corresponding trees, input and output files for all analyses, and any other relevant supplementary files are available in Zenodo, at https://doi.org/10.5281/zenodo.15952175.

Acknowledgements

We thank all reviewers for critically reviewing the manuscript and providing suggestions. This work was supported by the Brazilian federal institution “Coordenação de

Aperfeiçoamento de Pessoal de Nível Superior" (CAPES), through a PhD scholarship to MMAS (Finance Code 001).

REFERENCES

ALENCAR L. R., et al. Opportunity begets opportunity to drive macroevolutionary dynamics of a diverse lizard radiation. **Evolution Letters** 8: 623-637, 2024.

ANISIMOVA M., KOSIOL C. Investigating protein-coding sequence evolution with probabilistic codon substitution models. **Molecular Biology and Evolution** 26: 255-271, 2009.

ARCONES A., et al. Mitochondrial substitution rates estimation for divergence time analyses in modern birds based on full mitochondrial genomes. **Ibis** 163: 1463-1471, 2021.

AVILA L. J., MARTÍNEZ L. E., MORANDO M. Checklist of lizards and amphisbaenians of Argentina: an update, 2013.

AVISE J. C., et al. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. **Annu Rev Ecol Syst** 18: 489–522, 1987.

BALLARD J. W. O., WHITLOCK M. C. The incomplete history of mitochondria. **Molecular Ecology** 13: 729–744, 2004.

BALLARD J. W. O., RAND D. M. The population biology of mitochondrial DNA and its phylogenetic implications. **Annu. Rev. Ecol. Evol. Syst.** 36: 621-642, 2005.

BERGERON L. A., et al. Evolution of the germline mutation rate across vertebrates. **Nature** 615: 285-291, 2023.

BERNARDO P. H., et al. Extreme mito-nuclear discordance in a peninsular lizard: the role of drift, selection, and climate. **Heredity** 123: 359-370, 2019.

BININDA-EMONDS O. R. P. Fast genes and slow clades: comparative rates of molecular evolution in mammals. **Evolutionary Bioinformatics** 3: 59-85, 2007.

BLANKERS T., et al. Contrasting global-scale evolutionary radiations: phylogeny, diversification, and morphological evolution in the major clades of iguanian lizards. **Biological Journal of the Linnean Society** 108: 127-143, 2013.

BOFKIN L., GOLDMAN N. Variation in evolutionary processes at different codon positions. **Molecular Biology and Evolution** 24: 513-521, 2007.

BOROWIEC M. L. AMAS: a fast tool for alignment manipulation and computing of summary statistics. **PeerJ** 4: e1660, 2016.

BOUCKAERT R. R., DRUMMOND A. J. bModelTest: Bayesian phylogenetic site model averaging and model comparison. **BMC evolutionary biology** 17: 1-11, 2017.

BROUGHTON R. E., RENEAU P. C. Spatial covariation of mutation and nonsynonymous substitution rates in vertebrate mitochondrial genomes. **Molecular Biology and Evolution** 23: 1516-1524, 2006.

BUCKLEY T. R., SIMON C., CHAMBERS G. K. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. **Systematic Biology** 50: 67-86, 2001.

BUCKLEY T. R. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. **Systematic Biology** 51: 509-523, 2002.

CAMURUGI F., et al. Isolation by distance and past climate resistance shaped the distribution of genealogical lineages of a neotropical lizard. **Systematics and Biodiversity** 20, 2022. https://doi.org/10.1080/14772000.2022.2084470.

CARVALHO A. L. G., et al. Biogeography of the Lizard Genus Tropidurus Wied-Neuwied, 1825 (Squamata: Tropiduridae): Distribution, Endemism, and Area Relationships in South America. **PloS One**: 8, 2013.

CARVALHO A. L. G., et al. A highly polymorphic South American collared lizard (Tropiduridae: Tropidurus) reveals that open-dry refugia from South-western Amazonia staged allopatric speciation. **Zoological Journal of the Linnean Society** 201: 493–533, 2024. https://doi.org/10.1093/zoolinnean/zlad138.

CHAN L. M., et al. Defining spatial and temporal patterns of phylogeographic structure in Madagascar's iguanid lizards (genus Oplurus). **Molecular Ecology** 21: 3839-3851, 2012.

CHINTALAPATI M., MOORJANI P. Evolution of the mutation rate across primates. **Current opinion in genetics & development** 62: 58-64, 2020.

CONRAD J. L., & NORELL M. A. A complete Late Cretaceous iguanian (Squamata, Reptilia) from the Gobi and identification of a new iguanian clade. **American Museum Novitates**, 2007(3584), 1-47, 2007.

DRUMMOND A. J., RAMBAUT A. BEAST: Bayesian evolutionary analysis by sampling trees. **BMC Evolutionary Biology** 7: 1-8, 2007.

DUCHÊNE S., et al. Mitogenome phylogenetics: The impact of using single regions and partitioning schemes on topology, substitution rate and divergence time estimation. **PloS ONE** 6, 2011. https://doi.org/10.1371/journal.pone.0027138.

DUCHÊNE S., et al. The impact of calibration and clock-model choice on molecular estimates of divergence times. **Molecular Phylogenetics and Evolution** 78: 277-289, 2014.

EO S. H., DEWOODY J. A. Evolutionary rates of mitochondrial genomes correspond to diversification rates and to contemporary species richness in birds and reptiles. **Proceedings of the Royal Society B: Biological Sciences** 277: 3587-3592, 2010.

FENG S., et al. Dense Sampling of Bird Diversity Increases Power of Comparative Genomics. **Nature** 587: 252–257, 2020.

FONTANELLA F. M., et al. Molecular dating and diversification of the South American lizard genus Liolaemus (subgenus Eulaemus) based on nuclear and mitochondrial DNA sequences. **Zoological Journal of the Linnean Society** 164: 825-835, 2012.

FUNK D. J., OMLAND K. E. Species-Level Paraphyly and Polyphyly: Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. **Annual Review of Ecology, Evolution, and Systematics** 34: 397–423, 2003. Doi:10.1146/annurev.ecolsys.34.01

GABLE S. M., et al. The State of Squamate Genomics: Past, Present, and Future of Genome Research in the Most Speciose Terrestrial Vertebrate Order. In **Genes** (Vol. 14, Issue 7). Multidisciplinary Digital Publishing Institute (MDPI), 2023. https://doi.org/10.3390/genes14071387

GENEREUX D. P., et al. A Comparative Genomics Multitool for Scientific Discovery and Conservation. **Nature** 587: 240–245, 2020.

GRAY M. W., BURGER G., LANG B. F. Mitochondrial evolution. **Science** 283: 1476-1481, 1999.

HASSANIN A. Phylogeny of Arthropoda inferred from mitochondrial sequences: strategies for limiting the misleading effects of multiple changes in pattern and rates of substitution. **Molecular Phylogenetics and Evolution** 38: 100-116, 2006.

HEBERT P. D., et al. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. **Proceedings of the Royal Society of London. Series B: Biological Sciences** 270: S96-S99, 2003.

HIPSLEY C. A., MÜLLER J. Beyond fossil calibrations: realities of molecular clock practices in evolutionary biology. **Frontiers in genetics** 5: 138, 2014.

HO S. Y. Calibrating molecular estimates of substitution rates and divergence times in birds. **Journal of Avian Biology** 38: 409-414, 2007.

JING Y., et al. Influence of life-history traits on mitochondrial DNA substitution rates exceeds that of metabolic rates in teleost fishes. **Current Zoology**: zoae045, 2024.

JOHNS G. C., AVISE J. C. A comparative summary of genetic distances in the vertebrates from the mitochondrial cytochrome b gene. **Molecular Biology and Evolution** 15: 1481-1490, 1998.

KATOH K., STANDLEY D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. **Molecular Biology and Evolution** 30: 772-780, 2013.

KIMURA M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. **Journal of molecular evolution** 16: 111-120, 1980.

LEACHÉ A. D., & LINKEM C. W. Phylogenomics of horned lizards (Genus: Phrynosoma) using targeted sequence capture data. **Copeia** 103: 586-594, 2015.

LEMMON A. R., BROWN J. M., STANGER-HALL K., LEMMON E. M. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. **Systematic Biology** 58: 130-145, 2009.

LOSOS J. B. Lizards in an evolutionary tree: ecology and adaptive radiation of anoles (Vol. 10). Univ of California Press, 2011.

MACKIEWICZ P., et al. Phylogeny and evolution of the genus Cervus (Cervidae, Mammalia) as revealed by complete mitochondrial genomes. **Scientific Reports** 12: 16381, 2022.

MELLO B., SCHRAGO C. G. Assignment of calibration information to deeper phylogenetic nodes is more effective in obtaining precise and accurate divergence time estimates. **Evolutionary Bioinformatics** 10: EBO-S13908, 2014.

MUELLER R. L. Evolutionary rates, divergence dates, and the performance of mitochondrial genes in Bayesian phylogenetic analysis. **Systematic biology** 55: 289-300, 2006.

MOWER J. P., et al. Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants. **BMC Evolutionary Biology** 7, 2007. https://doi.org/10.1186/1471-2148-7-135

NABHOLZ B., et al. Strong variations of mitochondrial mutation rate across mammals- the longevity hypothesis. **Molecular Biology and Evolution** 25: 120-130, 2008. 10.1093/molbev/msm248.

NABHOLZ B., et al. Body mass-corrected molecular rate for bird mitochondrial DNA. **Molecular ecology** 25: 4438-4449, 2016.

NON A. L., et al. Identification of the most informative regions of the mitochondrial genome for phylogenetic and coalescent analyses. **Molecular Phylogenetics and Evolution** 44: 1164-1171, 2007.

OLAVE M., et al. Model-based approach to test hard polytomies in the Eulaemus clade of the most diverse South American lizard genus Liolaemus (Liolaemini, Squamata). **Zoological Journal of the Linnean Society** 174: 169-184, 2015.

PÄCKERT M., et al. Calibration of a molecular clock in tits (Paridae)—Do nucleotide substitution rates of mitochondrial genes deviate from the 2% rule? **Molecular Phylogenetics and Evolution** 44: 1-14, 2007.

PARKINSON C. L., et al. Multiple major increases and decreases in mitochondrial substitution rates in the plant family Geraniaceae. **BMC Evolutionary Biology** 5: 1–12, 2005. https://doi.org/10.1186/1471-2148-5-73

PATWARDHAN A., et al. Molecular markers in phylogenetic studies-a review. **Journal of Phylogenetics & Evolutionary Biology** 2: 131, 2014.

PARK E., et al. Estimation of divergence times in cnidarian evolution based on mitochondrial protein-coding genes and the fossil record. **Molecular Phylogenetics and Evolution** 62: 329-345, 2012.

PIANKA E. P., VITT L. J. Lizards: windows to the evolution of diversity (Vol. 5). Univ of California Press, 2003.

PLAZZI F., PUCCIO G., PASSAMONTI M. Comparative large-scale mitogenomics evidences clade specific evolutionary trends in mitochondrial DNAs of Bivalvia. **Genome Biology and Evolution** 8: 2544-2564, 2016.

PONS J., et al. Nucleotide substitution rates for the full set of mitochondrial protein-coding genes in Coleoptera. **Molecular Phylogenetics and Evolution** 56: 796–807, 2010. https://doi.org/10.1016/j.ympev.2010.02.007.

PORTELLI S. N., et al. Historical biogeographic reconstruction of the South American Liolaemus boulengeri group (Iguania: Liolaemidae). **South American Journal of Herpetology** 25: 41-56, 2022.

PRATES I., et al. Phylogenetic relationships of Amazonian anole lizards (Dactyloa): taxonomic implications, new insights about phenotypic evolution and the timing of diversification. **Molecular phylogenetics and evolution** 82: 258-268, 2015.

RAMBAUT A., DRUMMOND A. J., SUCHARD M. Tracer v1. 6 http://beast.bio.ed.ac.uk, 2007.

RAMBAUT A., DRUMMOND A. J. LogCombiner v2. 1.3. Institute of Evolutionary Biology,University of Edinburgh, 2014.

RASMUSSEN M. D., KELLIS M. Accurate gene-tree reconstruction by learning gene-and species specific substitution rates across multiple complete genomes. **Genome research** 17: 1932-1942, 2007.

RITCHIE A. M., et al. The impact of the tree prior on molecular dating of data sets containing a mixture of inter-and intraspecies sampling. **Systematic Biology** 66: 413-425, 2017.

ROE A. D., SPERLING F. A. Patterns of evolution of mitochondrial cytochrome c oxidase I and II DNA and implications for DNA barcoding. **Molecular Phylogenetics and evolution** 44: 325-345, 2007.

ROGERS T. F., et al. Using recent genetic history to inform conservation options of two Lesser Caymans iguana (Cyclura 107otali caymanensis) populations. **Conservation Genetics** 25: 711-724, 2024.

ROMÁN-PALACIOS C., et al. When did anoles diverge? An analysis of multiple dating strategies. **Molecular Phylogenetics and Evolution** 127: 655-668, 2018.

RUBINOFF D., HOLLAND B. Between Two Extremes: Mitochondrial DNA is neither the Panacea nor the Nemesis of Phylogenetic and Taxonomic Inference. **Systematic Biology** 54: 952–961, 2005. Doi:10.1080/10635150500234674

SACCONE C., et al. Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. **Gene** 238: 195-209, 1999.

SALLES, M. M. A., et al. Ancient introgression explains mitochondrial genome capture and mitonuclear discordance among South American collared Tropidurus lizards. **bioRxiv**, 2025-04, 2025. https://doi.org/10.1101/2025.04.25.650633.

SCARPETTA S. G. The first known fossil Uma: ecological evolution and the origins of North American fringe-toed lizards. **BMC Evolutionary Biology** 19: 178, 2019.

SCHENK J. J. Consequences of secondary calibrations on divergence time estimates. **PloS one** 11: e0148228, 2016.

SHERRATT E., et al. Amber fossils demonstrate deep-time stability of Caribbean lizard communities. **Proceedings of the National Academy of Sciences** 112: 9961-9966, 2015.

SIMÕES T. R., PYRON R. A. The squamate tree of life. **Bulletin of the Museum of Comparative Zoology** 163: 47-95, 2021.

SLOAN D. B., OXELMAN B., RAUTENBERG A., TAYLOR D. R. Phylogenetic analysis of mitochondrial substitution rate variation in the angiosperm tribe Sileneae. **BMC Evolutionary Biology** 9, 2009. https://doi.org/10.1186/1471-2148-9-260

SMITH S. A., BROWN J. W., WALKER J. F. So many genes, so little time: a practical approach to divergence-time estimation in the genomic era. **PloS one** 13: e0197433, 2018.

SU Z., et al. The impact of incorporating molecular evolutionary model into predictions of phylogenetic signal and noise. **Frontiers in Ecology and Evolution** 2: 11, 2014.

TOWNSEND T. M., et al. Eastward from Africa: palaeocurrent-mediated chameleon dispersal to the Seychelles islands. **Biology Letters** 7: 225-228, 2011.

UETZ P., et al. A quarter century of reptile and amphibian databases. **Herpetol. Rev.** 52: 246-255, 2021.

UETZ P., FREED P., AGUILAR R., REYES F., KUDERA J., HOŠEK J. (eds.). The Reptile Database, http://www.reptile-database.org, accessed January 31, 2025.

WARNOCK R. C., et al. Calibration uncertainty in molecular dating analyses: there is no substitute for the prior evaluation of time priors. **Proceedings of the Royal Society B: Biological Sciences** 282: 20141013, 2015.

WELCH J. J., BININDA-EMONDS O. R., BROMHAM L. Correlates of substitution rate variation in mammalian protein-coding sequences. **BMC Evolutionary Biology** 8: 1-12, 2008.

WERNECK F. P., et al. Biogeographic history and cryptic diversity of saxicolous Tropiduridae lizards endemic to the semiarid Caatinga. **BMC Evolutionary Biology** 15: 1–24, 2015. https://doi.org/10.1186/s12862-015-0368-3

WILLIAMS E. J. B., HURST L. D. Is the synonymous substitution rate in mammals gene-specific? **Molecular Biology and Evolution** 19: 1395-1398, 2002.

XU W., et al. The relationship between the rate of molecular evolution and the rate of genome rearrangement in animal mitochondrial genomes. **Journal of molecular evolution** 63: 375-392, 2006.

YAN H., HU Z., et al. PhyloAcc-GT: A Bayesian method for inferring patterns of substitution rate shifts on targeted lineages accounting for gene tree discordance. **Molecular Biology and Evolution** 40: msad195, 2023.

YAN L., XU W., ZHANG D., LI J. Comparative analysis of the mitochondrial genomes of flesh flies and their evolutionary implication. **International Journal of Biological Macromolecules** 174: 385–391, 2021. https://doi.org/10.1016/j.ijbiomac.2021.01.188

YANG Z., NIELSEN, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. **Molecular Biology and Evolution** 17: 32-43, 2000.

YANG Z. PAML 4: phylogenetic analysis by maximum likelihood. **Molecular biology and evolution** 24: 1586-1591, 2007.

YANG H., LI T., DANG K., BU W. Compositional and mutational rate heterogeneity in mitochondrial genomes and its effect on the phylogenetic inferences of Cimicomorpha (Hemiptera: Heteroptera). **BMC Genomics** 19, 2018. https://doi.org/10.1186/s12864-018-4650-9

ZARDOYA R., MEYER A. Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. **Molecular biology and evolution** 13: 933-942, 1996.

ZARZA E., REYNOSO V. H., EMERSON B. C. Diversification in the northern neotropics: mitochondrial and nuclear DNA phylogeography of the iguana Ctenosaura 109otaling109 and related species. **Molecular Ecology** 17: 3259-3275, 2008.

ZHENG Y., WIENS J. J. Do missing data influence the accuracy of divergence-time estimation with BEAST? **Molecular Phylogenetics and Evolution** 85: 41-49, 2015.

ZHENG Y., WIENS J. J. Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. **Molecular phylogenetics and evolution** 94: 537-547, 2016.

ZHU A., GUO W., JAIN K., MOWER J. P. Unprecedented heterogeneity in the synonymous substitution rate within a plant genome. **Molecular Biology and Evolution** 31: 1228–1236, 2014. https://doi.org/10.1093/molbev/msu079

Supporting information

Table S1. References regarding values for substitution rate priors adopted in the present study.

Gene	Values	References
COX1	0.01	Phrynosomatidae: Bernardo et al., 2019; *Tropidurus:* Camurugi et al., 2022
CYTB	0.019355 | 0.0113 | 0.0223	*Cyclura*: Rogers et al., 2024; Liolaemidae*:* Olave et al., 2015; *Liolaemus:* Fontanella et al., 2012; *Tropidurus:* Werneck et al., 2015;
ND1	0.013876	*Oplurus:* Chan et al., 2012
ND2	0.013	*Anolis:* Román-Palacios et al., 2018
ND4	0.0113 | 0.0078	*Cyclura*: Rogers et al., 2024; Iguaninae: Zarza et al., 2008
12s & 16s	0.006339 | 0.00576	Liolaemidae*:* Olave et al., 2015; *Liolaemus:* Fontanella et al., 2012; *Tropidurus:* Carvalho et al., 2024;

Table S2. Model selection results from BEAST, including only models within the 95% HPD interval for each gene. The four-digit model code represents how substitution rates are grouped, following the order of relative rates for A-C, A-G, A-T, C-G, C-T, and G-T. A complete list of model codes can be found here: https://taming-the-beast.org/tutorials/Substitution-model-averaging/.

Gene	Non-calibrated	Calibrated
	Models	
12s	GTR	GTR
16s	GTR	GTR
ATP6	HKY, GTR, K81$_{123324}$, TIM$_{123345}$, TVM$_{123425}$	GTR, K80, K81$_{123324}$, TIM$_{123345}$, TN93$_{121131}$, TIM$_{123345}$
ATP8	K80, K81$_{123321}$, K81$_{123324}$, TIM$_{123341}$, TIM$_{123345}$, TN93$_{121131}$	K80, K81$_{3321}$, K81$_{3324}$, TIM$_{3341}$, TIM$_{3345}$, TN93$_{1131}$
COX1	TIM$_{3341}$, TIM$_{3345}$, TN93$_{1131}$	TIM$_{3341}$, TIM$_{3345}$, TN93$_{1131}$
COX2	K80, K81$_{3321}$, K81$_{3324}$, TIM$_{3341}$, TN93$_{1131}$	K80, K81$_{3321}$, K81$_{3324}$, TIM$_{3341}$, TN93$_{1131}$
COX3	K80, K81$_{3321}$, K81$_{3324}$, TIM$_{3345}$, TN93$_{1131}$	K80, K81$_{3321}$, K81$_{3324}$, TIM$_{3341}$, TN93$_{1131}$
CYTB	GTR, TVM$_{3425}$	GTR, K81$_{3324}$, TVM$_{3425}$
ND1	GTR, K81$_{3324}$, TIM$_{3345}$, TVM$_{3425}$	GTR, K81$_{3324}$, TIM$_{3345}$, TVM$_{3425}$
ND2	K81$_{3324}$, TIM$_{3345}$, TVM$_{3425}$	K81$_{3324}$, TIM$_{3345}$, TIM$_{3345}$
ND3	K80, K81$_{3321}$, TIM$_{3341}$, TIM$_{3345}$	K80, K81$_{3321}$, TN93$_{1131}$, K81$_{3324}$, TIM$_{3341}$
ND4	GTR, TIM$_{3345}$, TN93$_{1131}$	GTR, TIM$_{3345}$
ND4L	GTR, K80, K81$_{3321}$, K81$_{3324}$, TIM$_{3341}$, TIM$_{3345}$	K80, K81$_{3321}$, K81$_{3324}$, TIM$_{3341}$, TIM$_{3345}$, TVM$_{3425}$
ND5	K81$_{3324}$, TIM$_{3345}$, TVM$_{3425}$	K81$_{3324}$, TIM$_{3345}$, TVM$_{3425}$
ND6	GTR, TVM$_{3425}$	GTR, TVM$_{3425}$

4. CAPÍTULO IV: ANCIENT INTROGRESSION EXPLAINS MITOCHONDRIAL GENOME CAPTURE AND MITONUCLEAR DISCORDANCE AMONG SOUTH AMERICAN COLLARED *Tropidurus* LIZARDS

Esta seção apresenta um dos artigos desenvolvidos ao longo do doutorado, já publicado. A formatação segue as normas da revista que o artigo foi aceito.

REFERÊNCIA: Salles, M. M. A., Carvalho, A. L. G., Leache, A. D., Martinez, N., Bauer, F., Motte, M., Espínola, V., Rodrigues, M. T., Piantoni, C., Colli, G. R., Werneck, F., Pie, M., Olivotto, A., Choueri, E., & Domingos, F. M. C. B. (2025). Ancient introgression explains mitochondrial genome capture and mitonuclear discordance among South American collared Tropidurus lizards. *Molecular Ecology*, 2025, doi.org/10.1111/mec.70130

Data availability

The data underlying this article, including phylogenetic datasets, corresponding trees, input and output files for all analyses, and any other relevant supplementary files, are available in Zenodo, at https://doi.org/10.5281/zenodo.15376373.

Author contributions

Matheus Salles: study design, analyses, lead writing. Fabricius Domingos and André Carvalho: study design, field sampling, genome sequencing, writing final draft. Nicolas Martinez, Frederick Bauer, Martha Motte, Viviana Espínola, Miguel T. Rodrigues, Carla Piantoni, Guarino R. Colli, Fernanda P. Werneck: field sampling efforts and/or provision of collection materials and infrastructure. Márcio Pie, André Olivotto, Erik Choueri, Adam Leache: lab work, genome sequencing. All authors read, commented, and approved the final draft of the manuscript.

Acknowledgments

We thank all the institutions that provided technical, material, and structural support during the sampling phase of this study, namely the Herpetological Collection of the University of Brasília (CHUNB), the Amphibians and Reptiles Collection of the National Institute of Amazonian Research (INPA-H), the Biosciences Institute of the University of São Paulo (IB-USP), the American Museum of Natural History (AMNH), the Museo Nacional de Historia Natural del Paraguay (MNHNP), the Mato Grosso Federal University (UFMT), Museo de Historia Natural Alcide d'Orbigny, and the Universidade Nacional de Salta (UNSa). We extend our gratitude to our esteemed colleagues Eliana Lizarraga, Felipe F. Curcio (UFMT), Karina Atkinson and Jean-Paul Brouard (Fundación Para La Tierra, Paraguay), Johanna López, Nestor Romero (and family) (Estancia Cerrados del Tagatiya), Pastor Enmanuel Pérez-Estigarribia (and family), Ricardo

Céspedes, Rodrigo Ayala, and Rolando Rivas for crucial assistance or support during fieldwork. We thank Angela Coda for allowing access to the Reserva Natural Cerrados del Tagatiya and Malvina Duarte for allowing access to the Reserva Natural Laguna Blanca. We also thank Junior Nadaline and Carlos Daniel Rivadeneira for their invaluable assistance with genetic sample processing and DNA sequencing. MTR also thanks all members of his lab for their help in the field. Finally, we extend our gratitude to Prof. Renato Pires Machado (UFPR) for his valuable contributions throughout the corresponding author's PhD project.

Funding

This work was funded by the Brazilian National Council for Scientific and Technological Development (CNPq #200798/2010-3), the São Paulo Research Foundation (FAPESP #2011/50146-6, #2016/08249-6, #2017/20235-3), the U.S. National Science Foundation [NSF #1855845], the Explorers Club, the Andrew Sabin Family Foundation, the American Museum of Natural History (Richard Gilder Graduate School), and the Brazilian funding agency 'Coordenação de Aperfeiçoamento de Pessoal de Nível Superior' (CAPES), which provided a PhD scholarship to the correspondent author. FPW thanks CNPq for a productivity fellowship (#307695/2023-9). GRC thanks CAPES, CNPQ, and Fundação de Apoio à Pesquisa do Distrito Federal (FAPDF) for the continuous support. FMCBD thanks CNPq for a productivity fellowship.

Conflict of Interest

The authors declare no conflicts of interest.

Benefit-Sharing Statement

A research collaboration was developed with scientists from the countries providing genetic samples; all collaborators are included as co-authors, the research results have been shared with the broader scientific community, and the research addresses a priority concern—the evolutionary history of organisms being studied. More broadly, our group is committed to international scientific partnerships, as well as institutional capacity building.

ABSTRACT

Mitonuclear discordance—evolutionary discrepancies between mitochondrial and nuclear DNA phylogenies—can arise from various factors, including introgression, incomplete lineage sorting, recent or ancient demographic fluctuations, sex-biased dispersal asymmetries, among others. Understanding this phenomenon is crucial for accurately reconstructing evolutionary histories, as failing to account for discordance can lead to misinterpretations of species boundaries, phylogenetic relationships, and historical biogeographic patterns. We investigate the evolutionary drivers of mitonuclear discordance in the *Tropidurus spinulosus* species group, which contains nine species of lizards inhabiting open tropical and subtropical environments in South America. Using a combination of population genetic and phylogenomic approaches applied to mitochondrial and nuclear data, we identified different instances of gene flow that occurred in ancestral lineages of extant species. Our results point to a complex evolutionary history marked by prolonged isolation between species, demographic fluctuations, and potential episodes of secondary contact with genetic admixture. These conditions likely facilitated mitochondrial genome capture while diluting signals of nuclear introgression. Furthermore, we found no strong evidence supporting incomplete lineage sorting or natural selection as primary drivers of the observed mitonuclear discordance. Therefore, the unveiled patterns are most consistent with neutral demographic processes, coupled with ancient mitochondrial introgression, as the main factors underlying the mismatch between nuclear and mitochondrial phylogenies in this system. Future research could further explore the role of other demographic processes, such as asymmetric sex-biased dispersal, in shaping these complex evolutionary patterns.

Key words: Introgression; Mitogenome; Molecular Evolution; Phylogenomics; Squamata.

Introduction

Mitochondrial and nuclear genomes evolve independently, through distinct processes, often leading to conflicting genetic patterns—a phenomenon known as mitonuclear discordance. Cases of mitonuclear discordance are increasingly reported across a wide range of animal species. Recent examples can be found among mammals (Good et al. 2015; Phuong et al. 2016; Seixas et al. 2018; Fedorov et al. 2022), birds (Andersen et al. 2021; DeRaad et al. 2023), reptiles (Firneno et al. 2021), amphibians (Zieliński et al. 2013; Dufresnes et al. 2020; Rancilhac et al. 2021), and insects (Hinojosa et al. 2019; Dong et al., 2023). Generally, conflicting patterns between nuDNA and mtDNA arise from processes such as introgression, demographic fluctuations, sex-biased dispersal asymmetries, incomplete lineage sorting (ILS), human introductions, or *Wolbachia* infection in insects (see reviews by Toews and Brelsford 2012; Bonnet et al. 2017; Deprés 2019). Determining whether mitonuclear discordance is driven by any of these factors requires a clear understanding of the genealogies among different loci, the demographic history of populations, and their historical introgression/migration patterns and rates. Notably, these complexities often present significant empirical challenges in resolving the underlying causes of discordance.

Most cases of mitonuclear discordance are attributed to mtDNA introgression (Toews and Brelsford 2012), although the extent of introgression varies widely among animal groups. For instance, Zozaya et al. (2024) showed that, despite frequent hybrid formation across multiple contact zones in Australian *Heteronotia* geckos, mtDNA exchange can be limited to narrow geographic areas, indicating that hybridization does not necessarily result in widespread mitochondrial replacement. When discordance is attributable to introgression, it is most often linked to processes such as gene flow across hybrid zones (Zieliński et al. 2013; Good et al. 2015; Dufresnes et al. 2020; Mao and Rossiter 2020) or mitochondrial capture (Andersen et al. 2021). The latter refers specifically to a form of complete introgression, where the mitochondrial genome of one species is entirely replaced by another's (Zieliński et al. 2013). Mitochondrial capture can occur at various stages of the speciation process and between distantly related species (Rancilhac et al. 2021). Additionally, mitochondrial introgression can take place with minimal or no accompanying nuDNA introgression (Zieliński et al. 2013; Good et al. 2015; Andersen et al. 2021; Rancilhac et al. 2021; Fedorov et al. 2022), making it challenging to detect introgressed nuclear and mitochondrial markers simultaneously (Bonnet et al. 2017; Seixas et al. 2018). Therefore, ILS is another phenomenon that cannot be overlooked in the discussion of mitonuclear discordance, given the challenges in distinguishing it from introgression (Andersen et al. 2021; DeRaad et al. 2023). While some studies have identified ILS as a primary driver of mitonuclear discordance, these remain relatively rare (e.g., Firneno et al. 2021).

Demographic factors, such as population size fluctuations, particularly at the periphery of a species' geographic range, can also contribute to mitonuclear discordance (Phuong et al. 2016; Dufresnes et al. 2020; Fedorov et al. 2022). In this scenario, neutral processes such as genetic drift can help in the fixation of mtDNA haplotypes at range edges of expanding populations (Bonnet et al. 2017). Even in situations where introgression involves both mtDNA and nuclear loci, mtDNA is expected to reach fixation faster, since it has a smaller effective population size and is non-recombining (Moore 1995). Following demographic expansions and secondary contact, nuclear genomes can fully recombine, thereby diluting historical introgression signals, whereas divergent, non-recombining mitochondrial haplotypes may persist.

Finally, selection can also play a role in generating mitonuclear discordance (Toews and Brelsford 2012). If a beneficial mitochondrial mutation arises but is initially incompatible with the nuclear genetic background of a population, selection may favor nuclear variants that restore mitonuclear compatibility. This compensatory selection could lead to a scenario where the advantageous mitochondrial mutation spreads widely, while the nuclear DNA follows a different evolutionary trajectory. Empirical evidence robustly demonstrates that natural selection can drive mtDNA introgression between populations through hybridization, even in the absence of corresponding nuDNA introgression, provided there is even a minor selective advantage (Excoffier et al. 2009; Zieliński et al. 2013; Good et al. 2015; Phuong et al. 2016). As a result, the mitochondrial genome with the highest fitness may introgress into another species, regardless of whether it originates from the resident or the invading population.

Here, we investigate the drivers of mitonuclear discordance in a biological system with known occurrences of this phenomenon, the lizard family Tropiduridae (e.g., *Microlophus*: Benavides et al. 2007; 2009; *Tropidurus*: Carvalho et al. 2016). Specifically, we explored the evolutionary history of nine lineages within the *T. spinulosus* group using a multi-locus genomic approach. This diverse group of lizards is distributed across open-vegetation tropical and subtropical environments in South America and predominantly inhabit rocky and arboreal habitats (Frost et al. 1998; Carvalho 2013, 2016; Carvalho et al. 2013). Although the taxonomy of the *T. spinulosus* group has been relatively stable, new species continue to be described (e.g., Carvalho 2016). Additionally, our research includes all taxa presently admitted to the group plus one undescribed species, which is currently undergoing formal description (Carvalho et al. in prep) and is referred to as *Tropidurus* sp. nov. in this study.

We integrated mitochondrial genomes, nuclear ultraconserved elements (UCEs), and single-nucleotide polymorphisms (SNPs) derived from UCEs to assess whether the mitochondrial genealogy corresponded to reciprocally monophyletic nuclear lineages. To determine the potential sources of the mitonuclear discordance, we examined the relative contributions of neutral and selective processes. Specifically, we first compared mtDNA and nuclear phylogenies to identify conflicts and potential introgression patterns, using methods that either account for ILS or do not. In a scenario where these approaches yield largely congruent results, ILS would not play a significant role in the observed mitonuclear discordance. Second, we evaluated whether mtDNA introgression was accompanied by detectable levels of nuDNA introgression, as most cases of discordance are attributed to processes such as gene flow across hybrid zones. We also assessed signals of positive selection on mtDNA to evaluate potential adaptive processes in shaping mitonuclear discordance. Finally, as demographic processes can contribute to emerging phylogenetic discordant patterns as well, we examined historical changes in population size to assess their role in shaping the observed discordance, as this would align with expectations under a non-adaptive scenario.

Methods

Sampling and Lab work

The phylogenomic approaches used in this study were based on anchored sequencing methods, specifically Ultraconserved Elements (UCE) as described by Faircloth et al. (2012). We sequenced a total of 43 individuals (Supplementary Material, Table S9), representing all species of the *Tropidurus spinulosus* group. RAPiD Genomics LLC (Gainesville, FL) was tasked with sample library preparation and target enrichment of UCEs using the tetrapod 5k probe set (Faircloth et al. 2012), followed by multiplexed paired-end (PE) sequencing (2 × 100 bp) of UCEs on an Illumina HiSeq 3000 PE100 platform. Our samples are mainly distributed around the Pantanal Basin, with some extending to dry environments of Paraguay and Argentina (Figure 1).

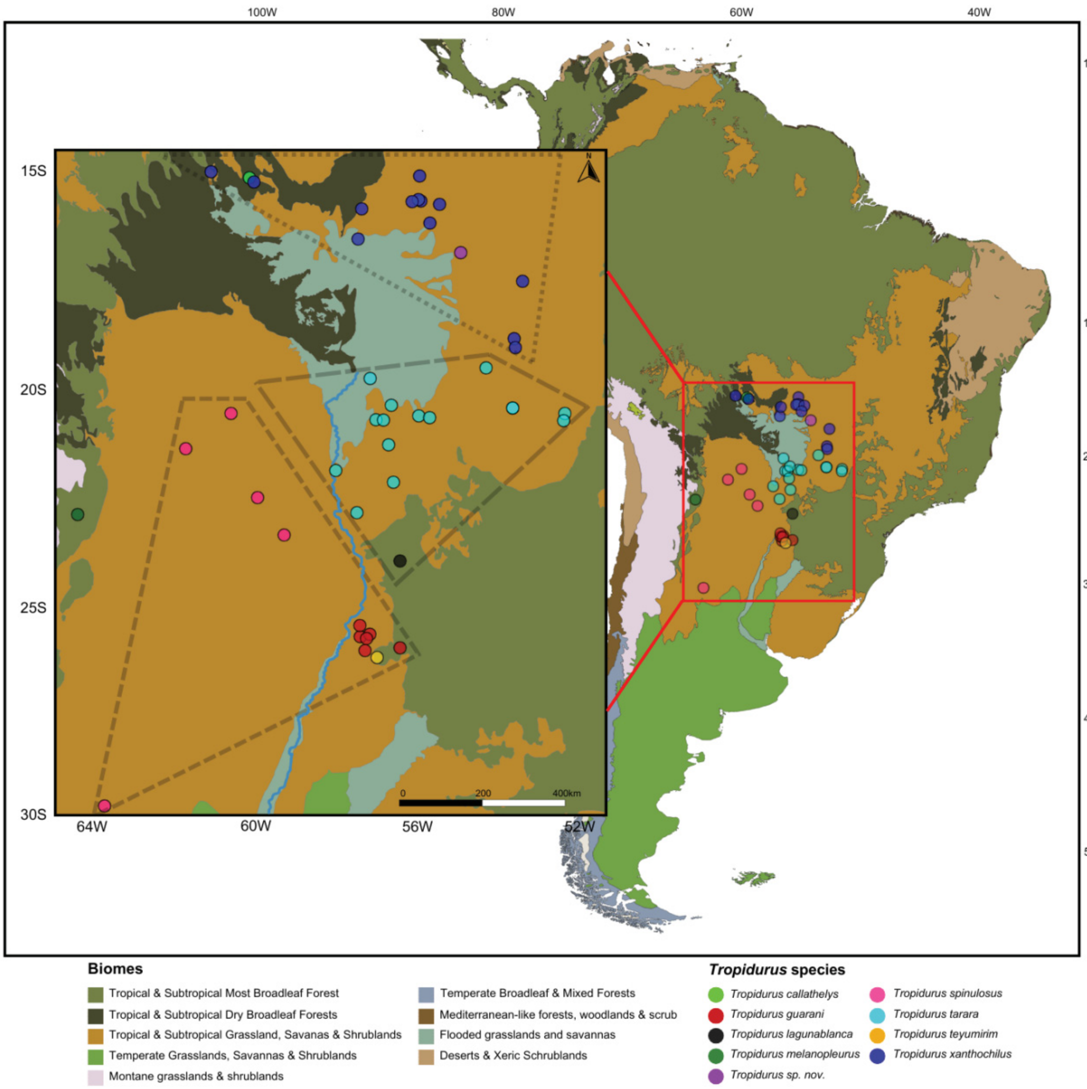

100W
80W
60W
40W
10N
0
10S
20S
30S
40S
50s
15S
20S
25S
30S
64W
60W
56W
52W
0
200
400km
N
Biomes
Tropical & Subtropical Most Broadleaf Forest
Tropical & Subtropical Dry Broadleaf Forests
Tropical & Subtropical Grassland, Savanas & Shrublands
Temperate Grasslands, Savannas & Shrublands
Montane grasslands & shrublands
Temperate Broadleaf & Mixed Forests
Mediterranean-like forests, woodlands & scrub
Flooded grasslands and savannas
Deserts & Xeric Schrublands
Tropidurus species
Tropidurus callathelys
Tropidurus guarani
Tropidurus lagunablanca
Tropidurus melanopleurus
Tropidurus sp. nov.
Tropidurus spinulosus
Tropidurus tarara
Tropidurus teyumirim
Tropidurus xanthochilus

Figure 1. Sample distribution with species assignments based on our concatenated phylogenetic analysis. The map at a larger scale shows the distribution of samples across various South American ecoregions, according to the classification of “Terrestrial Ecoregions of the World” adopted by the World Wildlife Fund”. For a detailed review of the resolution and delimitation differences in these categories, see Olson et al. (2001). The smaller-scale map highlights the specific region where the analyzed samples are distributed. Phylogenetically identified groups are marked with three distinct dashed line patterns, emphasizing their allopatric distribution: (1) the *T. spinulosus* + *T. guarani* + *T. teyumirim clade* (multi-species), (2) the *T. tarara clade* (with *T. lagunablanca* tentatively considered a synonym of *T. tarara*), and (3) the *T. xanthochilus* + *T. sp. nov.* clade. The map in the smaller box also shows part of the course of the Paraguay River (light blue), which separates the *T. spinulosus* lineages to the west from *T. guarani* and *T. tarara* to the east.

UCEs and mtDNA pipeline

We assembled reads using the Phyluce v1.7.1 pipeline (Faircloth 2016). Demultiplexed reads were cleaned to remove low-quality bases and adapter sequences in Trimmomatic (Bolger et al. 2014), using the wrapper program illumiprocessor in Phyluce. To accelerate assembly and render challenging datasets tractable, we normalized read depth to a minimum of 5 with the bbnorm.sh script from the BBTools suite (Bushnell 2018). Trimmed and normalized reads were inspected for quality and adapter contamination using FastQC v0.11.9 (Andrews 2010), and then assembled into contigs with SPAdes v3.15.5 (Bankevich et al. 2012). We used Phyluce to align assembled contigs back to their associated UCE loci, remove duplicate matches, create a taxon-specific database of contig-to-UCE matches, and extract UCE loci for all individuals. Specifically, contigs matching UCE loci were identified and extracted using the program LastZ v1.0 (Harris 2007) within Phyluce, both in incomplete and complete (75% and 95% of completeness) matrices (see Faircloth 2016). After probes and UCEs were matched, we aligned UCE contigs with MAFFT v7.471 (Katoh and Standley 2013) using specific customized settings (-globalpair, --maxiterate 1000, --adjustdirection), and trimmed the resulting alignments using the internal-trimming algorithm (Gblocks: Castresana 2000). As a final step, AMAS (Borowiec 2016) was used to compute final summary statistics for all alignments.

Due to the high abundance of mtDNA in samples and the less-than-perfect efficiency of target enrichment methods, mitochondrial data (including entire mitogenomes) are often generated as a byproduct of the UCE sequencing process (Allio et al. 2020). Thus, we extracted mtDNA from our contig assembly produced through Phyluce using MitoFinder (Allio et al. 2020), which can find and extract multiple mitochondrial genes or even entire mitogenomes from assembled sets of bulk sequences. In general terms, mitochondrial sequences were aligned following the same procedure applied to UCE data, as described above.

SNP calling workflow

We selected the sample with the largest number of UCE loci among all individuals ('MTR 29586', with 4800 loci) retrieved during the extraction step described above as the primary reference during the process of calling SNPs as recommended in the Zarza et. Al. (2016) pipeline to extract SNPs from UCE reads. Then, we followed the workflow developed by Harvey et al. (2016) to sequence capture data from population-level samples. The first step involves BWA, which we used to map raw reads from individuals to contigs (Li et al. 2009; Li 2013). Thus, SAM files were converted to BAM with SAMtools (Li et al. 2009). Alignments were checked for BAM format violations, read group header information was added, and PCR duplicates were marked for each individual using Picard (v. 1.106). The resulting BAM files for each individual in a lineage were merged into a single file with Picard, which was then indexed with SAMtools. The Genome Analysis Toolkit (GATK; v. 3.4-0; McKenna et al. 2010) was used to locate and realign around indels, which was followed by calling SNPs using the 'UnifiedGenotyper' tool in GATK. SNPs and indels were then annotated, and indels were masked. Finally, we used GATK to restrict our datasets to high-quality SNPs (Q30 filter) and performed read-backed phasing. After that, we used the R packages SNPfiltR (DeRaad 2022) and vcfR (Knaus and Grunwald 2017) to visually and iteratively filter our SNP dataset based on specific quality and completeness metrics. For analyses requiring unlinked SNPs, a linkage disequilibrium-based filtering approach was applied using PLINK (Purcell et al. 2007) with fixed parameters (--indep-pairwise 50 10 0.5). At the end of this process, two datasets were generated. The first, referred to as the '*unlinked SNP dataset I*', included 42 samples and excluded the *Tropidurus melanopleurus* sample, retaining only one sample external to the largest clade within the group, namely *T. callathelys* (used in D-suite analysis; 40,265 SNPs, 25.34% missing data). The second one, '*unlinked SNP dataset II',* with 37 samples, excluding the *T. melanopleurus, T. callathelys, T. teyumirim*, and *Tropidurus* sp. nov. samples, and assigning the *T. lagunablanca* sample as *T. tarara* (used in Stairway Plot analysis; 27,454 SNPs, 24.71 % missing data).

Gene and species tree analysis

Gene trees for all UCE loci and the mtDNA locus were inferred using IQ-TREE v2.2.6 (Minh et al. 2020), and support was inferred using 1000 ultrafast bootstraps (≥ 95 considered as strong support). For partitioned ML analyses, ModelFinder (Kalyaanamoorthy et al. 2017), part of the IQ-TREE package, was used to select the best-fit model for each partition, followed by tree reconstruction (-m TESTNEWMERGE option), allowing partitions to have different evolutionary rates (-spp option). Support was inferred using 1000 ultrafast bootstraps. Additionally, species trees were inferred for the UCE dataset using ASTRAL v5.15.4 (Zhang et al. 2018), which does this from gene trees and provides internal branch lengths in coalescent units of gene tree discordance, as well as branch support values in the form of local posterior probabilities (LPPs). In our ASTRAL analyses, we used the topology resulting from the partitioned approach produced through IQ-TREE as a reference. Finally, we used BPP v.4.6 (Flouri et al., 2018) to estimate a species tree for the *Tropidurus spinulosus* group, based on the full dataset of 820 UCE loci.

In the case of mitochondrial data, besides partitioning by gene, the alignment was also partitioned by third codon positions (in the case of coding sequences), and we did an additional analysis including two outgroup species to evaluate the monophyly of the *Tropidurus spinulosus* group — *Plica plica* and *T. torquatus* (Figure S6; GenBank accession numbers: AB218961 for *P. plica*, and KU245273, KU245300, KU245090 and KU245062 for *T. torquatus*). We present only the trees (both based on nuDNA and mtDNA) featuring species from the *T. spinulosus* group (our in-group) in the results section. In all cases, trees were manually rooted in *T. melanopleurus* using FigTree (Rambaut 2014), following the topology obtained in the previously described analysis.

Divergence time estimation

A time tree was inferred using two different approaches. The first was the RelTime-Branch Lengths method (Tamura et al., 2018) implemented in MEGA11 (Tamura et al., 2021), which infers divergence times from a previously estimated phylogenetic tree with branch lengths. In this case, the time-tree was computed using the UCE-based phylogenetic tree (complete matrix) with one calibration constraint, namely the node separating *Tropidurus callathelys* from all the other species in the group (and using *T. melanopleurus* as the outgroup). We used 10.74 Ma as the reference value for this constraint, following the estimates obtained by Zheng and Wiens (2016) regarding the node separating *T. callathelys* and the rest of the species in our ingroup. The method described in Tao et al. (2020) was used to estimate confidence intervals and set a normal distribution (mean = 10.74 and standard deviation = 0.5) on the node for which calibration densities were provided.

Alternatively, divergence time estimates were also estimated using MCMCTree v4.10 (dos Reis and Yang 2019), which uses an approximate likelihood approach. In this case, the 100 most clocklike UCE loci within our dataset were selected, particularly the loci with the least root-to-tip variance. Clock-likeness was assessed using SortaDate (Smith et al. 2018), which measures root-to-tip variance within gene trees and then sorts them from highest to lowest variance. The calibration in this case was performed by constraining the age of the same node described above to be between 10 and 11 Ma, using a fixed topology. Molecular clock estimates were obtained using the independent rates model, and we used the HKY85 model. The gamma prior on the mean substitution rate for partitions (rgene_gamma) was set to G (7, 9.75), which means a substitution rate of approximately 0.00717 substitutions/site/Ma (value available for the phylogenetically closest species to *Tropidurus* in the germline mutation rate study of Bergeron et al. 2023). The gamma rate variance (sigma2_gamma) was specified with G (1, 10). Two independent runs were performed, each consisting of 550.000 MCMC iterations, with 10.000 as burn-in, ‘nsample’ = 100.000, and ‘sampfreq’ = 5. The ESS of the MCMCTree runs were examined in Tracer v.1.7 (Rambaut et al. 2018) to determine convergence, and values > 200 were retained.

Gene flow and phylogenetic network estimation

Failing to consider the potential influence of gene flow and introgressive hybridization among species can negatively impact phylogenetic inference (e.g., Solís-Lemus and Ané 2016; Solís-Lemus et al. 2016). Therefore, given the mitonuclear discordance observed in the *Tropidurus spinulosus* clade, we evaluated the potential for reticulate evolution in our dataset using five different approaches: (i) the phylogenetic networks applying quartets (SnaQ) method implemented in PhyloNetworks (Solís-Lemus and Ané 2016; Solís-Lemus et al. 2017); (ii) the maximum pseudolikelihood estimate (MPL) method implemented in PhyloNet (Wen et al. 2018); (iii) the extension of the typical tree-based model to general networks, assuming no ILS and independent loci, available in the NetRAX program (Lutteropp et al. 2022); (iv) ABBA-BABA tests for detecting gene flow between pairs of species, through the D-suite program (Malinsky et al., 2021); and (v) the MSC-M model implemented in BPP v.4.6 (Flouri et al. 2018; 2023) to estimate migration rate between pair of species that showed signals of mitonuclear discordance. The first four methods serve as exploratory tools to identify potential reticulations or gene flow events, while the BPP-based approach directly quantifies migration rates. Regarding the phylogenetic network and the ABBA-BABA approaches, please refer to the Supplementary Material, section “*Details on phylogenetic networks methodology and results*”, for a detailed explanation of all reticulation tests.

To directly assess how gene flow differentially impacted nuclear and mitochondrial genomes, we used the migration model (MSC-M) available in BPP v.4.6 (Flouri et al. 2018; 2023). This allowed us to estimate the number of migrants per generation and the direction of gene flow for species showing the strongest mitonuclear discordance in our phylogenetic trees. This was done in two stages, either using the nuclear species tree or the mitochondrial genealogy as the guide tree. In the first case, we used only a portion of the nuDNA dataset—specifically, the 100 most clocklike UCEs selected using SortaDate (Smith et al. 2018). A second migration analysis was performed using that same dataset, but this time incorporating mitochondrial data alongside nuclear data. In other words, two analyses were conducted using the nuclear topology as the guide tree: one using nuclear markers only (100 UCE loci, 93,246 bp) and another combining nuclear data with mitochondrial data (100 UCE loci, 93,246 bp + 1 mitochondrial partition, 13,578 bp).

In both cases, we tested bidirectional migration scenarios with the following species: *Tropidurus xantochilus* ↔ *T. guarani*, *T. xanthochilus* ↔ *T. spinulosus,* and *T. xanthochilus* ↔ *T. tarara*. Scenarios involving ancestral populations or species of the following groups were also tested. For analyses using the mitochondrial genealogy as the guide tree, these included: (A) *T. tarara* and one *T. xanthochilus* group; (B) *T. guarani* and *T. spinulosus*; (C) *T. guarani, T. spinulosus*, and another *T. xanthochilus* group. When using the nuclear topology as the guide tree, the same scenarios with current species mentioned above were tested, while also considering the following scenarios involving ancestral nodes: (B) *T. guarani* and *T. spinulosus*; (D) *Tropidurus* sp. nov. and *T. xanthochilus*; I *T. tarara, T. guarani*, *T. spinulosus,* and *T. teyumirim*; (F) *T. guarani*, *T. spinulosus*, and *T. teyumirim.* Population sizes (θ) were also estimated and recorded for each species and ancestral node involved in the migration pairs. In all cases, we used an inverse gamma prior of G(3, 0.04) for the Θ parameter and G(3, 0.2) for the τ parameter, fixing the species model prior on the topologies described above. The gamma migration prior used was 0.1 ($\alpha = 1, \beta = 10$). We ran analyses for 4 x 10^5 MCMC generations, taking samples every five and using 5 x 10^3 burn-in generations. To check for consistency of results, we conducted two independent runs for each analysis.

Furthermore, we also used POPART v. 1.7 (Leigh et al. 2015) to generate a TCS haplotype network (Clement et al. 2002), based on three mitochondrial genes – cytochrome c oxidase I, II, and III (COX1, COX2, and COX3).

Tests for positive selection and demographic changes through time

To detect significant deviations from the hypothesis of neutral evolution in the mtDNA dataset, we tested for the evidence of positive selection using MEME (Mixed Effects Model of Evolution; Murrell et al. 2012), BUSTED (Branch-Site Unrestricted Statistical Test for Episodic Diversification; Murrell et al. 2015), aBSREL (adaptive Branch-Site Random Effects Likelihood; Smith et al. 2015), and FUBAR (Fast, Unconstrained, Bayesian Approximation; Murrell et al. 2013). MEME is designed to identify sites underlying episodic selection, BUSTED provides a gene-wide (not site-specific) test for positive selection by asking whether a gene has experienced positive selection at least one site on at least one branch, aBSREL focuses on tree branches and can be used to test if positive selection has occurred on a proportion of branches, while FUBAR uses a Bayesian approach to infer nonsynonymous (dN) and synonymous (dS) substitution rates on a per-site basis for a given coding alignment and corresponding phylogeny. All of these tests were performed utilizing the HYPHY package through the DataMonkey server, using default parameters, and the threshold of $p < 0.05$ for all four analyses. The same alignment of mitochondrial genes used for the phylogenetic analyses was utilized here, with the stop codons of each coding gene removed manually to perform the selection analyses. Also, Tajima's D was calculated to evaluate evidence of natural selection and population size changes in the mitogenome using the R function 'tajima.test', from the 'pegas' package (Paradis 2010). Mitochondrial nucleotide diversity was estimated for populations of *T. guarani*, *T. spinulosus*, *T. tarara*, and *T. xanthochilus* using the 'nuc.div' function, also from the 'pegas' package.

Because selection tests indicated neutrality in the mtDNA dataset (see results), we then inferred multilocus coalescent-based Extended Bayesian Skyline Plots (EBSP; Heled and Drummond 2008) implemented in BEAST v2.7.6 (Bouckaert et al. 2019) to estimate changes in effective population sizes over time. To infer the demographic history using the nuDNA dataset, we first generated a folded SFS file for the SNP dataset with easySFS (Gutenkunst et al. 2009). We then used Stairway Plot v. 2.18 (Liu and Fu 2020) to estimate historical effective population size changes of *Tropidurus xanthochilus*, *T. tarara*, *T. spinulosus*, and *T. guarani*. At this point, it is important to note that although EBSP, Stairway Plot, and Tajima's D were applied to monophyletic clusters below the species level, these methods are designed for unstructured samples drawn from local populations. Consequently, the presence of intraspecific population structure violates their underlying assumptions. In coalescent-based approaches, for instance, such structure often results in inflated estimates of historical effective population size (e.g., Heller et al. 2013). While we proceeded with these analyses, we acknowledge that their results may be biased and interpret them with extreme caution. Please refer to the Supplementary Material, section "Details on demographic inference methodology and results", for a detailed explanation of these tests.

Table 1 provides a summary of all analyses conducted in this study, along with their respective objectives and main results.

Table 1. Overview of the analytical workflow used in this study, detailing the rationale behind each step. SNP = single nucleotide polymorphism; MSA = multiple sequence alignment; nuc = nuclear; mt = mitochondrial.

Method	**Data**	**Analysis**	**Objectives**	**Main results**
IQ-TREE	MSA-nuc (100, 95, and 75% UCE matrices) MSA-mt (coding + non-coding regions)	Individual-level phylogenies	What are the phylogenetic relationships of sampled individuals? Do nuc and mt topologies agree?	Nuc: Three main clades detected: (*Tropidurus xanthochilus* + *T. sp. nov.*), (*T. tarara* + *T. lagunablanca*), (*T. teyumirim* + *T. guarani* + *T. spinulosus*), in addition to *T. melanopleurus* and *T. callathelys* as external lineages to the clade comprising all those species. Mt: Strong disagreement between mitochondrial and nuclear-based phylogenies.
ASTRAL	MSA-nuc (100, 95, and 75% UCE matrices)	Species-level phylogenies	Does the species tree agree with the relationships inferred in the individual-level phylogenies?	Overall, it agrees with the IQ-TREE nuclear phylogenies.
BPP	MSA-nuc (100% matrix)	Species tree		
RelTime-Branch Lengths	Nuclear phylogenetic tree (100% UCE complete matrix)	Divergence times	When did the diversification of the *T. spinulosus* group occur? Could this temporal pattern be linked to the species' demographic history, or might it help explain gene flow events throughout their evolution?	Major divergence events occurred from 20 Ma onward. The initial split occurred between 20 and 15 Ma, with *T. melanopleurus* diverging first. Around 10 Ma, *T. callathelys* separated from the remaining taxa, while most subsequent speciation events took place between 7.5 and 2.5 Ma.
MCMCTree	MSA-nuc (100 most clocklike UCE loci selected through SortaDate)			
PhyloNetworks (SnaQ)	Unrooted nuclear gene trees (100% UCE complete matrix) + concordance factors	Phylogenetic Network	Can we detect introgression in the nuclear genome? Is evolutionary	All phylogenetic network results suggest minimal or no introgression within the *T.*

PhyloNet	Rooted nuclear gene trees (100% UCE complete matrix)		history better explained by networks or bifurcating trees in this case?	*spinulosus* species group.
NetRAX	Unrooted nuclear gene trees (100% UCE complete matrix)			
POPART (TCS haplotype network)	Three mtDNA coding genes (COX1, COX2, and COX3)	Population structure	How is mtDNA structured in the group's lineages?	The clusters align with the groupings in the mt-phylogenetic tree, exhibiting the same pattern of discordance with the nuclear data.
D-suite (ABBA-BABA tests)	SNP dataset	Gene flow	Is gene flow occurring between species within the *Tropidurus spinulosus* group?	Minimal interspecific nuclear admixture, suggesting that introgressive hybridization has had limited impact on the nuclear genomes of these species.
BPP (MSC-M model)	MSA-nuc (100 most clocklike UCE loci selected through SortaDate) + MSA-mt (coding + non-coding regions)	Multiple parameter estimation (mainly ancestral population sizes and migration rates)	Evaluate the differential impact of gene flow on nuclear and mitochondrial genomes. Determine which phylogeny more accurately represents the history of population divergence. Specifically, assess whether the mitochondrial genealogy reflects species history, with discordance in the nuclear tree attributable to other processes, or whether the nuclear phylogeny instead captures the true species relationships,	The nuclear phylogeny likely reflects the true species relationships and the actual history of population divergence. In contrast, the mitochondrial genealogy appears discordant, potentially due to processes such as introgression and historical gene flow.

			with the mitochondrial tree being discordant due to factors like introgression and gene flow.	
HYPHY	MSA-mt (coding regions)	Selection tests	Is there evidence of positive selection in the mitogenomes of *Tropidurus spinulosus* lineages, supporting an adaptive-based process leading to mitonuclear discordance?	Overall, there is no strong evidence of adaptive processes influencing mtDNA, with positive selection detected at only a few sites in the mitogenome.
Skyline Plot (BEAST)	MSA-mt (three partitions, comprising six genes)	Demography tests	Estimate changes in effective population sizes over time	Demographic expansions in all analyzed species, particularly from 70 kya onward.
Stairway Plot	SNP dataset			

Results

UCE and mtDNA phylogenomic datasets

We obtained a final dataset including 820 UCE loci for 43 individuals of nine lineages of the *Tropidurus spinulosus* species group. Our complete matrix comprised alignments with 725,505 bp after internal-trimming. Matrices with 75% and 95% of completeness generated alignments with 2,269 and 1,613 loci, 1.98M and 1.40M bp, respectively. Nuclear phylogenetic analyses using both the partitioned approach in IQ-TREE, based on complete and incomplete matrices (Figures S1-S3), and the BPP species tree approach (Figure 2), indicated that our samples can be grouped into three main clades: (*Tropidurus* sp. nov. + *T. xanthochilus), (T. lagunablanca + T. tarara), (T. guarani + T. spinulosus + T. teyumirim)*, in addition to *T. callathelys and T. melanopleurus* as external lineages to the clade comprising the remaining species. The trees inferred using ASTRAL, which account for ILS, produced similar patterns (Figures S4-S6). Compared to the UCE dataset, the mtDNA dataset (alignment with 40 samples, 15 genes, and 13,578 bp) produced markedly different results regarding topology, indicating a clear case of mitonuclear discordance (Figure 2; Figures S7-S8). Notably different from the nuclear trees, our mitochondrial genealogy shows that while most *T. xanthochilus* samples were grouped as sister to the clade containing *T. spinulosus* and *T.*

guarani, two *T. xanthochilus* samples, the *T. sp. nov.* and the *T. teyumirim* samples, were all recovered within the *T. tarara* + *T. lagunablanca* clade. Furthermore, despite *T. xanthochilus* and *T. sp. nov.* forming a clade based on nuclear data, they do not appear closely related in the mitochondrial dataset. Similarly, most groups observed in these two phylogenetic trees are reflected in the haplotype network inferred for the *T. spinulosus* group, showing the same pattern of discordance compared to the nuclear data (Figure S9).

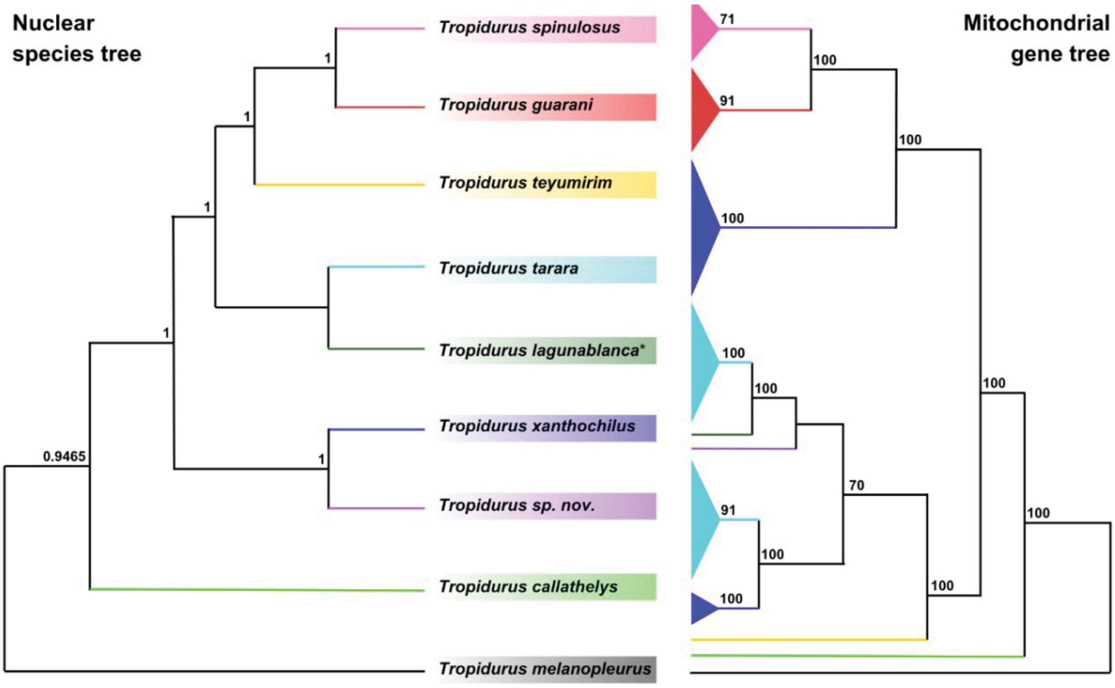


Figure 2. Comparison between the nuclear species tree and the mitochondrial gene tree for the group *Tropidurus spinulosus*. The left panel shows the species tree inferred from multilocus nuclear data using BPP, with branch support values indicated next to nodes (posterior probabilities). The right panel displays the mitochondrial gene tree inferred through IQ-TREE, with branch support shown as bootstrap values (in this case, we only show values above 70). Colored clades correspond to the same taxa in both trees. Collapsed branches in the mitochondrial tree represent intraspecific variation (for the fully detailed tree, see the Supplementary Material). *All nuclear gene trees recovered *T. lagunablanca* nested within *T. tarara* samples, indicating that these lineages likely represent a single evolutionary unit. Accordingly, the *T. lagunablanca* sample was assigned to *T. tarara* in the BPP analysis, so that they represent a single branch in the estimated species tree. However, for a clearer comparison with the mitochondrial genealogy, here we present the two lineages as separate branches within the same clade.

Divergence dating with nuclear data indicates that the initial split within the *Tropidurus spinulosus* group occurred between 20 and 15 million years ago (Ma), separating *T. melanopleurus* from the other lineages. The split involving *T. callathely*s occurred approximately 10 Ma ago, and most subsequent divergence events among the remaining species of the group took place between 7.5 and 2.5 Ma (Figure 3; Figure S10). In most cases, the divergence times estimated using RelTime fall within the same 95%

highest posterior density (HPD) intervals provided by MCMCTree, but some differences can be noticed (Figures S10-11).

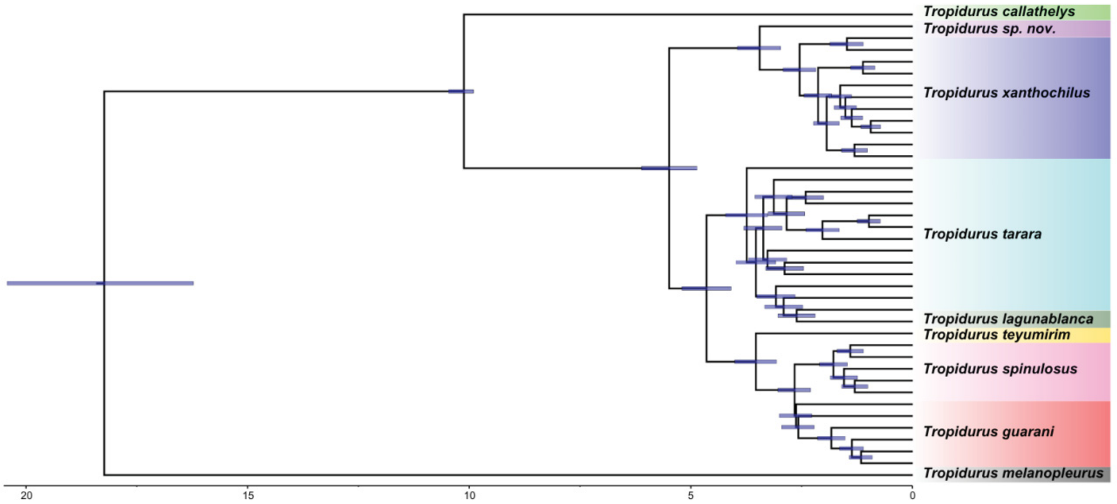


Figure 3. Divergence times (Ma) estimated for the *Tropidurus spinulosus* species group using the UCE dataset. The 95% HPD interval of the posterior estimates (blue shaded bars), as estimated through the MCMCTree program, is shown above each node of the tree. The scale bar is in Ma. Specimen names are collapsed (for the fully detailed tree, see the Supplementary Material).

Phylogenetic networks and gene flow

Analyses of the UCE dataset using PhyloNetworks revealed that models incorporating up to three reticulation events provided a better fit to our data compared to strictly bifurcating trees, where gene tree discordance is attributed solely to ILS. However, these events are mostly intraspecific and do not involve the species with discordant relationships; thus, Phylonetworks does not indicate that reticulation is the primary source of mitonuclear discordance (Figures S12-S15; Table S1). Analyses using the other phylogenetic network approaches corroborated the pattern of minimal reticulations between species observed in our UCE data (Figures S16-S18), and ABBA-BABA tests for detecting gene flow also do not indicate significant nuclear admixture within the *Tropidurus spinulosus* species group (Table S2). Please refer to the Supplementary Material, section "*Details on Phylogenetic Networks methodology and results*," for a detailed explanation of all reticulation tests and respective results.

Analyses in BPP under the MSC-M model with nuclear data indicate minimal or no gene flow among extant species, but inclusion of ancestral populations yields higher migration rate estimates (Figure 4; Tables S3-S4), particularly those associated with the observed phylogenetic discordance. This pattern is robust to the choice of different guide trees: whether constrained by the mitochondrial topology or by the partitioned nuclear topology, although the former produces relatively higher migration rate values. Notably, both guide tree inferences implicate *Tropidurus xanthochilus*, or its ancestors, in these ancient gene flow events, suggesting that historical mitochondrial capture involving this lineage may underlie the main observed discordances between nuclear and mitochondrial phylogenies. Also, when combining the nuclear dataset alongside the complete mitochondrial genome, higher rate estimates were generally produced than those obtained using the nuclear dataset only. In this case, migration rates using the combined dataset (with mitochondrial and nuclear partitions) were higher than the estimates using solely the nuclear data (Table S4). In summary, all migration scenarios evaluated, including information derived from the mitochondrial data (whether based on topology or mitogenome alignments), consistently resulted in higher inferred migration rates compared to nuclear-only analyses, demonstrating that the inclusion of mtDNA in the analysis strengthens the detection of historical gene flow in this system.

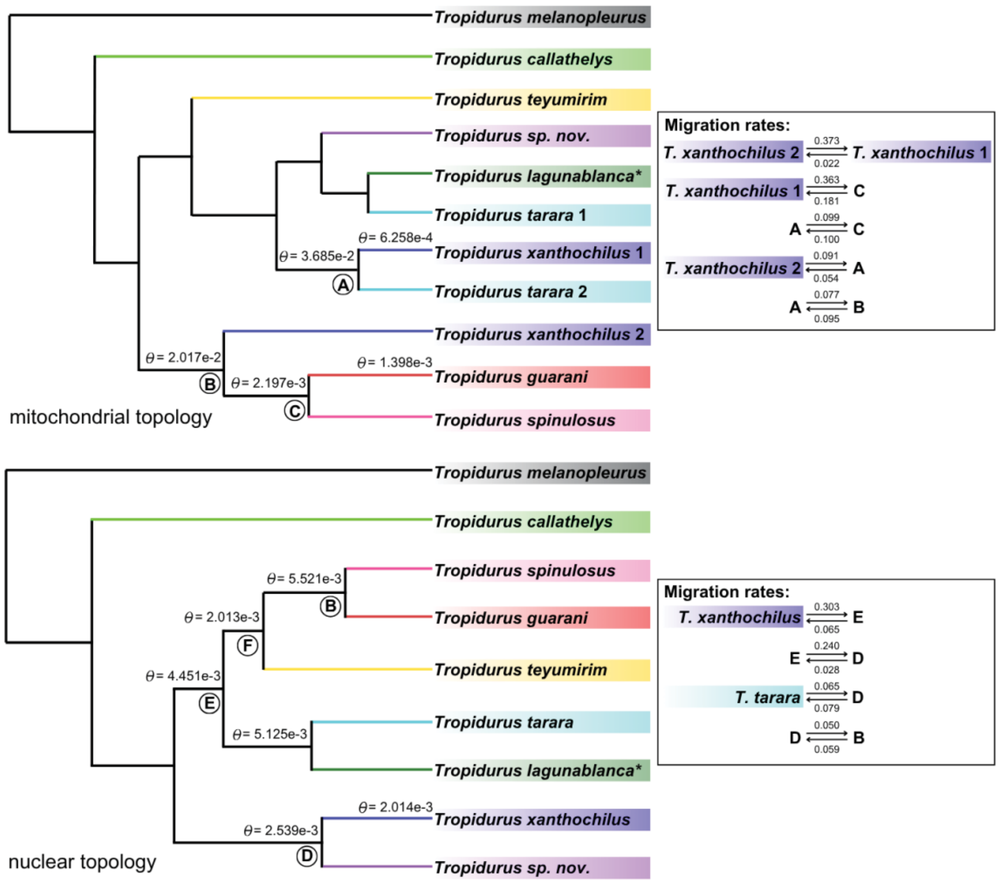


Figure 4. Migration (MSC-M) model for *Tropidurus spinulosus* species, displaying only the scenarios with the highest estimated migration rates. For a complete list of all tested migration scenarios and their corresponding rate values, please refer to the Supplementary Material. Population size (theta, θ) values for each species or ancestral node involved in gene flow events are shown above the branches, with the remaining theta values also provided in the Supplementary Material. The upper panel presents the analysis using the mitochondrial topology as the guide tree; the lower panel shows the same analysis based on the nuclear topology, in this case using both nuclear and mitochondrial data as input. In both trees, the circled letters below the nodes denote the ancestral lineages included in the migration tests.

Estimates of population size (θ) for both current species and ancestral populations further support the previously described scenario (Figure 4; Tables S5-S6), for instance, indicating a directional gene flow event in which the larger ancestral population of the *T. guarani* + *T. spinulosus* clade might have captured *T. xanthochilus* mitochondrial genomes.

Selection

Regarding the mtDNA dataset, our analyses indicate minimal evidence of positive selection acting on the mitogenomes of *Tropidurus spinulosus* lineages. Codon-based tests detected limited signatures of episodic positive or diversifying selection: MEME identified nine sites (two in ATP6, one in ATP8, one in COX1, two in COX2, and one in COX3) with p-values ≤ 0.05, while FUBAR detected two sites (one in COX1 and one in COX2) with posterior probabilities ≥ 0.95. In contrast, both BUSTED and aBSREL analyses did not reveal significant episodic selection along any branches in the mitogenomic phylogeny, even after formal testing of 36 branches with correction for multiple comparisons. Furthermore, we did not observe any species with extreme nucleotide diversity (π) values, which could indicate selection on mitochondrial lineages; instead, the diversity values were nearly identical across all populations (Table S7). Notably, the significantly negative Tajima's D ($p \leq 0.05$, Table S8) indicates an excess of rare variants, a pattern that can arise from both demographic expansion and selection. Subsequent analyses support a scenario of recent expansion for the main populations, suggesting that this signal is evidence of historical demographic changes rather than selection acting on the mitogenomes.

Demographic history

As previously demonstrated, BPP estimates of population size reveal differences among ancestral populations likely involved in the ancient introgression events underlying the observed discordance; these differences also support the inferred direction of gene flow (please refer to the *Migration tests* in the *BPP* section of the Supplementary Material as well). Furthermore, EBSP results indicated a marked increase in mitochondrial effective population size across all populations, particularly within the last 100,000 years (Figure S19). In the nuclear dataset, Stairway Plot analyses showed a consistent pattern of population expansion beginning around 60,000 years ago, followed by relatively stable sizes thereafter (Figure S20).

Discussion

Our genomic analysis of the *Tropidurus spinulosus* species group reveals a complex evolutionary history characterized by: (a) significant discordance between nuclear and mitochondrial phylogenetic groupings, with very low levels of nuclear introgression; (b) at least two major mitochondrial capture events, specifically involving ancient *T. xanthochilus* populations and non-sister groups such as the *T. spinulosus* + *T. guarani* clade and *T. tarara*; (c) no signs of strong positive selection acting on mtDNA; and (d) demographic evidence of recent population expansions in both nuclear and mitochondrial genomes across various species in the group, along with differences in the sizes of ancestral populations that may have reinforced directionality of introgression events. Given these findings, the mitonuclear discordance observed in this study likely originated from admixture among ancient populations of the *T. spinulosus* species group that experienced long-term isolation, followed by secondary contact.

Methodological challenges in investigating introgression as a cause of mitonuclear discordance

Some caution is warranted when interpreting our results, especially given the nature of our dataset. UCEs, by definition, represent highly conserved genomic regions shared across evolutionarily distant taxa (Faircloth et al. 2012). While this might suggest that such data are primarily suited for addressing 'deep' phylogenetic questions, multiple studies have demonstrated their utility at the population and individual levels (Smith et al. 2014; Raposo do Amaral et al. 2018). Additionally, the introgression and gene flow tests we used can capture events on different time scales. For instance, the ABBA-BABA test assumes that multiple substitutions at a given site are rare, as an excess of substitutions can distort patterns of site discordance. This assumption tends to break down in deeply divergent taxa (Hibbins and Hahn 2022), making the ABBA-BABA test more suitable for detecting recent introgression events. We detected only one significant case of interspecific gene flow (both in the case of the PhyloNetworks analysis and the ABBA-BABA tests; however, it warrants caution (as discussed below). Conversely, phylogenetic networks, which analyze discordance between gene trees and other concordant factors, are less impacted by multiple substitutions at individual sites, making them better suited for detecting older introgression events (Hibbins and Hahn 2022).

Therefore, as we used different methodologies capable of detecting signals across distinct time scales, our results cannot be considered dependent on the nature of the used markers, even considering that the accuracy of phylogenetic network methods may also depend on the number of gene trees used. In this regard, previous studies found that a similar number of gene trees compared to our dataset is sufficient to reliably recover an accurate network topology, even though the exact direction of reticulation events may remain uncertain depending on the evolutionary scenario (Solís-Lemus and Ané 2016). Moreover, the consistency of recovered topologies across multiple analyses, with varying levels of data completeness, suggests that the majority of our UCE gene trees are resolved and phylogenetically informative. Therefore, although other genomic markers may offer greater sensitivity for detecting introgression, the consistency of our results across multiple analytical frameworks—each suited to detecting gene flow over different temporal scales—supports the conclusion of minimal or no nuclear introgression within the *Tropidurus spinulosus* group. In any case, future comparative studies are warranted to evaluate how the choice of genomic markers may influence the detection of complex evolutionary processes such as introgression.

Decoupled histories: mitonuclear discordance with minimal nuclear introgression

Reports of mitonuclear discordance are not unexpected, given that mitochondrial and nuclear genomes evolve semi-independently. In this context, while previous studies have documented nuclear introgression associated with mitonuclear discordance (e.g., Zieliński et al. 2013; Myers et al. 2022), our findings point to a different scenario. We found little evidence of substantial nuclear genome introgression, and even the two specific results that could indicate historical hybridization within the group—a significant nuclear reticulation event involving *Tropidurus callathelys* (Figure S14) and a significant D-statistic involving *T. teyumirim* (Table S2)—should be interpreted with caution. First, there is no evidence suggesting contemporary hybridization within the *T. spinulosus* group. The known biology and distribution of these specific lineages do not suggest the occurrence of hybridization events between them (Carvalho et al. 2013; Carvalho 2013; 2016), although this may reflect a distinction between contemporary gene flow and the ability of our methods to detect introgression over evolutionary timescales. Second, the extreme sampling disparity (n=1 for both *T. teyumirim* and *T.*

callathelys versus larger samples for other taxa) could have contributed to false positives in ABBA-BABA tests, which have reduced sensitivity at low admixture levels (Durand et al. 2011), and PhyloNetworks analyses due to stochastic effects.

ILS is another phenomenon that can contribute to mitonuclear discordance, and its signals can easily be mistaken for those caused by introgression (Andersen et al. 2021; DeRaad et al. 2023). In our case, if some samples of a species coalesced deep in the tree, before this lineage diverged from its common ancestor, a significant portion should have coalesced if ILS were the predominant factor driving the discordance with mitochondrial data. However, even our phylogenetic network analyses, which explicitly account for ILS (particularly PhyloNetworks and PhyloNet), did not detect substantial nuclear introgression events. Moreover, the agreement between partitioned (based on IQ-TREE) and ASTRAL analyses of nuclear markers also suggests that ILS has a limited impact on the phylogenetic signal (Figures S1-S6), as ASTRAL is designed to recover the correct topology even under high levels of ILS (Zhang et al. 2018).

Finally, it is essential to acknowledge the lack of methods specifically designed to detect mitochondrial genome introgression with the same level of precision available for nuclear markers. The approach adopted in this study—particularly the BPP-based framework using both mitochondrial and nuclear topologies as guide trees and integrating the mitochondrial partition with multiple nuclear loci (Figure 4)—was designed to illustrate how gene flow may differ between nuclear and mitochondrial genomes due to their distinct inheritance patterns and evolutionary dynamics. Furthermore, we sought to highlight the value of a genome-wide analytical perspective for elucidating the role of introgressive hybridization in animal speciation—an approach still underrepresented in many biological systems (Good et al. 2015). Notably, recent studies have shown that, even in contact zones with frequent hybridization, genetic exchange can remain confined to very narrow geographic ranges, indicating that mitochondrial transfer does not necessarily lead to widespread introgression (Zozaya et al. 2024). This contrast illustrates how the relative permeability of nuclear and mitochondrial genomes to gene flow can vary markedly among squamate clades, underscoring the utility of our BPP-based framework for revealing a distinct pattern in *Tropidurus*.

Specifically, we discovered that while our nuclear data largely dismisses the occurrence of prominent introgression within our species group, the mitochondrial genome presents a contrasting picture. Migration rate estimates increase when mitochondrial DNA is included—either as an additional partition or by using the mitochondrial topology as the guide tree (Tables S3-S4). Specifically, while BPP analyses of 100 UCE loci produced minimal migration rate estimates (under the nuclear-guided topology), incorporating the mitogenome into the dataset increased the estimated values despite the much larger size of the nuclear dataset. Notably, higher migration rate estimates appeared only in scenarios involving ancestral populations (regardless of whether the nuclear or mitochondrial topology was used as the guide tree), producing more realistic values (Figure 4). These findings not only underscore the analytical value of including mitochondrial data in gene flow estimates but also support the conclusion that the nuclear phylogenies more accurately reflect species history in the *T. spinulosus* system.

Proximate explanations of mitonuclear discordance in the Tropidurus spinulosus species group

Evidence supporting a demography-related hypothesis regarding introgression typically falls into three main categories (e.g., Hinojosa et al. 2019; Palacios et al. 2023; Shen et al. 2025): (i) prolonged periods of isolation; (ii) population growth leading to secondary contact; and (iii) a biogeographic signal related to the discordance. Below, we discuss how our findings align with this framework, as the primary cause of the mitonuclear discordance reported in this study is mitochondrial capture associated with the demography of the lineages within the *Tropidurus spinulosus* group. Specifically, our results indicate ancient gene flow events occurring in populations that were likely isolated for extended periods before undergoing secondary contact, although we detected limited evidence of nuclear introgression. Additionally, both nuclear and mitochondrial datasets reveal widespread historical fluctuations in population size (Figures S19-S20; Tables S5-S6). Similar cases have been documented in lizards (e.g., Myers et al. 2022), where geographic isolation, followed by range and demographic expansions, led to secondary contact and gene flow between previously diverged groups.

First, our divergence time estimates support a gradual and prolonged separation among species in the *Tropidurus spinulosus* group, with divergence events between species occurring from 20 Ma onward. Specifically, the initial split occurred between 20 and 15 Ma, with *T. melanopleurus* diverging first. Around 10 Ma, *T. callathelys* separated from the remaining taxa, while most subsequent speciation events took place between 8 and 2.5 Ma (Figure 3). Related to that, it is essential to note that the evolutionary history of South America’s dry diagonal has been shaped by multiple Neogene (23 Ma onward) geological and climatic processes, with individual taxa responding uniquely to local conditions (Guillory et al. 2024). Deep divergences among Squamates, including those within the *T. spinulosus* group, likely stem from Miocene events such as the uplift of the Central Brazilian Plateau and marine incursions in the southern dry diagonal (Guillory et al., 2024; Bezerra et al., 2025).

If such geological events led to prolonged geographic isolation of lineages, their gradual cessation would have increased opportunities for secondary contact and potential introgression, particularly when coupled with demographic and range expansions. In such a scenario, periods of secondary contact may have led to the formation of hybrid zones with varying levels of interbreeding, followed by contrasting patterns of gene flow between the nuclear and mitochondrial genomes. During periods of geographic and/or demographic expansion, nuclear genomes—subject to recombination and meiotic segregation—undergo widespread admixture and tend to homogenize over time, diluting distinct traces of past isolation and secondary contact. In contrast, divergent mitochondrial haplotypes, which are non-recombining and inherited solely from the females, may persist due to genetic drift, particularly if introgression occurred early in the population expansion or when the population sizes were relatively low.

The directionality of gene flow events is another factor that could have played a crucial role in facilitating neutral mitochondrial capture. Because genetic drift is more pronounced in small populations, neutral mutations in mitogenomes are more likely to become fixed (Moore 1995; Després, 2019). Consequently, unidirectional introgression from a larger, resident population into a smaller, invading one would be more probable during the early stages of contact. If the ancestral populations of any species in the *Tropidurus spinulosus* group were more range-restricted, such conditions would further facilitate the fixation of captured mitochondrial variants through drift. The repeated instances of discordance involving *T. xanthochilus*, due to ancient introgression, may suggest that this lineage had a smaller population size compared to others. Our ancestral population size estimates are consistent with this scenario. For instance, in the model where migration rates are higher (in this case, using the mitochondrial topology as the guide tree; Tables S3-S4), Θ estimates of *T. xanthochilus* are considerably lower than those of the ancestor involved in the gene flow event—in this case M (*T. xanthochilus*→ancestor (*T. guarani*, *T. spinulosus*)) = 0.3629, with Θ values of 0.0006258 and 0.02017, respectively (Table S5). A similar pattern was observed when the nuclear topology was used as the guide tree, with the ancestors of the *T. tarara* + *T. spinulosus* + *T. guarani* clade exhibiting higher Θ values than those of the ancestral *T. xanthochilus* populations (Table S6). Under this scenario, it would be plausible to suggest that the mitogenomes of some *T. xanthochilus* individuals could have been incorporated by the larger ancestral populations of the *T. guarani* + *T. spinulosus* clade, and also by ancestral populations of *T. tarara*.

It is also noteworthy to emphasize that two of our demographic analyses (EBSP and Stairway Plot) may capture only a narrower temporal window than the full span of historical gene-flow events. Furthermore, both approaches assume panmictic, unstructured populations (an assumption likely violated by our sampling of divergent lineages), which can bias demographic reconstructions (Heller et al. 2013). Even so, it is reasonable to assume that relatively recent demographic shifts during the Pleistocene (2.5 Ma–12 ka) may have contributed to the observed mitonuclear discordance, either by facilitating the fixation of mitochondrial variants or by reducing detectable signatures of nuclear introgression. Our results suggest that population expansions might have begun at least 100 ka ago in the mitochondrial dataset (Figure S19) and around 60 ka in the nuclear dataset (Figure S20), with relative stability thereafter. Similar late-Pleistocene demographic fluctuations shaping phylogeographic patterns have been documented in other squamate lineages (Guillory et al. 2024). Finally, the temporal mismatch between mitochondrial and nuclear estimates likely reflects the distinct mutation rates applied to each dataset; as with many non-model organisms, specific substitution rate estimates for *Tropidurus* species are unavailable, adding uncertainty to absolute time estimates. For these reasons, we interpret our demographic inferences with caution, recognizing that both population structure and rate uncertainty may influence the magnitude and timing of inferred events.

In addition, the two particular instances of mtDNA capture involving *Tropidurus xanthochilus* are also associated with biogeographic signals: one possibly occurring to the west of *T. xanthochilus*' distribution, involving ancestral populations of the *T. spinulosus* + *T. guarani* clade, and another to the south, involving *T. tarara*. *Tropidurus spinulosus* is found west of the Paraguay River, from north-central Argentina and northwestern Paraguay to southeastern and central Bolivia, covering both the Chaco and dry forest zones (Frost et al. 1998; Carvalho 2013). Although distributional data remain limited, the known ranges of *T. spinulosus* and *T. xanthochilus* are allopatric. The nearest known populations of *T. spinulosus* are located at least 350 km south of the type locality of *T. xanthochilus,* near the '*Serranía de Huanchaca*' (Harvey and Gutberlet, 1998). Even so, Harvey and Gutberlet (1998) proposed that *T. xanthochilus* and *T. spinulosus* might be parapatric in the area where the forests of the Tarvo and Paraguá Rivers merge with the semideciduous *'Chiquitania'* dry forest. On the other hand, *T. tarara* is known from the northern portion of Eastern Paraguay, east of the Paraguay River, to the Brazilian state of Mato Grosso do Sul (near the southern limit of the distribution of *T. xanthochilus*).

Future studies (whether empirical or based on modelling scenarios) will be crucial to assess the potential overlap (past or present) in the distributions of *Tropidurus xanthochilus* with *T. spinulosus* + *T. guarani* clade, and *T. tarara* as well. Such overlap would provide additional biogeographical evidence for mitochondrial capture among these lineages. Environmental changes can increase opportunities for hybridization and introgression through neutral demographic processes (Excoffier et al. 2009; Phuong et al. 2016), with mitonuclear discordance of this nature often linked to climatic instability and its effects on demographic dynamics (Phuong et al. 2016). Phylogenetic niche modeling, which combines past and present climate data with occurrence records and time-calibrated phylogenies to reconstruct historical geographic ranges (e.g., Guillory and Brown, 2021), may be beneficial for testing these hypotheses in the future.

Finally, our findings suggest that mitochondrial capture in the *Tropidurus spinulosus* group likely occurred during a period when the current lineages were not yet fully differentiated and still overlapped in both space and time. This interpretation is supported by our migration rate results, which show that the highest estimated values were associated with scenarios involving ancestral populations (Figure 4). Notably, this aligns with the "gray zone of speciation" concept—a transitional phase where populations shift from widespread admixture to complete genetic isolation (Roux et al. 2016; Burbrink et al. 2021), facilitated by pre- and/or post-zygotic barriers (Stankowski and Ravinet, 2021). The complexity of our evolutionary scenario also suggests that a bibliographic synthesis indicating the major evolutionary drivers and consequences of introgression and genomic capture across taxonomic groups and biogeographic regions providing research avenues and broad hypothesis-testing frameworks would be of great interest to the evolutionary biology community.

Mitonuclear discordance beyond neutral processes: a case for natural selection?

Our analyses did not reveal strong evidence of positive selection acting on the mitochondrial genome of the *Tropidurus spinulosus* species group. Site-model tests across the phylogeny identified only a few individual codons under selection, indicating that these methods captured limited signals of episodic or pervasive positive selection acting on the mitogenome. However, these approaches are designed to detect selection at the level of specific codons or phylogenetic branches, rather than to assess the rapid fixation of mitochondrial haplotypes following introgression (see the HyPhy documentation for a detailed explanation). In such a scenario, an entire haplotype could be fixed in a new species, a process that might not be directly detectable by the methods adopted here.

Nonetheless, complementary analyses, including nucleotide diversity estimates and Tajima's D test results, do not provide support for natural selection either. First, comparable π levels across the lineages involved in the discordance (Table S7) are not indicative of selective forces strongly acting on the mitogenomes of *Tropidurus spinulosus* species. Similarly, the negative Tajima's D values (Table S8), while theoretically compatible with both selection and population expansion, are here interpreted as indicative of historical demographic changes in light of the additional evidence previously discussed. Furthermore, in lizards—where mitochondrial function is less constrained by high metabolic demands compared to endotherms (Chung and Schulte 2020)—the selective pressure for rapid fixation of an introgressed mitogenome is likely lower. Nonetheless, while selective factors are unlikely to be the primary drivers of mitonuclear discordance in *T. spinulosus* lineages, adaptive explanations may still be underrepresented in the literature (e.g., Mao and Rossiter 2020), which may partly reflect the inherent challenges in reliably detecting selection on mtDNA (Bonnet et al. 2017).

Sex-biased dispersal in Tropiduridae and its potential role in mitonuclear discordance

Finally, future investigations could look further into the role of mitonuclear incompatibilities and mating systems in explaining contrasting signals between nuDNA and mtDNA. Asymmetries between sexes in mating behavior, offspring production, and dispersal are well-known demographic factors contributing to mitonuclear discordance (Dessi et al. 2022). For example, in species where males exhibit greater dispersal, nuDNA is expected to show higher genetic homogeneity across distant populations, as male-mediated gene flow facilitates the spreading and mixing of their genetic material. Conversely, if females are more philopatric (i.e., remain near their birthplace), mtDNA may display greater population structure, preserving historical patterns of isolation. As a result, while nuDNA may indicate a single, cohesive population over a broad geographic range, mtDNA could reveal distinct subpopulations shaped by limited female dispersal.

This process can lead to geographically structured mitochondrial lineages that persist despite widespread nuclear gene flow (see reviews in Toews and Brelsford 2012; Bonnet et al. 2017). Currently, empirical data on the dispersal behavior of *Tropidurus* species are lacking. However, among close relatives of Tropiduridae (e.g., Phrynosomatidae), females exhibit strong philopatry and possibly territoriality, whereas males may disperse further from contact zones, facilitating nuclear gene flow but limiting mitochondrial inheritance (Qi et al. 2013). Although limited studies have been conducted on squamate reptiles, some evidence indicates that sex-biased differences in behavior and fitness can influence dispersal patterns. Therefore, we recommend future research to investigate whether sex-related dispersal contributes to the mitonuclear discordance observed in *Tropidurus*.

REFERENCES

ALLIO, R., et al. MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. **Molecular Ecology Resources**, 20: 892–905. 2020.

ANDERSEN, M. J., et al. Complex histories of gene flow and a mitochondrial capture event in a non-sister pair of birds. **Molecular Ecology**, 30: 2087–2103. 2021. https://doi.org/10.1111/mec.15856

ANDREWS, S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. 2010.

BANKEVICH, A., et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. **Journal of Computational Biology**, 19: 455–477. 2012.

BENAVIDES, E., BAUM, R., MCCLELLAN, D., SITES JR., J. W. Molecular phylogenetics of the lizard genus *Microlophus* (Squamata: Tropiduridae): aligning and retrieving indel signals from nuclear introns. **Systematic Biology**, 56: 776–797. 2007.

BENAVIDES, E., BAUM, R., SNELL, H. M., SNELL, H. L., SITES JR., J. W. Island biogeography of Galápagos lava lizards (Tropiduridae: *Microlophus*): species diversity and colonization of the archipelago. **Evolution**, 63: 1606–1626. 2009.

BERGERON, L. A., et al. Evolution of the germline mutation rate across vertebrates. **Nature**, 615: 285–291. 2023.

BEZERRA, C. H., et al. Biogeographical Origins of Caatinga Squamata Fauna. **Journal of Biogeography**, 52: 521–531. 2025.

BOLGER, A. M., LOHSE, M., USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, 30: 2114–2120. 2014.

BONNET, T., et al. A reassessment of explanations for discordant introgressions of mitochondrial and nuclear genomes. **Evolution**, 71: 2140–2158. 2017.

BOROWIEC, M. L. AMAS: a fast tool for alignment manipulation and computing of summary statistics. **PeerJ**, 4: e1660. 2016.

BOUCKAERT, R., et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. **PloS Computational Biology**, 15: e1006650. 2019.

BURBRINK, F. T., GEHARA, M., MCKELVY, A. D., MYERS, E. A. Resolving spatial complexities of hybridization in the context of the gray zone of speciation in North American ratsnakes (*Pantherophis obsoletus* complex). **Evolution**, 75: 260–277. 2021.

BUSHNELL, B. BBTools: a suite of fast, multithreaded bioinformatics tools designed for analysis of DNA and RNA sequence data. **Joint Genome Institute**. 2018.

CARVALHO, A. L. G. On the distribution and conservation of the South American lizard genus *Tropidurus* Wied-Neuwied, 1825 (Squamata: Tropiduridae). **Zootaxa**, 3640: 42–56. 2013.

CARVALHO, A. L. G. Three new species of the *Tropidurus spinulosus* group (Squamata: Tropiduridae) from eastern Paraguay. **American Museum Novitates**, 3853: 1–44. 2016. Doi:10.1206/3853.1

CARVALHO, A. L. G., DE BRITTO, M. R., FERNANDES, D. S. Biogeography of the lizard genus *Tropidurus* Wied-Neuwied, 1825 (Squamata: Tropiduridae): distribution, endemism, and area relationships in South America. **PloS One**, 8. 2013.

CARVALHO, A. L. G., et al. A new *Tropidurus* (Tropiduridae) from the semiarid Brazilian Caatinga: evidence for conflicting signal between mitochondrial and nuclear loci affecting the phylogenetic reconstruction of South American collared lizards. **American Museum Novitates**, 3852: 1–68. 2016.

CASTRESANA, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. **Molecular Biology and Evolution**, 17: 540–552. 2000.

CHUNG, D. J., SCHULTE, P. M. Mitochondria and the thermal limits of ectotherms. **Journal of Experimental Biology**, 223: jeb227801. 2020.

CLEMENT, M., et al. TCS: estimating gene genealogies. In: *Parallel and Distributed Processing Symposium, International* (Vol. 2, pp. 7-pp). **IEEE Computer Society**. 2002.

DERAAD, D. A. SNPfiltR: an R package for interactive and reproducible SNP filtering. **Molecular Ecology Resources**, 22: 2443–2453. 2022.

DERAAD, D. A., et al. Mitonuclear discordance results from incomplete lineage sorting, with no detectable evidence for gene flow, in a rapid radiation of *Todiramphus* kingfishers. **Molecular Ecology**, 32: 4844–4862. https://doi.org/10.1111/mec.17080. 2023.

DESPRÉS, L. One, two or more species? Mitonuclear discordance and species delimitation. **Molecular Ecology**, 28: 3845–3847. https://doi.org/10.1111/mec.15211. 2019.

DESSI, M. C., et al. The role of sex-biased dispersion in promoting mitonuclear discordance in *Partamona helleri* (Hymenoptera: Apidae: Meliponini). **Biological Journal of the Linnean Society**, 136: 423–435. 2022.

DONG, X., et al. Mitochondrial introgression and mito-nuclear discordance obscured the closely related species boundaries in *Cletus* Stål from China (Heteroptera: Coreidae). **Molecular Phylogenetics and Evolution**, 184. https://doi.org/10.1016/j.ympev.2023.107802. 2023.

DOS REIS, M., YANG, Z. Bayesian molecular clock dating using genome-scale datasets. In: **Evolutionary genomics: Statistical and computational methods**. New York, NY: Springer New York, 2019. P. 309-330.

DUFRESNES, C., et al. Are glacial refugia hotspots of speciation and cytonuclear discordances? Answers from the genomic phylogeography of Spanish common frogs. **Molecular Ecology**, 29: 986–1000. https://doi.org/10.1111/mec.15368. 2020.

DURAND, E. Y., PATTERSON, N., REICH, D., SLATKIN, M. Testing for ancient admixture between closely related populations. **Molecular Biology and Evolution**, 28: 2239–2252. 2011.

EXCOFFIER, L., FOLL, M., PETIT, R. J. Genetic consequences of range expansions. **Annual Review of Ecology, Evolution, and Systematics**, 40: 481–501. 2009.

FAIRCLOTH, B. C., et al. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. **Systematic Biology**, 61: 717–726. 2012.

FAIRCLOTH, B. C. PHYLUCE is a software package for the analysis of conserved genomic loci. **Bioinformatics**, 32: 786–788. 2016.

FEDOROV, V. B., et al. Conflicting nuclear and mitogenome phylogenies reveal ancient mitochondrial replacement between two North American species of collared lemmings (*Dicrostonyx groenlandicus*, *D. hudsonius*). **Molecular Phylogenetics and Evolution**, 168. 2022.

FIRNENO, T. J., et al. Delimitation despite discordance: evaluating the species limits of a confounding species complex in the face of mitonuclear discordance. **Ecology and Evolution**, 11: 12739–12753. https://doi.org/10.1002/ece3.8018. 2021.

FLOURI, T., JIAO, X., RANNALA, B., YANG, Z. Species tree inference with BPP using genomic sequences and the multispecies coalescent. **Molecular Biology and Evolution**, 35: 2585–2593. https://doi.org/10.1093/molbev/msy147. 2018.

FLOURI, T., JIAO, X., HUANG, J., RANNALA, B., YANG, Z. Efficient Bayesian inference under the multispecies coalescent with migration. **Proceedings of the National Academy of Sciences**, 120: e2310708120. 2023.

FROST, D. R., CRAFTS, H. M., FITZGERALD, L. A., TITUS, T. A. Geographic variation, species recognition, and molecular evolution of cytochrome oxidase I in the *Tropidurus spinulosus* complex (Iguania: Tropiduridae). **Copeia**, 839–851. 1998.

GOOD, J. M., VANDERPOOL, D., KEEBLE, S., BI, K. Negligible nuclear introgression despite complete mitochondrial capture between two species of chipmunks. **Evolution**, 69: 1961–1972. https://doi.org/10.1111/evo.12712. 2015.

GUILLORY, W. X., BROWN, J. L. A new method for integrating ecological niche modeling with phylogenetics to estimate ancestral distributions. **Systematic Biology**, 70: 1033–1045. 2021.

GUILLORY, W. X., et al. Geoclimatic drivers of diversification in the largest arid and semi-arid environment of the Neotropics: perspectives from phylogeography. **Molecular Ecology**, e17431. 2024.

GUTENKUNST, R. N., HERNANDEZ, R. D., WILLIAMSON, S. H., BUSTAMANTE, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP data. **PloS Genetics**, 5: e1000695. 2009.

HARRIS, R. S. Improved pairwise alignment of genomic DNA. **The Pennsylvania State University**. 2007.

HARVEY, M. B., GUTBERLET JR, R. L. Lizards of the genus *Tropidurus* (Iguania: Tropiduridae) from the Serranía de Huanchaca, Bolivia: new species, natural history, and a key to the genus. **Herpetologica**, 493–520. 1998.

HARVEY, M. G., et al. Sequence capture versus restriction site associated DNA sequencing for shallow systematics. **Systematic Biology**, 65: 910–924. 2016.

HELED, J., DRUMMOND, A. J. Bayesian inference of population size history from multiple loci. **BMC Evolutionary Biology**, 8: 1–15. 2008.

HELLER, R., et al. The confounding effect of population structure on Bayesian skyline plot inferences of demographic history. **PloS ONE**, 8: e62992. https://doi.org/10.1371/journal.pone.0062992. 2013.

HIBBINS, M. S., HAHN, M. W. Phylogenomic approaches to detecting and characterizing introgression. **Genetics**, 220. 2022.

HINOJOSA, J. C., et al. A mirage of cryptic species: genomics uncover striking mitonuclear discordance in the butterfly *Thymelicus sylvestris*. **Molecular Ecology**, 28: 3857–3868. https://doi.org/10.1111/mec.15153. 2019.

KALYAANAMOORTHY, S., MINH, B. Q., WONG, T. K., VON HAESELER, A., JERMIIN, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. **Nature Methods**, 14: 587–589. 2017.

KATOH, K., STANDLEY, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. **Molecular Biology and Evolution**, 30: 772–780. 2013.

KNAUS, B. J., GRÜNWALD, N. J. vcfr: a package to manipulate and visualize variant call format data in R. **Molecular Ecology Resources**, 17: 44–53. 2017.

LEIGH, J. W., et al.. POPART: full-feature software for haplotype network construction. **Methods in Ecology & Evolution**, 6. 2015.

LI, H., et al. The sequence alignment/map format and SAMtools. **Bioinformatics**, 25: 2078–2079. 2009.

LI, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. **arXiv**, 1303.3997v2 [q-bio.GN]. 2013.

LIU, X., FU, Y. X. Stairway Plot 2: demographic history inference with folded SNP frequency spectra. **Genome Biology**, 21: 280. 2020.

LUTTEROPP, S., et al.. NetRAX: accurate and fast maximum likelihood phylogenetic network inference. **Bioinformatics**, 38: 3725–3733. 2022.

MALINSKY, M., et al. Dsuite—fast D-statistics and related admixture evidence from VCF files. **Molecular Ecology Resources**, 21: 584–595. 2021.

MAO, X., ROSSITER, S. J. Genome-wide data reveal discordant mitonuclear introgression in the intermediate horseshoe bat (*Rhinolophus affinis*). **Molecular Phylogenetics and Evolution**, 150: 106886. 2020.

MCKENNA, A., et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. **Genome Research**, 20: 1297–1303. 2010.

MINH, B. Q., et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. **Molecular Biology and Evolution**, 37: 1530–1534. 2020.

MOORE, W. S. Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees. **Evolution**, 49: 718–726. 1995.

MURRELL, B., et al. Detecting individual sites subject to episodic diversifying selection. **PloS Genetics**, 8: e1002764. 2012.

MURRELL, B., et al. FUBAR: a fast, unconstrained Bayesian approximation for inferring selection. **Molecular Biology and Evolution**, 30: 1196–1205. 2013.

MURRELL, B., et al. Gene-wide identification of episodic selection. **Molecular Biology and Evolution**, 32: 1365–1371. 2015.

MYERS, E. A., et al. Interspecific gene flow and mitochondrial genome capture during the radiation of Jamaican *Anolis* lizards (Squamata: Iguanidae). **Systematic Biology**, 71: 501–511. 2022.

OLSON, D. M., et al. Terrestrial ecoregions of the world: a new map of life on Earth. **BioScience**, 51: 933–938. 2001.

PALACIOS, C., CAMPAGNA, L., PARRA, J. L., CADENA, C. D. Mito-nuclear discordance in the phenotypically variable Andean hummingbirds *Coeligena bonapartei* and *Coeligena helianthea* (Trochilidae). **Biological Journal of the Linnean Society**, 139: 145–157. 2023.

PARADIS, E. pegas: an R package for population genetics with an integrated–modular approach. **Bioinformatics**, 26: 419–420. 2010.

PHUONG, M. A., BI, K., MORITZ, C. Range instability leads to cytonuclear discordance in a morphologically cryptic ground squirrel species complex. **Molecular Ecology**, 26: 4743–4755. https://doi.org/10.1111/mec.14238. 2016.

PURCELL, S., et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. **The American Journal of Human Genetics**, 81: 559–575. 2007.

QI, Y., et al. Genetic evidence for male-biased dispersal in the Qinghai toad-headed agamid *Phrynocephalus vlangalii* and its potential link to individual social interactions. **Ecology & Evolution**, 3: 1219–1230. 2013.

RAMBAUT, A. FigTree v.1.4.2: tree drawing tool. Available at: http://tree.bio.ed.ac.uk/software/figtree/. 2014.

RAMBAUT, A., DRUMMOND, A. J., XI, D., BAELE, G., SUCHARD, M. A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. **Systematic Biology**, 67: 901–904. 2018.

RANCILHAC, L., et al. Phylotranscriptomic evidence for pervasive ancient hybridization among Old World salamanders. **Molecular Phylogenetics and Evolution**, 155. https://doi.org/10.1016/j.ympev.2020.106967. 2021.

RAPOSO DO AMARAL, F., et al. Recent chapters of Neotropical history overlooked in phylogeography: Shallow divergence explains phenotype and genotype uncoupling in *Antilophia* manakins. **Molecular Ecology**, 27: 4108–4120. 2018.

ROUX, C., et al. Shedding light on the grey zone of speciation along a continuum of genomic divergence. **PloS Biology**, 14. 2016.

SEIXAS, F. A., et al. The genomic impact of historical hybridization with massive mitochondrial DNA introgression. **Genome Biology**, 19. https://doi.org/10.1186/s13059-018-1471-8. 2018.

SHEN, C. C., et al. Exploring Mitonuclear Discordance: Ghost Introgression From an Ancient Extinction Lineage in the *Odorrana swinhoana* Complex. **Molecular Ecology**, e17763. 2025.

SMITH, B. T., et al. Target Capture and Massively Parallel Sequencing of Ultraconserved Elements (UCEs) for Comparative Studies at Shallow Evolutionary Time Scales. **Systematic Biology**, 64: 83–95. Doi:10.1093/sysbio/syt061. 2014.

SMITH, M. D., et al. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. **Molecular Biology and Evolution**, 32: 1342–1353. 2015.

SMITH, S. A., et al. So many genes, so little time: a practical approach to divergence-time estimation in the genomic era. **PloS One**, 13: e0197433. 2018.

SOLÍS-LEMUS, C., ANÉ, C. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. **PloS Genetics**, 12: e1005896. 2016.

SOLÍS-LEMUS, C., YANG, M., ANÉ, C. Inconsistency of species tree methods under gene flow. **Systematic Biology**, 65: 843–851. 2016.

SOLÍS-LEMUS, C., BASTIDE, P., ANÉ, C. PhyloNetworks: a package for phylogenetic networks. **Molecular Biology and Evolution**, 34: 3292–3298. 2017.

STANKOWSKI, S., RAVINET, M. Defining the speciation continuum. **Evolution**, 75: 1256–1273. 2021.

TAMURA, K., QIQING, T., KUMAR, S. Theoretical Foundation of the RelTime Method for Estimating Divergence Times from Variable Evolutionary Rates. **Molecular Biology and Evolution**, 35: 1770–1782. 2018.

TAMURA, K., STECHER, G., KUMAR, S. MEGA 11: Molecular Evolutionary Genetics Analysis Version 11. **Molecular Biology and Evolution**. https://doi.org/10.1093/molbev/msab120. 2021.

TAO, Q., TAMURA, K., MELLO, B., KUMAR, S. Reliable Confidence Intervals for RelTime Estimates of Evolutionary Divergence Times. **Molecular Biology and Evolution**, 37: 280–290. 2020.

TOEWS, D. P., BRELSFORD, A. The biogeography of mitochondrial and nuclear discordance in animals. **Molecular Ecology**, 21: 3907–3930. 2012.

WEN, D., YU, Y., ZHU, J., NAKHLEH, L. Inferring phylogenetic networks using PhyloNet. **Systematic Biology**, 67: 735–740. 2018.

ZARZA, E., et al. Hidden histories of gene flow in highland birds revealed with genomic markers. **Molecular Ecology**, 25: 5144–5157. 2016.

ZHANG, C., et al. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. **BMC Bioinformatics**, 19: 15–30. 2018.

ZHENG, Y., WIENS, J. J. Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. **Molecular Phylogenetics and Evolution**, 94: 537–547. 2016.

ZIELIŃSKI, P., et al. No evidence for nuclear introgression despite complete mtDNA replacement in the Carpathian newt (*Lissotriton montandoni*). **Molecular Ecology**, 22: 1884–1903. 2013.

ZOZAYA, S. M., MACOR, S. A., SCHEMBRI, R., HIGGIE, M., HOSKIN, C. J., O'HARA, K., … MORITZ, C. Contact zones reveal restricted introgression despite frequent hybridization across a recent lizard radiation. **Evolution**, 79: 411–422. 2025.

SUPPLEMENTARY MATERIAL AVAILABLE ONLINE, INCLUDING DATASETS, CODE AND SCRIPTS:

https://www.biorxiv.org/content/10.1101/2025.04.25.650633v4.supplementary-material

5. CAPÍTULO V: GENÔMICA EVOLUTIVA E BIOGEOGRAFIA DO GRUPO *Tropidurus spinulosus* (SQUAMATA, TROPIDURIDAE)

RESUMO

A biodiversidade sul-americana foi moldada por uma complexa interação entre processos geológicos e climáticos, mas as contribuições relativas de eventos específicos na evolução de sua fauna nativa ainda são pouco compreendidas. Investigamos os padrões evolutivos e os processos de especiação do grupo de espécies *Tropidurus spinulosus* (Squamata: Tropiduridae), amplamente distribuído pelos habitats abertos do centro-sul da América do Sul, como modelo para desvendar esses fatores. Para isso, integramos dados de filogenômica, inferência de tempos de divergência, reconstrução de caracteres ancestrais, projeções paleoclimáticas e seleção de modelos demográficos. Sequenciamos loci de elementos ultraconservados (UCE) de representantes de todas as espécies reconhecidas, além de uma linhagem ainda não descrita. As análises filogenômicas revelaram que as relações dentro do grupo são mais complexas do que se reconhecia anteriormente, indicando a necessidade de uma possível revisão taxonômica. Por exemplo, *T. lagunablanca* está inserido dentro de *T. tarara*, sugerindo que ambas devem ser sinonimizadas. Além disso, *T. guarani* mostrou-se parafilético em relação a *T. spinulosus*, evidenciando fronteiras ainda indefinidas entre esses táxons. A estimativa dos tempos de divergência situou as separações mais antigas do grupo no Mioceno, coincidindo com incursões marinhas no continente. Esses resultados apontam a vicariância induzida por eventos marinhos como o principal fator inicial de isolamento de linhagens, especialmente em *T. melanopleurus*, a espécie que divergiu mais precocemente no grupo. Eventos geológicos subsequentes também influenciaram a especiação e os padrões de estrutura populacional. A subsidência da Bacia do Pantanal e a reorganização hidrográfica associada (que culminaram no estabelecimento do curso moderno do rio Paraguai durante a transição Plio–Pleistocênica) reforçaram a compartimentalização populacional e contribuíram para a atual alopatria de espécies-irmãs (por exemplo, *T. spinulosus* a oeste e *T. guarani* a leste). As reconstruções de caracteres ancestrais indicam um ancestral saxícola, com pelo menos três transições independentes para o hábito arborícola (*T. spinulosus*, *T. tarara* e *T. xanthochilus*), sugerindo que mudanças de micro-habitat complementaram o isolamento geográfico ao promover divergência ecológica e reprodutiva. As reconstruções de distribuições ancestrais combinadas com modelos demográficos revelam ainda expansões de distribuição no Pleistoceno, que remodelaram as distribuições e possibilitaram episódios de contato secundário e introgressão. Em conjunto, nossos resultados descrevem um processo em camadas temporais: a vicariância miocênica induzida por incursões marinhas foi responsável pela separação inicial das linhagens; o particionamento hidrográfico no Plio–Pleistoceno reforçou o isolamento espacial; e as oscilações climáticas pleistocênicas modularam as dinâmicas demográficas e histórias de introgressão. Esse arcabouço integrativo destaca como eventos geológicos e climáticos do Neógeno e do Quaternário atuaram conjuntamente na geração da diversidade herpetofaunística contemporânea e oferece previsões explícitas e testáveis para estudos biogeográficos comparativos de vertebrados co-distribuídos nos ecossistemas abertos do sul da América do Sul.

Palavras-chave: especiação; filogenômica; Neógeno; Pleistoceno; Rio Paraguai; transgressões marinhas; vicariância.

ABSTRACT

South American biodiversity has been shaped by a complex interplay of geological and climatic processes, yet the relative contributions of specific events on the evolution of its native fauna remain poorly resolved. We investigated the evolutionary patterns and speciation processes of the *Tropidurus spinulosus* (Squamata: Tropiduridae) species group, widespread across open habitats of south-central South America, as a model to disentangle these drivers by integrating phylogenomics, divergence-time inference, ancestral trait reconstruction, paleoclimatic projections, and demographic model selection. We sequenced ultraconserved elements (UCE) from representatives of all recognized species plus one undescribed lineage. Phylogenomic analyses revealed that relationships within the group are more complex than previously recognized, warranting potential taxonomic revision. For instance, *T. lagunablanca* is nested within *T. tarara*, suggesting that it should be synonymized. Moreover, *T. guarani* is paraphyletic with respect to *T. spinulosus*, highlighting unresolved boundaries between these taxa. Divergence-time estimation placed the earliest splits within the group in the Miocene, coincident with marine incursions in the continent. These results implicate marine-driven vicariance as an initial driver of lineage isolation, particularly in *T. melanopleurus*, the earliest-diverging species in the group. Subsequent geological events further influenced speciation and patterns of population structure in the group. The subsidence of the Pantanal Basin and associated hydrographic reorganization, culminating in the establishment of the modern Paraguay River during the Plio–Pleistocene transition, reinforced population compartmentalization and contributed to the current allopatry of sister taxa (e.g., *T. spinulosus* west vs. *T. guarani* east). Ancestral state reconstructions indicate a saxicolous ancestor with at least three independent transitions to arboreality (in *T. spinulosus*, *T. tarara*, and *T. xanthochilus*), suggesting that microhabitat shifts complemented geographic isolation in promoting ecological and reproductive divergence. Reconstructions of ancestral distributions combined with demographic models further reveal Pleistocene range expansions that reshaped distributions and facilitated episodes of secondary contact and introgression. Taken together, our results describe a temporally layered process: Miocene vicariance driven by marine incursions was responsible for initial lineage splitting; Plio–Pleistocene hydrographic partitioning reinforced spatial isolation; and Pleistocene climatic oscillations modulated demographic dynamics and introgressive histories. This integrative framework highlights how Neogene and Quaternary earth-history events jointly generated contemporary herpetofaunal diversity and provides explicit, testable predictions for comparative biogeographic studies of co-distributed vertebrates in the southern open ecosystems of South America.

Key words**:** marine transgressions; Neogene; Paraguay River; Pleistocene; phylogenomics; speciation; vicariance.

5.1 INTRODUCTION

5.1.1 BIOTIC DIVERSIFICATION IN SOUTH AMERICA AND THE ROLE OF GEOCLIMATIC EVENTS

The evolutionary history of South American biodiversity has been shaped by diverse geological and climatic processes, particularly since the Neogene (~23–2.5 Ma; Rull, 2011; Hoorn et al., 2023; Jaramillo, 2023), with taxa responding differently to local conditions (Guillory et al., 2024). This period was characterized by intense Andean orogeny, with successive uplift events that fundamentally reshaped South America's landscapes and climate, leaving a lasting impact on the biotic composition of the Neotropics (Gregory-Wodzicki, 2000; Antonelli et al., 2009; Hoorn et al., 2010). A large amount of evidence has illuminated the main processes defining South American biological diversity (Antonelli et al., 2018; Rull, 2020; Palma-Silva et al., 2022), with finer-scale phylogeographic investigations testing multiple hypotheses of population dynamics and diversification (Turchetto-Zolet et al., 2013; Esquerré et al., 2019; Guillory et al., 2024). However, the exact mechanisms linking geological and climatic events to current biodiversity patterns remain debated, particularly due to limited data availability in terms of taxonomic representation and regional coverage.

Nevertheless, molecular and paleontological evidence is sufficient to suggest that the origins of many South American Squamata lineages likely date back to the Miocene period (~23–5 Ma). Empirical evidence from various taxa has progressively reinforced the importance of events that occurred in this period in shaping South American herpetofauna diversification (Guarnizo et al., 2016; Pereira and Schrago, 2017; Sheu et al., 2020; Bezerra et al., 2025). Specifically, geological events such as the uplift of the Brazilian Central Plateau (BCP) and marine incursions had a major role in this context (Guillory et al., 2024; Bezerra et al., 2025). On the other hand, although Neogene events contributed to deep divergences and broad biodiversity patterns, studies indicate that Quaternary processes, particularly during the Pleistocene (~2.5 Ma–12 Ka), also played a significant role, particularly in shaping specific environmental conditions, which directly influenced population structure, genetic diversity, and the historical demography of squamate species (Ledo et al., 2020; Marques-Souza et al., 2020; Fonseca et al., 2024). Therefore, it is important to consider the influence of both older and more recent geoclimatic events on the evolutionary history of our focal group, *Tropidurus* Wied, 1825.

5.1.2 THE *Tropidurus spinulosus* GROUP: DISTRIBUTION, ECOLOGY AND KNOWLEDGE GAPS

With approximately 31 recognized species (Carvalho 2016; Carvalho et al. 2024; Ferreira et al., 2025), *Tropidurus* lizards occupy tropical and subtropical open habitats across South America, particularly along the Dry Diagonal, Amazonian savanna enclaves, and portions of the Atlantic Forest (Ávila-Pires, 1995; Carvalho, 2013; Carvalho et al., 2013; Werneck et al., 2015). Most species are sedentary and show consistent substrate associations, living on trees, rocky outcrops, or sandy soils (Rodrigues, 1987; Carvalho, 2013; Carvalho et al., 2013). These contrasting environments impose distinct mechanical demands: stable rock surfaces facilitate rapid acceleration, whereas loose sand reduces traction and increases locomotor costs, often selecting for enhanced sprint performance under predation and thermal pressures (Lejeune et al., 1998; Rocha, 1998). Empirical comparisons show that sister species can diverge quickly in locomotor performance despite minimal skeletal differentiation, particularly when they occupy different substrates (Kohlsdorf et al., 2001). Such functional trade-offs, together with repeated transitions in habitat use documented across the genus (Vitt, 1991; Ellinger et al., 2001; Kohlsdorf et al., 2001, 2008; Kohlsdorf et al., 2004; Grizante et al., 2010), highlight how ecological specialization has shaped their morphological and performance traits. This combination of ecological heterogeneity, morphological diversity, and broad geographic distribution makes *Tropidurus* a powerful model for studying ecologically driven divergence (e.g., Kohlsdorf et al., 2001; Grizante et al., 2010).

In this context, the evolutionary history of *Tropidurus* has been investigated using diverse types of data. Early phylogenetic analyses were mostly based on morphological traits, including osteological characters, external and hemipenial morphology, and scale patterns (Frost, 1992; Harvey and Gutberlet, 2000; but see Frost et al., 1998 for the first molecular study of this group, and Frost et al., 2001 for the first study incorporating molecular evidence in a combined analysis). Since then, molecular approaches, especially phylogeographic studies, have greatly expanded our understanding on the evolution of the genus (Werneck et al., 2015; Carvalho et al., 2016, 2018; Domingos et al., 2017; Carvalho et al., 2024; Ferreira et al., 2025). Despite these advances, many studies have focused on sampling near type localities, limiting the representation of broader population diversity. Additionally, the wide distribution of many species (Carvalho, 2013) underscores the need for more comprehensive population-level sampling. Recent studies have clarified important phylogenetic issues within two of its four unranked species groups, the *Tropidurus semitaeniatus* (Werneck et al., 2015; Ferreira et al., 2025) and *T. torquatus* (Carvalho et al., 2016, 2018, 2024) groups. However, the evolutionary history of the remaining two clades—particularly the *Tropidurus spinulosus* group, with the *T. bogerti* group being monotypic—remains relatively less understood.

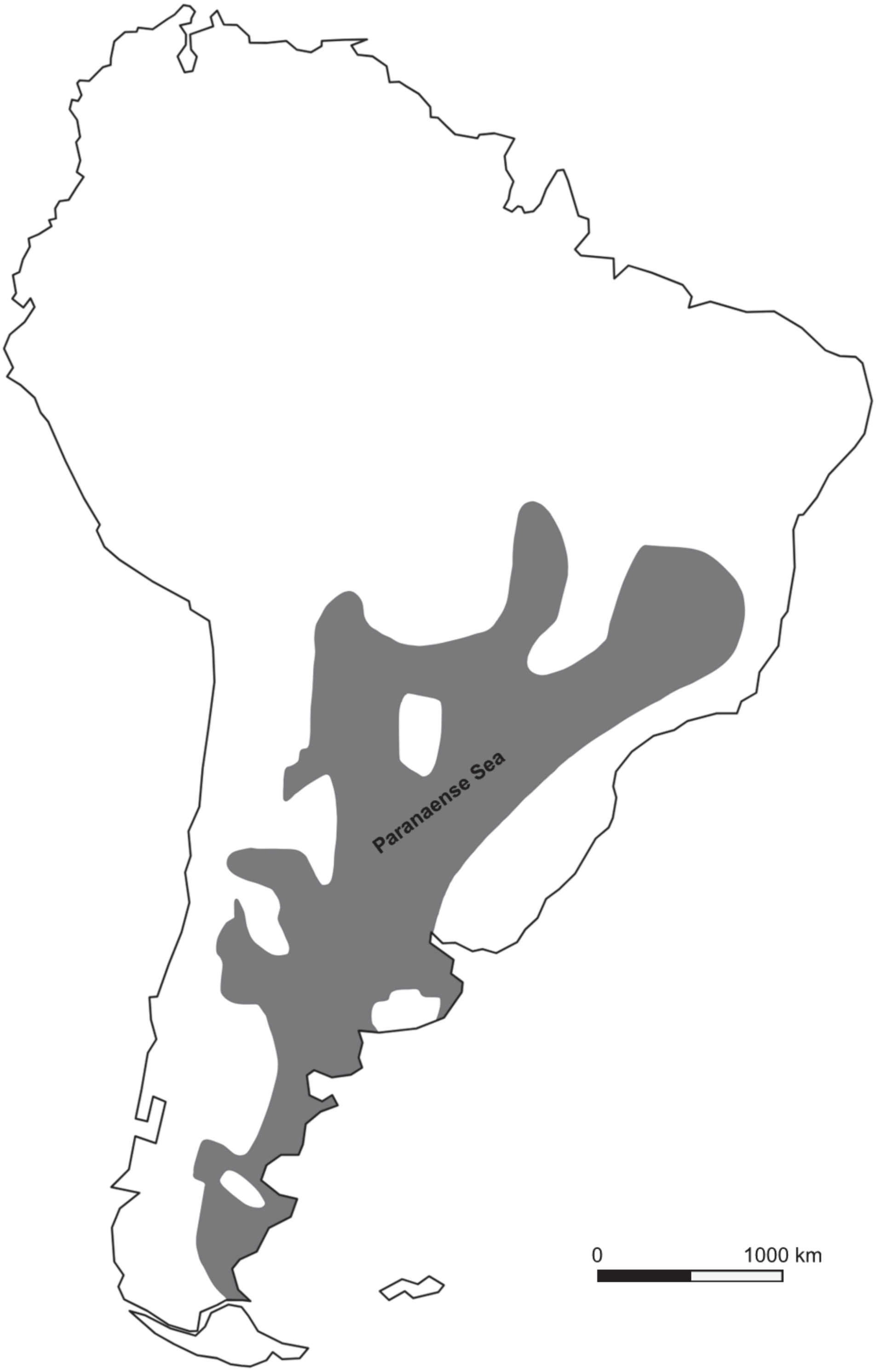
In particular, the *Tropidurus spinulosus* group is widely distributed across tropical and subtropical open environments in South America, predominantly in rocky and arboreal habitats (Álvarez et al., 1994; Frost et al., 1998; Harvey et al., 1998; Carvalho, 2016). Initially recognized by Frost et al. (2001) as comprising five species, the group currently includes at least eight recognized taxa: *T. callathelys*, *T. guarani*, *T. lagunablanca*, *T. melanopleurus*, *T. spinulosus*, *T. tarara*, *T. teyumirim*, and *T. xanthochilus* (Frost et al., 1998; Harvey and Gutberlet, 1998; Frost et al., 2001; Carvalho, 2016). Additionally, there is evidence of undescribed species, one of which is currently undergoing formal description (Carvalho, in preparation; here referred to as *T. sp. nov.*). Rivadeneira (2008) also mentioned two *T. melanopleurus* subspecies that may represent valid species (*T. melanopleurus melanopleurus* and *T. melanopleurus pictus*). The diversification timeline of *Tropidurus* remains poorly resolved (Carvalho et al., 2024), and this is particularly true for the *T. spinulosus* group, in which sparse molecular sampling and a shortage of reliable calibration points preclude precise divergence-time estimates. This scenario highlights the need for integrative studies combining broader population sampling, improved molecular clock analyses, and detailed phylogenetic and biogeographic reconstructions to provide a clearer picture of their evolution.

5.1.3 GEOLOGICAL CONTEXT FOR *Tropidurus spinulosus* DIVERSIFICATION: ANCIENT MARINE INCURSIONS, THE PANTANAL, THE BRAZILIAN SHIELD AND THE PARAGUAY RIVER

Among the major geoclimatic events that shaped the evolution of our focal group, marine incursions represent a key process. Conceptually, marine incursions, or transgressions, occur when ocean waters flood continental interiors due to global sea-level rise and regional tectonic subsidence, and have reshaped terrestrial habitats on every continent with varying extent and timing (Hoorn, 1993; Lovejoy et al., 2006; Bloom et al., 2011). By inundating lowland environments, these events render vast areas inhospitable to land-dwelling organisms, forcing populations into upland refugia or isolated pockets with favorable local conditions. Over prolonged periods, such habitat fragmentation can drive genetic divergence and ultimately allopatric speciation, and highland zones often serve as biodiversity reservoirs that generate high endemism and species richness (Bloom et al., 2011; Yang et al., 2013). Despite their potential importance, marine transgressions have received limited empirical scrutiny in Neotropical terrestrial systems, hindering the formulation of explicit hypotheses. Nevertheless, pioneering studies have provided potential explanations regarding marine-driven diversification. For instance, Nores (2004), examining avian distribution patterns across the lowlands of northern South America, proposed that a sea-level rise of approximately 100 m would have inundated extensive lowland areas, isolating upland "islands" within a flooded landscape and thereby promoting vicariant speciation. This hypothesis is supported by patterns of bird endemism that correspond closely to the 100 m paleocontours. Molecular phylogenetic studies across diverse groups in the Neotropical region (e.g., birds: Aleixo 2004; fishes: Cooke et al. 2012) have further provided at least partial support for vicariant patterns expected under an ancient incursion scenario, reinforcing the role of marine transgressions as dynamic drivers of terrestrial diversification.

In the context of the present study, geological evidence shows that during the Oligocene–Miocene transition, and throughout the Miocene, global sea levels rose significantly, triggering at least two major Atlantic marine incursions in southern South America beginning around 18–13 Ma, and then from 10–5 Ma, approximately (Miller et al., 2005; Hernández et al., 2005; del Río et al., 2018). Concurrent uplift of the Andes and subsidence of foreland basins further amplified regional marine flooding (Gregory-Wodzicki, 2000; Garzione et al., 2008; Cuitiño et al., 2012). These transgressions, collectively known as the "Mar Paranaense", inundated eastern Argentina, western Uruguay, southern Paraguay, and southeastern Bolivia, with sediments of the Yecua Formation documenting incursions into central South America around 10 Ma (Räsänen et al., 1995; Ortiz-Jaureguizar & Cladera, 2006) (FIGURE 1). Flooding of the Paraná Basin region likely submerged much of the Chaco biome, as indicated by sedimentological and fossil records (Hernández et al., 2005; Hulka et al., 2006). Furthermore, northwestern extensions of the "Mar Paranaense" may have connected with contemporaneous incursions in the eastern Andes, the Guyana and Brazilian basins, and even Amazonian seaways during the late Miocene (Uba et al., 2009). By the Mio–Pliocene transition (around 5 Ma), the "Mar Paranaense" had regressed, giving rise to extensive plains across northern Patagonia, central and northern Argentina, and Uruguay, as well as lowland areas adjacent to the eastern Andean foothills in Bolivia and Peru. This regression, possibly driven by the Andean “Quechua” diastrophism marked the onset of the Southern Plains Era, during which climatic cooling at the end of the Oligocene fostered the replacement of forests by savannas and helped delineate the biogeographic subregions of southern South America (Uliana & Biddle, 1988; Ortiz-Jaureguizar, 1998; Barreda & Palazzesi, 2007). In sum, evidence from sedimentology, paleontology, phylogenetics, and biogeography underscores marine incursions as underappreciated but pivotal forces in shaping the distribution, isolation, and diversification of terrestrial lineages across South America, likely including *Tropidurus* lizards.

FIGURE 1 – Possible extension of the Miocene (~ 12 Ma) transgression of the Paranaense Sea. Modified from Ramos & Alemán (2000).



SOURCE: the author & Laura Laino da Costa (2025).

Beyond historical marine incursions, other environmental factors also seem to shape the distribution of *Tropidurus spinulosus* species, yet the relationships among populations in the Pantanal Basin and its surroundings remain poorly resolved. Previous studies have not thoroughly examined the distribution of these species around the floodplain and adjacent regions from a historical biogeographic perspective.

The Pantanal, which harbors lower biodiversity relative to adjacent biomes (see review in Junk & Cunha, 2018), has been the focus of herpetofaunal studies examining how geological history has influenced differences in species richness and composition between the Pantanal depression and the surrounding plateaus (Nogueira et al., 2011; Strüssmann et al., 2011; Piatti et al., 2019). Specifically, the Pantanal constitutes a seasonally flooded ecoregion surrounded by a mosaic of Cerrado and other adjacent formations (FIGURE 2A), creating contact zones and potential ecological barriers to species dispersal. The formation of both the basin and the bordering highlands (plateaus) is likely tied to Andean uplift events around 2.5 Ma (Ussami et al., 1999)—a geological process that played a key role in shaping the region's complex landscape and, in turn, influenced species distributions, range boundaries, and demographic dynamics, potentially limiting connectivity among Cerrado populations near the Pantanal margins. In this context, the peripheral distribution of the *Tropidurus spinulosus* group around the Pantanal may be partially attributable to ancient geological transformations and the specific habitats preferred by these lizards. Alternatively, the Pantanal floodplain may act as an environmental filter, excluding taxa not adapted to its cyclical regimes of flooding and drought (Junk et al., 2006; Strüssmann et al., 2011).

To the northeast and east, the plateaus surrounding the Pantanal Floodplain are contiguous with the broader Brazilian Shield—a stable Precambrian cratonic block extending from the Amazon lowlands in the north, to the La Plata estuary in the south (within the Chaco-Paraná Basin), and bounded by the Madeira and Paraguay river lowlands to the west, and the Atlantic coast to the east (Cordani, 1988; Buckup, 2011; FIGURE 2B). Although 157otaling157ics activity continued through the Cenozoic (Ribeiro, 2006; Buckup, 2011), the Shield's cratonic basement has remained relatively stable in contrast to the more dynamic Andean foreland basins. The Brazilian Shield also underlies much of central and southeastern Brazil, forming the BCP, which reaches elevations of up to 1,700 m across the Cerrado biome (Sano et al., 2019; Lira-Martins et al., 2022) (FIGURE 2B). Between approximately 3 and 2.5 Ma, peripheral depressions surrounding the BCP underwent subsidence while the plateau itself attained its modern elevation, increasing regional topographic and ecological heterogeneity (Ab'Sáber, 1983; Del'Arco & Bezerra, 1989). Geological events in these areas have been linked to lineage divergences in multiple groups since at least 18 Ma (see review in Guillory et al., 2024).

Finally, along the southern margin of the Pantanal, the Bodoquena Plateau—an Upper Proterozoic block of the Corumbá Group formed by tectonic subsidence—further restricts water outflow and reinforces the region's seasonal flooding dynamics (Souza & Souza, 2010). Together, these multiple geological boundaries have created a mosaic of elevational barriers that may have acted both as refugia and environmental filters.

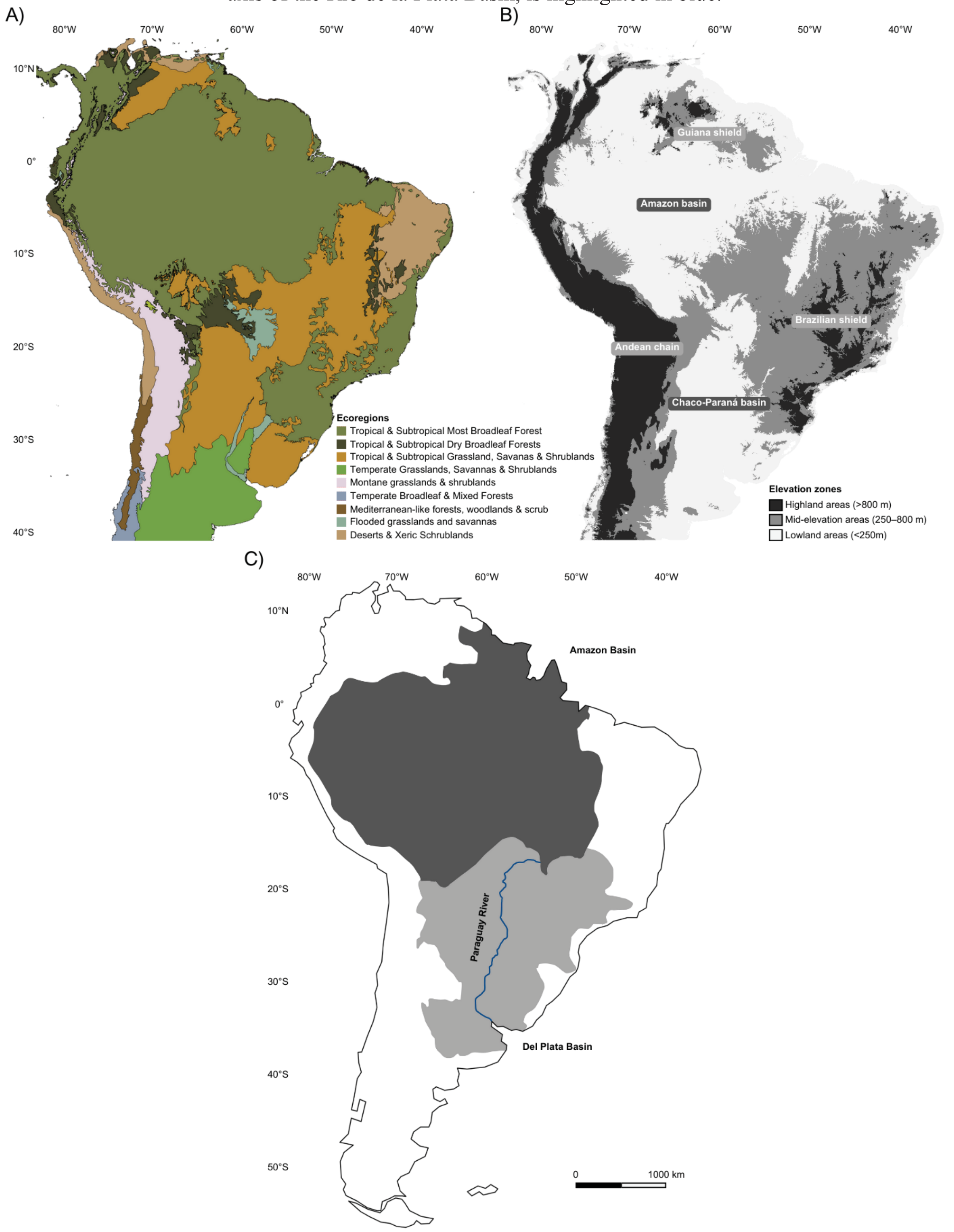
Furthermore, being the world's largest tropical floodplain, the Pantanal biome is consistently shaped by river dynamics, seasonal flooding, and ecological influences from surrounding biomes (Hamilton, 1996; Alho, 2008). In hydrological terms, the Pantanal Basin forms the upper portion of the Paraguay River Basin and is part of the vast Río de la Plata system—here referred to as the Del Plata Basin (Brea & Zucol, 2011; Assine et al., 2015). Covering approximately 3.17×10^6 km² (~1.22×10^6 mi²), the Del Plata Basin drains extensive areas of southeastern Bolivia, southern and central Brazil, most of Paraguay and Uruguay, and northern Argentina—making it the second-largest drainage system in South America after the Amazon Basin (Brea & Zucol, 2011; Assine et al., 2015; FIGURE 2C).

Functionally, the Pantanal Basin acts as a major hydrological regulator within this system: during the rainy season, widespread flooding retains large volumes of water, delaying the discharge of the Paraguay River and helping to prevent its flood peak from coinciding with that of the Paraná River (Barros et al., 2004; Brea & Zucol, 2011; Assine et al., 2015). Of particular relevance to the present study, the Paraguay River originates in Mato Grosso, Brazil, and flows southward through the Pantanal biome before merging with the Paraná River near Corrientes, Argentina, and ultimately discharging into the La Plata estuary (Brea & Zucol, 2011; Assine et al., 2015; FIGURE 2C). The historical course of the Paraguay River is closely tied to the geological evolution of the Pantanal Basin itself, whose formation was primarily driven by the Andean orogeny. Successive phases of uplift and flexural subsidence since the Miocene have progressively shaped the basin's present-day configuration (Assine & Soares, 2004; Brea & Zucol, 2011; Assine et al., 2015). During the Late Pliocene (~2.5 Ma), the reactivation of forebulge faults deepened the Pantanal depression, enabling the developing basin to capture tributaries that had previously drained into the upper Paraná and Tocantins rivers (Ussami et al., 1999; Assine & Soares, 2004; Assine et al., 2015; Brea & Zucol, 2011; Carvalho & Albert, 2011). Additional stream-capture events between the Amazon and Del Plata basins, driven by continued Andean uplift, progressively redirected the Paraná and Paraguay river systems into their modern courses by the Plio–Pleistocene transition (Brea

& Zucol, 2011; Carvalho & Albert, 2011). Further geomorphological and sedimentary evidence demonstrates that the Paraguay River and its tributaries have repeatedly shifted course through time due to processes such as avulsion, megafan lobe switching, channel deflection along floodplain margins, and episodic stream capture (Assine & Silva, 2009; Assine et al., 2015; Pupim et al., 2017). For example, a ~90° channel deflection occurs where the Paraguay River enters the Pantanal floodplain, marking a transition from confined meandering to a distributary pattern with active depositional lobes (Assine & Silva, 2009). These dynamics underscore that the Pantanal floodplain is not a static landscape but a highly mobile system, where hydrological and sedimentary processes continuously remodel drainage networks and floodplain habitats.

Taken together, the geological history of the region (shaped by Andean uplift, forebulge reactivation, and historical stream capture) and the more recent Quaternary to Holocene fluvial dynamics (including avulsion, megafan activity, channel deflection, and backwater flooding), provide a coherent mechanistic framework linking landscape evolution to habitat reorganization in the Pantanal. These processes have produced spatially and temporally variable patterns of connectivity, confinement, and inundation across the basin, periodically generating barriers and corridors that influence the dispersal and persistence of terrestrial and freshwater organisms (Borba et al., 2013; Dorado-Rodrigues et al., 2025). As such, the same geomorphic and hydrological mechanisms responsible for the modern floodplain dynamics likely contributed to the historical structuring of connectivity, refugia, and isolation that shaped the distribution, genetic structure, and diversification of terrestrial taxa in and around the Pantanal Basin—including species of the *Tropidurus spinulosus* group. However, it is important to note that precisely reconstructing the Paraguay River's course changes is challenging because it requires integrating multiple lines of evidence (paleochannel mapping and sedimentary records, dating methods, remote-sensing imagery, and geomorphological modelling) each with limited spatial and temporal resolution and their own sources of uncertainty (reworking, erosion, dating errors), which makes producing unambiguous, time-resolved maps of past channels difficult.

FIGURE 2. Geographic and geomorphological context of central South America, including its major ecoregions, geomorphological units, and hydrographic basins. A) Ecoregions of South America; for a detailed review of the resolution and delimitation differences in these ecoregions, see Olson et al. (2001). B) Major geomorphological units of South America. Lowland areas (below 250 m) are highlighted in light grey, contrasting with the surrounding highlands and plateaus (in darker colors), including the Andean chain, Guiana and Brazilian shields, and the Chaco–Paraná and Amazon basins. C) Map of South America showing the two largest river basins, the tropical Amazon Basin (dark gray) and the subtropical Río de la Plata Basin (light gray). The Paraná–Paraguay River system, the main north–south hydrographic axis of the Río de la Plata Basin, is highlighted in blue.


A)
80°W
70°W
60°W
50°W
40°W
10°N
0°
10°S
20°S
30°S
40°S
Ecoregions
Tropical & Subtropical Most Broadleaf Forest
Tropical & Subtropical Dry Broadleaf Forests
Tropical & Subtropical Grassland, Savanas & Shrublands
Temperate Grasslands, Savannas & Shrublands
Montane grasslands & shrublands
Temperate Broadleaf & Mixed Forests
Mediterranean-like forests, woodlands & scrub
Flooded grasslands and savannas
Deserts & Xeric Schrublands
B)
80°W
70°W
60°W
50°W
40°W
Guiana shield
Amazon basin
Brazilian shield
Andean chain
Chaco-Paraná basin
Elevation zones
Highland areas (>800 m)
Mid-elevation areas (250–800 m)
Lowland areas (<250m)
C)
80°W
70°W
60°W
50°W
40°W
10°N
0°
10°S
20°S
30°S
40°S
50°S
Amazon Basin
Paraguay River
Del Plata Basin
0
1000 km

SOURCE: the author & Laura Laino da Costa (2025).

Finally, it is important to note that large rivers can impose long-term barriers to dispersal and gene flow, promoting population differentiation and speciation when their courses remain stable over evolutionary timescales. One of the first proponents of this mechanism was the British naturalist Alfred R. Wallace (Wallace, 1852). Although this "riverine barrier" model is best documented in the tropical forests like the Amazon (e.g., Boubli et al., 2015; Ribas et al., 2022), subtropical river systems (such as the Paraná–Paraguay river system) may also play a role in shaping patterns of species divergence, indicating at least partial barrier effects (Kopuchian et al., 2020). Under this model, synchronous divergence times across different species pairs on opposite riverbanks would imply a vicariant event linked to the river's formation. In contrast, asynchronous splits or heterogeneous genetic distances would imply more recent dispersal or ongoing gene flow, modulated by species-specific dispersal abilities and ecological preferences. Within this framework, the *Tropidurus spinulosus* group displays a conspicuous east–west distribution relative to the Paraguay River. For instance, *T. guarani* occupies lowland regions east of the river, while its sister species, *T. spinulosus*, is found westward, primarily throughout the Dry Chaco in north-central Argentina, northwestern Paraguay, and southeastern to central Bolivia (Frost et al., 1998; Carvalho, 2013). This pattern suggests that the Mio–Pliocene tectonic-related reconfiguration of the Paraguay Basin, along with the establishment of the Paraguay River as a major hydrological axis, may have facilitated the divergence between sister taxa in our focal group (particularly by restricting gene flow) or at minimum contributed to the delineation of current species range boundaries.

5.1.4 AIMS AND HYPOTHESIS

The evolution of the South American biota has been strongly influenced by geological and climatic events since the Neogene, but the effects of these processes on the diversification of many groups (including lizards of the genus *Tropidurus*) remain poorly understood. In this study, we investigated how the geoclimatic history of South America has shaped diversification patterns in the *Tropidurus spinulosus* group, which is broadly distributed across open habitats in the south-central portion of the continent. To this end, we integrate genomic data, temporal analyses, and ecological methods to address the following questions:

1. What are the phylogenetic relationships among all currently recognized species and populations within the *T. spinulosus* group?
2. Which major geoclimatic events and processes (e.g., Andean uplift, Pantanal subsidence, elevational gradients, Miocene marine incursions) are temporally associated with speciation events within the group?
3. To what extent have ecological divergence (e.g., in substrate use) and geographic barriers (e.g., the Paraguay River) contributed to speciation?

Given the extensive temporal span of the *Tropidurus spinulosus* group's evolutionary history, it is not expected that a single mechanism will account for all speciation events in the group. Over millions of years, multiple processes and mechanisms might have acted at different times and scales, each contributing to lineage isolation and adaptation across specific parts of the group phylogeny. Therefore, rather than favoring a single hypothesis (such as marine vicariance, elevational gradients, or ecological shifts) a more realistic framework acknowledges that distinct speciation events likely resulted from varying combinations of these processes, depending on the temporal, geographic, and ecological context. An outline of our hypothesis, along with their respective main expectations is summarized in TABLE 1.

TABLE 1. Summary of the three main hypotheses proposed in this study to explain speciation processes within the *Tropidurus spinulosus* group. Each hypothesis emphasizes different mechanisms, and is characterized by specific expectations regarding timing, demographic signatures, phylogeographic patterns, ecological correlates, and potential for gene flow.

	H0: Vicariant-Marine	**H1: Elevational Gradient**	**H2: Microhabitat-Niche Shift**
Primary mechanism	Population/species fragmentation due to marine incursion events	Orogenic uplift and/or climate oscillations	Shifts between contrasting microhabitats (rocks *vs*. trees)
Expected timing	Miocene: initial divergence events ~18 Ma; later phases of species divergence between 10–5 Ma	Orogenic: 5–2.5 Ma; Climate-driven: < 2.5 Ma	Variable, depending on local substrate availability (Miocene–Pleistocene)
Demographic signature	18–13 Ma & 10–5 Ma: marked contractions during marine incursions Immediately after retreat of marine incursions (~ 5 Ma onwards): rapid expansion and possible secondary contact	Orogenic (5–2.5 Ma): decline in population size concomitant with tectonic uplift, followed by demographic stability Climate-driven (< 2.5 Ma): repeated cycles of contraction during glaciations and expansion during	In times of emergence of new substrates (Miocene–Pleistocene transition): founding event followed by rapid population expansion; No prolonged decline in population sizes after

		interglaciations, with peaks at each glacial maximum	establishment in the new niche
Phylogeographic pattern	Sister species on opposite sides of ancient marine corridors (E–W, N–S)	Sister species in adjacent, non-overlapping altitudinal zones	Sympatric or parapatric sister species dwelling on distinct substrates
Ecological expectation	Geographical isolation by water courses; no direct relationship with substrate type or altitude	Isolation by altitude, no need to change habitat	Isolation by ecological specialization (substrate); divergent selection even in spatial proximity
Gene flow	Interrupted by marine barriers; possible secondary contact after the retreat of continental marine incursions	Limited by enduring topographical barriers; potentially cyclical in the climate scenario	Restricted by strong divergent selection associated with substrate specialization, even in the absence of other physical barriers

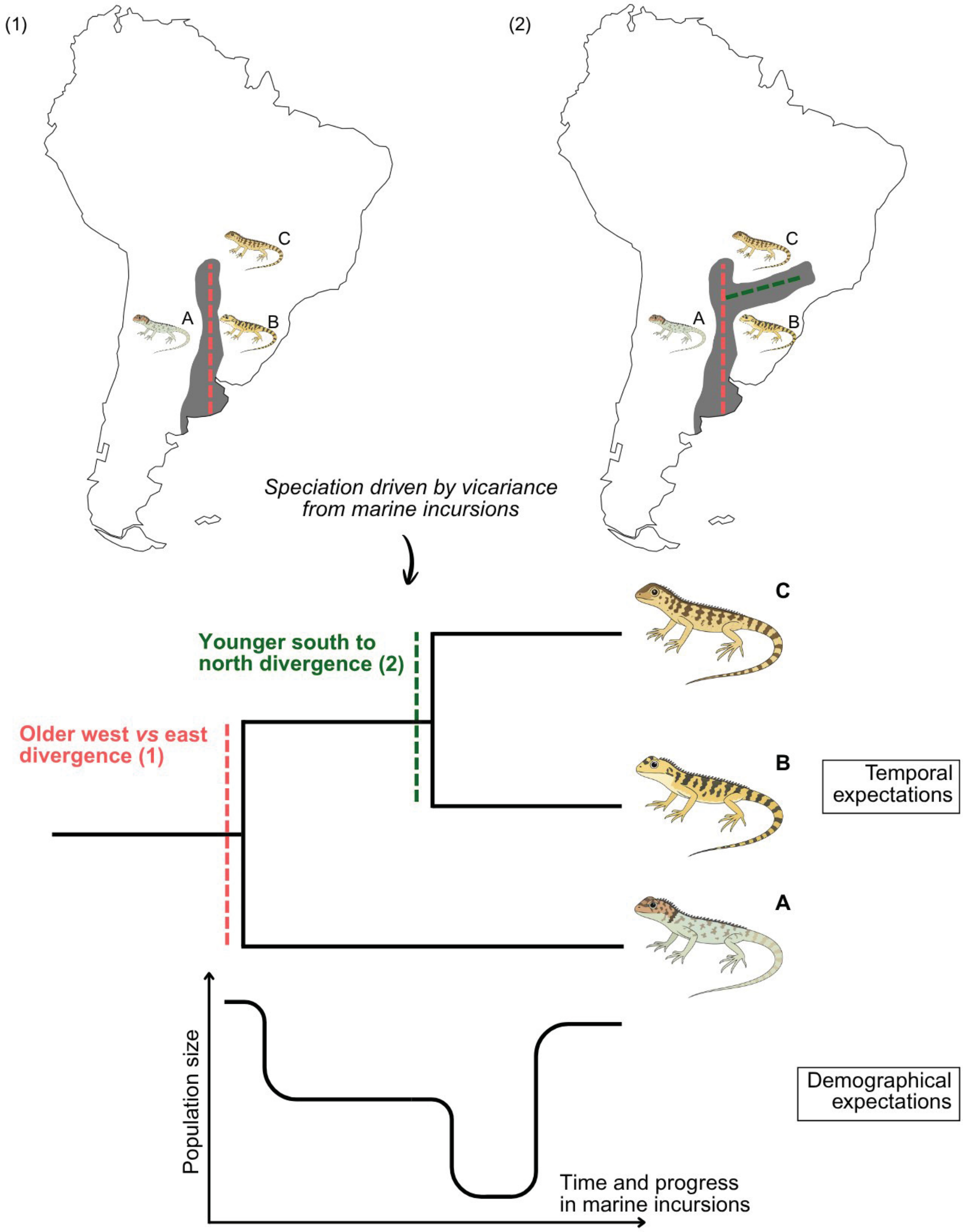
SOURCE: the author (2025).

A detailed version of our (non-mutually exclusive) hypotheses can be found bellow:

Null Hypothesis = Hypothesis 0 (H0): Vicariant-Marine Hypothesis

During the late Miocene and Pliocene, repeated marine transgressions (particularly the expansion and retreat of the "Paranaense Sea") may have fragmented the once-continuous range of ancestral *Tropidurus spinulosus* group lineages, displacing populations into upland refugia or isolating them in areas where local conditions remained suitable. Under this scenario, the earliest divergences within the clade likely occurred when populations became isolated on opposite sides of the advancing sea (i.e., eastern vs. western margins), especially during the initial phase of the incursion around ~18 Ma. Subsequent divergence events would have unfolded along the northern and southern margins of the "Paranaense Sea", driven by later phases of marine expansion between approximately 10 and 5 Ma (FIGURE 3). Demographically, these marine incursions are expected to have triggered periods of population contraction, followed by expansion and potential episodes of secondary contact as terrestrial corridors reemerged.

FIGURE 3. H0 graphic representation.


(1)
(2)
A
B
C
Speciation driven by vicariance from marine incursions
Younger south to north divergence (2)
Older west *vs* east divergence (1)
C
B
A
Temporal expectations
Population size
Time and progress in marine incursions
Demographical expectations

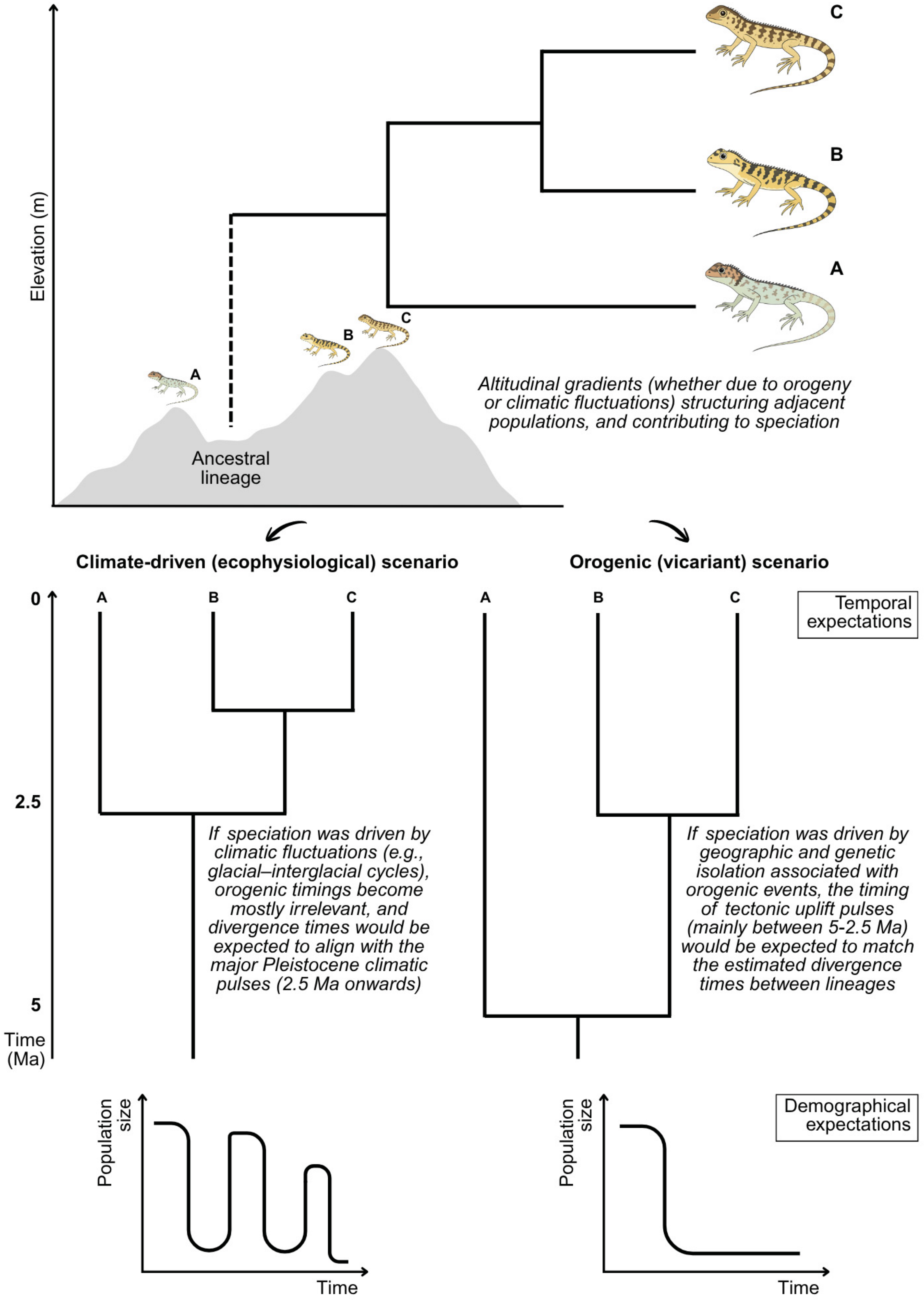
SOURCE: the author & Laura Laino da Costa (2025).

Hypothesis 1 (H1): Elevational Gradient Hypothesis

The current distribution of the *Tropidurus spinulosus* species group suggests that diversification within the group may have been shaped by lineage structuring along emergent altitudinal and environmental gradients. These gradients may have resulted from either Neogene tectonic uplift of highland plateaus in South America (ca. 5–2.5 Ma) or from Pleistocene glacial–interglacial cycles that repeatedly altered climatic conditions (<2.5 Ma). Under this scenario, formerly continuous lowland populations may have become isolated in different elevational zones. Consequently, we would expect sister taxa to occupy adjacent, non-overlapping altitude ranges, without a consistent geographic gradient in divergence times along east–west or north–south axes (FIGURE 4A). Within this framework, two alternative (but not mutually exclusive) scenarios can be considered (FIGURE 4B):

- **Climate-driven (ecophysiological) scenario:** If repeated glacial–interglacial oscillations imposed strong physiological constraints, molecular-clock estimates for divergence times should cluster around major Pleistocene transitions (<2.5 Ma). Demographically, lineages would exhibit evidence of cyclical range contractions into cooler or moister high-elevation refugia during unfavorable phases, followed by downslope expansions when conditions ameliorated (interglacial periods), manifesting as recurrent bottleneck–expansion signatures.
- **Orogenic (vicariant) scenario:** Conversely, if tectonic uplift was the primary isolating mechanism, divergence times should match the main pulses of Neogene orogeny in South America (~5–2.5 Ma), albeit some variance may occur due to local topographic heterogeneity. In one perspective, speciation would be driven directly by these uplift events, with divergence times matching pulses of mountain building. Alternatively, uplift would predate divergence but still provides the geographic isolation necessary for later speciation. In both cases, demographic reconstructions should reveal an initial decline in effective population size following orogeny-induced isolation and little evidence of repeated contraction–expansion cycles, since enduring elevational barriers maintain lineage segregation (rather than cyclic climatic shifts).

FIGURE 4. H1 graphic representation.

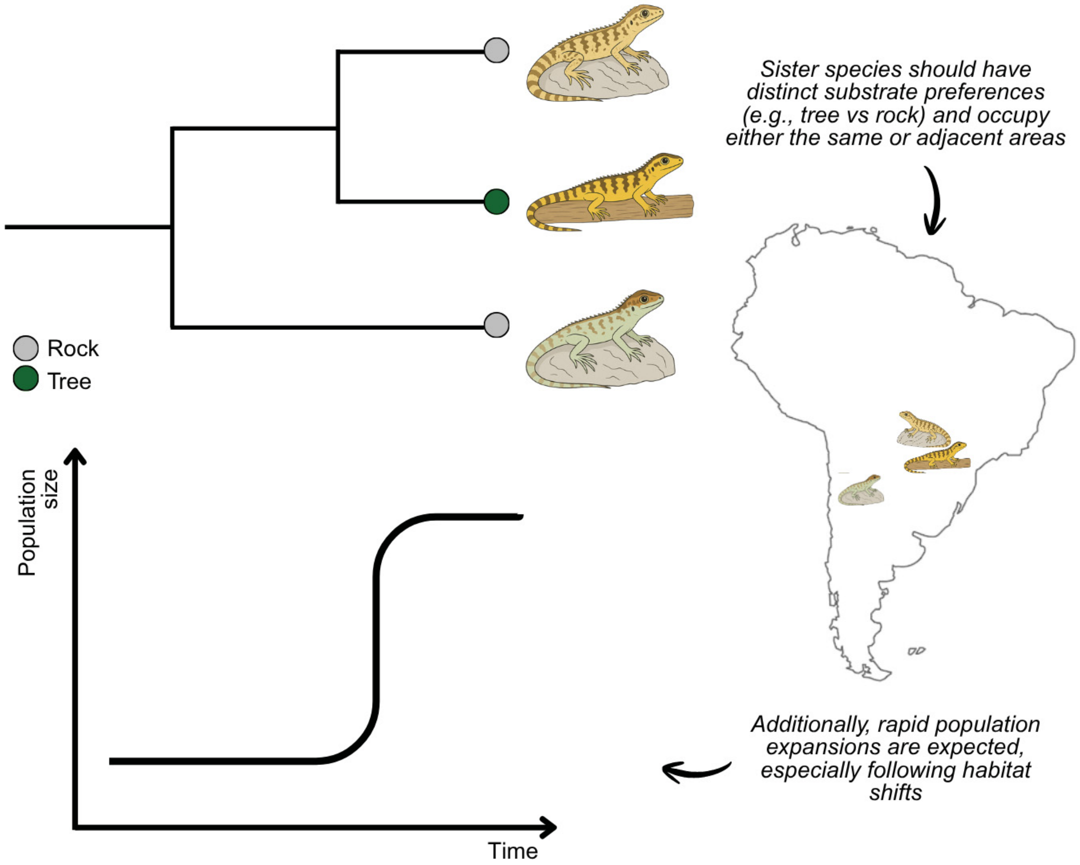

C
B
A
Elevation (m)
A
B
C
Ancestral lineage
Altitudinal gradients (whether due to orogeny or climatic fluctuations) structuring adjacent populations, and contributing to speciation
Climate-driven (ecophysiological) scenario
Orogenic (vicariant) scenario
0
2.5
5
Time (Ma)
A
B
C
A
B
C
Temporal expectations
If speciation was driven by climatic fluctuations (e.g., glacial–interglacial cycles), orogenic timings become mostly irrelevant, and divergence times would be expected to align with the major Pleistocene climatic pulses (2.5 Ma onwards)
If speciation was driven by geographic and genetic isolation associated with orogenic events, the timing of tectonic uplift pulses (mainly between 5-2.5 Ma) would be expected to match the estimated divergence times between lineages
Population size
Time
Population size
Time
Demographical expectations

SOURCE: the author (2025).

Hypothesis 2 (H2): Microhabitat-Shift Hypothesis

A third alternative posits that speciation events within the *Tropidurus spinulosus* group were driven by shifts between distinct structural microhabitats (specifically, rock outcrops *versus* trees) rather than by geological, geographical, or elevational factors per se. Under this scenario, when ancestral *T. spinulosus* lineages colonized novel substrates, strong divergent selection on ecomorphological traits could have led to the emergence of reproductive barriers even in sympatric or parapatric contexts, ultimately promoting or contributing to speciation (FIGURE 5). Divergence events would thus be temporally heterogeneous, coinciding with the local emergence or expansion of suitable exposed rock (or rock outcrops) or habitats with arboreal elements (even if sparse) throughout the Miocene–Pleistocene. In practical terms, this dynamic may have unfolded in two main ways. In one scenario, substrate shifts would have occurred during episodes of geographic expansion, with lineages colonizing newly available environments, leading to changes in habitat use and rapid population growth. Alternatively, substrate shifts may have originated within already isolated populations, where vicariant events established barriers to gene flow prior to habitat transitions. Despite these different mechanisms, both scenarios are expected to produce similar demographic signatures, in which no population bottlenecks are associated with habitat shifts, but rather founder events followed by rapid expansion within the newly colonized niche, and long-term restrictions to gene flow between substrate specialists. Phylogeographically, sister species would be expected to occur in the same or adjacent regions but occupy contrasting substrates, with no consistent latitudinal or elevational pattern.

FIGURE 5. H2 graphic representation.


Rock
Tree
Sister species should have distinct substrate preferences (e.g., tree vs rock) and occupy either the same or adjacent areas
Population size
Time
Additionally, rapid population expansions are expected, especially following habitat shifts

SOURCE: the author (2025).

5.2 MATERIAL AND METHODS

5.2.1 DISTRIBUTION RECORDS OF THE *Tropidurus spinulosus* GROUP

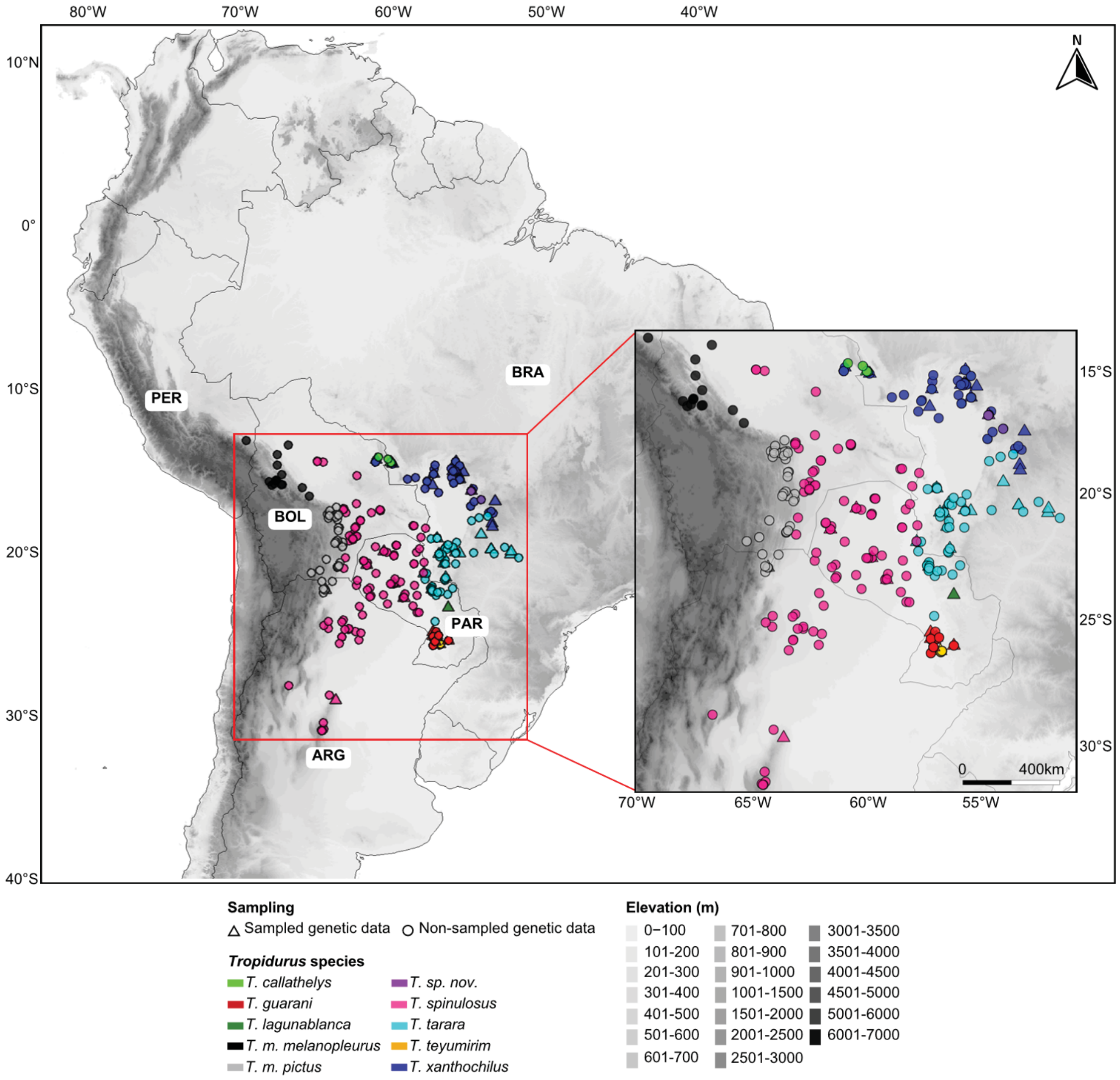
Early studies addressing the distribution, morphology, and taxonomy of the *Tropidurus spinulosus* group were conducted by Álvarez et al. (1994) and Frost et al. (1998). A comparably comprehensive review of museum specimens and literature was not carried out until Carvalho (2013, 2015, 2016). Here, we present an updated distribution map for the *T. spinulosus* group, incorporating new collection sites and applying the current taxonomy (per Cavalho, 2016) to correct population names and document novel occurrences.

Our dataset comprises 404 *Tropidurus spinulosus* records (see Online Documentation and the Supplementary Material). This covers all eight valid species for the group—*T. callathelys, T. guarani*, *T. lagunablanca*, *T. melanopleurus, T. spinulosus*, *T. tarara, T. teyumirim, T. xanthochilus*—and one yet-to-be-described taxon (*T. sp. nov.*). All records were drawn from Carvalho (2013, 2016) and post-2013 field expeditions led by Carvalho et al., which used the same methodology of their original study for voucher examination, collection catalog review, and species identification. Whenever available, we used specimen-associated geographic coordinates; otherwise, we georeferenced localities without coordinates via Google Earth (Google Inc., 2017). To ensure spatial accuracy, we excluded records identified only to province or state. All coordinates were converted to decimal degrees and imported into R (R Core Team, 2022) to generate the distributional maps. Elevation data (in meters), when not available from specimen records or collection labels, were obtained based on geographic coordinates by extracting point elevations from AWS Terrain Tiles using the 'elevatr' package (Hollister et al., 2017).

To quantitatively assess elevational distribution and differentiation among species, we performed statistical analyses on the compiled elevation data. As the data violated assumptions of normality and homoscedasticity for parametric tests, we used non-parametric methods. We first applied a Kruskal-Wallis test to evaluate overall differences in elevation ranges across all species and subspecies. For significant results, we then conducted Dunn's post-hoc tests for pairwise comparisons, applying a Bonferroni correction to account for multiple comparisons. Additionally, to investigate potential altitudinal replacement between sister lineages, we examined species co-occurrence across 100-meter elevation bands up to 1,000 meters and specifically tested for significant elevational differentiation between sister species pairs.

The resulting map from our dataset is shown in FIGURE 6, together with sampled genetic data (as described in the next section).

FIGURE 6. Distribution of the *Tropidurus spinolosus* species group plotted over an altitudinal map (right) of South America. Country borders in South America are also shown on the map, with particular emphasis on those encompassing records of species from the group. ARG = Argentina; BOL = Bolivia; BRA = Brazil; PAR = Paraguai; PER = Peru.



SOURCE: the author (2025).

5.2.2 GENETIC SAMPLING AND LAB WORK

The phylogenomic approaches used in this study were based on anchored sequencing methods, specifically Ultraconserved Elements (UCE) as described by Faircloth et al. (2012). We sequenced a total of 43 individuals (Supplementary Material), representing all species of the *Tropidurus spinulosus* group. RAPiD Genomics LLC (Gainesville, FL) was tasked with sample library preparation and target enrichment of UCEs using the tetrapod 5k probe set (Faircloth et al., 2012), followed by multiplexed paired-end (PE) sequencing (2 × 100 bp) of UCEs on an Illumina HiSeq 3000 PE100 platform.

5.2.3 UCEs PIPELINE

We assembled the reads using the Phyluce v1.7.1 pipeline (Faircloth, 2016). Demultiplexed reads were cleaned to remove low quality bases and adapter sequences in Trimmomatic (Bolger et al., 2014), using the wrapper program Illumiprocessor in Phyluce. To accelerate assembly and render difficult datasets tractable, we normalized read depth at a minimum of 5 using the script bbnorm.sh available in the BBTools suite (Bushnell, 2018). Trimmed and normalized reads were inspected for quality and adapter contamination using FastQC v0.11.9 (Andrews, 2010), and then assembled into contigs with SPAdes v3.15.5 (Bankevich et al., 2012). We used Phyluce to align assembled contigs back to their associated UCE loci, remove duplicate matches, create a taxon-specific database of contig-to-UCE matches and extract UCE loci for all individuals. Specifically, contigs matching UCE loci were identified and extracted using the program LastZ v1.0 (Harris, 2007) within Phyluce, both in incomplete (75% and 95% of completeness) and complete matrices (100% of completeness) (see Faircloth 2016). After probes and UCEs were matched, we aligned UCE contigs with MAFFT v7.471 (Katoh and Standley, 2013) using specific customized settings (-globalpair, --maxiterate 1000, --adjustdirection), and trimmed the resulting alignments using the internal-trimming algorithm (Gblocks: Castresana, 2000). As a final step, AMAS (Borowiec, 2016) was used to compute final summary statistics for all alignments.

5.2.4 GENE AND SPECIES TREE ANALYSIS

Gene trees for every UCE locus were inferred using IQ-TREE v2.2.6 (Minh et al., 2020), and support was inferred using 1000 ultrafast bootstraps (≥ 95 considered as strong support). For partitioned ML analyses, ModelFinder (Kalyaanamoorthy et al. 2017), part of the IQ-TREE package, was used to select the best fit model for each partition followed by tree reconstruction (-m TESTNEWMERGE option), allowing partitions to have different evolutionary rates (-spp option). Support was inferred using 1000 ultrafast bootstraps. Additionally, species trees were inferred for the UCE dataset using ASTRAL v5.15.4 (Zhang et al., 2018), which does this from gene trees (like the one we inferred through IQ-TREE) and provides internal branch lengths in coalescent units of gene tree discordance, as well as branch support values in the form of local posterior probabilities (LPPs). In our ASTRAL analyses, we used the topology resulting from the partitioned approach produced through IQ-TREE as a reference. For the species tree estimation of the *Tropidurus spinulosus* group, we selected the 100% completeness matrix of 820 UCE loci and analyzed it using BPP v.4.6 (Flouri et al., 2018).

5.2.5 DIVERGENCE TIME ESTIMATION

A time-tree was inferred using three different approaches. The first one was the RelTime-Branch Lengths method (Tamura et al., 2018), available in MEGA11 (Tamura et al., 2021), which uses a phylogenetic tree previously estimated branch lengths to infer divergence dates. In this case, the time-tree was computed using the UCE-based phylogenetic tree (complete matrix) with one calibration constraint, namely the node separating *Tropidurus callathelys* from all the other species in the group. We used 10.74 Ma as the reference value for this constraint following the estimates obtained by Zheng and Wiens (2016) regarding the node separating *T. callathelys* and the rest of the species in our ingroup. The method described in Tao et al. (2020) was used to estimate confidence intervals and set a normal distribution on the node for which calibration densities were provided, and a normal distribution was used with the mean centered on 10.74 and standard deviation = 0.5.

Alternatively, divergence times (both for the individual-level phylogeny and the species tree), were also estimated using MCMCTree v4.10 (dos Reis and Yang, 2019), which implements an approximate likelihood method. In this case, the 100 most clocklike UCE loci within our dataset were selected, particularly the loci with the least root-to-tip variance. Clock-likeness was assessed using the program SortaDate (Smith et al., 2018), which measures root-to-tip variance within gene trees and then sorts them from highest to lowest variance. For this calibration, we constrained the age of the same node described above using a secondary calibration point of >10 and <11 Ma, obtained from our own divergence time estimates, and enforced a fixed topology. Molecular clock estimates were obtained using the independent rates model and we used the HKY85 model. The gamma prior on the mean substitution rate for partitions (*rgene_gamma*) was set to G (7, 9.75), which means a substitution rate of approximately 0.00717 substitutions/site/Ma—value available for the phylogenetically closest species to *Tropidurus* in the germline mutation rate study of Bergeron et al. (2023). The gamma rate variance (*sigma2_gamma*) was specified with G(1, 10). In this case, analyses were performed both on the individual-level phylogenetic tree, including all samples, and on the species tree level. Two independent runs for each analysis were performed, with each consisting of 550.000 MCMC iterations, with 10.000 as burn-in, ‘nsample’ = 100.000, and ‘sampfreq’ = 5. The

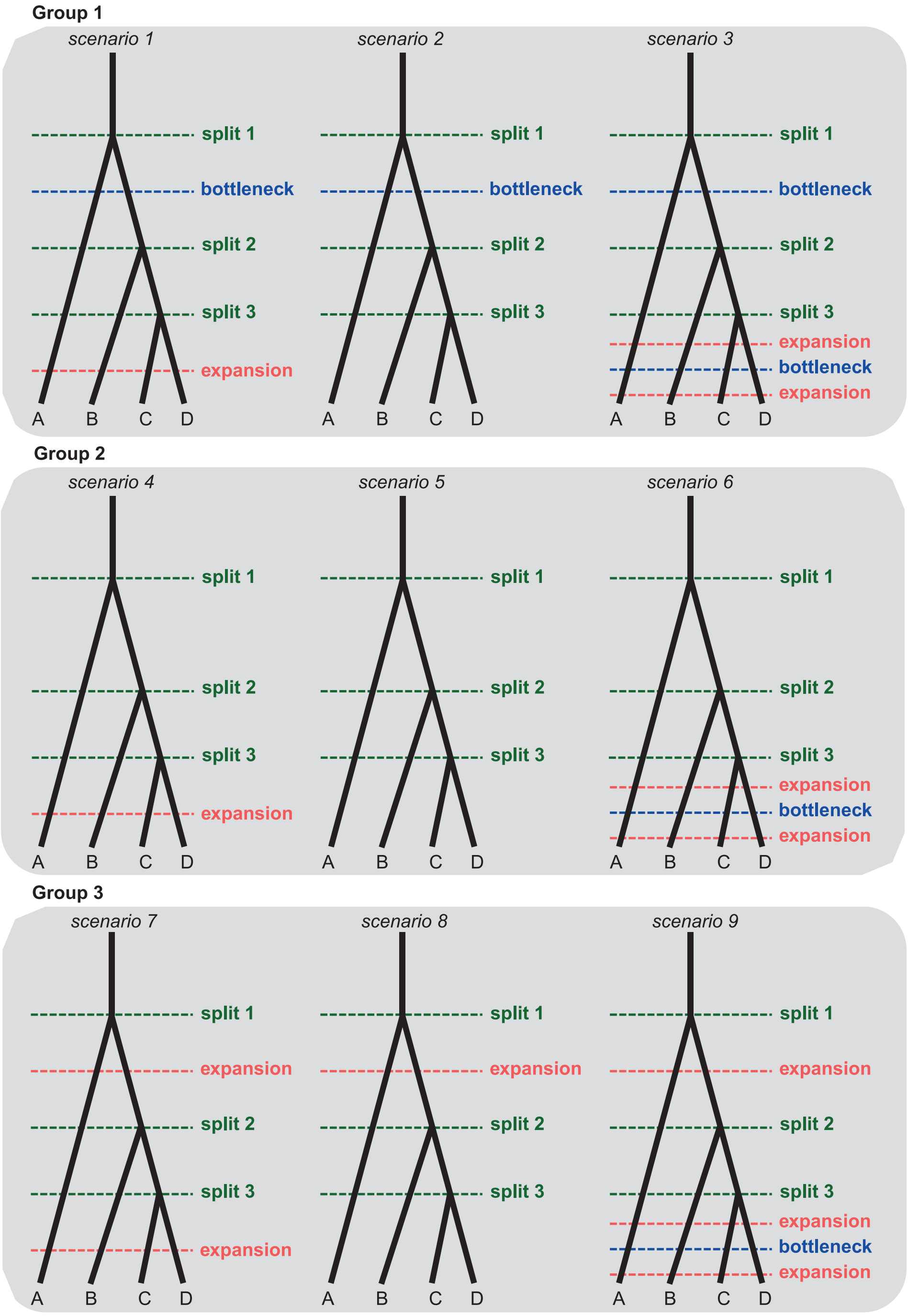
ESS of the MCMCTree runs were examined in Tracer v.1.7 (Rambaut et al., 2018) to determine convergence, and values > 200 were retained.

5.2.6 DEMOGRAPHIC MODEL SELECTION

We employed the Approximate Bayesian Computation with Random Forest framework (ABC-RF; Pudlo et al., 2016; Raynal et al., 2019) implemented in the R package DIYABC-RF v.1.2.1 (Collin et al., 2021) to evaluate alternative demographic scenarios for the evolutionary history of the *Tropidurus spinulosus* group. ABC-RF infers posterior probabilities of competing models from coalescent simulations of genetic data, while the DIYABC-RF workflow allows the simultaneous comparison of multiple scenarios. Its RF classifier is specifically designed for multi-class model selection and retains high accuracy even with many competing alternatives (Collin et al., 2021). Summary statistics synthesize both simulated and observed SNP datasets, and model support is assessed through RF classification votes, with posterior probabilities estimated following Pudlo et al. (2016).

As our primary question about the historical demography of the *Tropidurus spinulosus* group focuses on the occurrence of expansions and contractions events, we first reduced the number of potential scenarios by fixing the phylogenetic relationships among the four species with the most comprehensive population sampling in our dataset (*T. guarani*, *T. tarara*, *T. spinulosus*, and *T. xanthochilus*). This topology was based on both the species tree and partitioned gene tree analyses. In this context, we considered three sets of three scenarios each (a total of nine scenarios), which varied in the number and type of demographic events (FIGURE 7); TABLE 2 shows those scenarios and the corresponding sets of hypotheses. For all nine scenarios, simulation parameters were initially sampled from broad prior distributions informed by existing knowledge of the group. Preliminary analyses indicated that analytical convergence improved when a generation time of 0.5 years was used. This value was selected for methodological reasons, as biologically more realistic values (e.g., 1 year) resulted in poor performance of the model during prior-checking tests, with simulated data distributions failing to encompass the empirical data. While this adjustment was necessary for convergence, we emphasize that it is a pragmatic choice and future studies should seek to better constrain the generation time for this group. A detailed description of all prior distributions is provided in the Supplementary Material (see section '*Detailed DIYABC-RF Workflow*'). RF analyses were first applied across all scenarios to identify the model set best fitting the empirical data. A second round of model selection was then performed within this best-fitting group, using narrower prior distributions to improve model discrimination, with additional simulations conducted under these refined priors.

FIGURE 7. Alternative demographic scenarios for the *Tropidurus spinulosus* group evaluated with ABC-RF. Nine scenarios, grouped into three sets (Groups 1–3), vary in the timing and type of demographic events: population splits (green), bottlenecks (blue), and expansions (red). Analyses first identified the best-fitting scenario group, followed by model selection within that group.



SOURCE: the author (2025).

TABLE 2. Alternative demographic scenarios for the *Tropidurus spinulosus* group, summarizing combinations of demographic events, associated hypotheses, and their evolutionary interpretations.

Scenario	Combination of demographic events	Most plausible hypothesis	Short description
1	Decrease → Increase	H0 & H2	Miocene marine incursions causing early contractions, with expansions after corridor reestablishment; substrate shifts followed by expansions in the Pleistocene.
2	Decrease → Stability	H0 & H1.A	Contraction events from marine incursions and/or Neogene orogeny followed by long-term isolation and relative demographic stability after uplift barriers.
3	Decrease → Cycles	H0 & H1.B	Early contraction phases followed by Pleistocene climatic cycles causing repeated bottlenecks, expansions, and potential secondary contacts.
4	Stability → Cycles	H1.A & H2	Prolonged stability interrupted by late expansions linked to ecological niche shifts or the establishment of populations following Neogene orogeny.
5	Long-standing stability	H1.A	Isolation is linked to Neogene orogeny with persistent barriers maintaining long-term demographic stability, independent of climatic cycles.
6	Stability → Cycles	H1.B	Initial stability followed by Pleistocene climatic oscillations shaping demographic fluctuations along elevational/climatic gradients.
7	Increase → Increase	H2	Early colonization of new microhabitats (rock outcrops, arboreal substrates) followed by continuous demographic expansion through Miocene–Pleistocene times.
8	Increase → Stability	H2	Founder events during habitat shifts followed by local demographic stabilization in newly colonized niches.
9	Increase → Cycle	H1.B & H2	Initial ecological-related expansions later affected by Pleistocene climatic fluctuations, combining divergence associated with substrate preference and environmental instability.

SOURCE: the author (2025).

Specifically, we simulated 10,000 to 20,000 genetic datasets per scenario using DIYABC-RF, matching the properties of the observed dataset in terms of the number of loci and the proportion of missing data. The observed dataset consisted of one SNP per locus from each of the 820 UCE loci present in all individuals (complete matrix). In accordance with DIYABC requirements, we excluded SNPs that were entirely missing in one population or monomorphic across all species. Throughout the simulation process, we used principal component analysis (PCA) to verify that the observed dataset fell within the range of simulated data, ensuring the suitability of conditions for subsequent RF analysis. Both simulated and observed datasets were summarized using the full set of 130 summary statistics proposed by DIYABC-RF for SNP markers. Additionally, linear discriminant analysis (LDA) combinations of summary statistics were combined for model choice (Estoup et al., 2012).

The comparison of posterior probabilities of the competing scenarios at each step was carried out using the RF classification procedure implemented in DIYABC-RF (Pudlo et al., 2016). The RF algorithm, a machine learning method, constructs decision trees from random subsets of the training data (bootstrapping), using the summary statistics as predictor variables. In each bootstrap iteration, simulations not included in the tree construction (out-of-bag simulations) are used to calculate the misclassification error rate (prior error rate), providing a direct and robust estimate of the cross-validation error rate (Pudlo et al., 2016). For each step, we generated classification forests with 500 and 1,000 trees to verify the convergence of the results. The RF procedure outputs the number of “votes” (trees) allocated to each scenario, allowing the selection of the one with the highest count and posterior probability. The overall classification performance (consequently, the power to discriminate among scenarios) was also evaluated using the confusion matrix. This matrix quantifies the accuracy with which simulated datasets are assigned to their true scenario, allowing the approximation of Type I and Type II error rates. Collectively, confidence in the ABC-RF scenario choice was assessed through the prior error rate, the confusion matrix, and the associated error rates. To ensure robustness and convergence of the results, the entire RF analysis was repeated three times for both model selection steps.

The full ABC-RF workflow, from simulations to the estimation of parameter posterior distributions, was carried out within the DIYABC-RF graphical interface. This included computing summary statistics, model checking, and scenario comparisons.

5.2.7 CLIMATE DATA EXTRACTION, PALEOCLIMATE LAYERS, AND ANCESTRAL DISTRIBUTION MODELING

We modeled the ancestral distribution of the *Tropidurus spinulosus* species group using a phyloclimatic framework as implemented in the R package *machuruku* (Guillory & Brown, 2021), which integrates occurrence records, a time-calibrated species tree (obtained through the MCMCTree approach), and paleoclimatic layers to estimate both extant and ancestral distributions across key time slices. Initially, occurrence data for each taxon were compiled and spatially thinned with a 10 km buffer using *machuruku*'s built-in routines to reduce spatial autocorrelation. We then assembled 19 candidate bioclimatic variables from PaleoClim (Brown et al., 2018) at 2.5 arc-minute resolution and removed those with high spatial correlation, known artifacts, or lack of coverage across all time slices.

The remaining ten variables (BIO1, BIO4, BIO10, BIO11, BIO12, BIO13, BIO14, BIO15, BIO16, and BIO17) were standardized (mean = 0, SD = 1) and then used to compute a principal component analysis (PCA) based on their values in the modern climate dataset. The first three principal components were retained, and the rotation derived from the current climate was projected onto all paleoclimatic layers (CUR, LIG, MIS19, MPWP, M2: present, 0.13, 0.787, 3.205, and 3.3 Ma, respectively) to ensure comparability and transferability of environmental axes across time slices. For model calibration, we tested two accessibility hypotheses: ecoregion-based calibrations with a 25 km buffer and hull-based calibrations (convex/concave) with a 200 km buffer, following recommended workflows. After preliminary analyses, we retained the convex hull–based calibration (200 km buffer, constructed considering the whole geographic data available for the *Tropidurus spinulosus* clade) for all downstream analyses because it provided the most realistic representation of the accessible area for our system.

Following data preparation, we used *machuruku*'s tip-level response fitting approach to summarize each taxon's climatic relationship as a skew-normal response distribution. This approach extracts climate values at occurrence locations (or generates supplementary points within a minimum convex polygon when sample sizes are low) and fits a modified Bioclim model (Booth et al., 2014) for each taxon and predictor, parameterized by location (mean), scale (standard deviation), and skew. Then, *machuruku* specifically employs the ace function from the *ape* R package (Paradis et al., 2019) to reconstruct ancestral states for the location, scale, and skew parameters under a Brownian motion (BM) model. This model assumes that trait variation accumulates proportionally with time, providing a probabilistic reconstruction of the ancestral climatic tolerances at each node.

To translate ancestral parameter estimates into spatial suitability hypotheses, *machuruku* reconstructs response curves for every combination of median, lower, and upper confidence limits of the skew-normal parameters, then computes a suitability raster for each combination across all paleoclimatic time slices, and finally sums and rescales these rasters to generate conservative, uncertainty-aware suitability maps. All raster arithmetics and projections were performed within *machuruku*'s vectorized routines, which leverage implementation-level optimizations to maintain computational efficiency across multiple nodes and timeslices. Final outputs were exported and visually inspected using *machuruku*'s plotting functions to compare tip- and node-level response curves and evaluate spatiotemporal patterns. Throughout the whole workflow, we documented all decisions regarding occurrence filtering, predictor selection, calibration area definition, and uncertainty propagation to maximize reproducibility and ensure that ancestral distribution estimates explicitly incorporated parameter uncertainty and the limitations of paleoclimatic data (Online resources include all scripts and documentation necessary to reproduce the analyses).

5.2.8 ANCESTRAL MICROHABITAT RECONSTRUCTION

To test the Microhabitat-Shift Hypothesis (H2), which posits that shifts in substrate preference and habitat use are temporally and phylogenetically aligned with cladogenetic events in the *Tropidurus spinulosus* group, we reconstructed a Tropiduridae phylogeny and mapped ecological transitions onto it. This approach enabled us to evaluate the congruence between lineage diversification and ecological shifts, providing insights into the role of ecological diversification in speciation within this clade. Specifically, ancestral character reconstruction (ASR) analyses were conducted using a mitochondrial phylogeny comprising all nine species of the *T. spinulosus* group plus 11 additional Tropiduridae species (TABLE 3). Our taxon selection and reliance on mitochondrial data were determined by the availability of genetic data and well-supported topological information, with ingroup sequences sourced from Salles et al. (2025) and additional *Tropiduridae* data, along with the reference topology, from Carvalho et al. (2024). Regarding ancestral states, here we considered two major substrate categories: rocks and trees. Furthermore, substrate use was modeled as a continuum, acknowledging that some species may exploit more than one substrate type. Ecological indices were constructed to estimate the proportion of substrate use by populations within a species (TABLE 3).

TABLE 3. Tropiduridae species included in ASR analyses. Species from the *Tropidurus spinulosus* group are highlighted in bold.

Species	States % (Rock, Tree)	References
Eurolophosaurus nanuzae	1, 0	Rodrigues (1981); Grizante et al. (2010)
Microlophus quadrivittatus	1, 0	Mella (2022)
Plica plica	0, 1	Vitt (1991); Grizante et al. (2010)
Strobilurus torquatus	0, 1	Rodrigues et al. (1989); Grizante et al. (2010)
Tropidurus bogerti	1, 0	Carvalho (2013)
Tropidurus callathelys	1, 0	Harvey & Gutberlet (1998); Carvalho (2013)
Tropidurus erythrocephalus	1, 0	Grizante et al. (2010)
Tropidurus guarani	1, 0	Carvalho (2013)
Tropidurus itambere	1, 0	Van Sluys (1993); Grizante et al. (2010)
Tropidurus melanopleurus	1, 0	Perez-Mellado & de la Riva (1993)
Tropidurus mucujensis	1, 0	Grizante et al. (2010)
Tropidurus semitaeniatus	1, 0	Vitt (1993); Grizante et al. (2010)
Tropidurus sp. nov.	1, 0	Carvalho (in prep.)
Tropidurus spinulosus	0.1, 0.9	Vitt (1991); Vitt (1997); Grizante et al. (2010); Carvalho (2013)
Tropidurus tarara	0, 1	Carvalho et al. (2016)
Tropidurus teyumirim	1, 0	Carvalho et al. (2016)
Tropidurus xanthochilus	0, 1	Harvey & Gutberlet (1998); Carvalho (2013)
Uracentron azureum	0, 1	Ellinger et al. (2001); Grizante et al. (2010)
Uranoscodon superciliosus	0, 1	Vitt et al. (1997)

SOURCE: the author (2025).

These estimates were based on published information about the proportion of populations whose individuals are known to use a given substrate. Values ranged from zero (no individuals in any population use that substrate type) to one (all populations consistently use that substrate type), assigned separately for rocks and trees. For example, a population strictly associated with trunks and branches would yield an index of zero for rocks and one for trees, whereas a more generalist species with both arboreal and rock-dwelling populations might present indices of 0.5 for each category. These ecological assignments were corroborated by field observations of multiple Tropiduridae species by A.L.G. Carvalho. While such indices may be biased by population-level differences, we minimized this limitation by prioritizing available published habitat use data used in previous studies on the same subject, or that matched the ecological profiles described in the literature.

Additionally, it is important to mention that trunks and branches were grouped into a single “tree” category because, although these microhabitats can impose distinct biomechanical demands, they both represent structurally complex arboreal environments

(Grizante et al., 2010). Treating them as a unified category therefore allowed us to retain the focus of our analytical approach on the habitat shift hypothesis (between species that use either rocks or trees, irrespective of their microhabitat use within such trees) in order to reduce redundancy and retains degrees of freedom for each category in the comparative analyses. Moreover, sandy habitats were not considered as a substrate category in our framework because sand-dwelling lineages within Tropiduridae represent a potentially more derived trait of a phylogenetically restricted subset of species (e.g., species in the *Tropidurus torquatus* species group), with no direct relevance to the evolutionary history of the *T. spinulosus* group. Therefore, including sand or other substrate categories would have added noise rather than clarity to our comparative framework, given its absence in the clade of interest.

We used *phytools* to conduct all ASRs analyses, specifically with the 'ancr' and 'simmap' functions. We first performed model selection to estimate the best transition models among both the built-in models in *phytools*: equal rates (ER), symmetrical (SYM) and all rates different (ARD). We set transition rates for a given value for rate heterogeneity with maximum likelihood using the fitMk function in *phytools* (Revell, 2024). We then estimated the ancestral states with the 'ancr' function using model averaging Akaike Information Criterion (AIC) scores across all models. Then, we used the best model as the character transition rate for both marginal ancestral character estimation and stochastic character mapping (Huelsenbeck & Bollback, 2001) as implemented in the *make.simmap* function (Revell, 2024). This approach uses repeated simulation of character evolution across a tree to produce a posterior distribution of character states at all points on the phylogeny. In this case, we ran 1,000 simulations of evolution of each character to produce the posterior distributions.

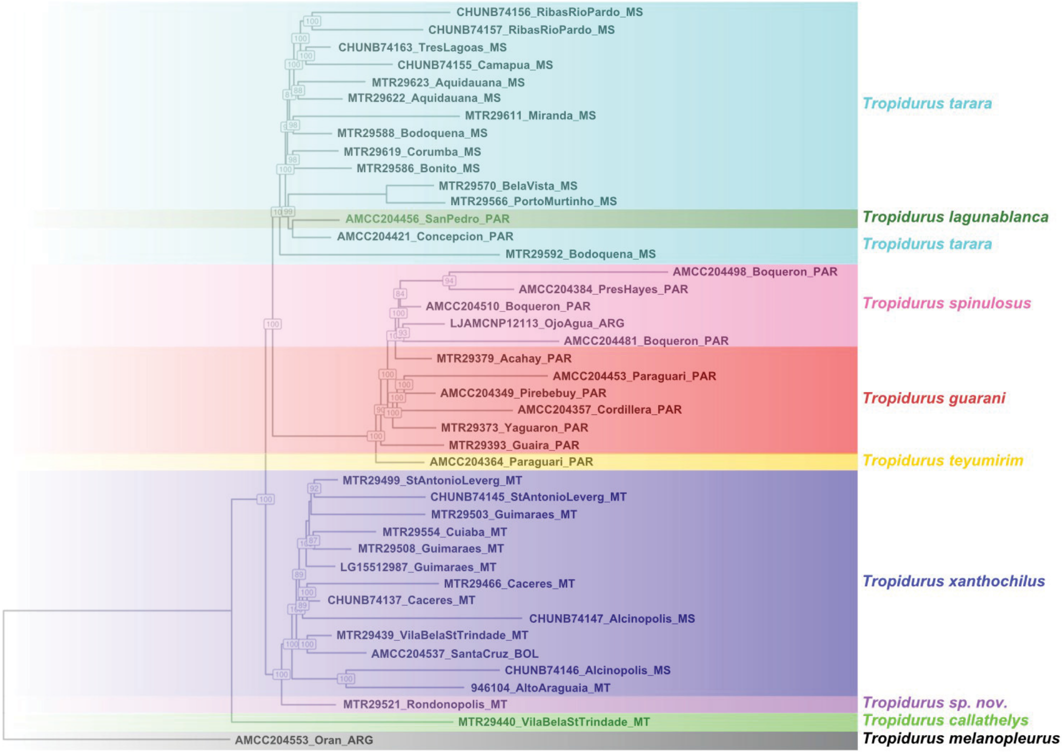
For these analyses, our dataset included 15 mitochondrial genes (13 protein-coding genes and 2 rRNAs), partitioned by gene and by codon position for coding sequences. Maximum likelihood analyses were performed in IQ-TREE v2.2.6 (Minh et al., 2020) with partition-specific substitution models selected using ModelFinder (Kalyaanamoorthy et al., 2017) under the -m TESTNEWMERGE option, allowing rate heterogeneity among partitions with the -spp option. Node support was evaluated with 1,000 ultrafast bootstrap replicates, considering values ≥ 95% as strong. Because ASR requires an ultrametric tree, we converted the resulting non-ultrametric tree using the *force.ultrametric* function in the *phytools* 2.1.1 package (Revell, 2024).

5.3 RESULTS

5.3.1 PHYLOGENOMIC RESULTS

After processing our data using the Phyluce pipeline (Faircloth 2016), we obtained a final dataset including 820 UCE loci for 43 individuals of nine lineages of the *Tropidurus spinulosus* species group. Our complete matrix comprised alignments with 725,505 bp after internal-trimming. Matrices with 75% and 95% of completeness were also generated, which, after internal-trimming, generated alignments with 2,269 loci 181otaling 1.98M bp, and 1,613 loci 181otaling 1.40M bp, respectively. Because the resulting topologies were virtually identical across all matrices, we chose to present results based on the 75% completeness matrix, which includes the largest number of loci, thereby maximizing the information content of our phylogenetic inference. However, topologies inferred from the other matrices are provided in the Supplementary Material for reference (FIGURES S1-S10). The nuclear phylogenetic analyses resulting from those matrices, using both a partitioned approach in the IQ-TREE software (Minh et al., 2020), and a species tree approach based on BPP (Flouri et al., 2018) (Supplementary Material), indicated that our samples can be grouped into three main clades: (*T. sp. nov.* + *T. xanthochilus), (T. lagunablanca* + *T. tarara), (T. guarani* + *T. spinulosus* + *T. teyumirim)*, in addition to *T. callathelys* as the sister lineage to these three, and *T. melanopleurus* as sister to all other lineages (FIGURE 8). Trees inferred using ASTRAL (Zhang et al., 2018), which accounts for incomplete lineage sorting (ILS), yielded congruent topologies (Supplementary Material). Notably, both analytical frameworks recovered *T. guarani* as paraphyletic, suggesting that it may represent either a paraphyletic species or a species complex in need of taxonomic revision. Furthermore, *T. lagunablanca* was found nested within samples of *T. tarara*, indicating that these two lineages likely represent a single evolutionary entity.

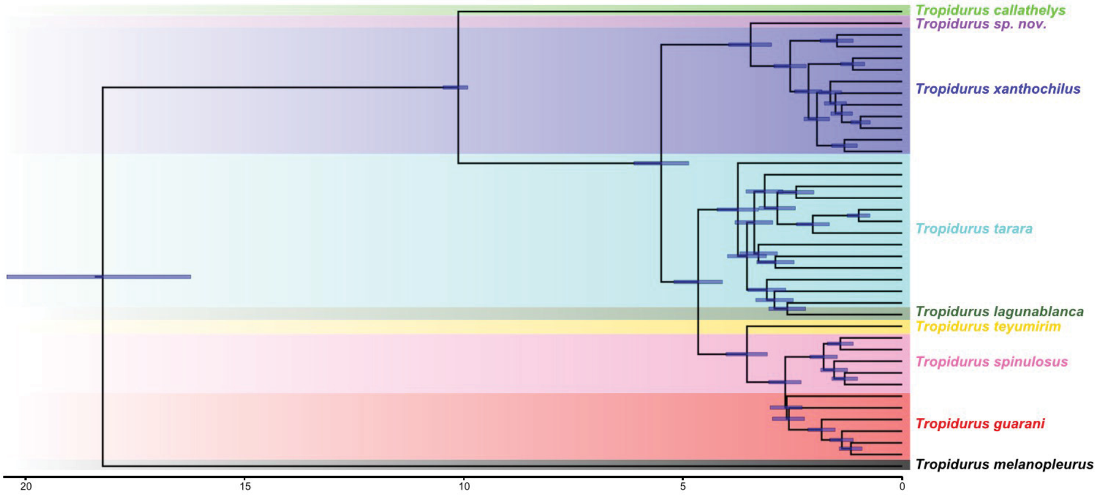
FIGURE 8. Partitioned phylogenetic analysis of the 2,269 UCE loci, matrix with 75% of completeness (via IQ-TREE). Numbers inside boxes refer to bootstrap support values (only nodes with support > 80 are shown).



SOURCE: the author (2025).

Divergence dating based on nuclear data indicates that the initial split within the *Tropidurus spinulosus* group occurred between 20–15 million years ago (Ma), separating *T. melanopleurus* from the other lineages. Subsequently, *T. callathely*s diverged around 10 Ma ago, and most subsequent divergence events among the remaining species in the group took place between 7.5–2.5 Ma (FIGURE 9; Supplementary Material). In most cases, divergence times estimated using RelTime (Tamura et al., 2018) fall within the same 95% highest posterior density (HPD) intervals provided by MCMCTree (dos Reis and Yang, 2019), although notable differences are also present. For instance, the clade comprising *T. guarani*, *T. spinulosus*, and *T. teyumirim* originated around 4 Ma according to the MCMCTree analysis, while RelTime estimates the divergence to be more recent, at approximately 3 Ma. In contrast, some node ages show notable discrepancies between the two methods. The divergence between *T. sp. nov.* and *T. xanthochilus* is estimated at approximately 2.5–3.5 Ma in the MCMCTree analysis, whereas RelTime dates the same split to about 7 Ma. Likewise, the origin of the *T. tarara* clade is estimated at ~3.75 Ma by MCMCTree but nearly 7.5 Ma by the RelTime analysis. The estimates provided by the species tree dated using MCMCTree provide intermediate values between the two analyses (Supplementary Material). In the absence of fossils and relevant ingroup secondary calibrations, and considering the fact that germline mutation rates are not available for our group, we opted to provide the comparisons among methods for the sake of clarity and in order to provide the actual evidence we were able to build from the available data.

FIGURE 9. Divergence times (Ma) estimated for the *Tropidurus spinulosus* species group using the UCE dataset. The 95% HPD interval of the posterior estimates (blue shaded bars), as estimated through the MCMCTree program, is shown above each node of the tree. The scale bar is in Ma.

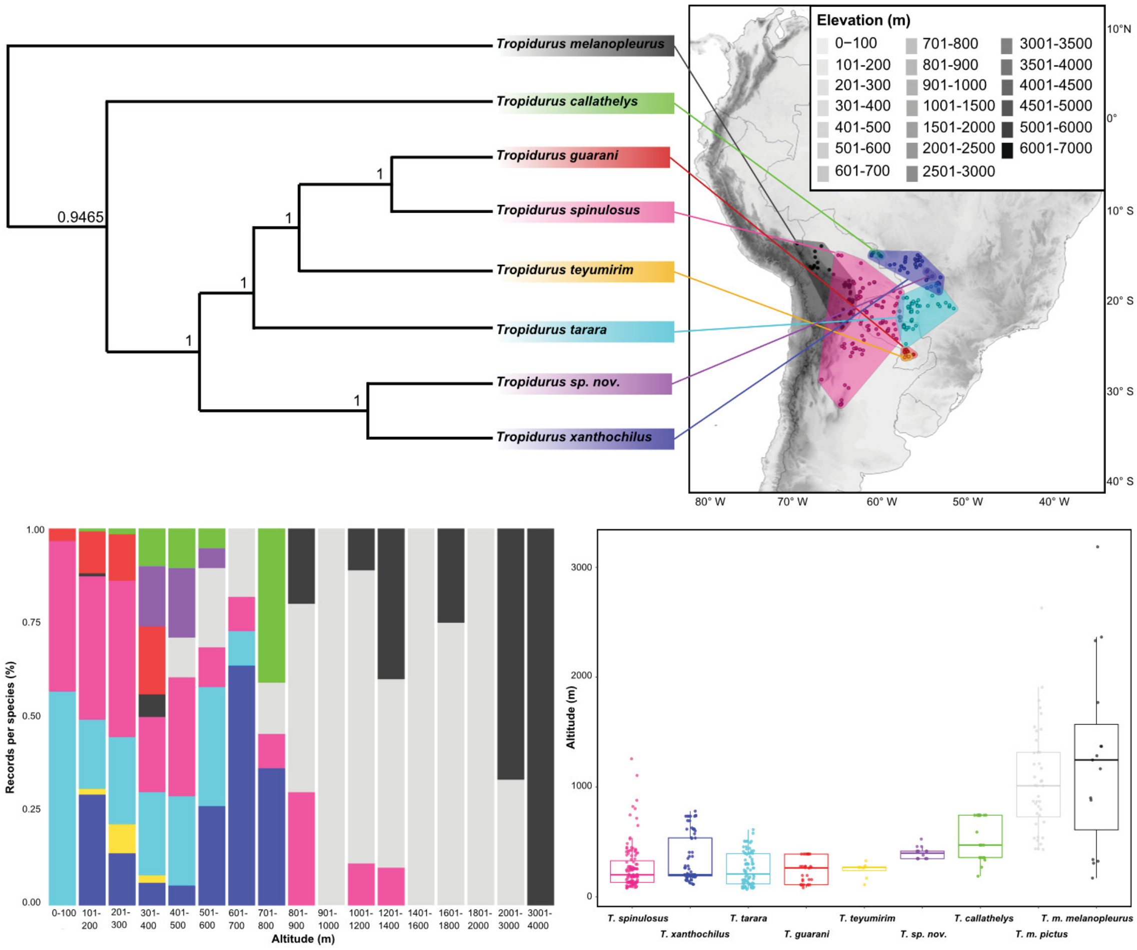


SOURCE: the author (2025).

5.3.2 SPATIO-ALTITUDINAL DISTRIBUTION PATTERNS

Our distribution records reveal that most species in the group predominantly occur below 1,000 m elevation, with *Tropidurus melanopleurus* representing the main exception (FIGURE 10). While the two *T. melanopleurus* subspecies (*T. m. melanopleurus* and *T. m. pictus*) exhibit overlapping elevational ranges, *T. m. melanopleurus* shows a stronger association with higher elevations, particularly above 2,000 meters. *Tropidurus spinulosus*, though occasionally recorded above 1,000 meters, primarily inhabits lowland areas below 500 meters, with higher elevation occurrences likely representing outliers. These high-elevation records correspond exclusively to populations in the Córdoba region, which are ecologically distinctive for being the only populations in the species that predominantly use rock outcrops. While genetic data from these populations are currently unavailable, their unique ecology and disjunct distribution suggest they may represent a distinct, cryptic lineage, a hypothesis that warrants future taxonomic investigation. Statistical analyses confirm significant elevational differentiation among species: Kruskal-Wallis tests (α = 0.05; non-parametric, as our data violated assumptions of parametric tests) revealed significant differences in elevation ranges, with Dunn's post-hoc tests (Bonferroni-corrected) showing significant differentiation in 16 of 36 pairwise comparisons between species. Notably, 10 of these significant comparisons involved at least one *T. melanopleurus* subspecies. Furthermore, when examining 100-meter elevation bands up to 1,000 meters, we found no evidence of sister species occupying distinct altitudinal zones, and no sister lineages showed significant elevational differentiation.

FIGURE 10. Altitudinal variation in the *Tropidurus spinulosus* group. The top-left panel shows the multilocus nuclear phylogeny inferred with BPP, with posterior probability values indicated at each node. The top-right panel displays species distributions overlaid on a South American elevation map; shaded polygons highlight areas of occurrence but are not intended as formal range estimates. The bottom-left panel presents the proportion of occurrence records per species across elevation bins. The bottom-right panel compares elevation distributions among species using box-and-jitter plots. In the two bottom panels, the subspecies of *T. melanopleurus* are shown separately; in the top panels, they are combined due to the availability of molecular data from only one individual.

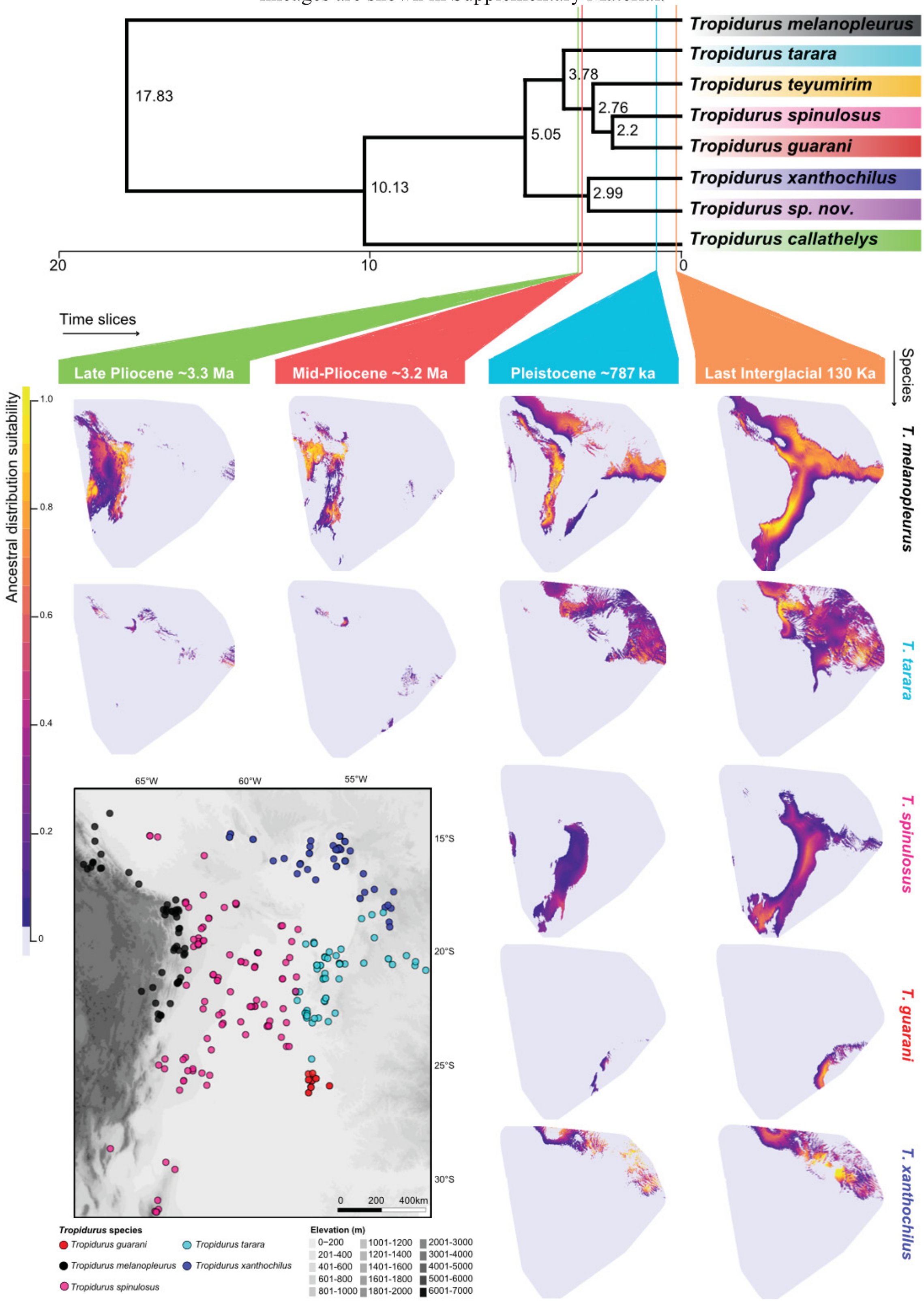


SOURCE: the author (2025).

5.3.3 CLIMATIC NICHE EVOLUTION AND ANCESTRAL DISTRIBUTION RECONSTRUCTION

Modeled present-day distributions partially align with the known geographic ranges of species within the *Tropidurus spinulosus* group but also indicate broad areas of suitable climatic conditions beyond their currently recognized occurrences (Supplementary Material, see section 'Ancestral climatic niche modelling results'). Based on these projections, ancestral distribution reconstructions are presented in FIGURE 11 and Supplementary Material. These estimates were generally more restricted compared to present-day ranges, especially during the Late and Mid-Pliocene (3.3–3.2 Ma), as observed for *T. melanopleurus* and especially *T. tarara*. It is important to note that not all extant lineages were present during these earlier periods. Over the past 787 Ka years, however, projections indicate substantial shifts and expansions in suitable habitats. Most lineages experienced eastward range expansions, although some projections, such as those for *T. melanopleurus*, appear inconsistent with the species' current distributions. The Last Glacial Maximum (130 Ka) marked another major expansion event, corresponding to the period of highest overall environmental suitability across all species.

FIGURE 11. Illustrative ancestral climatic niche models showing temporal variations in geographic displacements and range contractions/expansions among representative lineages within the *Tropidurus spinulosus* group, across four different time windows. The gradient bar in the top-left part of the figure represents the probability of occurrence of the lineages in their predicted areas. The box in the bottom-left part of the figure indicates the known distribution of the lineages. Predicted areas of the remaining lineages are shown in Supplementary Material.



SOURCE: the author (2025).

5.3.4 DEMOGRAPHIC MODEL SELECTION RESULTS

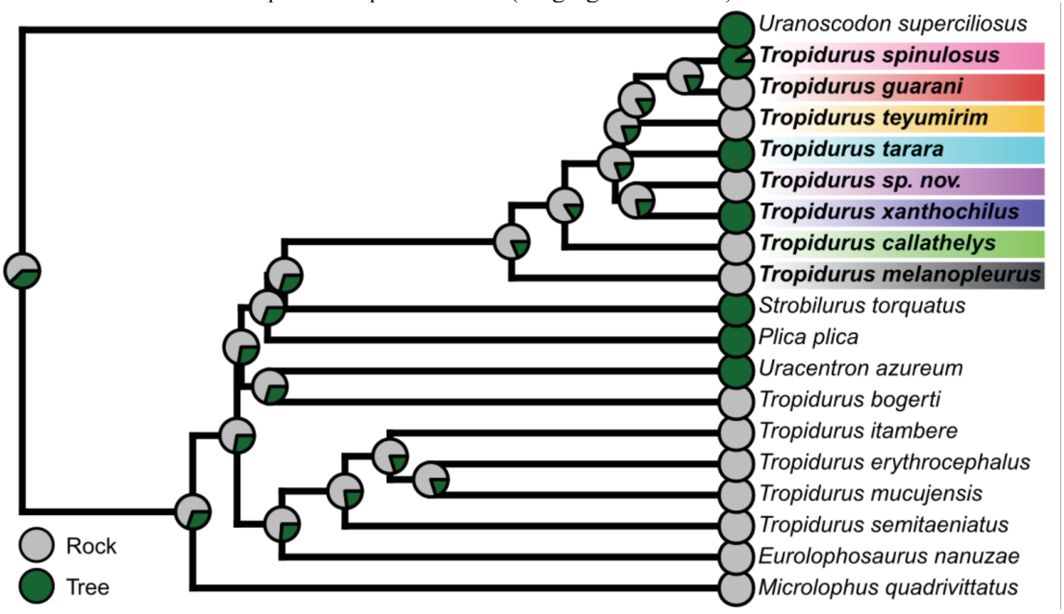
We further investigated the evolutionary history of *Tropidurus spinulosus* species, particularly their demographic history, in the ABC-RF analyses. At first, we assessed which sub-groups within our set of nine demographic scenarios best represented our UCE data set. In all simulations, principal component analysis (PCA) pre-scenario checks indicated that the observed dataset aligned well with the simulated dataset, suggesting that the analysis conditions were suitable for random forest analysis.

Through the first step of our simulation framework (using broad prior distributions and all nine scenarios), we identified the three best-fitting scenarios (scenarios 1, 4, and 7 in FIGURE 7) as those receiving the highest proportion of RF votes (cumulative proportion of 47.8% with 500 votes, and 45.7% with 1,000 votes). At this stage, the RF classifier showed limited ability to distinguish among these three models, as all received similar proportions of votes (approximately 10–20% each, regardless of whether 500 or 1,000 votes were used). Notably, these three scenarios consistently suggested relatively recent population growth in the evolutionary history of the species, occurring after the most recent divergence event considered in our modeling. In contrast, scenarios incorporating population declines or contraction–expansion cycles after the most recent divergence event received substantially less support. Following these results, we conducted additional ABC-RF analyses using new simulations with narrower prior distributions for these three scenarios. However, the RF classifier again showed limited discriminatory power. Even with 2,000 votes, the difference in support among the three models did not exceed 20%, with the best-performing model (Scenario 7) reaching only 0.4 posterior probability, indicating that our data probably does not provide enough power for model discrimination, given the statistical similarity of the three models.

5.3.5 ANCESTRAL MICROHABITAT RECONSTRUCTION RESULTS

ASR analyses indicate that the most recent common ancestor of the *Tropidurus spinulosus* species group was rock-dwelling, reflecting a conserved preference for lithic habitats among early Tropiduridae (FIGURE 12). From this rock-dwelling origin, at least three independent shifts to arboreal habitats occurred: one along the branch leading to the clade containing *T. spinulosus* and *T. tarara*, and another on the branch of *T. xanthochilus*. In both cases, their sister taxa remained rock-bound, demonstrating that these arboreal transitions were not isolated, ancient events but rather emerged separately. Early splits involved mostly rock-dwelling lineages, and substrate shifts did not coincide with initial cladogenesis events within the species group.

FIGURE 12. Phylogenetic tree showing ancestral state reconstruction of microhabitat preference (rock *vs.* tree) in Tropiduridae, with emphasis on the *Tropidurus spinulosus* species group (species in bold). Pie charts at the tips represent the observed states for each species, as listed in Table 1. Pie charts at internal nodes indicate the posterior probabilities (ranging from 0 to 1) for each ancestral state.



SOURCE: the author (2025).

5.4 DISCUSSION

5.4.1 NOTES ON THE SYSTEMATICS AND DISTRIBUTION OF THE *Tropidurus spinulosus* GROUP

Our phylogenetic results, particularly from the nuclear DNA dataset, are broadly consistent with previous systematic studies. Earlier morphological and molecular work (Frost et al., 1998) identified *Tropidurus spinulosus* and "*T. guarani"* as forming a clade, while subsequent analyses (Carvalho, 2016) revealed distinct allopatric morphotypes within populations previously assigned to *T. guarani,* leading to the formal description of *T. lagunablanca*, *T. tarara*, and *T. teyumirim.* Notably, our analyses confirm this pattern and further indicate that *T. guarani* is not monophyletic, suggesting a more complex evolutionary history for this taxon, that could either represent a paraphyletic species or a species complex in need of revision.

Tropidurus guarani is distributed east of the Paraguay River, whereas *T. spinulosus* occurs to the west, mainly within the Gran Chaco. With a broader distribution, *T. spinulosus* ranges from north-central Argentina and western Paraguay into southeastern and central Bolivia, where it occupies both xeric Chaco scrub and adjacent dry forests (Frost et al., 1998; Carvalho, 2013). This distributional pattern highlights the role of the Paraguay River as a biogeographic barrier. This holds true for at least a few other organisms such as snakes (Arzamendia & Giraudo, 2009) and birds (Kopuchian et al., 2020), indicating that the Paraná–Paraguay River system may have shaped regional biogeographic patterns by acting as a semi-permeable barrier that limited dispersal and contributed to the delineation of distinct ecoregions. Notably, the estimated divergence time between *T. guarani* and *T. spinulosus* (~2.5 Ma) coincides with the establishment of the river's modern course (Brea & Zucol, 2011; Carvalho & Albert, 2011; Schaefer, 2011), further supporting its role in promoting lineage divergence.

Our multilocus phylogenies further confirm that *Tropidurus tarara*—originally described from seasonally flooded savannas in northeastern Paraguay and recently confirmed in several localities east of the Paraguay River in the State of Mato Grosso do Sul, Brazil—occupies a distinct evolutionary position within the *T. spinulosus* species group. Contrary to earlier classifications (e.g., Alvarez et al., 1994; Frost et al., 1998) that assigned the populations later described by Carvalho (2016) as *T. tarara* to the nominal species *T. guarani*, our analyses robustly recover *T. tarara* as the sister lineage to a clade containing both *T. guarani* (*sensu lato*) and *T. spinulosus*, underscoring its independent evolutionary history despite geographic proximity to both lineages (Frost et al., 1998; Carvalho, 2016). Furthermore, our data indicate that *T. lagunablanca*, only known from its type locality at the Reserva Natural Laguna Blanca, Department of San Pedro, Paraguay (Carvalho, 2016), fails to resolve as a monophyletic unit and instead clusters within the *T. tarara* lineage. These findings suggest that *T. lagunablanca* does not represent a distinct evolutionary entity. In light of these results—and considering the recently confirmed overlap in morphological diagnostic traits (A.L.G. Carvalho, unpublished data)— we synonymize *T. lagunablanca* Carvalho, 2016 with *T. tarara* Carvalho, 2016. A more detailed assessment of this synonymization, including a comparative morphological analysis of *Tropidurus lagunablanca* Carvalho, 2016 (herein recognized as a junior synonym of *Tropidurus tarara* Carvalho, 2016) and critical analysis of the taxonomic literature, will be presented in a separate study by A.L.G. Carvalho.

Tropidurus xanthochilus was long known only from its type locality near the Serranía de Huanchaca in eastern Bolivia (Harvey & Gutberlet, 1998). However, our analyses reveal that *T. xanthochilus* is more widely distributed, occurring throughout the semi-deciduous forests of eastern Bolivia and the southwestern portion of the Brazilian State of Mato Grosso. Although Harvey and Gutberlet (1998), based on the few Bolivian records available at the time, noted that the nearest confirmed *T. spinulosus* group populations occurred at least 350 km (= *T. spinulosus*) to the south of *T. xanthochilus*, they nonetheless hypothesized a potential parapatric boundary along the transition between the Tarvo and Paraguá forest enclaves, which remain to be confirmed. Currently, what we know for a fact is that the species' type locality is within the extreme western limit of its range, with the bulk of its distribution located eastward, in Brazil. Our phylogenetic analyses also confirm *T. xanthochilus* as a divergent lineage relative to the clade comprising *T. guarani*, *T. spinulosus*, *T. tarara*, and *T. teyumirim*, underscoring its unique evolutionary trajectory. Furthermore, the taxon referred to here as *Tropidurus sp. nov.*, currently in the process of formal description (Carvalho et al., in prep.), is recovered as the sister taxon to *T. xanthochilus*. Although only a single individual of *Tropidurus sp. nov.* was included in our analyses, this sampling limitation does not undermine the inference of its distinctiveness. The two taxa occur sympatrically yet are morphologically diagnosable and exhibit contrasting habitat preferences (Nonato, 2018), providing strong evidence that they represent independent evolutionary lineages rather than variants of a single species. Moreover, independent species-delimitation analyses using BPP (Salles et al., 2025)—a framework known to yield reliable results even when sampling is restricted to one specimen per lineage—also supported the separation of *Tropidurus sp. nov.* and *T. xanthochilus* as distinct species. Together, these lines of evidence substantially reinforce their recognition as separate, well-delimited lineages.

The Serranía de Huanchaca region harbors both *Tropidurus callathelys* and *T. xanthochilus* in sympatry, despite these taxa not being sister species. *Tropidurus callathelys* is restricted to rocky outcrops along cliffs, borders, and waterfalls within the Serranía, whereas *T. xanthochilus* is strictly arboreal and occurs more broadly across the northeastern edge of the Pantanal Basin in the Brazilian states of Mato Grosso and Mato Grosso do Sul (Harvey & Gutberlet, 1998; Carvalho, 2013). Their co-occurrence in the same region but in distinct habitats provides compelling evidence against niche conservatism as a primary driver of diversification within the *T. spinulosus* group. Instead, the marked ecological divergence between these phylogenetically distant lineages highlights ecological plasticity and niche differentiation as key elements shaping the group's evolutionary history. This pattern of contrasting niches facilitating coexistence aligns with the broader finding that the combined role of dispersal and niche evolution underpins diversification in Neotropical lizards (see Sheu et al., 2020).

Finally, our data indicates that broad elevational gradients were not the primary drivers of speciation within the *T. spinulosus* clade. For instance, our phylogenetic framework places *T. callathelys* as the sister taxon to all members of the *T. spinulosus* group, except *T. melanopleurus*, which is the earliest diverging species. These two taxa exhibit allopatric distributions, with *T. callathelys* confined to eastern Bolivia and *T. melanopleurus* distributed across southern Peru and the Andean foothills of Bolivia and northern Argentina (Carvalho, 2013). While *T. melanopleurus* and, to a lesser extent, *T. spinulosus* are the only species in the group with occurrences above 1,000 m elevation, statistical comparisons revealed no significant elevational differentiation among sister species (FIGURE 10). This pattern persists despite the existence of major altitudinal gradients shaped by Neogene orogeny and Pleistocene climatic fluctuations. Together, these results indicate that broad elevational or orogenic events did not drive speciation within the group. However, localized elevational heterogeneity may still have facilitated population differentiation within specific areas. Therefore, local ecological patterns (such as fine-scale niche evolution, the availability of specific microhabitats, vegetation coverage, soil type, and thermal microenvironments), whether related to elevational gradients or not, would complement the broader geoclimatic processes considered in this study.

5.4.2 DIVERGENCE TIMES RESULTS

Divergence-time estimates reveal a long and temporally heterogeneous history for the *Tropidurus spinulosus* group. The deepest split, separating *T. melanopleurus* from the rest of the clade, dates to the early–middle Miocene (~20–15 Ma). This was followed by the divergence of *T. callathelys* at ~10 Ma, while most subsequent speciation events clustered between ~7.5 and 2.5 Ma (FIGURE 9; Supplementary Material). Specifically, the early Miocene divergence of *T. melanopleurus* aligns temporally with the initial phase of the Paranaense Sea incursion (~18–13 Ma; Hernández et al., 2005; Miller et al., 2005; del Río et al., 2018), which likely fragmented lowland habitats and promoted initial lineage isolation in the group. The subsequent emergence of *T. callathelys* near the Middle Miocene (~10 Ma) likewise coincides with later transgressive phases of the Paranaense Sea (12–6 Ma; del Río et al., 2018), which probably isolated upland populations east of the Chaco–Paraná Basin and set the stage for later north–south differentiation.

The extended interval between the divergence of *Tropidurus melanopleurus* and the later burst of diversification in the group implies a more complex history than a simple, gradual accumulation of species. This pattern is consistent with two nonexclusive scenarios: (1) a substantially older origin for the group than previously assumed, as divergence times estimated for other *Tropidurus* species groups such as *T. semitaeniatus* range from ~3 to 6 Ma (Werneck et al., 2015), or (2) the loss of multiple, now-undetected sister lineages (i.e., extinct lineages). In either case, *T. melanopleurus* emerges as a deeply diverged relict, and *T. callathelys* appears as another early-branching lineage that precedes a speciation pulse later on. Collectively, these patterns point to a temporally heterogeneous diversification history, punctuated by intervals of stasis or lineage loss, followed by several episodes of cladogenesis.

Extinction therefore remains a plausible contributor to both the temporal gap and the structural phylogenetic asymmetry observed in our tree. However, because no reliable fossil record exists for *Tropidurus* and most comparative or macroevolutionary approaches capable of estimating extinction require substantially denser taxon sampling than currently available, we cannot robustly quantify the magnitude or timing of these losses. As a result, our phylogenetic inferences should be interpreted with appropriate caution. Future work incorporating broader taxonomic sampling (including closely related congeners and geographically isolated populations), along with diversification models that explicitly accommodate extinction and incomplete sampling, will be essential for assessing how lineage loss may have influenced the evolutionary history reconstructed here.

Either way, while broad geoclimatic processes help explain the timing of deep splits and radiations, they do not exclude the influence of fine-scale ecological and altitudinal factors operating locally within populations. Later divergence events (~7.5–4 Ma) coincide not only with the Paranaense Sea regression, but also with late Neogene tectonic activity, including Andean uplift pulses and the reorganization of drainage basins (Garzione et al., 2008); all processes that likely altered habitat connectivity and elevational gradients, thus facilitating differentiation among populations/species. In addition, regressive phases of marine incursions, together with progressive aridification of the Chaco after ~5 Ma (Uliana & Biddle, 1988; Ortiz-Jaureguizar, 1998; Ortiz-Jaureguizar & Cladera, 2006), probably opened dispersal corridors that enabled secondary contact among previously isolated lineages. These dynamics likely contributed to the mitonuclear discordance observed within the group (Salles et al., 2025). Incongruence between the nuclear and mitochondrial genomes, as observed here, is commonly reported in squamates and often attributed to introgressive mitochondrial capture. In line with the established role of Pleistocene range dynamics in South America (Prado et al., 2012; Ledo et al. 2020), our results identify these same processes as likely contributors to the mitonuclear discordance in the *Tropidurus spinulosus* group.

Finally, although MCMCTree and RelTime produce discrepant absolute ages for some splits involving *Tropidurus* sp. nov., *T. xanthochilus*, and *T. tarara*, both approaches agree that the *T. guarani–T. spinulosus–T. teyumirim* clade diversified around ~3 Ma. Notably, the estimated divergence between *T. guarani* and *T. spinulosus* (~2.5 Ma) coincides with the establishment of the modern Paraguay River course (Brea & Zucol, 2011; Carvalho & Albert, 2011; Schaefer, 2011), highlighting the likely role of this river in promoting species divergence within the group, as previously discussed. Furthermore, this timing corresponds to geomorphological changes surrounding the Brazilian Central Plateau—including deepening peripheral depressions and progressive uplift to approximately 1,700 m (Ab'Sáber, 1983; Del'Arco & Bezerra, 1989)—implicating localized tectonic uplift and differential basin subsidence at the Pantanal–Cerrado transition in the creation of elevational barriers and novel dispersal routes that might have influenced these speciation events.

To further refine the diversification scenario within the *Tropidurus spinulosus* group, future studies should focus on particular populations within each of these more recently diverged species. A promising approach would be to apply high-resolution genomic data at a finer, population-level scale, combined with detailed analyses of local geological and topographic features. This framework would allow testing how fine-scale topographic variability—operating at the level of individual mountain ranges, river valleys, and other geological features, rather than the broader plateau scale considered here—has influenced gene flow, genetic diversity, and population structure. For example, the paraphyly observed in *T. guarani* underscores the potential importance of fine-scale altitudinal heterogeneity (populations structured on isolated plateaus with distinct geological histories) in shaping the complex evolutionary trajectory of the group. By correlating genetic divergence with specific geomorphological characteristics, such studies could clarify the precise mechanisms of vicariance and dispersal that have contributed to population fragmentation and incipient speciation within the clade.

5.4.3 ANCESTRAL DISTRIBUTION RECONSTRUCTION AND BIOGEOGRAPHIC HISTORY OF *Tropidurus spinulosus* SPECIES

The glacial–interglacial cycles of the Pleistocene had a profound impact on the genetic dynamics of South American vertebrates (Turchetto-Zolet et al., 2013; Guillory et al., 2024), and our results suggest that the *Tropidurus spinulosus* species group was similarly affected. In higher-latitude regions, glacial periods typically drove range contractions into refugial areas, followed by re-expansion into newly available habitats during interglacials (Ledo et al., 2020). While forest-dependent taxa generally conform to this classic contraction–expansion pattern, taxa adapted to open habitats (such as the lineages within the *T. spinulosus* group) often exhibit more idiosyncratic and spatially variable responses (Werneck et al., 2012; Miranda et al., 2019). It is important to note, however, that the notion of "ice-free refugia" is not directly applicable to the geological context considered in this study, as the central regions of South America examined here were never covered by continental glaciers. Although Pleistocene glacial cycles exerted broad climatic effects, extensive ice formations were largely restricted to North America, the southern tip of South America, and the highest elevations of the Andes (Glasser et al., 2008; Rodbell et al., 2022; Darvill, 2024). Within this framework, paleoclimatic and landscape dynamics (particularly expansions and contractions of savannas, forests, and wetlands) likely represented the primary drivers of ecological and evolutionary change, shaping patterns of population structure, connectivity, and diversification in lineages inhabiting these areas.

In this context, our ancestral distribution projections indicate a broad climatic envelope for *Tropidurus* across South America from ~730 Ka onward, with population-scale expansions intensifying in the last ~130 Ka (FIGURE 11). To some extent, these Pleistocene range expansions align with our null hypothesis (H0), suggesting they became possible once land corridors became available after extended isolation imposed by marine transgressions (from ~3 Ma onwards). However, it is important to note that these projections of climatic suitability represent a potential, not realized, distribution. For instance, a clear limitation of our models is that they do not incorporate non-climatic factors that fundamentally constrain the realized niche of these lizards. For the *T. spinulosus* group, these limitations include: (a) microhabitat specificity, in which saxicolous taxa require exposed rock outcrops with crevices and thermoregulatory sites, while arboreal taxa like this depend on areas with sparse vegetation that allow high solar exposure and easy access to basking sites; (b) geomorphological context with presence of specific landforms (e.g., escarpments, plateau margins, fluvial terraces) that generate the required substrates; (c) dispersal barriers, such as landscape matrices (e.g., wetlands, savannas) that act as conduits or barriers, the arrangement of habitat "stepping stones," and hard barriers like major rivers, extensive unsuitable lowlands, or persistent deposits from marine incursions; and (d) biotic interactions, such as competition from earlier-diverging congeners, predation pressure, and parasitism.

Consequently, while inferred climatic overlap—such as the Pleistocene co-occurrence of *Tropidurus xanthochilus*, *T. spinulosus*, and *T. tarara*—provides a plausible framework for potential secondary contact, actual gene flow would have required the simultaneous availability of both suitable climate and these specific ecological and geographical contexts. Furthermore, ancestral distribution models based on paleoclimatic data face the additional limitation of coarse resolution, which tends to inflate apparent suitable areas (Blois et al., 2025). These combined factors (the omission of non-climatic constraints and coarse resolution) explain mismatches between our projections and the empirical record, such as the suitable climate projected east of the real range of *T. melanopleurus*, where the absence of the specific microhabitats used by the species, mainly vertical rocky walls and riverbank cliffs, together with potential geographic barriers, likely prevented actual colonization. Therefore, we treat our climatic-based projections as hypotheses of potential range. To move from potential to realized distribution and more precisely evaluate scenarios of past introgression, future work must integrate broader geomorphological data, fossil data, and historical land-cover reconstructions to create fine-scale habitat masks for ancestral distribution modelling.

In addition, our current paleoclimate reconstructions (derived from PaleoClim data; Brown et al., 2018) are limited to only four time-intervals (3.2 Ma, 787 Ka, 130 Ka, and present), leaving critical gaps (especially between 3.2 Ma and 787 Ka) that hinder robust inference for pre-Pleistocene periods. Although tools such as Paleo Generate (Folk et al., 2023) can extrapolate PaleoClim variables to other time windows, their accuracy remains limited, having been validated for only a single variable (e.g., mean annual temperature, BIO1) and, provisionally, for a few others. In addition, platforms such as Oscillayers (Gamisch, 2019) provide continuous paleoclimate layers but have not gained widespread adoption, partly due to concerns (including from PaleoClim developers, see Brown et al., 2020) regarding their interpolation mechanisms and uncertain validation. These methodological limitations highlight a fundamental trade-off between expanding temporal coverage and maintaining the reliability of the reconstruction. With this in mind, we used more restrictive time windows rather than deal with even more methodological limitations. However, it is important that future studies address emerging frameworks, particularly those that aim to provide more integrative approaches (such as RRPhylogeography, see Mondanaro et al., 2025), even considering that these remain under development. Future work incorporating additional paleoclimatic layers, fossil occurrences and more accurate dispersal modeling techniques will be essential to refine ancestral-range estimates and disentangle the relative roles of abiotic forcing versus biotic constraints in this group's evolutionary history.

5.4.4 DEMOGRAPHIC HISTORY OF THE *Tropidurus spinulosus* SPECIES GROUP

Our study employed an ABC-RF framework to evaluate alternative demographic scenarios for the *Tropidurus spinulosus* species group. This approach is powerful for discriminating among complex models; however, our analysis revealed strong support for multiple scenarios, all characterized by relatively recent population expansion, while also highlighting significant challenges in precisely distinguishing demographic events. Although this outcome does not point to a single, unequivocal history, it offers profound insights into the evolutionary forces that shaped this group and allows us to evaluate the demographic expectations of our proposed hypotheses (TABLE 1).

The most consistent result from our model selection is the robust support for scenarios involving population expansion following the most recent divergence event within our dataset (Scenarios 1, 4, and 7; FIGURE 7). Empirically, this corresponds to expansions across the group after the divergence between *Tropidurus guarani* and *T. spinulosus*, which we estimate to approximately 2.5 Ma. This recurrent signal suggests that Pleistocene population growth is a fundamental feature of the group's evolutionary history. Crucially, this finding aligns with the demographic signature expected under H0 (Vicariant-Marine Hypothesis), which predicts expansions following the retreat of continental seas, and is also consistent with the expansion phase following a founding event under H2 (Ecological-Niche Shift Hypothesis).

The re-establishment of land corridors after Plio-Pleistocene marine regressions (~3 Ma onwards) likely provided the initial opportunity for range expansion. Subsequent Pleistocene climatic fluctuations, particularly the drier, cooler glacial periods, further reinforced these expansions by contracting humid forests and promoting the spread of open, arid habitats like savannas and dry woodlands (Ab'Sáber, 1998; Werneck, 2011)—

the preferred ecosystems for these heliophilous lizards. This timing also aligns with the establishment of the modern course of the Paraguay River, a major biogeographic barrier that separates sister species such as *T. guarani* (east of the river) and *T. spinulosus* (west, in the Gran Chaco). Originating on the Mato Grosso plateau and flowing south through the Pantanal, the Paraguay River attained its modern course during the Plio–Pleistocene, following Late Pliocene forebulge reactivation (~2.5 Ma) that deepened the Pantanal depression and redirected tributaries formerly draining into the upper Paraná and Tocantins; subsequent stream-capture and avulsion events during this period consolidated the river's present configuration (Ussami et al., 1999; Assine & Soares, 2004; Brea & Zucol, 2011; Carvalho & Albert, 2011).

As previously highlighted, large river systems across South America have repeatedly structured populations of vertebrates during the Neogene and Quaternary (Arzamendia & Giraudo, 2009; Kopuchian et al., 2020; Cassemiro et al., 2023). The Paraguay River fits within this broader pattern, acting not only as a physical barrier but also interacting with geological processes that shaped regional drainage reorganization. Thus, the divergence between *Tropidurus spinulosus* and *T. guarani* reflects a biogeographic mechanism recurrently documented across different taxa. In this context, the westward expansion of *T. spinulosus* may have been facilitated by the spread of xeric Chaco scrub and dry forests, habitats it was already preadapted to exploit following its earlier divergence from *T. guarani*. Thus, our models indicate that these combined geoclimatic changes created a broader, albeit compartmentalized, landscape that enabled rapid range expansions of previously isolated lineages. In sum, this process provides a clear mechanism for the phylogeographic pattern of expansion and potential secondary contact expected under both H0 and H2.

A key result is the limited ability of the RF classifier to decisively discriminate among the top-supported expansion scenarios. Methodologically, this underscores the challenge of distinguishing between scenarios that produce similar genetic summary statistics. Yet, this ambiguity can also be interpreted as a reflection of complex biological reality rather than a methodological failure. The similar genetic signatures suggest these demographic events may have been near-simultaneous, a scenario in which Pleistocene drivers triggered concurrent population growth in multiple lineages, on both sides of the Paraguay River, potentially under both post-incursion (H0) and ecological (H2) processes. Therefore, our demographic inference is greatly strengthened when interpreted synthetically. Our ABC-RF analysis provides evidence that expansions occurred, while our proposed hypotheses offer a plausible mechanism for *how* they likely unfolded.

5.4.5 MICROHABITAT PREFERENCE THROUGH THE EVOLUTIONARY HISTORY OF TROPIDURIDAE

Within the realm of H2, which proposes that diversification in the *Tropidurus spinulosus* species group may have been driven by repeated shifts in structural microhabutat, ancestral state reconstructions indicate that the group's most recent common ancestor was rock-dwelling (FIGURE 12). This result points to a conserved lithophilic preference shared across early Tropiduridae. From this ancestral saxicolous condition, at least three independent transitions to arboreal habitats occurred: one in *T. tarara*, another in *T. spinulosus*, and the last one in *T. xanthochilus*. In the latter two cases, the respective sister species (*T. guarani* and *T. sp. nov.*) retained the ancestral saxicolous condition, reinforcing the view that these arboreal shifts arose independently in response to local ecological conditions rather than persisting as remnants of earlier transitions. These repeated ecological shifts provide a compelling case study of how niche evolution, when combined with new dispersal opportunities, as previously discussed, can drive the diversification of a Neotropical lizard clade (also, see Sheu et al., 2020).

Morphological data provide strong evidence for adaptive divergence linked to substrate use in our focal group. In rocky habitats, for example, Tropidurinae species tend to have narrower foot soles, which may enhance grip and control on inclined surfaces (Grizante et al., 2010). Nonetheless, *Tropidurus* lizards generally exhibit lower morphological disparity than other iguanian clades (e.g., *Anolis*: Mahler et al., 2010), suggesting that generalized body plans may be adequate across structurally distinct environments (Kohlsdorf et al., 2001). Empirical comparisons between sister species reveal that locomotor performance can diverge rapidly, even without substantial skeletal differences, especially when species occupy contrasting substrates (Kohlsdorf et al., 2001). Such patterns reflect trade-offs in foot and limb traits tied to substrate specialization. Repeated morphological convergence among independently derived arboreal species further highlights how structural niche shifts promote fine-scale ecological specialization, as distinct habitats impose specific mechanical demands (Lejeune et al., 1998; Rocha, 1998). Phylogenetic evidence supports this view, showing that performance traits have evolved in association with substrate transitions and their associated functional trade-offs (Grizante et al., 2010). However, earlier comparative analyses relied on phylogenies that were taxonomically incomplete and topologically inconsistent with current, more robust hypotheses. Consequently, many fine-scale inferences from those studies require re-evaluation in light of updated phylogenetic frameworks.

It is important to emphasize, however, that observed shifts in substrate preference should not be conflated with primary speciation mechanisms, as local adoption of rupicolous or arboreal habits can represent secondary ecological responses to local opportunity or to unrecognized lineages rather than the initial driver of lineage splitting. Conversely, in much of the Chaco the near absence of rock outcrops suitable to *T. spinulosus* may likely prevent establishment of genuine saxicolous populations even if climatic suitability existed. Furthermore, the two pairs of sister species exhibiting contrasting substrate preferences—*T. xanthochilus* (arboreal) + *T. sp. nov.* (saxicolous), and *T. spinulosus* (arboreal) + *T. guarani* (saxicolous)—likely differ in their evolutionary contexts. For example, the divergence between *T. spinulosus* and *T. guarani* can be best interpreted in a vicariant framework: their largely allopatric distributions separated by the Paraguay River indicate that geographic isolation might have preceded ecological divergence, and the subsequent colonization of arboreal habitats by *T. spinulosus* on the western side of the river appears to have been a secondary opportunity that facilitated range expansion rather than the initial speciation trigger. In other words, ecological divergence in this species pair likely accentuated differentiation after isolation rather than initiated it. Our demographic modeling, which explicitly tested for post-split expansion signatures, provides further evidence for this process, indicating that rapid expansion followed habitat shifts in both *T. guarani* and *T. spinulosus*. The same causal relationship, with substrate shift following, or driving, speciation, may also apply to the other sister pair (*T.* sp. nov. – *T. xanthochilus*), although this remains to be explicitly tested.

It is not possible to address the temporal dimension of substrate preference evolution across all *Tropidurus spinulosus* species based on our results. This is because, particularly in cases of independent arboreal habit acquisition, we find significant discrepancies in divergence time estimates—for example, *T. tarara* is dated at ~ 3.75 Ma by MCMCTree but nearly 7.5 Ma in RelTime analyses. Moreover, our ASR analyses rely on mitochondrial data, which derive substantially different branch lengths compared to nuclear data. Either way, in all instances, it is evident that these transitions were followed by major demographic expansions observed in these species, which took place largely within the last 200 Ka (see section 3.3.3; Salles et al., 2025). Taken together, these temporal and demographic patterns suggest that niche shifts often occurred while populations were smaller and that such shifts were likely consolidated by subsequent demographic expansions into newly available habitats. Thus, substrate transitions could have facilitated the fixation and morphological differentiation of traits related to substrate preference after lineage splitting rather than acting as the primary speciation trigger in all cases.

Furthermore, it is important to highlight that these shifts may be associated with the emergence of novel habitats as Miocene marine incursions receded, especially from 5 Ma onward. Notably, evidence of secondary contact and potential ancestral gene flow involves these same species (Salles et al., 2025), which share the same structural niche. This implies that niche similarity (i.e., shared arboreality) may have facilitated genetic exchange during periods of geographic overlap, underscoring the ecological influence on their evolutionary paths. Nevertheless, the directionality and timing of ecological shifts relative to geographic isolation likely varies among species pairs, and resolving whether habitat shifts were causes or consequences of speciation will require targeted phylogeographic and demographic tests at finer population levels.

Finally, although informative, our results reflect only one ecological axis captured in our analyses, namely structural microhabitat use. This proxy provides a meaningful window into habitat structure and its mechanical demands, but it represents just a subset of the broader ecological landscape potentially relevant to diversification in the group. Other ecological dimensions (such as microclimatic requirements, and behavioral specializations) were not assessed here and may similarly contribute to lineage divergence or to the reinforcement of differentiation after isolation. These additional lines of evidence should therefore be incorporated into future comparative and phylogeographic studies to more fully evaluate the ecological context of diversification within the *Tropidurus spinulosus* group.

5.5 CONCLUSION

The diversification of the *Tropidurus spinulosus* species group reflects a complex interplay of geological, climatic, and ecological forces, acting over millions of years. Our divergence-time estimates place the earliest divergence events between ~18–7.5 Ma, mostly coinciding with transgressions events of the Paranaense Sea. These marine incursions fragmented once-continuous populations along eastern and western South America, driving initial vicariant divergence through cycles of isolation, range contraction, and secondary contact as land corridors reemerged. While this vicariance scenario explains the group's foundational splits, it cannot fully account for subsequent speciation events.

Later geoclimatic events, such as Neogene orogenic pulses and Pleistocene glacial–interglacial cycles, may have introduced elevational gradients across the group's distribution. However, we found no consistent altitudinal segregation among sister taxa, suggesting that these factors were not the primary drivers of diversification. Pleistocene climatic oscillations, particularly over the last ~787 Ka (with major expansions around ~130 Ka), had a more localized impact on the evolutionary history of the *Tropidurus spinulosus* group, shaping patterns of demographic change, admixture, and introgression. Climatic fluctuations that occurred in this period, particularly the drier, cooler glacial periods, might have reinforced these expansions by contracting humid forests and promoting the spread of open, dry habitats like savannas and dry woodlands, which are among the preferred ecosystems for several species in this group. Notably, overlapping ranges and ecological niches among the arboreal species *T. spinulosus*, *T. tarara*, and *T. xanthochilus* likely facilitated ancient introgression and related processes. However, these appear to reflect secondary genetic dynamics rather than the main forces behind species divergence.

Furthermore, our results indicate that the Paraguay River system has shaped biogeographic patterns by acting as a semi-permeable barrier, restricting dispersal and contributing to the delineation of species limits. This is particularly evident in the divergence between *Tropidurus guarani* and *T. spinulosus*, sister lineages with allopatric distributions on opposite sides of the river. Their estimated divergence time (~2.5 Ma) coincides with the establishment of the river's modern course, supporting its role as a biogeographic boundary. The allopatric distributions of these species further suggest that geographic isolation preceded ecological divergence in the group. Specifically, ancestral-state reconstructions support a rock-dwelling most recent common ancestor for the clade and identify three independent transitions to arboreality, in *T. tarara*, *T. spinulosus*, and *T. xanthochilus*. While these transitions may have reinforced divergence later in evolutionary history (by contributing to reproductive isolation or adaptive differentiation), they may also have occurred after the initial divergence events. In this view, ecological adaptations would represent subsequent trait shifts that added complexity to the speciation continuum rather than serving as its primary drivers.

In *T. spinulosus* (arboreal) and *T. guarani* (saxicolous), allopatric distributions across the Paraguay River suggest that ecological differentiation followed geographic isolation, making it secondary to divergence. In other words, the later colonization of arboreal habitats by *T. spinulosus* on the river's western side likely facilitated range expansion rather than initiating speciation. By contrast, *T. xanthochilus* (arboreal) and *Tropidurus sp. nov.* (saxicolous), sympatry, recent divergence, and ecological distinctiveness suggest that substrate preference may have directly contributed to speciation. Either way, assuming these transitions occurred early in each lineage's history, they would predate the major demographic expansions observed in these species, most of which unfolded within the last 2.5 Ma. Although divergence time estimates vary across analyses, the prevailing pattern is that species origins are older than this, whereas large-scale demographic expansions took place mainly during the Pleistocene. This temporal framework suggests that niche shifts took place when population sizes were still small, followed by expansions into novel habitats—a dynamic that can facilitate the fixation of new ecological strategies. Collectively, these cases highlight that the contribution of ecological transitions to speciation is lineage-specific and contingent on biogeographic context and timing.

In sum, no single model (vicariance mediated by marine incursions, altitudinal isolation, niche shifts, or others) fully explains the group's trajectory. The marine incursion framework, while presenting limitations (e.g., range and landscape uncertainty), remains productive for biogeographic research on South American squamates. Ultimately, in *Tropidurus spinulosus*, diversification was initially shaped by ancient marine dynamics, then modulated by Neogene orogenetic processes and fluvial reorganization, with subsequent Pleistocene climate fluctuations refining demographic and phylogeographic patterns. Substrate shifts to arboreality reinforced niche divergence in specific lineages but may have operated as secondary processes, as they arose independently, possibly postdated cladogenetic events, and reflected local adaptations rather than primary speciation mechanisms. Future studies integrating more accurate paleoclimatic reconstructions, fossils, trait evolution, and fine-scale dispersal modeling will be essential to unravel how these intertwined forces generated one of South America's most distinctive lizard radiations.

5.6 REFERENCES

AB'SÁBER, A. N. O domínio dos cerrados: introdução ao conhecimento. **Revista do Serviço Público**, v. 40, n. 4, 1983.

AB'SÁBER, A. N. Participação das depressões periféricas e superfícies aplainadas na compartimentação do planalto brasileiro: considerações finais e conclusões. **Revista do Instituto Geológico**, v. 19, n. 1-2, p. 51-69, 1998.

ALEIXO, A. Historical diversification of a terra-firme forest bird superspecies: a phylogeographic perspective on the role of different hypotheses of Amazonian diversification. **Evolution**, v. 58, n. 6, p. 1303–1317, 2004.

ALHO, C. J. R. Biodiversity of the Pantanal: response to seasonal flooding regime and to environmental degradation. **Brazilian Journal of Biology**, v. 68, p. 957-966, 2008.

ÁLVAREZ, B. B.; CEI, J. M.; SCOLARO, J. A. A new subspecies of Tropidurus spinulosus (Cope, 1862) from the subtropical wet mesic Paraguayan region (Reptilia Squamata Tropiduridae). **Tropical Zoology**, v. 7, n. 1, p. 161–179, 1994.

ANDREWS, S. FastQC: A quality control tool for high throughput sequence data [software]. 2010.

ANTONELLI, A.; et al. Tracing the impact of the Andean uplift on Neotropical plant evolution. **Proceedings of the National Academy of Sciences**, v. 106, n. 24, p. 9749–9754, 2009.

ANTONELLI, A.; et al. Conceptual and empirical advances in Neotropical biodiversity research. **PeerJ**, v. 6, p. e5644, 2018.

ARZAMENDIA, V.; GIRAUDO, A. R. Influence of large South American rivers of the Plata Basin on distributional patterns of tropical snakes: a panbiogeographical analysis. **Journal of Biogeography**, v. 36, n. 9, p. 1739-1749, 2009.

ASSINE, M. L.; et al. Geology and geomorphology of the Pantanal basin. In: DITT, C.; PIEN, E. (org.). Dynamics of the Pantanal wetland in South America. p. 23–50, 2015.

ASSINE, M. L.; SILVA, A. Contrasting fluvial styles of the Paraguay River in the northwestern border of the Pantanal wetland, Brazil. **Geomorphology**, v. 113, n. 3-4, p. 189-199, 2009.

ASSINE, M. L.; SOARES, P. C. Quaternary of the Pantanal, west-central Brazil. **Quaternary International**, v. 114, n. 1, p. 23–34, 2004.

ÁVILA-PIRES, T. C. S. Lizards of brazilian amazonia (Reptilia: Squamata). **Zoologische verhandelingen**, 1995.

BANKEVICH, A.; et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. **Journal of Computational Biology**, v. 19, n. 5, p. 455–477, 2012.

BARREDA, V.; PALAZZESI, L. Patagonian vegetation turnovers during the Paleogene-early Neogene: origin of arid-adapted floras. **The Botanical Review**, v. 73, p. 31–50, 2007.

BARROS, V.; et al. The major discharge events in the Paraguay River: magnitudes, source regions, and climate forcings. **Journal of Hydrometeorology**, v. 5, n. 6, p. 1161–1170, 2004.

BERGERON, L. A.; et al. Evolution of the germline mutation rate across vertebrates. **Nature**, v. 615, n. 7951, p. 285–291, 2023.

BEZERRA, C. H.; et al. Biogeographical origins of Caatinga Squamata fauna. **Journal of Biogeography**, v. 52, n. 3, p. 521–531, 2025.

BLOIS, J. L. et al. Paleobiogeographic insights gained from ecological niche models: progress and continued challenges. **Paleobiology**, v. 51, n. 1, p. 8-28, 2025.

BLOOM, Devin D.; et al. The biogeography of marine incursions in South America. In: ALBERT, J. S.; REIS, R. E. (org.). Historical biogeography of Neotropical freshwater fishes. p. 137–144, 2011.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114–2120, 2014.

BOOTH, T. H.; et al. BIOCLIM: the first species distribution modelling package, its early applications and relevance to most current MAXENT studies. **Diversity and distributions**, v. 20, n. 1, p. 1-9, 2014.

BORBA, R. S.; et al. Genetic structure of the ornamental tetra fish species *Piabucus melanostomus* Holmberg, 1891 (CHARACIDAE, IGUANODECTINAE) in the Brazilian Pantanal wetlands inferred by mitochondrial DNA sequences. **Biota Neotropica**, v. 13, p. 42-46, 2013.

BOROWIEC, M. L. AMAS: a fast tool for alignment manipulation and computing of summary statistics. **PeerJ**, v. 4, p. e1660, 2016.

BOUBLI, J. P.; et al. Spatial and temporal patterns of diversification on the Amazon: a test of the riverine hypothesis for all diurnal primates of Rio Negro and Rio Branco in Brazil. **Molecular Phylogenetics and Evolution**, v. 82, p. 400–412, 2015.

BREA, M.; ZUCOL, A. F. The Paraná-Paraguay basin: geology and paleoenvironments. In: ALBERT, J. S.; REIS, R. E. (org.). Historical biogeography of Neotropical freshwater fishes. p. 69–88, 2011.

BROWN, J. L.; et al. PaleoClim, high spatial resolution paleoclimate surfaces for global land areas. **Scientific Data**, v. 5, n. 1, p. 1–9, 2018.

BROWN, Jason L.; HILL, Daniel J.; HAYWOOD, Alan M. A critical evaluation of the Oscillayers methods and datasets. **Global Ecology and Biogeography**, v. 29, n. 8, p. 1435–1442, 2020.

BUCKUP, Paulo A. The eastern Brazilian shield. In: ALBERT, J. S.; REIS, R. E. (org.). Historical biogeography of Neotropical freshwater fishes. p. 203–209, 2011.

BUSHNELL, B. BBTools: a suite of fast, multithreaded bioinformatics tools designed for analysis of DNA and RNA sequence data [software]. Joint Genome Institute, 2018.

CARVALHO, T. P.; ALBERT, J. S. The Amazon-Paraguay divide. In: ALBERT, J. S.; REIS, R. E. (org.). Historical biogeography of Neotropical freshwater fishes. p. 193–202, 2011.

CARVALHO, A. L. G. On the distribution and conservation of the South American lizard genus Tropidurus Wied-Neuwied, 1825 (Squamata: Tropiduridae). **Zootaxa**, v. 3640, n. 1, p. 42–56, 2013.

CARVALHO, A. L. G.; et al. Biogeography of the lizard genus Tropidurus Wied-Neuwied, 1825 (Squamata: Tropiduridae): distribution, endemism, and area relationships in South America. **PLOS One**, v. 8, n. 3, p. e59736, 2013.

CARVALHO, A. L. G.; et al. A new Tropidurus (Tropiduridae) from the semiarid Brazilian Caatinga: evidence for conflicting signal between mitochondrial and nuclear loci affecting the phylogenetic reconstruction of South American collared lizards. **American Museum Novitates**, v. 2016, n. 3852, p. 1–68, 2016.

CARVALHO, A. L. G. Three new species of the Tropidurus spinulosus group (Squamata: Tropiduridae) from eastern Paraguay. **American Museum Novitates**, v. 2016, n. 3853, p. 1-44, 2016.

CARVALHO, A. L. G; et al. A new collared lizard (Tropidurus: Tropiduridae) endemic to the Western Bolivian Andes and its implications for seasonally dry tropical forests. **American Museum Novitates**, v. 2018, n. 3896, p. 1-56, 2018.

CARVALHO, A. L. G.; et al. A highly polymorphic South American collared lizard (Tropiduridae: Tropidurus) reveals that open–dry refugia from Southwestern Amazonia staged allopatric speciation. **Zoological Journal of the Linnean Society**, v. 201, n. 2, p. 493–533, 2024.

CASTRESANA, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. **Molecular Biology and Evolution**, v. 17, n. 4, p. 540–552, 2000.

CASSEMIRO, F. A. S.; et al. Landscape dynamics and diversification of the megadiverse South American freshwater fish fauna. **Proceedings of the National Academy of Sciences**, v. 120, n. 2, p. e2211974120, 2023.

COLLIN, F. D. et al. Extending Approximate Bayesian Computation with Supervised Machine Learning to infer demographic history from genetic polymorphisms using

DIYABC Random Forest. **Molecular Ecology Resources**, v. 21, n. 8, p. 2598–2613, 2021.

COOKE, G M.; CHAO, N L.; BEHEREGARAY, L B. Marine incursions, cryptic species and ecological diversification in Amazonia: the biogeographic history of the croaker genus Plagioscion (Sciaenidae). **Journal of Biogeography**, v. 39, n. 4, p. 724–738, 2012.

CORDANI, U. G.; et al. The growth of the Brazilian Shield. **Episodes Journal of International Geoscience**, v. 11, n. 3, p. 163–167, 1988.

CUITIÑO, J. I.; et al. High-resolution isotopic ages for the early Miocene “Patagoniense” transgression in Southwest Patagonia: stratigraphic implications. **Journal of South American Earth Sciences**, v. 38, p. 110–122, 2012.

DARVILL, C. Late Pleistocene glaciation in South America. In: Encyclopedia of Quaternary Science. 2024.

DEL'ARCO, J. F.; BEZERRA, P. E. L. Geologia. In: DUARTE, A. C. (org.). Geografia do Brasil-Região Centro-Oeste. Rio de Janeiro: FIBGE-Diretoria de Geociências, 1989. p. 35–51.

DEL RÍO, C. J.; et al. Dating late Miocene marine incursions across Argentina and Uruguay with Sr-isotope stratigraphy. **Journal of South American Earth Sciences**, v. 85, p. 312–324, 2018.

DOMINGOS, F. M. C. B.; et al. In the shadows: phylogenomics and coalescent species delimitation unveil cryptic diversity in a Cerrado endemic lizard (Squamata: Tropidurus). **Molecular Phylogenetics and Evolution**, v. 107, p. 455–465, 2017.

DORADO-RODRIGUES, T. F.; et al. Herpetofauna from a protected area situated in a biogeographic transition zone in Central South America. **Biota Neotropica**, v. 25, n. 1, p. e20241681, 2025.

DOS REIS, M.; YANG, Z. Bayesian molecular clock dating using genome-scale datasets. In: EMELIEFOU, K.; STERLING, A. (org.). Evolutionary Genomics: Statistical and Computational Methods. p. 309–330, 2019.

ELLINGER, N.; SCHLATTE, G.; JEROME, N.; HÖDL, W. Habitat use and activity patterns of the neotropical arboreal lizard *Tropidurus* (*=Uracentron*) *azureus werneri* (Tropiduridae). **Journal of Herpetology**, v. 35, p. 395–402, 2001.

ESQUERRÉ, D.; et al. How mountains shape biodiversity: the role of the Andes in biogeography, diversification, and reproductive biology in South America's most species-rich lizard radiation (Squamata: Liolaemidae). **Evolution**, v. 73, n. 2, p. 214–230, 2019.

ESTOUP, Arnaud et al. Estimation of demo-genetic model probabilities with Approximate Bayesian Computation using linear discriminant analysis on summary statistics. **Molecular ecology resources**, v. 12, n. 5, p. 846-855, 2012.

FAIRCLOTH, B. C.; et al. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. **Systematic Biology**, v. 61, n. 5, p. 717–726, 2012.

FAIRCLOTH, B. C. PHYLUCE is a software package for the analysis of conserved genomic loci. **Bioinformatics**, v. 32, n. 5, p. 786–788, 2016.

FERREIRA, E.; et al. Microendemism can be the rule in the Brazilian Caatinga: evidence from flat lizards of the Tropidurus semitaeniatus group (Squamata: Tropiduridae). **Systematics and Biodiversity**, in press, 2025.

FLOURI, T.; et al. Species tree inference with BPP using genomic sequences and the multispecies coalescent. **Molecular Biology and Evolution**, v. 35, n. 10, p. 2585–2593, 2018.

FOLK, R. A.; et al. Identifying climatic drivers of hybridization with a new ancestral niche reconstruction method. **Systematic Biology**, v. 72, n. 4, p. 856–873, 2023.

FONSECA, E. M.; et al. Genetic structure and landscape effects on gene flow in the Neotropical lizard *Norops brasiliensis* (Squamata: Dactyloidae). **Heredity**, v. 132, n. 6, p. 284–295, 2024.

FROST, D. R. Phylogenetic analysis and taxonomy of the Tropidurus group of lizards (Iguania, Tropiduridae). **American Museum Novitates**, no. 3033, 1992.

FROST, D. R.; et al. Geographic variation, species recognition, and molecular evolution of cytochrome oxidase I in the Tropidurus spinulosus complex (Iguania: Tropiduridae). **Copeia**, p. 839–851, 1998.

FROST, D. R.; et al. Phylogenetics of the lizard genus Tropidurus (Squamata: Tropiduridae: Tropidurinae): direct optimization, descriptive efficiency, and sensitivity analysis of congruence between molecular data and morphology. **Molecular Phylogenetics and Evolution**, v. 21, n. 3, p. 352–371, 2001.

GAMISCH, A. Oscillayers: a dataset for the study of climatic oscillations over Plio-Pleistocene time-scales at high spatial-temporal resolution. **Global Ecology and Biogeography**, v. 28, n. 11, p. 1552–1560, 2019.

GARZIONE, C. N.; et al. Rise of the Andes. **Science**, v. 320, n. 5881, p. 1304–1307, 2008.

GLASSER, N. F.; et al. The glacial geomorphology and Pleistocene history of South America between 38 S and 56 S. **Quaternary Science Reviews**, v. 27, n. 3-4, p. 365-390, 2008.

GOOGLE INC. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: planetary-scale geospatial analysis for everyone [software]. 2017.

GREGORY-WODZICKI, K. M. Uplift history of the Central and Northern Andes: a review. **Geological Society of America Bulletin**, v. 112, n. 7, p. 1091–1105, 2000.

GRIZANTE, M. B.; et al. Morphological evolution in Tropidurinae squamates: an integrated view along a continuum of ecological settings. **Journal of Evolutionary Biology**, v. 23, n. 1, p. 98–111, 2010.

GUARNIZO, C. E.; et al. Cryptic lineages and diversification of an endemic anole lizard (Squamata, Dactyloidae) of the Cerrado hotspot. **Molecular Phylogenetics and Evolution**, v. 94, p. 279–289, 2016.

GUILLORY, W. X.; BROWN, J. L. A new method for integrating ecological niche modeling with phylogenetics to estimate ancestral distributions. **Systematic Biology**, v. 70, p. 1033–1045, 2021.

GUILLORY, W. X.; et al. Geoclimatic drivers of diversification in the largest arid and semi-arid environment of the Neotropics: perspectives from phylogeography. **Molecular Ecology**, v. 33, n. 14, p. e17431, 2024.

HAMILTON, S. K.; SIPPEL, S. J.; MELACK, J. M. Inundation patterns in the Pantanal wetland of South America determined from passive microwave remote sensing. **Archiv für Hydrobiologie**, v. 137, n. 1, p. 1-23, 1996.

HARVEY, M. B.; GUTBERLET, R. L. Lizards of the genus *Tropidurus* (Iguania: Tropiduridae) from the Serranía de Huanchaca, Bolivia: new species, natural history, and a key to the genus. **Herpetologica**, v. 54, n. 3, p. 493–520, 1998.

HARVEY, M. B.; GUTBERLET, R. L. A phylogenetic analysis of the tropidurine lizards (Squamata: Tropiduridae), including new characters of squamation and epidermal microstructure. **Zoological Journal of the Linnean Society**, v. 128, n. 2, p. 189–233, 2000.

HARRIS, R. S. Improved pairwise alignment of genomic DNA. The Pennsylvania State University, 2007.

HERNÁNDEZ, R. M.; et al. Age, distribution, tectonics, and eustatic controls of the Paranense and Caribbean marine transgressions in southern Bolivia and Argentina. **Journal of South American Earth Sciences**, v. 19, n. 4, p. 495–512, 2005.

HOLLISTER, Jeffrey; et al. elevatr: access elevation data from various APIs. R package version 0.1, v. 3, 2017.

HOORN, Carina. Marine incursions and the influence of Andean tectonics on the Miocene depositional history of northwestern Amazonia: results of a palynostratigraphic study. **Palaeogeography, Palaeoclimatology, Palaeoecology**, v. 105, n. 3–4, p. 267–309, 1993.

HOORN, C.; et al. Amazonia through time: Andean uplift, climate change, landscape evolution, and biodiversity. **Science**, v. 330, n. 6006, p. 927–931, 2010.

HOORN, C.; et al. Neogene history of the Amazonian flora: a perspective based on geological, palynological, and molecular phylogenetic data. **Annual Review of Earth and Planetary Sciences**, v. 51, n. 1, p. 419–446, 2023.

HUELSENBECK, J. P.; BOLLBACK, J. P. Empirical and hierarchical Bayesian estimation of ancestral states. **Systematic biology**, v. 50, n. 3, p. 351-366, 2001.

HULKA, C.; et al. Depositional setting of the middle to late Miocene Yecua Formation of the Chaco Foreland Basin, southern Bolivia. **Journal of South American Earth Sciences**, v. 21, n. 1–2, p. 135–150, 2006.

JARAMILLO, C. The evolution of extant South American tropical biomes. **New Phytologist**, v. 239, n. 2, p. 477–493, 2023.

JUNK, W. J.; et al. Biodiversity and its conservation in the Pantanal of Mato Grosso, Brazil. **Aquatic Sciences**, v. 68, n. 3, p. 278–309, 2006. DOI: 10.1007/s00027-006-0851-4.

JUNK, W. J.; DA CUNHA, C. N. The Pantanal: a brief review of its ecology, biodiversity, and protection status. In: MILLER, R. L.; et al. (org.). The Wetland Book. Dordrecht: Springer, p. 797–811, 2018.

KALYAANAMOORTHY, S.; et al. ModelFinder: fast model selection for accurate phylogenetic estimates. **Nature Methods**, v. 14, n. 6, p. 587–589, 2017.

KATOH, K.; STANDLEY, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. **Molecular Biology and Evolution**, v. 30, n. 4, p. 772–780, 2013.

KOHLSDORF, T.; GARLAND JR, T.; NAVAS, C. A. Limb and tail lengths in relation to substrate usage in Tropidurus lizards. **Journal of Morphology**, v. 248, p. 151–164, 2001.

KOHLSDORF, T.; et al. Locomotor performance of closely related Tropidurus species: relationships with physiological parameters and ecological divergence. **Journal of Experimental Biology**, v. 207, n. 7, p. 1183–1192, 2004.

KOHLSDORF, T.; GRIZANTE, M. B.; NAVAS, C. A.; HERREL, A. Head shape evolution in Tropidurinae lizards: does locomotion constrain diet? **Journal of Evolutionary Biology**, v. 21, n. 5, p. 781–790, 2008.

KOPUCHIAN, C.; et al. A test of the riverine barrier hypothesis in the largest subtropical river basin in the Neotropics. **Molecular Ecology**, v. 29, n. 12, p. 2137-2149, 2020.

LEDO, R. M. D.; et al. Pleistocene expansion and connectivity of mesic forests inside the South American Dry Diagonal supported by the phylogeography of a small lizard. **Evolution**, v. 74, n. 9, p. 1988–2004, 2020.

LEJEUNE, T. M.; WILLEMS, P. A.; HEGLUND, N. C. Mechanics and energetics of human locomotion on sand. **Journal of Experimental Biology**, v. 201, p. 2071–2080, 1998.

LIRA-MARTINS, D.; et al. Soil properties and geomorphic processes influence vegetation composition, structure, and function in the Cerrado domain. **Plant and Soil**, v. 476, n. 1–2, p. 549–588, 2022. DOI: 10.1007/s11104-022-05517-y.

LOVEJOY, N. R.; ALBERT, J. S.; CRAMPTON, W. G. R. Miocene marine incursions and marine/freshwater transitions: evidence from Neotropical fishes. **Journal of South American Earth Sciences**, v. 21, n. 1–2, p. 5–13, 2006.

MAHLER, D. L.; et al. Ecological opportunity and the rate of morphological evolution in the diversification of Greater Antillean anoles. **Evolution**, v. 64, n. 9, p. 2731–2735, 2010.

MARQUES-SOUZA, S.; et al. Hidden in the DNA: how multiple historical processes and natural history traits shaped patterns of cryptic diversity in an Amazon leaf-litter lizard Loxopholis osvaldoi (Squamata: Gymnophthalmidae). **Journal of Biogeography**, v. 47, n. 2, p. 501–515, 2020.

MELLA, J. Preferencia de microhabitat por Microlophus quadrivittatus (Reptilia: Squamata: Tropiduridae) en la costa de Iquique: diferencias sexuales, ontogenéticas, estacionales y ambientales. **Boletín Museo Nacional de Historia Natural**, v. 71, n. 2, p. 23-39, 2022.

MILLER, K. G.; et al. The Phanerozoic record of global sea-level change. **Science**, v. 310, n. 5752, p. 1293–1298, 2005.

MINH, B. Q.; et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. **Molecular Biology and Evolution**, v. 37, n. 5, p. 1530–1534, 2020.

MIRANDA, N. E. O. et al. Diversification of the widespread neotropical frog Physalaemus cuvieri in response to Neogene-Quaternary geological events and climate dynamics. **Molecular phylogenetics and evolution**, v. 132, p. 67-80, 2019.

MONDANARO, A.; et al. RRphylogeography: a new method to find the area of origin of species and the history of past contacts between species. **Methods in Ecology and Evolution**, in press, 2025.

NOGUEIRA, C.; et al. Vicariance and endemism in a Neotropical savanna hotspot: distribution patterns of Cerrado squamate reptiles. **Journal of Biogeography**, v. 38, n. 10, p. 1907–1922, 2011.

NONATO, G.A.S. Aspectos da história natural de duas populações de Tropidurus gr. spinulosus (Squamata, Tropiduridae) no Mato Grosso, Brasil. Dissertação. Instituto de Biociências da Universidade Federal do Mato Grosso. 2018.

NORES, M. The implications of Tertiary and Quaternary sea level rise events for avian distribution patterns in the lowlands of northern South America. **Global Ecology and Biogeography**, v. 13, n. 2, p. 149–161, 2004.

OLSON, D. M.; et al. Terrestrial ecoregions of the world: a new map of life on Earth. **BioScience**, v. 51, p. 933–938, 2001.

ORTIZ-JAUREGUIZAR, E. Paleoecologia y evolucion de la fauna de mamiferos de America del Sur durante la "Edad de las planicies Australes" (Mioceno superior-Plioceno superior). In: **Estudios Geologicos**, v. 54, 1998.

ORTIZ-JAUREGUIZAR, E.; CLADERA, G. A. Paleoenvironmental evolution of southern South America during the Cenozoic. **Journal of Arid Environments**, v. 66, n. 3, p. 498–532, 2006. DOI: 10.1016/j.jaridenv.2006.01.007.

PALMA-SILVA, C.; et al. Drivers of exceptional neotropical biodiversity: an updated view. **Botanical Journal of the Linnean Society**, v. 199, n. 1, p. 1–7, 2022.

PARADIS, E.; et al. Package ‘ape’. Analyses of phylogenetics and evolution, version, v. 2, n. 4, p. 47, 2019.

PEREIRA, A. G.; SCHRAGO, C. G. Arrival and diversification of mabuyine skinks (Squamata: Scincidae) in the Neotropics based on a fossil-calibrated timetree. **PeerJ**, v. 5, p. e3194, 2017.

PEREZ-MELLADO, V; DE LA RIVA, I. Sexual size dimorphism and ecology: the case of a tropical lizard, *Tropidurus melanopleurus* (Sauria: Tropiduridae). **Copeia**, p. 969-976, 1993.

PIATTI, L.; et al. Snake diversity in floodplains of central South America: is flood pulse the principal driver? **Acta Oecologica**, v. 97, p. 34–41, 2019.

PRADO, C. P.; HADDAD, C. F., & ZAMUDIO, K. R.. Cryptic lineages and Pleistocene population expansion in a Brazilian Cerrado frog. Molecular Ecology, v.21, n.4, 921–941, 2012. https://doi. org/10.1111/j.1365-294X.2011.05409.x

PUDLO, P., MARIN, J.M., ESTOUP, A., CORNUET, J.M., GAUTIER, M., & ROBERT, C.P. Reliable ABC model choice via random forests. **Bioinformatics**, v.32, 859–866, 2016, https://doi.org/10.1093/bioinformatics/btv684.

PUPIM, F. N.; ASSINE, M. L.; SAWAKUCHI, A. O. Late Quaternary Cuiabá megafan, Brazilian Pantanal: channel patterns and paleoenvironmental changes. **Quaternary International**, v. 438, p. 108-125, 2017.

R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL: https://www.R-project.org/

RAMBAUT, A.; et al. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. **Systematic Biology**, v. 67, n. 5, p. 901–904, 2018.

RAMOS, V. A.; ALEMÁN, A. Tectonic evolution of the Andes. In **Tectonic evolution of Soutth America**, CORDANI, U. G., MILANI, E. J., THOMAZ FILHO, A., CAMPOS, D. A. (eds). 31st International Geological Congress: Rio de Janeiro, 635–685, 2000.

RÄSÄNEN, M.; et al. Palaeogeographical implications of the Miocene Quendeque Formation (Bolivia) and tidally-influenced strata in southwestern Amazonia. **Palaeogeography, Palaeoclimatology, Palaeoecology**, v. 243, p. 23–41, 2007.

RAYNAL, L., MARIN, J.M., PUDLO, P., RIBATET, M., ROBERT, C.P., & ESTOUP, A. 2019. ABC random forests for Bayesian parameter inference. **Bioinformatics**, v. 35, 1720–1728.

REVELL, Liam J. phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). **PeerJ**, v. 12, p. e16505, 2024.

RIBAS, C. C.; FRITZ, S. C.; BAKER, P. A. The challenges and potential of geogenomics for biogeography and conservation in Amazonia. **Journal of Biogeography**, v. 49, n. 10, p. 1839–1847, 2022.

RIBEIRO, A. C. Tectonic history and the biogeography of the freshwater fishes from the coastal drainages of eastern Brazil: an example of faunal evolution associated with a divergent continental margin. **Neotropical Ichthyology**, v. 4, p. 225–246, 2006.

RIVADENEIRA, E. Y. D. Variação geográfica do complexo Tropidurus melanopleurus (Sauria: Tropiduridae). Monografia (Graduação em Biologia) – Faculdade de Ciências Puras e Naturais, Universidad Mayor de San Andrés, La Paz, Bolivia. 2008.

RODBELL, D. T.; et al. 700,000 years of tropical Andean glaciation. **Nature**, v. 607, n. 7918, p. 301-306, 2022.

ROCHA, P. L. B. Uso e partição de recursos pelas espécies de lagartos das dunas do Rio São Francisco, Bahia (Squamata). Tese (Doutorado em Zoologia) – Instituto de Biociências, Universidade de São Paulo, São Paulo, Brazil. 1998.

RODRIGUES, M. T. Sistemática, ecologia e zoogeografia dos Tropidurus do grupo torquatus ao sul do Rio Amazonas (Sauria, Iguanidae). **Arquivos de Zoologia**, v. 31, n. 3, p. 105–230, 1987.

RODRIGUES, M. T.; YASSUDA, Y. Y.; KASAHARA, S. Notes on the ecology and karyotypic description of Strobilurus torquatus (Sauria, Iguanidae). **Rev. bras. genét**, p. 747-59, 1989.

RULL, V. Neotropical biodiversity: timing and potential drivers. **Trends in Ecology & Evolution**, v. 26, n. 10, p. 508–513, 2011.

RULL, V. Neotropical diversification: historical overview and conceptual insights. In: Neotropical diversification: Patterns and processes. p. 13–49, 2020.

SALLES, M. M. A.; et al. Ancient Introgression Explains Mitochondrial Genome Capture and Mitonuclear Discordance Among South American Collared *Tropidurus* Lizards. **Molecular Ecology**, p. e70130, 2025.

SANO, E. E.; et al. Cerrado ecoregions: a spatial framework to assess and prioritize Brazilian savanna environmental diversity for conservation. **Journal of Environmental Management**, v. 232, p. 818–828, 2019.

SCHAEFER, S. A. The Andes: riding the tectonic uplift. In: Historical biogeography of Neotropical freshwater fishes. p. 259–278, 2011.

SHEU, Y.; et al. The combined role of dispersal and niche evolution in the diversification of Neotropical lizards. **Ecology and Evolution**, v. 10, n. 5, p. 2608–2625, 2020.

SMITH, S. A.; BROWN, J. W.; WALKER, J. F. So many genes, so little time: a practical approach to divergence-time estimation in the genomic era. **PLOS One**, v. 13, n. 5, p. e0197433, 2018.

SMITH, M. L.; CARSTENS, B. C. Process-based species delimitation leads to identification of more biologically relevant species. **Evolution**, v. 74, n. 2, p. 216–229, 2020.

SOUZA, C. A.; SOUZA, J. B. Pantanal Mato-Grossense: origem, evolução e as características atuais. **Revista Eletrônica da Associação dos Geógrafos Brasileiros**, 2010.

STRÜSSMANN, C.; et al. Diversity, ecology, management and conservation of amphibians and reptiles of the Brazilian Pantanal: a review. In: JUNK, W. J.; DA SILVA, C. J.; WANTZEN, K. M. (orgs.). The Pantanal: Ecology, biodiversity and sustainable management of a large neotropical seasonal wetland. Pensoft Publishers, Sofia-Moscow, p. 497–521, 2011.

TAMURA, K.; et al. Theoretical foundation of the RelTime method for estimating divergence times from variable evolutionary rates. **Molecular Biology and Evolution**, v. 35, n. 7, p. 1770–1782, 2018.

TAMURA, K.; et al. MEGA11: molecular evolutionary genetics analysis version 11. **Molecular Biology and Evolution**, v. 38, n. 7, p. 3022–3027, 2021.

TAO, Q.; et al. Reliable confidence intervals for RelTime estimates of evolutionary divergence times. **Molecular Biology and Evolution**, v. 37, n. 1, p. 280–290, 2020.

TURCHETTO-ZOLET, A. C.; et al. Phylogeographical patterns shed light on evolutionary processes in South America. **Molecular Ecology**, v. 22, n. 5, p. 1193–1213, 2013.

UBA, C. E.; et al. Isotopic, paleontologic, and ichnologic evidence for late Miocene pulses of marine incursions in the central Andes. **Geology**, v. 37, p. 827–830, 2009.

ULIANA, M. A.; BIDDLE, K. T.; et al. Mesozoic-Cenozoic paleogeographic and geodynamic evolution of southern South America. **Revista Brasileira de Geociências**, v. 18, n. 2, p. 172–190, 1988.

USSAMI, N.; SHIRAIWA, S.; DOMINGUEZ, J. M. L. Basement reactivation in a sub-Andean foreland flexural bulge: the Pantanal wetland, SW Brazil. **Tectonics**, v. 18, n. 1, p. 25–39, 1999.

VAN SLUYS, M. Food habits of the lizard *Tropidurus itambere* (Tropiduridae) in southeastern Brazil. **Journal of herpetology**, v. 27, n. 3, p. 347-351, 1993.

VITT, L. J. An introduction to the ecology of Cerrado lizards. **Journal of Herpetology**, v. 25, p. 79–90, 1991.

VITT, L. J. Ecology of isolated open-formation Tropidurus (Reptilia: Tropiduridae) in Amazonian lowland rain forest. **Canadian Journal of zoology**, v. 71, n. 12, p. 2370-2390, 1993.

VITT, L. J.; ZANI, P. A.; AVILA-PIRES, T. C. S. Ecology of the arboreal tropidurid lizard Tropidurus (Plica) umbra in the Amazon Region. **Canadian Journal of Zoology**, v. 75, p. 1876–1882, 1997.

WALLACE, A. R. On the monkeys of the Amazon. **Annals and Magazine of Natural History**, v. 14, n. 84, p. 451–454, 1854.

WERNECK, F. P. The diversification of eastern South American open vegetation biomes: historical biogeography and perspectives. **Quaternary Science Reviews**, v. 30, n. 13-14, p. 1630-1648, 2011.

WERNECK, F. P. et al. Deep diversification and long-term persistence in the South American ‘dry diagonal’: integrating continent-wide phylogeography and distribution modeling of geckos. **Evolution**, v. 66, n. 10, p. 3014-3034, 2012.

WERNECK, F. P.; et al. Biogeographic history and cryptic diversity of saxicolous Tropiduridae lizards endemic to the semiarid Caatinga. **BMC Evolutionary Biology**, v. 15, p. 1–24, 2015.

YANG, L.; HOU, Z.; LI, S. Marine incursion into East Asia: a forgotten driving force of biodiversity. **Proceedings of the Royal Society B: Biological Sciences**, v. 280, n. 1757, p. 20122892, 2013.

ZHANG, C.; et al. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. **BMC Bioinformatics**, v. 19, p. 15–30, 2018.

ZHENG, Y.; WIENS, J. J. Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. **Molecular Phylogenetics and Evolution**, v. 94, p. 537–547, 2016.

5.7 DATA AND CODE

The data underlying this article, including phylogenetic datasets, corresponding trees, input and output files for all analyses, and any other relevant supplementary files, are available in Zenodo at https://doi.org/10.5281/zenodo.17289523.

5.8 SUPPLEMENTARY MATERIAL

Detailed DIYABC-RF Workflow

This supplementary section provides a detailed account of the workflow implemented in DIYABC-RF v1.2.1 (Collin et al., 2021) for the Approximate Bayesian Computation using Random Forests (ABC-RF; Pudlo et al., 2016; Raynal et al., 2019). The primary aim of this analysis was to evaluate alternative demographic scenarios describing the evolutionary history of the *Tropidurus spinulosus* species group. While the main methodological rationale and general pipeline are presented in the main Methods section, the present document expands on specific simulation settings, prior distributions, model constraints, and verification steps adopted during the analyses.

Input Data Preparation

The empirical dataset consisted of one SNP per locus derived from 820 UCE loci present across all individuals (complete UCE matrix, as described in the main text). Following DIYABC requirements, SNPs missing in their entirety in a population or monomorphic across all species were excluded. The dataset was formatted according to DIYABC-RF input conventions, defining four populations corresponding to the focal species: *T. guarani*, *T. tarara*, *T. spinulosus*, and *T. xanthochilus*.

Scenario Design

Nine alternative demographic scenarios were constructed, grouped into three sets (Groups 1–3) that differed in the number, timing, and type of demographic events—population splits (green), bottlenecks (blue), and expansions (red)—as illustrated in FIGURE 7 of the main text. Each group represents a conceptual demographic model linking specific evolutionary hypotheses (see TABLE 2 of the main text). Scenarios were parameterized using broad priors, allowing the simulation of complex demographic histories, including sequential or overlapping demographic events.

Simulation Settings

For each scenario, between 10,000 and 20,000 datasets were simulated, matching the empirical dataset in terms of number of loci (820) and proportion of missing data. Simulations followed a coalescent model framework, with parameters sampled from prior distributions detailed below. To ensure biological realism while maintaining analytical convergence, a generation time of 0.5 years was used—this pragmatic adjustment allowed the simulated data distributions to encompass the empirical data during prior-checking tests.

Prior Distributions

Broad uniform (UN) and log-uniform (LU) prior distributions were defined for population sizes (N) and divergence times (t), respectively. These priors were designed to encompass a wide range of plausible demographic conditions.

- All priors related population sizes followed a uniform distribution: [10, 100000]
- All priors related to divergence times followed a log uniform distribution: [10000, 3500000]

These parameters were defined in the simulation header file with specific conditional relationships (see below) to ensure temporal consistency and biologically meaningful constraints among population size changes.

Conditional Constraints

Conditions were implemented in the DIYABC header file to enforce logical and temporal dependencies among parameters. These constraints ensured, for instance, that expansions follow bottlenecks and that demographic transitions occur in chronological order. The full list of constraints was as follows:

- N4_expt8 > N4_btnt7
- N3_expt8 > N3_btnt7
- N2_expt8 > N2_btnt7
- N1_expt8 > N1_btnt7
- N4_btnt7 < N2_expt6
- N3_btnt7 < N2_expt6
- N2_btnt7 < N2_expt6
- N1_btnt7 < N1_expt6
- N1_expt6 > N1_btnt3
- N2_expt6 > N2_btnt3
- N1_expt6 > N1_stbt3
- N2_expt6 > N2_stbt3
- N1_expt6 > N1_inct3
- N2_expt6 > N2_inct3
- N1_stbt3 ≤ Na
- N2_stbt3 ≤ Na
- N1_btnt3 < Na
- N2_btnt3 < Na
- N1_inct3 > Na
- N2_inct3 > Na
- t1 > t2 > t3 > t4 > t5 > t6 > t7 > t8

Model Validation and Data Checking

Before model selection, prior predictive checks were conducted to confirm that simulated datasets encompassed the empirical data. This was done by visual inspection of principal component analyses (PCA) of the summary statistics distributions. All observed datasets fell within the simulated data cloud, validating the prior parameter ranges.

Summary Statistics

Both simulated and observed datasets were summarized using the full set of 130 summary statistics available for SNP markers in DIYABC-RF. These include measures of within-population diversity (e.g., heterozygosity, allele frequencies), pairwise differentiation (e.g., Fst, Nei’s distance), and shared allele frequencies. Additionally, Linear Discriminant Analysis (LDA) combinations of summary statistics were used for model choice, following Estoup et al. (2012).

Random Forest Model Selection

Model selection was performed in two hierarchical steps:

1. Global comparison: all nine scenarios were compared to identify the best-fitting model group.
2. Within-group comparison: a refined selection among the three scenarios of the best-fitting group was performed, using narrower priors and additional simulations.

Each RF analysis was run with 500 and 1,000 trees to verify convergence.

Posterior probabilities were estimated following Pudlo et al. (2016), and model choice was based on the number of votes (trees) assigned to each scenario.

Output Interpretation

Posterior probabilities and votes for each scenario were extracted from the RF results. Scenarios with the highest vote proportions were retained as the most supported. The confusion matrix and prior error rate jointly informed the reliability of the classification, while PCA plots confirmed the overlap between observed and simulated data distributions.

Summary of Analysis Pipeline

1. Definition of empirical dataset and scenario structure.
2. Assignment of priors and logical constraints.
3. Generation of 10,000–20,000 simulations per scenario.
4. Prior-checking via PCA.
5. Calculation of 130 summary statistics per dataset.
6. Random Forest training (500–1,000 trees).
7. Model selection in two steps (global → within-group).
8. Assessment of classification performance.
9. Extraction of posterior probabilities and scenario validation.

All steps were executed using the graphical interface of DIYABC-RF, which integrated the simulation, summary statistic computation, model checking, and Random Forest analyses.

Phylogenetic inference and divergence time results: Figures S1-S10

FIGURE S1. Partitioned phylogenetic analysis of 2269 UCE loci, matrix with 75% of completeness (via IQ-TREE). Numbers inside boxes refer to bootstrap support values (only nodes with support > 80 are shown).

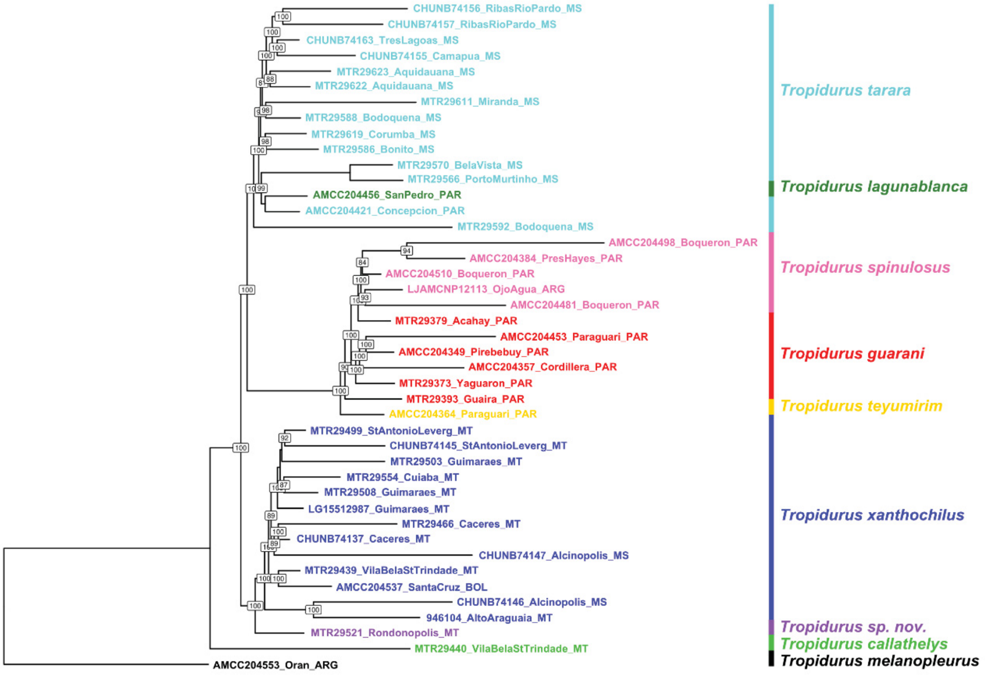


FIGURE S2. Partitioned phylogenetic analysis of 820 UCE loci (complete matrix), produced via IQ-TREE. Numbers next to nodes refer to bootstrap support values.

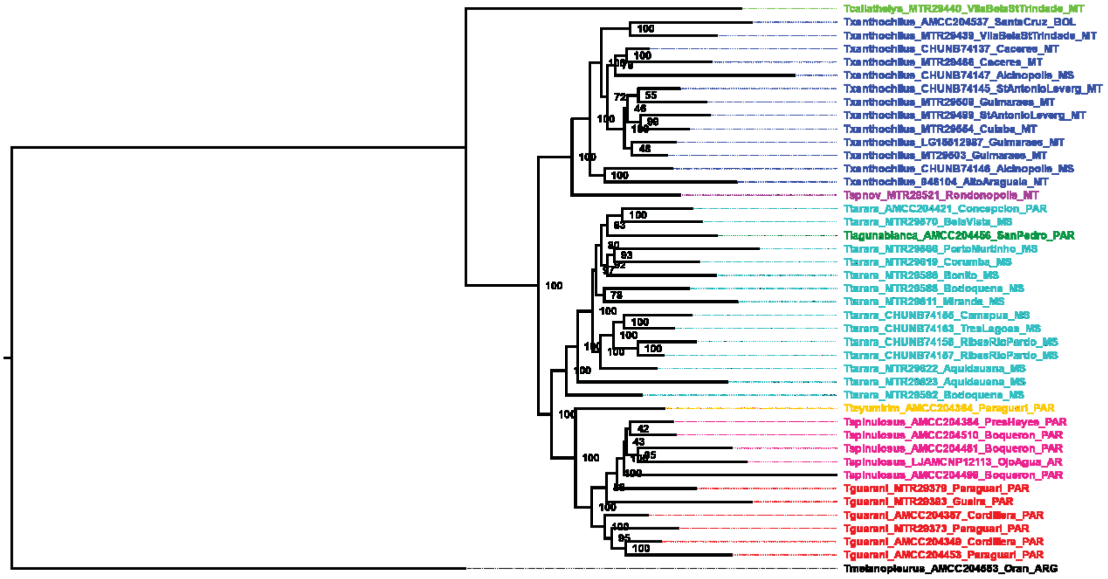

Tcallathelys_MTR29440_VilaBelaStTrindade_MT
Txanthochilus_AMCC204537_SantaCruz_BOL
Txanthochilus_MTR29439_VilaBelaStTrindade_MT
Txanthochilus_CHUNB74137_Caceres_MT
Txanthochilus_MTR29466_Caceres_MT
Txanthochilus_CHUNB74147_Alcinopolis_MS
Txanthochilus_CHUNB74145_StAntonioLeverg_MT
Txanthochilus_MTR29508_Guimaraes_MT
Txanthochilus_MTR29499_StAntonioLeverg_MT
Txanthochilus_MTR29554_Cuiaba_MT
Txanthochilus_LG15512987_Guimaraes_MT
Txanthochilus_MT29503_Guimaraes_MT
Txanthochilus_CHUNB74146_Alcinopolis_MS
Txanthochilus_946104_AltoAraguaia_MT
Tspnov_MTR29521_Rondonopolis_MT
Ttarara_AMCC204421_Concepcion_PAR
Ttarara_MTR29570_BelaVista_MS
Tlagunablanca_AMCC204456_SanPedro_PAR
Ttarara_MTR29566_PortoMurtinho_MS
Ttarara_MTR29619_Corumba_MS
Ttarara_MTR29586_Bonito_MS
Ttarara_MTR29588_Bodoquena_MS
Ttarara_MTR29611_Miranda_MS
Ttarara_CHUNB74155_Camapua_MS
Ttarara_CHUNB74163_TresLagoas_MS
Ttarara_CHUNB74156_RibasRioPardo_MS
Ttarara_CHUNB74157_RibasRioPardo_MS
Ttarara_MTR29622_Aquidauana_MS
Ttarara_MTR29623_Aquidauana_MS
Ttarara_MTR29592_Bodoquena_MS
Tteyumirim_AMCC204364_Paraguari_PAR
Tspinulosus_AMCC204384_PresHayes_PAR
Tspinulosus_AMCC204510_Boqueron_PAR
Tspinulosus_AMCC204481_Boqueron_PAR
Tspinulosus_LJAMCNP12113_OjoAgua_AR
Tspinulosus_AMCC204498_Boqueron_PAR
Tguarani_MTR29379_Paraguari_PAR
Tguarani_MTR29393_Guaira_PAR
Tguarani_AMCC204357_Cordillera_PAR
Tguarani_MTR29373_Paraguari_PAR
Tguarani_AMCC204349_Cordillera_PAR
Tguarani_AMCC204453_Paraguari_PAR
Tmelanopleurus_AMCC204553_Oran_ARG

FIGURE S3. Partitioned phylogenetic analysis of 1613 UCE loci (matrix with 95% of completeness), produced via IQ-TREE. Numbers next to nodes refer to bootstrap support values.

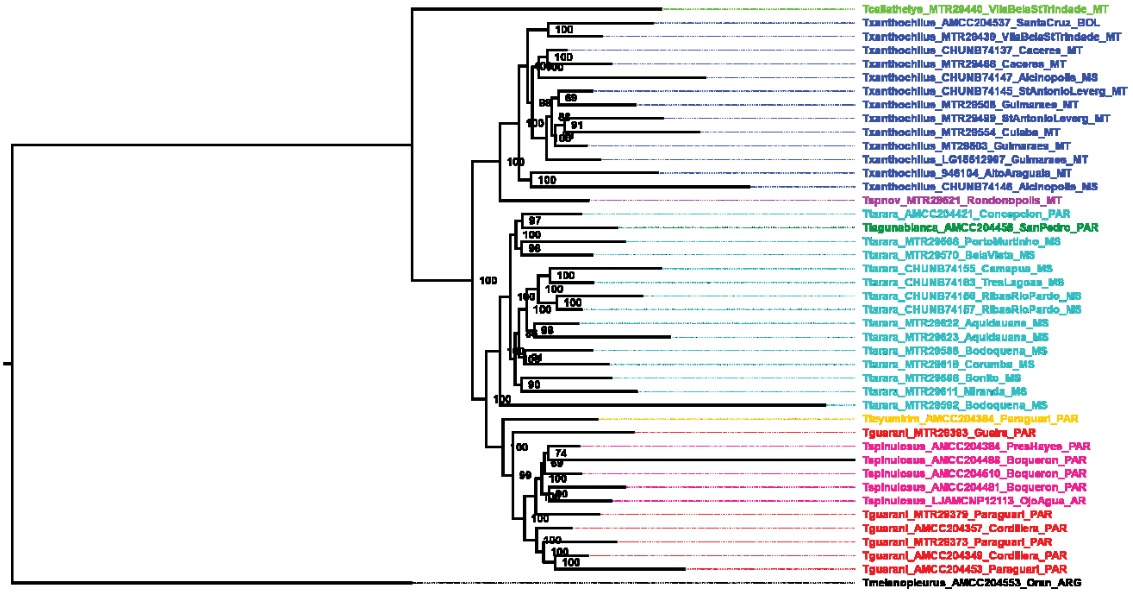

Tcallathelys_MTR28440_VilaBelaStTrindade_MT
Txanthochilus_AMCC204537_SantaCruz_BOL
Txanthochilus_MTR28439_VilaBelaStTrindade_MT
Txanthochilus_CHUNB74137_Caceres_MT
Txanthochilus_MTR28486_Caceres_MT
Txanthochilus_CHUNB74147_Alcinopolis_MS
Txanthochilus_CHUNB74145_StAntonioLeverg_MT
Txanthochilus_MTR28508_Guimaraes_MT
Txanthochilus_MTR28489_StAntonioLeverg_MT
Txanthochilus_MTR28554_Cuiaba_MT
Txanthochilus_MT28503_Guimaraes_MT
Txanthochilus_LG15512967_Guimaraes_MT
Txanthochilus_946104_AltoAraguaia_MT
Txanthochilus_CHUNB74146_Alcinopolis_MS
Tspnov_MTR28621_Rondonopolis_MT
Ttarara_AMCC204421_Concepcion_PAR
Tlagunablanca_AMCC204456_SanPedro_PAR
Ttarara_MTR29566_PortoMurtinho_MS
Ttarara_MTR29570_BelaVista_MS
Ttarara_CHUNB74155_Camapua_MS
Ttarara_CHUNB74163_TresLagoas_MS
Ttarara_CHUNB74158_RibasRioPardo_MS
Ttarara_CHUNB74157_RibasRioPardo_MS
Ttarara_MTR29622_Aquidauana_MS
Ttarara_MTR29623_Aquidauana_MS
Ttarara_MTR29585_Bodoquena_MS
Ttarara_MTR29619_Corumba_MS
Ttarara_MTR29596_Bonito_MS
Ttarara_MTR29611_Miranda_MS
Ttarara_MTR29592_Bodoquena_MS
Tteyumirim_AMCC204384_Paraguari_PAR
Tguarani_MTR28393_Guaira_PAR
Tspinulosus_AMCC204394_PresHayes_PAR
Tspinulosus_AMCC204498_Boqueron_PAR
Tspinulosus_AMCC204610_Boqueron_PAR
Tspinulosus_AMCC204481_Boqueron_PAR
Tspinulosus_LJAMCNP12113_OjoAgua_AR
Tguarani_MTR28379_Paraguari_PAR
Tguarani_AMCC204357_Cordillera_PAR
Tguarani_MTR28373_Paraguari_PAR
Tguarani_AMCC204349_Cordillera_PAR
Tguarani_AMCC204453_Paraguari_PAR
Tmelanopleurus_AMCC204553_Oran_ARG

FIGURE S4. Nuclear inferred from multilocus nuclear data using BPP, with branch support values indicated next to nodes (posterior probabilities). *All nuclear gene trees recovered *T. lagunablanca* nested within *T. tarara* samples, indicating that these lineages likely represent a single evolutionary unit. Accordingly, the *T. lagunablanca* sample was assigned to *T. tarara* in the BPP analysis, so that they represent a single branch in the estimated species tree.

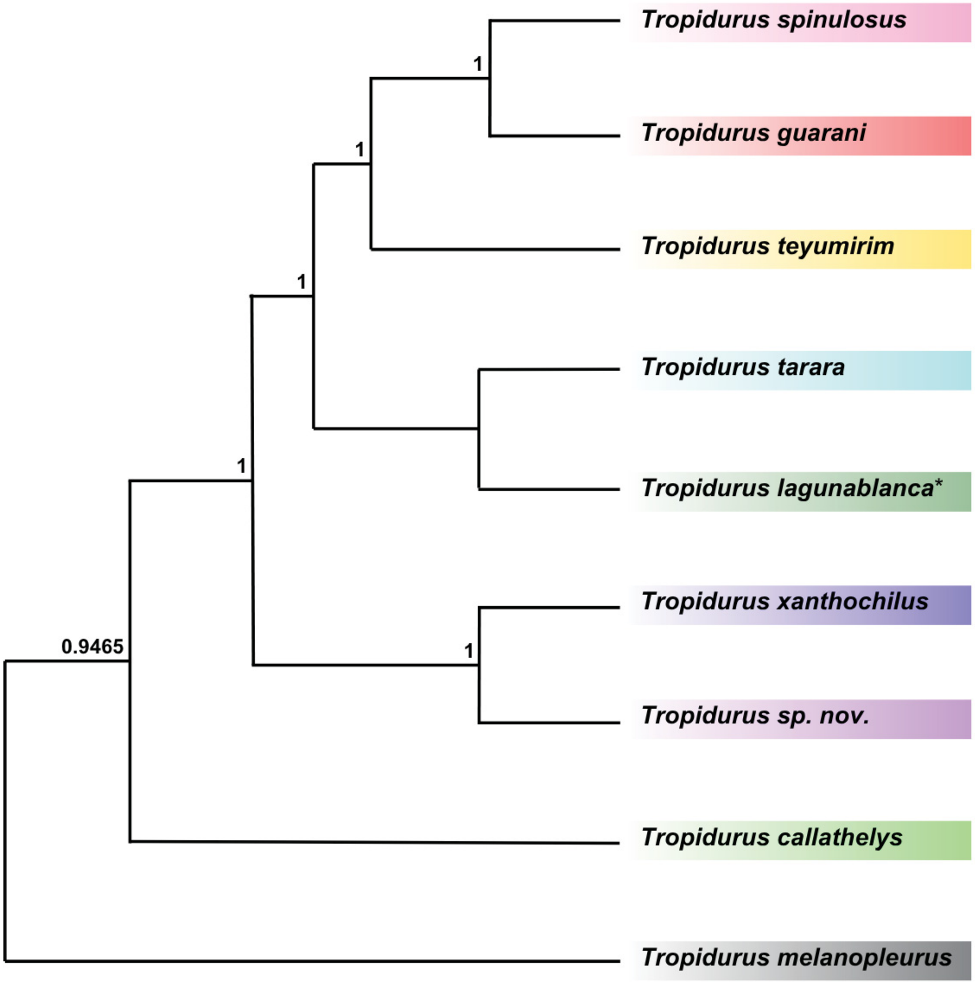


FIGURE S5. Results produced from the species tree analysis of the 820 UCE loci (complete matrix), produced via ASTRAL. Numbers next to nodes refer to bootstrap support values.

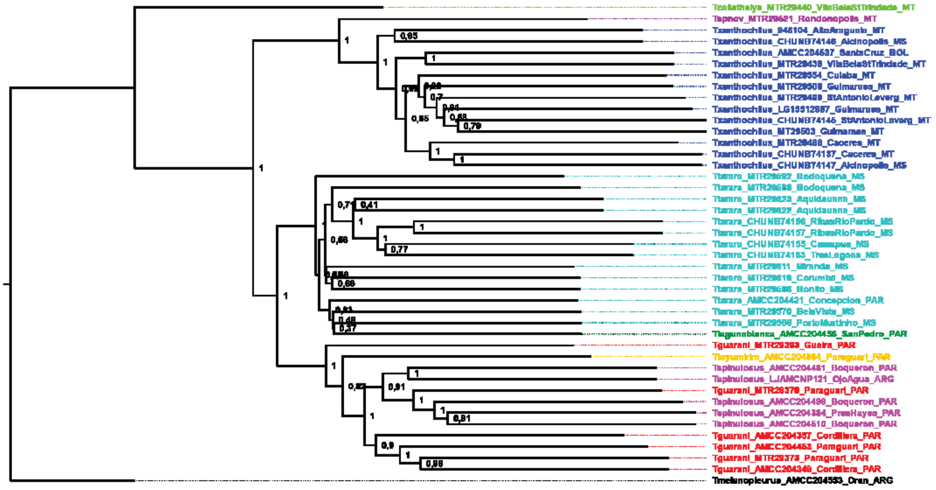

Tcallathelys_MTR29440_VilaBelaStTrindade_MT
Tspnov_MTR29521_Rondonopolis_MT
Txanthochilus_946104_AltoAraguaia_MT
Txanthochilus_CHUNB74146_Alcinopolis_MS
Txanthochilus_AMCC204537_SantaCruz_BOL
Txanthochilus_MTR29439_VilaBelaStTrindade_MT
Txanthochilus_MTR29554_Cuiaba_MT
Txanthochilus_MTR29508_Guimaraes_MT
Txanthochilus_MTR29499_StAntonioLeverg_MT
Txanthochilus_LG15512987_Guimaraes_MT
Txanthochilus_CHUNB74145_StAntonioLeverg_MT
Txanthochilus_MT29503_Guimaraes_MT
Txanthochilus_MTR29488_Caceres_MT
Txanthochilus_CHUNB74137_Caceres_MT
Txanthochilus_CHUNB74147_Alcinopolis_MS
Ttarara_MTR29592_Bodoquena_MS
Ttarara_MTR29588_Bodoquena_MS
Ttarara_MTR29623_Aquidauana_MS
Ttarara_MTR29622_Aquidauana_MS
Ttarara_CHUNB74156_RibasRioPardo_MS
Ttarara_CHUNB74157_RibasRioPardo_MS
Ttarara_CHUNB74155_Camapua_MS
Ttarara_CHUNB74163_TresLagoas_MS
Ttarara_MTR29611_Miranda_MS
Ttarara_MTR29619_Corumba_MS
Ttarara_MTR29585_Bonito_MS
Ttarara_AMCC204421_Concepcion_PAR
Ttarara_MTR29570_BelaVista_MS
Ttarara_MTR29566_PortoMurtinho_MS
Tlagunablanca_AMCC204456_SanPedro_PAR
Tguarani_MTR29393_Guaira_PAR
Tteyumirim_AMCC204364_Paraguari_PAR
Tspinulosus_AMCC204481_Boqueron_PAR
Tspinulosus_LJAMCNP121_OjoAgua_ARG
Tguarani_MTR29379_Paraguari_PAR
Tspinulosus_AMCC204496_Boqueron_PAR
Tspinulosus_AMCC204384_PresHayes_PAR
Tspinulosus_AMCC204510_Boqueron_PAR
Tguarani_AMCC204357_Cordillera_PAR
Tguarani_AMCC204453_Paraguari_PAR
Tguarani_MTR29373_Paraguari_PAR
Tguarani_AMCC204349_Cordillera_PAR
Tmelanopleurus_AMCC204553_Oran_ARG
1
0,95
1
1
0,98
0,98
0,7
0,95
0,91
0,83
0,79
1
1
1
0,71
0,41
1
1
0,98
1
0,77
0,68
0,81
0,48
0,37
1
1
0,82
0,91
1
1
1
0,81
1
0,9
1
0,98

FIGURE S6. Results produced from the species tree analysis of 1613 UCE loci (matrix with 95% of completeness), produced via ASTRAL. Numbers next to nodes refer to bootstrap support values.

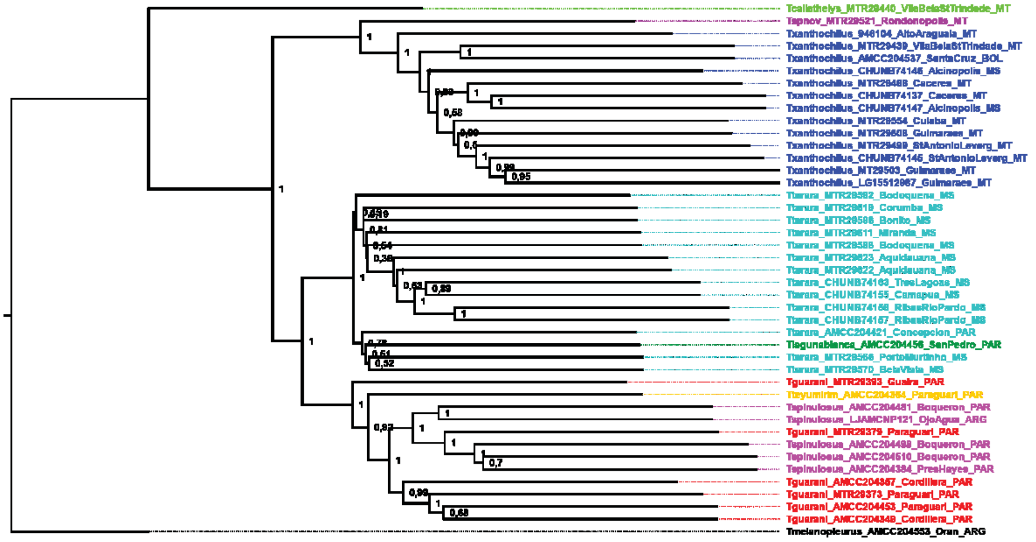
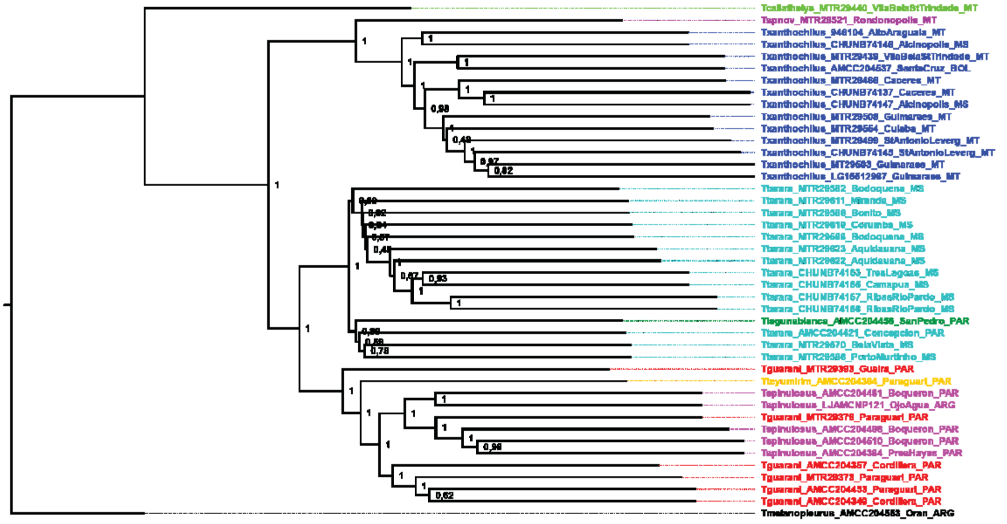

Tcallathelys_MTR29440_VilaBelaStTrindade_MT
Tspnov_MTR29521_Rondonopolis_MT
Txanthochilus_946104_AltoAraguaia_MT
Txanthochilus_MTR29439_VilaBelaStTrindade_MT
Txanthochilus_AMCC204537_SantaCruz_BOL
Txanthochilus_CHUNB74146_Alcinopolis_MS
Txanthochilus_MTR29488_Caceres_MT
Txanthochilus_CHUNB74137_Caceres_MT
Txanthochilus_CHUNB74147_Alcinopolis_MS
Txanthochilus_MTR29554_Cuiaba_MT
Txanthochilus_MTR29508_Guimaraes_MT
Txanthochilus_MTR29499_StAntonioLeverg_MT
Txanthochilus_CHUNB74145_StAntonioLeverg_MT
Txanthochilus_MT29503_Guimaraes_MT
Txanthochilus_LG15512987_Guimaraes_MT
Ttarara_MTR29592_Bodoquena_MS
Ttarara_MTR29619_Corumba_MS
Ttarara_MTR29585_Bonito_MS
Ttarara_MTR29611_Miranda_MS
Ttarara_MTR29588_Bodoquena_MS
Ttarara_MTR29623_Aquidauana_MS
Ttarara_MTR29622_Aquidauana_MS
Ttarara_CHUNB74163_TresLagoas_MS
Ttarara_CHUNB74155_Camapua_MS
Ttarara_CHUNB74156_RibasRioPardo_MS
Ttarara_CHUNB74157_RibasRioPardo_MS
Ttarara_AMCC204421_Concepcion_PAR
Tlagunablanca_AMCC204456_SanPedro_PAR
Ttarara_MTR29566_PortoMurtinho_MS
Ttarara_MTR29570_BelaVista_MS
Tguarani_MTR29393_Guaira_PAR
Tteyumirim_AMCC204364_Paraguari_PAR
Tspinulosus_AMCC204481_Boqueron_PAR
Tspinulosus_LJAMCNP121_OjoAgua_ARG
Tguarani_MTR29379_Paraguari_PAR
Tspinulosus_AMCC204496_Boqueron_PAR
Tspinulosus_AMCC204510_Boqueron_PAR
Tspinulosus_AMCC204384_PresHayes_PAR
Tguarani_AMCC204357_Cordillera_PAR
Tguarani_MTR29373_Paraguari_PAR
Tguarani_AMCC204453_Paraguari_PAR
Tguarani_AMCC204349_Cordillera_PAR
Tmelanopleurus_AMCC204553_Oran_ARG
1
1
1
1
0,93
1
0,58
0,99
0,5
1
0,99
0,95
1
0,79
0,51
0,54
0,36
1
0,62
0,93
1
1
1
0,78
0,51
0,32
1
0,82
1
1
1
1
0,7
0,96
1
0,68

FIGURE S7. Results produced from the species tree analysis of 2269 UCE loci (matrix with 75% of completeness), produced via ASTRAL. Numbers next to nodes refer to bootstrap support values.


Tcallathelys_MTR29440_VilaBelaStTrindade_MT
Tspnov_MTR29521_Rondonopolis_MT
Txanthochilus_946104_AltoAraguaia_MT
Txanthochilus_CHUNB74146_Alcinopolis_MS
Txanthochilus_MTR29439_VilaBelaStTrindade_MT
Txanthochilus_AMCC204537_SantaCruz_BOL
Txanthochilus_MTR29488_Caceres_MT
Txanthochilus_CHUNB74137_Caceres_MT
Txanthochilus_CHUNB74147_Alcinopolis_MS
Txanthochilus_MTR29508_Guimaraes_MT
Txanthochilus_MTR29554_Cuiaba_MT
Txanthochilus_MTR29499_StAntonioLeverg_MT
Txanthochilus_CHUNB74145_StAntonioLeverg_MT
Txanthochilus_MT29503_Guimaraes_MT
Txanthochilus_LG15512987_Guimaraes_MT
Ttarara_MTR29592_Bodoquena_MS
Ttarara_MTR29611_Miranda_MS
Ttarara_MTR29585_Bonito_MS
Ttarara_MTR29619_Corumba_MS
Ttarara_MTR29588_Bodoquena_MS
Ttarara_MTR29623_Aquidauana_MS
Ttarara_MTR29622_Aquidauana_MS
Ttarara_CHUNB74163_TresLagoas_MS
Ttarara_CHUNB74155_Camapua_MS
Ttarara_CHUNB74157_RibasRioPardo_MS
Ttarara_CHUNB74156_RibasRioPardo_MS
Tlagunablanca_AMCC204456_SanPedro_PAR
Ttarara_AMCC204421_Concepcion_PAR
Ttarara_MTR29570_BelaVista_MS
Ttarara_MTR29566_PortoMurtinho_MS
Tguarani_MTR29393_Guaira_PAR
Tteyumirim_AMCC204364_Paraguari_PAR
Tspinulosus_AMCC204481_Boqueron_PAR
Tspinulosus_LJAMCNP121_OjoAgua_ARG
Tguarani_MTR29379_Paraguari_PAR
Tspinulosus_AMCC204496_Boqueron_PAR
Tspinulosus_AMCC204510_Boqueron_PAR
Tspinulosus_AMCC204384_PresHayes_PAR
Tguarani_AMCC204357_Cordillera_PAR
Tguarani_MTR29373_Paraguari_PAR
Tguarani_AMCC204453_Paraguari_PAR
Tguarani_AMCC204349_Cordillera_PAR
Tmelanopleurus_AMCC204553_Oran_ARG
1
1
1
1
1
1
1
0,98
1
0,48
1
0,97
0,82
1
0,89
0,82
0,84
0,57
0,48
1
0,87
0,93
1
1
1
0,96
0,89
0,78
1
1
1
1
1
1
0,96
1
1
1
0,62

In most cases, the divergence times estimated using RelTime (FIGURE S8; Tamura et al. 2018) fall within the same 95% highest posterior density (HPD) intervals provided by MCMCTree (FIGURE S9; dos Reis and Yang 2019), but important differences also occur. For instance, the clade comprising *T. guarani*, *T. spinulosus*, and *T. teyumirim* originates around 4 Ma according to the MCMCTree analysis, while the divergence is estimated to have occurred more recently based on the RelTime approach, around 3 Ma. In contrast, some node ages show notable discrepancies between the two methods. The divergence between *T. sp. nov.* and *T. xanthochilus* is estimated at approximately 3.5 Ma in the MCMCTree analysis, whereas RelTime dates the same split to about 7 Ma. Likewise, the origin of the *T. tarara* clade is estimated at ~3.75 Ma by MCMCTree but nearly 7.5 Ma in the RelTime analysis. The estimates provided by the species tree dated using MCMCTree provide intermediate values between the two analyses (FIGURE S10).

FIGURE S8. Divergence times (Ma) estimated for the *Tropidurus spinulosus* species group using the UCE dataset. The 95% HPD interval of the posterior estimates (blue shaded bars), as estimated through the MCMCTree program, is shown above each node of the tree. The scale bar is in Ma.

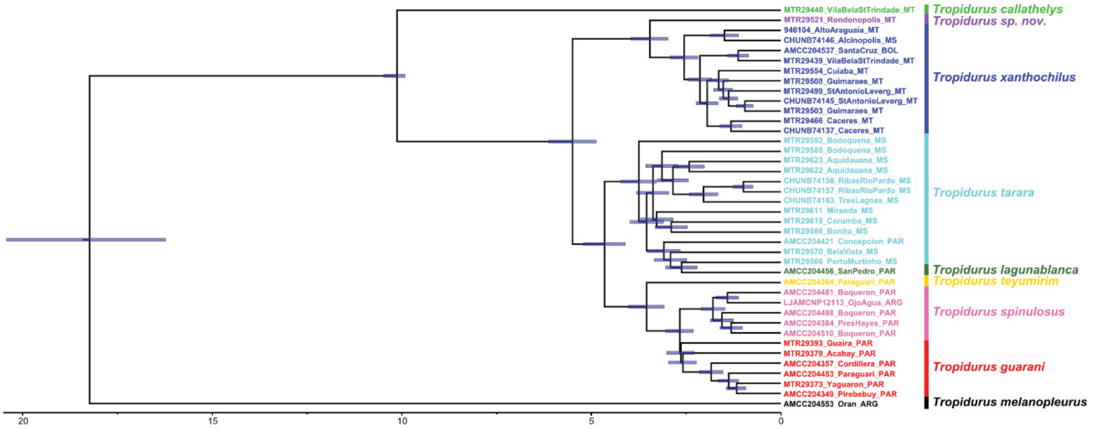


FIGURE S9. A timetree inferred by applying the RelTime-Branch Length method. The timetree was computed using 1 calibration constraint, as described in the main text. The scale bar is in Ma.

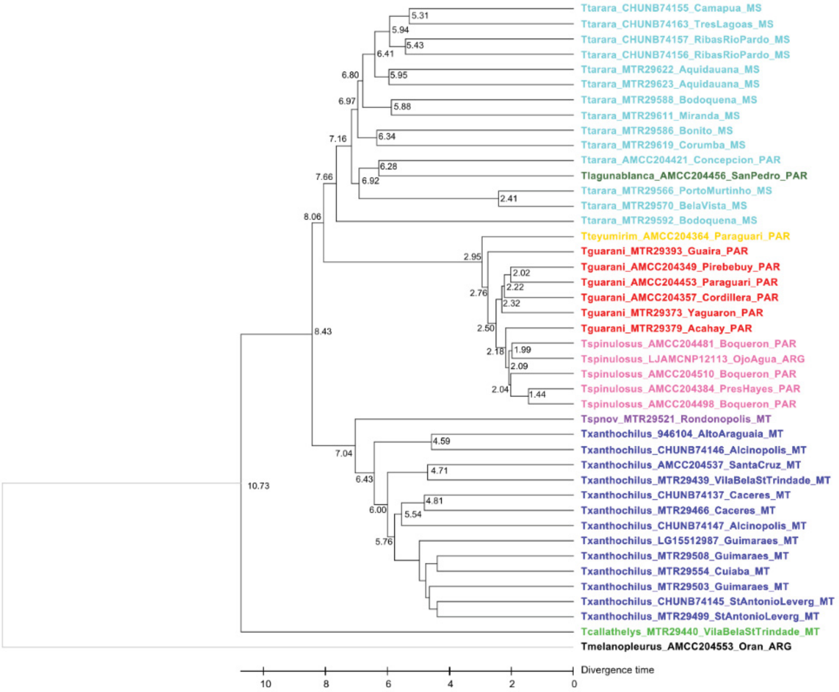
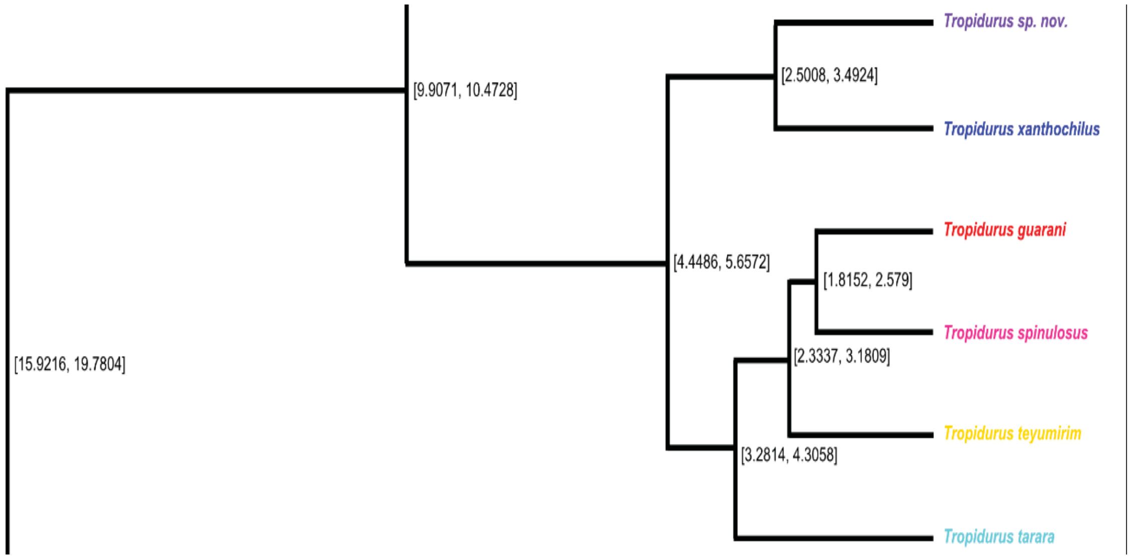

Ttarara_CHUNB74155_Camapua_MS
Ttarara_CHUNB74163_TresLagoas_MS
Ttarara_CHUNB74157_RibasRioPardo_MS
Ttarara_CHUNB74156_RibasRioPardo_MS
Ttarara_MTR29622_Aquidauana_MS
Ttarara_MTR29623_Aquidauana_MS
Ttarara_MTR29588_Bodoquena_MS
Ttarara_MTR29611_Miranda_MS
Ttarara_MTR29586_Bonito_MS
Ttarara_MTR29619_Corumba_MS
Ttarara_AMCC204421_Concepcion_PAR
Tlagunablanca_AMCC204456_SanPedro_PAR
Ttarara_MTR29566_PortoMurtinho_MS
Ttarara_MTR29570_BelaVista_MS
Ttarara_MTR29592_Bodoquena_MS
Tteyumirim_AMCC204364_Paraguari_PAR
Tguarani_MTR29393_Guaira_PAR
Tguarani_AMCC204349_Pirebebuy_PAR
Tguarani_AMCC204453_Paraguari_PAR
Tguarani_AMCC204357_Cordillera_PAR
Tguarani_MTR29373_Yaguaron_PAR
Tguarani_MTR29379_Acahay_PAR
Tspinulosus_AMCC204481_Boqueron_PAR
Tspinulosus_LJAMCNP12113_OjoAgua_ARG
Tspinulosus_AMCC204510_Boqueron_PAR
Tspinulosus_AMCC204384_PresHayes_PAR
Tspinulosus_AMCC204498_Boqueron_PAR
Tspnov_MTR29521_Rondonopolis_MT
Txanthochilus_946104_AltoAraguaia_MT
Txanthochilus_CHUNB74146_Alcinopolis_MT
Txanthochilus_AMCC204537_SantaCruz_MT
Txanthochilus_MTR29439_VilaBelaStTrindade_MT
Txanthochilus_CHUNB74137_Caceres_MT
Txanthochilus_MTR29466_Caceres_MT
Txanthochilus_CHUNB74147_Alcinopolis_MT
Txanthochilus_LG15512987_Guimaraes_MT
Txanthochilus_MTR29508_Guimaraes_MT
Txanthochilus_MTR29554_Cuiaba_MT
Txanthochilus_MTR29503_Guimaraes_MT
Txanthochilus_CHUNB74145_StAntonioLeverg_MT
Txanthochilus_MTR29499_StAntonioLeverg_MT
Tcallathelys_MTR29440_VilaBelaStTrindade_MT
Tmelanopleurus_AMCC204553_Oran_ARG
5.31
5.94
5.43
6.41
6.80
5.95
6.97
5.88
7.16
6.34
7.66
6.28
6.92
2.41
8.06
2.95
2.02
2.22
2.76
2.32
2.50
1.99
2.18
2.09
2.04
1.44
8.43
4.59
7.04
4.71
6.43
10.73
4.81
6.00
5.54
5.76
Divergence time
10
8
6
4
2
0

FIGURE S10. A timetree inferred by applying the MCMCTree method, but at the species level. The 95% HPD interval of the posterior estimates, as estimated through the MCMCTree program, is shown next to each node of the tree.



Ancestral climatic niche modelling results: Figures S11-12

FIGURE S11. Present SDMs for species within the *Tropidurus spinulosus* group, based on a phylogenetic niche modelling approach. The gradient bar next to each figure represents the probability of occurrence of the lineages in their predicted areas.

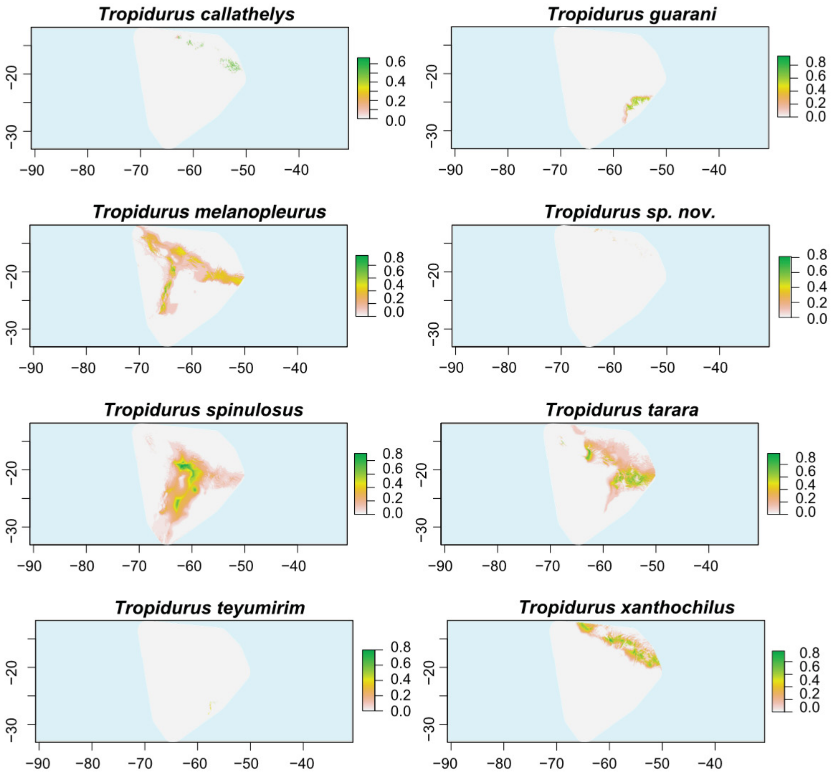
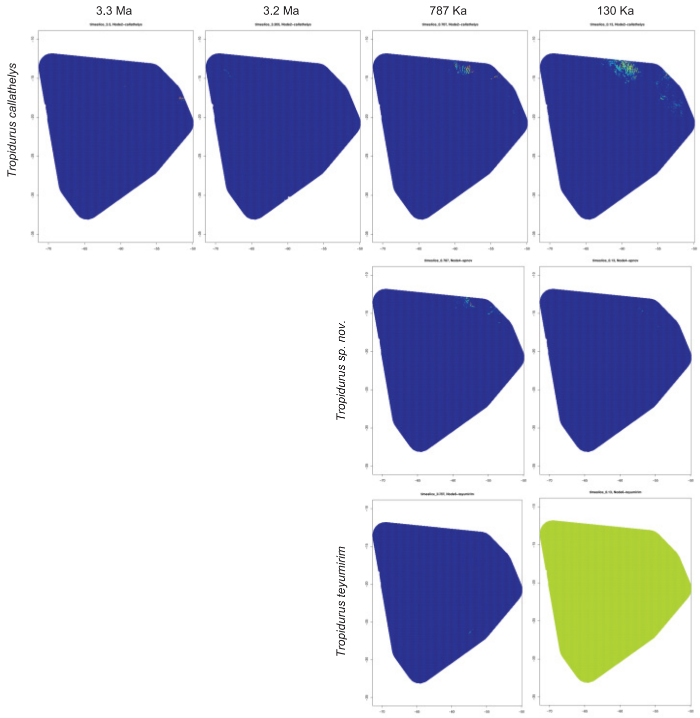


FIGURE S12. Ancestral climatic projections for *Tropidurus spinulosus* species which were not presented in the main text.



Dataset comprising all *Tropidurus spinulosus* records used in this study: Table S1

Table S1. Sampling localities and geographical coordinates for the 43 specimens included in this study, alongside with its genetic assignment in the BEAST analysis. The complete table, containing all records, can be found here.

Sample ID	Species assignment	Location	Longitude	Latitude
MTR 29623	*Tropidurus tarara*	Aquidauana, Mato Grosso do Sul, BRA	-55.5543	-20.4830
MTR 29622	*Tropidurus tarara*	Aquidauana, Mato Grosso do Sul, BRA	-55.8217	-20.4445
MTR 29619	*Tropidurus tarara*	Corumbá, Mato Grosso do Sul, BRA	-57.0361	-19.5738
MTR 29611	*Tropidurus tarara*	Miranda, Mato Grosso do Sul, BRA	-56.5084	-20.1928
MTR 29592	*Tropidurus tarara*	Bodoquena, Mato Grosso do Sul, BRA	-56.8965	-20.5319
MTR 29588	*Tropidurus tarara*	Bodoquena, Mato Grosso do Sul, BRA	-56.7138	-20.5424
MTR 29586	*Tropidurus tarara*	Bonito, Mato Grosso do Sul, BRA	-56.5812	-21.1074
MTR 29570	*Tropidurus tarara*	Bela Vista, Mato Grosso do Sul, BRA	-56.4635	-21.9755
MTR 29566	*Tropidurus tarara*	Porto Murtinho, Mato Grosso do Sul, BRA	-57.8963	-21.7065
MTR 29554	*Tropidurus xanthochilus*	Cuiabá, Mato Grosso, BRA	-55.9886	-15.4575
MTR 29521	*Tropidurus sp. nov.*	Rondonópolis, Mato Grosso, BRA	-54.7676	-16.6510
MTR 29508	*Tropidurus xanthochilus*	Chapada dos Guimarães, Mato Grosso, BRA	-55.8313	-15.4072
MTR 29503	*Tropidurus xanthochilus*	Chapada dos Guimarães, Mato Grosso, BRA	-55.7723	-15.4490
MTR 29499	*Tropidurus xanthochilus*	Santo Ant. do Leverger, Mato Grosso, BRA	-55.5380	-15.9643
MTR 29466	*Tropidurus xanthochilus*	Cáceres, Mato Grosso, BRA	-57.2507	-15.6263
MTR 29440	*Tropidurus callathelys*	Vila Bela da St. Trindade, Mato Grosso, BRA	-60.0703	-14.9111
MTR 29439	*Tropidurus xanthochilus*	Vila Bela da St. Trindade, Mato Grosso, BRA	-59.9546	-15.0086
MTR 29393	*Tropidurus guarani*	Villarrica, Guairá, PAR	-56.2890	-25.8323
MTR 29379	*Tropidurus guarani*	Acahay, Paraguarí, PAR	-57.1733	-25.8942
MTR 29373	*Tropidurus guarani*	Yaguarón, Paraguarí, PAR	-57.2941	-25.5675
LJAM-CNP 12113	*Tropidurus spinulosus*	Ojo de Água, ARG	-63.6961	-29.5004
LG1551/2987	*Tropidurus xanthochilus*	Chapada dos Guimarães, Mato Grosso, BRA	-55.8000	-14.8666
CHUNB74163	*Tropidurus tarara*	Três Lagoas, Mato Grosso do Sul, BRA	-52.1941	-20.5516
CHUNB74157	*Tropidurus tarara*	Ribas Rio Pardo, Mato Grosso do Sul, BRA	-53.4730	-20.2532
CHUNB74156	*Tropidurus tarara*	Ribas Rio Pardo, Mato Grosso do Sul, BRA	-53.4730	-20.2532
CHUNB74155	*Tropidurus tarara*	Camapuã, Mato Grosso do Sul, BRA	-54.1364	-19.3246
CHUNB74147	*Tropidurus xanthochilus*	Alcinópolis, Mato Grosso do Sul, BRA	-53.4039	-18.8577
CHUNB74146	*Tropidurus xanthochilus*	Alcinópolis, Mato Grosso do Sul, BRA	-53.4325	-18.6444
CHUNB74145	*Tropidurus xanthochilus*	Santo Ant. do Leverger, Mato Grosso, BRA	-55.3031	-15.5231
CHUNB74137	*Tropidurus xanthochilus*	Cáceres, Mato Grosso, BRA	-57.3437	-16.3297
AMCC 204553	*Tropidurus melanopleurus*	Águas Blancas, Oran, ARG	-64.3696	-22.7274
AMCC 204537	*Tropidurus xanthochilus*	Santa Cruz, BOL	-61.0344	-14.7665
AMCC 204510	*Tropidurus spinulosus*	Boquerón, PAR	-60.5332	-20.3763
AMCC 204498	*Tropidurus spinulosus*	Boquerón, PAR	-61.6625	-21.2040

AMCC 204481	*Tropidurus spinulosus*	Loma Plata, Boquerón, PAR	-59.8611	-22.3435
AMCC 204456	*Tropidurus lagunablanca*	Santa Rosa del Araguay, San Pedro, PAR	-56.2946	-23.8120
AMCC 204453	*Tropidurus guarani*	Paraguarí, PAR	-57.1296	-25.6068
AMCC 204421	*Tropidurus tarara*	Concepcíon, PAR	-57.3693	-22.6923
AMCC 204384	*Tropidurus spinulosus*	Rio Verde, Presidente Hayes, PAR	-59.2027	-23.2140
AMCC 204364	*Tropidurus teyumirim*	Paraguarí, PAR	-56.8702	-26.0499
AMCC 204357	*Tropidurus guarani*	San Bernardino, Cordillera, PAR	-57.3027	-25.3071
AMCC 204349	*Tropidurus guarani*	Pirebebuy, Cordillera, PAR	-57.0456	-25.5144
946104	*Tropidurus xanthochilus*	Alto Araguaia, Mato Grosso, BRA	-53.2207	-17.3136

EPÍLOGO

Esta tese buscou integrar abordagens metodológicas e empíricas para compreender fenômenos e processos diversos em Biologia Evolutiva. Ao longo dos capítulos, foram investigados desde fundamentos conceituais sobre o uso de aprendizado de máquina em delimitação de espécies até reconstruções históricas detalhadas baseadas em dados genômicos e modelos paleoclimáticos para lagartos do grupo *Tropidurus spinulosus*. A sequência dos capítulos reflete, portanto, um percurso que vai do desenvolvimento de ferramentas analíticas e reflexões metodológicas à aplicação concreta de ferramentas de inteligência artificial para problemas em Biologia Evolutiva até a elucidação da história evolutiva de um grupo neotropical de lagartos.

O primeiro capítulo discutiu criticamente o papel do aprendizado de máquina na delimitação de espécies, um campo em rápida expansão dentro da biologia evolutiva. A revisão mostrou que esta ferramenta tem potencial para lidar com conjuntos de dados extensos e multidimensionais, oferecendo alternativas eficientes para problemas complexos como a detecção de limites entre espécies. Entretanto, também evidenciou que o entusiasmo em torno dessas abordagens deve ser acompanhado por cautela metodológica. Modelos de aprendizado de máquina dependem de conjuntos de treino e teste representativos e podem ser sensíveis a pressupostos implícitos. Assim, o capítulo propôs boas práticas para o uso responsável do aprendizado de máquina em delimitação de espécies, enfatizando a importância da simulação de cenários realistas, da avaliação de adequação de modelos e da combinação entre predições automatizadas e interpretações biológicas fundamentadas.

No segundo capítulo, essas considerações teóricas foram testadas de maneira preliminar, mas ao mesmo tempo aprofundada, por meio de experimentos de simulação que avaliaram a capacidade de classificadores supervisionados de diferenciar distintos modelos de delimitação de espécies. Os resultados demonstraram que, mesmo quando treinados sob modelos de substituição simples, esses algoritmos conseguem generalizar adequadamente para dados com níveis moderados de heterogeneidade entre loci, apresentando boa acurácia na identificação de padrões de divergência e fluxo gênico. Os resultados reforçam que o uso de aprendizado de máquina em inferências evolutivas deve ser acompanhado por diagnósticos de adequação dos modelos evolutivos avaliados por simulações sob múltiplas condições, garantindo um equilíbrio entre eficiência computacional e realismo biológico.

O terceiro capítulo abordou outro desafio metodológico: a modelagem da heterogeneidade molecular em inferências de taxa de substituição baseadas em genomas mitocondriais de Squamata. As estimativas obtidas indicaram taxas médias de substituição na faixa de 0,006 a 0,02 substituições por sítio por milhão de anos para Pleurodonta, mas com ampla variação entre genes e posições de códon. Em suma, o estudo demonstrou a necessidade de particionamentos adequados e calibrações consistentes com a biologia do grupo, reforçando que inferências temporais robustas dependem não apenas da quantidade de dados, mas da qualidade da modelagem evolutiva.

Os capítulos seguintes aplicaram praticamente todo esse arcabouço metodológico ao estudo empírico do grupo *Tropidurus spinulosus*, um conjunto de espécies amplamente distribuídas em formações abertas da América do Sul centro-meridional. A integração de dados genômicos revelou uma filogenia nuclear bem resolvida, porém parcialmente discordante em relação ao sinal mitocondrial. Essa discordância mitonuclear, frequentemente interpretada como um desafio ao entendimento da história evolutiva dos organismos, foi aqui examinada sob diferentes hipóteses biogeográficas e demográficas. As análises integradas de genealogia, estrutura populacional e modelagem demográfica indicaram que a incongruência entre os genomas nuclear e mitocondrial resulta principalmente de eventos históricos de introgressão mitocondrial, possivelmente associados a contatos secundários em períodos de instabilidade climática. Em outras palavras, os padrões genéticos observados refletem uma história evolutiva permeada por episódios de hibridização e captura mitocondrial, e não apenas processos de divergência alopátrica estrita.

Essa interpretação foi reforçada pela análise temporal e geográfica conduzida no quinto capítulo, que reconstruiu a diversificação do grupo em diferentes escalas temporais. As estimativas de tempo situam as divergências mais profundas no Mioceno, um período marcado por eventos geológicos e marinhos que remodelaram a paisagem da América do Sul. Posteriormente, durante o Plio–Pleistoceno, a reorganização hidrográfica (em especial o estabelecimento do curso moderno do rio Paraguai) contribuiu para o isolamento e a diferenciação de linhagens, funcionando como uma importante barreira biogeográfica. Já no Pleistoceno, as flutuações climáticas promoveram eventos de expansão demográfica, criando oportunidades para contato secundário e introgressão. Assim, a história evolutiva do grupo de espécies *Tropidurus spinulosus* reflete uma interação complexa entre vicariância neógena e dinâmica quaternária, em um contexto marcado pela permeabilidade de barreiras e pela persistência de fluxo gênico episódico.

Em conjunto, os resultados desta tese demonstram que compreender os processos de diversificação em sistemas biológicos requer abordagens verdadeiramente integrativas, capazes de extrair evidências biológicas robustas a partir de múltiplas fontes de dados. A conjugação entre simulações, aprendizado de máquina, filogenômica e modelagem ecológica não apenas aprimora o potencial analítico das ferramentas disponíveis, mas também permite reinterpretar a história evolutiva de grupos amplamente distribuídos e ecologicamente diversos, como *Tropidurus spinulosus*. Os avanços metodológicos aqui propostos têm potencial para aplicação em outros sistemas biológicos, especialmente naqueles em que os limites entre espécies são difusos ou permeáveis.

Esta tese reforça, ainda, que a evolução raramente segue trajetórias lineares ou discretas. A história do grupo de espécies *Tropidurus spinulosus* exemplifica como processos em múltiplas escalas temporais (desde deriva genética e o fluxo gênico até mudanças tectônicas e climáticas) interagem de forma contingente para gerar padrões complexos de diversidade ecológica e discordância genômica. Ao integrar ferramentas analíticas modernas e uma perspectiva evolutiva sob diferentes escalas, este trabalho contribui para o entendimento de como a história geológica e climática da América do Sul se reflete, em detalhe, na evolução de sua fauna, oferecendo um caminho metodológico robusto para futuras investigações sobre a formação da biodiversidade neotropical.

Por fim, esta tese reforça o papel central da ciência aberta como componente intrínseco da pesquisa contemporânea em Biologia Evolutiva. Ao tornar públicos todos os dados, rotinas analíticas e fluxos de trabalho utilizados, este trabalho se compromete com a transparência e com a possibilidade de verificação independente dos resultados. Essa postura, além de promover reprodutibilidade, amplia o alcance e o potencial transformador da pesquisa ao permitir que outros grupos utilizem, critiquem ou expandam os achados aqui apresentados. Em um cenário marcado pela crise de reprodutibilidade em diversas áreas do conhecimento, a construção de uma ciência mais colaborativa e acessível depende de gestos concretos como esse, que democratizam o conhecimento, mitigam barreiras à validação independente e fortalecem a confiança na prática científica.

REFERÊNCIAS

ABADI, S.; AZOURI, D.; PUPKO, T.; MAYROSE, I. Model selection may not be a mandatory step for phylogeny reconstruction. Nature Communications, v. 10, n. 1, p. 934, 2019.

ALENCAR L. R., et al. Opportunity begets opportunity to drive macroevolutionary dynamics of a diverse lizard radiation. Evolution Letters 8: 623-637, 2024.

ALLIO, R., et al. MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. Molecular Ecology Resources, 20: 892–905. 2020.

ANDERSEN, M. J., et al. Complex histories of gene flow and a mitochondrial capture event in a non-sister pair of birds. Molecular Ecology, 30: 2087–2103. 2021. https://doi.org/10.1111/mec.15856

ANDREWS, S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. 2010.

ANGERMUELLER, C., et al. Deep learning for computational biology. Molecular Systems Biology, 12, 2016.

ANISIMOVA M., KOSIOL C. Investigating protein-coding sequence evolution with probabilistic codon substitution models. Molecular Biology and Evolution 26: 255-271, 2009.

ANTONELLI, A; et al. Molecular studies and phylogeography of Amazonian tetrapods and their relation to geological and climatic models. Amazonia, landscape and species evolution: a look into the past, v. 386, p. 404, 2010.

ANTONELLI, A.; ZIZKA, A.; CARVALHO, F. A. SCHARN, R., BACON, C. D., SILVESTRO, D., & CONDAMINE, F. L. Amazonia is the primary source of Neotropical biodiversity. Proc Natl Acad Sci USA, v. 115, p. 6034–6039, 2018.

ARCONES A., et al. Mitochondrial substitution rates estimation for divergence time analyses in modern birds based on full mitochondrial genomes. Ibis 163: 1463-1471, 2021.

ARENAS, M. Simulation of molecular data under diverse evolutionary scenarios. PloS Computational Biology, 8, 2012.

ARENAS, M. Trends in substitution models of molecular evolution. Frontiers in Genetics, v. 6, p. 319, 2015.

ARNAB, S. P., et al. Uncovering footprints of natural selection through spectral analysis of genomic summary statistics. Molecular Biology and Evolution, 40, 2023.

ARNOLD, M. L. Natural hybridization as an evolutionary process. Annual Review of Ecology and Systematics, 23, 237–261, 1992.

AVILA L. J., MARTÍNEZ L. E., MORANDO M. Checklist of lizards and amphisbaenians of Argentina: an update, 2013.

AVISE J. C., et al. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Annu Rev Ecol Syst 18: 489–522, 1987.

BAELE, G. et al. BEAST X for Bayesian phylogenetic, phylogeographic and phylodynamic inference. Nature Methods, p. 1-4, 2025.

BALLARD J. W. O., WHITLOCK M. C. The incomplete history of mitochondria. Molecular Ecology 13: 729–744, 2004.

BALLARD J. W. O., RAND D. M. The population biology of mitochondrial DNA and its phylogenetic implications. Annu. Rev. Ecol. Evol. Syst. 36: 621-642, 2005.

BANKEVICH, A., et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology, 19: 455–477. 2012.

BEAUMONT, M. A., et al. Approximate Bayesian computation in population genetics. Genetics, 162, 2025–2035, 2002.

BENAVIDES, E., BAUM, R., MCCLELLAN, D., SITES JR., J. W. Molecular phylogenetics of the lizard genus Microlophus (Squamata: Tropiduridae): aligning and retrieving indel signals from nuclear introns. Systematic Biology, 56: 776–797. 2007.

BENAVIDES, E., BAUM, R., SNELL, H. M., SNELL, H. L., SITES JR., J. W. Island biogeography of Galápagos lava lizards (Tropiduridae: Microlophus): species diversity and colonization of the archipelago. Evolution, 63: 1606–1626. 2009.

BERGERON L. A., et al. Evolution of the germline mutation rate across vertebrates. Nature 615: 285-291, 2023.

BERNARDO P. H., et al. Extreme mito-nuclear discordance in a peninsular lizard: the role of drift, selection, and climate. Heredity 123: 359-370, 2019.

BERTORELLE, G., et al. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. Molecular Ecology, 19, 2609–2625, 2010. Doi:10.1111/j.1365-294X.2010.04690.x

BEZERRA, C. H., et al. Biogeographical Origins of Caatinga Squamata Fauna. Journal of Biogeography, 52: 521–531. 2025.

BININDA-EMONDS O. R. P. Fast genes and slow clades: comparative rates of molecular evolution in mammals. Evolutionary Bioinformatics 3: 59-85, 2007.

BLANKERS T., et al. Contrasting global-scale evolutionary radiations: phylogeny, diversification, and morphological evolution in the major clades of iguanian lizards. Biological Journal of the Linnean Society 108: 127-143, 2013.

BLEIDORN, C. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. Systematics and biodiversity, v. 14, n. 1, p. 1-8, 2016.

BLISCHAK, P. D., et al. Chromosome-scale inference of hybrid speciation and admixture with convolutional neural networks. Molecular Ecology Resources, 21, 2676–2688, 2021. https://doi.org/10.1111/1755-0998.13355

BOLGER, A. M., LOHSE, M., USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30: 2114–2120. 2014.

BONNET, T., et al. A reassessment of explanations for discordant introgressions of mitochondrial and nuclear genomes. Evolution, 71: 2140–2158. 2017.

BOOKER, W. W., et al. This population does not exist: learning the distribution of evolutionary histories with generative adversarial networks. Genetics, 224(2), iyad063, 2023.

BOFKIN L., GOLDMAN N. Variation in evolutionary processes at different codon positions. Molecular Biology and Evolution 24: 513-521, 2007.

BOROWIEC, M. L. AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ, 4: e1660. 2016.

BOROWIEC, M. L., et al. Deep learning as a tool for ecology and evolution. Methods in Ecology and Evolution, 13, 1640–1660, 2022.

BORTOLUS, A. Error cascades in the biological sciences: the unwanted consequences of using bad taxonomy in ecology. AMBIO: A Journal of the Human Environment, 37, 114–118, 2008.

BOUCKAERT R. R., DRUMMOND A. J. bModelTest: Bayesian phylogenetic site model averaging and model comparison. BMC evolutionary biology 17: 1-11, 2017.

BOUCKAERT, R., et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PloS Computational Biology, 15: e1006650. 2019.

BOUSMALIS, K., et al. Unsupervised pixel-level domain adaptation with generative adversarial networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3722–3731, 2017.

BREITMAN, M. F., et al. A new species of Enyalius (Squamata, Leiosauridae) endemic to the Brazilian Cerrado. Herpetologica, 74, 355–369, 2018.

BROUGHTON R. E., RENEAU P. C. Spatial covariation of mutation and nonsynonymous substitution rates in vertebrate mitochondrial genomes. Molecular Biology and Evolution 23: 1516-1524, 2006.

BUCKLEY T. R., SIMON C., CHAMBERS G. K. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. Systematic Biology 50: 67-86, 2001.

BUCKLEY T. R. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. Systematic Biology 51: 509-523, 2002.

BURBRINK, F. T., & RUANE, S. Contemporary philosophy and methods for studying speciation and delimiting species. Ichthyology & Herpetology, 109, 874–894, 2021.

BURBRINK, F. T., GEHARA, M., MCKELVY, A. D., MYERS, E. A. Resolving spatial complexities of hybridization in the context of the gray zone of speciation in North American ratsnakes (Pantherophis obsoletus complex). Evolution, 75: 260–277. 2021.

BUSHNELL, B. BBTools: a suite of fast, multithreaded bioinformatics tools designed for analysis of DNA and RNA sequence data. Joint Genome Institute. 2018.

CALLIER, V. Machine learning in evolutionary studies comes of age. Proceedings of the National Academy of Sciences, 119, 2022.

CAMURUGI F., et al. Isolation by distance and past climate resistance shaped the distribution of genealogical lineages of a neotropical lizard. Systematics and Biodiversity 20, 2022. https://doi.org/10.1080/14772000.2022.2084470.

CARSTENS, B. C., et al. How to fail at species delimitation. Molecular Ecology, 22, 4369–4383, 2013.

CARSTENS, B.C., PELLETIR, T.A., REID, N.M. & SATLER, J.D. 2013. How to fail at species delimitation. Molecular Ecology, v. 22, p. 4369–83.

CARVALHO, A. L. G. On the distribution and conservation of the South American lizard genus *Tropidurus* Wied-Neuwied, 1825 (Squamata: Tropiduridae). Zootaxa, v. 3640, n. 1, p. 42-56, 2013.

CARVALHO, A. L. G. 2016. Three New Species of the *Tropidurus* spinulosus Group (Squamata: Tropiduridae) from Eastern Paraguay. American Museum Novitates, 3853(3853): 1–44, doi:10.1206/3853.1.

CARVALHO, A. L. G., DE BRITTO, M. R., FERNANDES, D. S. Biogeography of the lizard genus *Tropidurus* Wied-Neuwied, 1825 (Squamata: Tropiduridae): distribution, endemism, and area relationships in South America. PloS One, 8. 2013.

CARVALHO, A. L. G., et al. A new *Tropidurus* (Tropiduridae) from the semiarid Brazilian Caatinga: evidence for conflicting signal between mitochondrial and nuclear loci affecting the phylogenetic reconstruction of South American collared lizards. American Museum Novitates, 3852: 1–68, 2016.

CARVALHO, A. L.G.; SENA M. A.; PELOSO, P. L.V.; MACHADO, F. A.; MONTESINOS, R.; SILVA H. R.; CAMPBELL G.; RODRIGUES M. T. A. New *Tropidurus* (Tropiduridae) from the Semiarid Brazilian Caatinga: Evidence for

Conflicting Signal between Mitochondrial and Nuclear Loci Affecting the Phylogenetic Reconstruction of South American Collared Lizards. American Museum Novitates, v. 3852, p. 1–68, 2016.

CARVALHO, A. L. G., et al. A highly polymorphic South American collared lizard (Tropiduridae: Tropidurus) reveals that open-dry refugia from South-western Amazonia staged allopatric speciation. Zoological Journal of the Linnean Society 201: 493–533, 2024. https://doi.org/10.1093/zoolinnean/zlad138.

CASTRESANA, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution, 17: 540–552. 2000.

CERCA, J., et al. Removing the bad apples: A simple bioinformatic method to improve loci-recovery in de novo RADseq data for non-model organisms. Methods in Ecology and Evolution, 12, 805–817, 2021.

CHADHA, A., et al. Deepfake: An overview. Proceedings of Second International Conference on Computing, Communications, and Cyber-Security, 557–566, Springer, Singapore, 2021.

CHAN L. M., et al. Defining spatial and temporal patterns of phylogeographic structure in Madagascar’s iguanid lizards (genus Oplurus). Molecular Ecology 21: 3839-3851, 2012.

CHICCO, D. Ten quick tips for machine learning in computational biology. BioData Mining, 10, 1–17, 2017. https://doi.org/10.1186/s13040-017-0155-3

CHINTALAPATI M., MOORJANI P. Evolution of the mutation rate across primates. Current opinion in genetics & development 62: 58-64, 2020.

CHRISTIN, S., et al. Applications for deep learning in ecology. Methods in Ecology and Evolution, 10, 1632–1644, 2019.

CHUNG, D. J., SCHULTE, P. M. Mitochondria and the thermal limits of ectotherms. Journal of Experimental Biology, 223: jeb227801. 2020.

CLEMENT, M., et al. TCS: estimating gene genealogies. In: Parallel and Distributed Processing Symposium, International (Vol. 2, pp. 7-pp). IEEE Computer Society. 2002.

COCHRAN, K., et al. Domain adaptive neural networks improve cross-species prediction of transcription factor binding. Genome Research, 32, 512–523, 2022.

COLLIN, F. D., et al. Extending approximate Bayesian computation with supervised machine learning to infer demographic history from genetic polymorphisms using DIYABC Random Forest. Molecular Ecology Resources, 21, 2598–2613, 2021. https://doi.org/10.1111/1755-0998.13413.

CONRAD J. L., & NORELL M. A. A complete Late Cretaceous iguanian (Squamata, Reptilia) from the Gobi and identification of a new iguanian clade. American Museum Novitates, 2007(3584), 1-47, 2007.

CSILLÉRY, K., et al. Approximate Bayesian computation (ABC) in practice. Trends in Ecology & Evolution, 25, 410–418, 2010.

DAYRAT, B. Towards integrative taxonomy. Biological Journal of the Linnean Society, 85, 407–417, 2005.

DEGNAN, J. H., & ROSENBERG, N. A. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology & Evolution, 24, 332–340, 2009.

DEL AMPARO, R.; ARENAS, M. Consequences of substitution model selection on protein ancestral sequence reconstruction. Molecular Biology and Evolution, v. 39, n. 7, msac144, 2022.

DEL AMPARO, R.; ARENAS, M. Influence of substitution model selection on protein phylogenetic tree reconstruction. Gene, v. 865, p. 147336, 2023.

DERAAD, D. A. SNPfiltR: an R package for interactive and reproducible SNP filtering. Molecular Ecology Resources, 22: 2443–2453. 2022.

DERAAD, D. A., et al. Mitonuclear discordance results from incomplete lineage sorting, with no detectable evidence for gene flow, in a rapid radiation of Todiramphus kingfishers. Molecular Ecology, 32: 4844–4862. https://doi.org/10.1111/mec.17080. 2023.

DERKARABETIAN, S., et al. A demonstration of unsupervised machine learning in species delimitation. Molecular Phylogenetics and Evolution, 139, 2019. https://doi.org/10.1016/j.ympev.2019.106562

DERKARABETIAN, S., et al. Using natural history to guide supervised machine learning for cryptic species delimitation with genetic data. Frontiers in Zoology, 19, 1–15, 2022.

DESPRÉS, L. One, two or more species? Mitonuclear discordance and species delimitation. Molecular Ecology, 28: 3845–3847. https://doi.org/10.1111/mec.15211. 2019.

DESSI, M. C., et al. The role of sex-biased dispersion in promoting mitonuclear discordance in Partamona helleri (Hymenoptera: Apidae: Meliponini). Biological Journal of the Linnean Society, 136: 423–435. 2022.

DIKE, H. U., et al. Unsupervised learning based on artificial neural network: A review. IEEE International Conference on Cyborg and Bionic Systems (CBS), 322–327, 2018.

DOMINGOS, F. M., et al. Out of the deep: cryptic speciation in a Neotropical gecko (Squamata, Phyllodactylidae) revealed by species delimitation methods. Molecular Phylogenetics and Evolution, 80, 113–124, 2014.

DOMINGOS, F.M.C.B.; BOSQUE, R.J.; CASSIMIRO, J.; COLLI, G.R; RODRIGUES, M.T.; SANTOS, M.G.; BEHEGARAY, L.B. Out of the deep: Cryptic speciation in a Neotropical gecko (Squamata, Phyllodactylidae) revealed by species delimitation methods. Molecular Phylogenetics and Evolution, v. 80, p. 113–24, 2014.

DOMINGOS, P. A few useful things to know about machine learning. Communications of the ACM, 55, 78–87, 2012.

DONG, X., et al. Mitochondrial introgression and mito-nuclear discordance obscured the closely related species boundaries in Cletus Stål from China (Heteroptera: Coreidae). Molecular Phylogenetics and Evolution, 184. https://doi.org/10.1016/j.ympev.2023.107802. 2023.

DOS REIS, M., YANG, Z. Bayesian molecular clock dating using genome-scale datasets. In: Evolutionary genomics: Statistical and computational methods. New York, NY: Springer New York, 2019. P. 309-330.

DUAN, L., et al. Species delimitation of the liquorice tribe (Leguminosae: Glycyrrhizeae) based on phylogenomic and machine learning analyses. Journal of Systematics and Evolution, 61, 22–41, 2023. https://doi.org/10.1111/jse.12902.

DRUMMOND A. J., RAMBAUT A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology 7: 1-8, 2007.

DUCHÊNE S., et al. Mitogenome phylogenetics: The impact of using single regions and partitioning schemes on topology, substitution rate and divergence time estimation. PloS ONE 6, 2011. https://doi.org/10.1371/journal.pone.0027138.

DUCHÊNE S., et al. The impact of calibration and clock-model choice on molecular estimates of divergence times. Molecular Phylogenetics and Evolution 78: 277-289, 2014.

DUFRESNES, C., et al. Are glacial refugia hotspots of speciation and cytonuclear discordances? Answers from the genomic phylogeography of Spanish common frogs. Molecular Ecology, 29: 986–1000. https://doi.org/10.1111/mec.15368. 2020.

DURAND, E. Y., PATTERSON, N., REICH, D., SLATKIN, M. Testing for ancient admixture between closely related populations. Molecular Biology and Evolution, 28: 2239–2252. 2011.

ECHAVE, J., SPIELMAN, S. & WILKE, C. Causes of evolutionary rate variation among protein sites. Nat Rev Genet, v. 17, 109–121, 2016.

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

EDWARDS, S. V., et al. Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. Molecular Phylogenetics and Evolution, 94, 447–462, 2016.

EDWARDS, S. V., et al. Reticulation, divergence, and the phylogeography–phylogenetics continuum. Proceedings of the National Academy of Sciences, 113, 8025–8032, 2016.

ELY, C. V., et al. Implications of poor taxonomy in conservation. Journal for Nature Conservation, 36, 10–13, 2017.

EO S. H., DEWOODY J. A. Evolutionary rates of mitochondrial genomes correspond to diversification rates and to contemporary species richness in birds and reptiles. Proceedings of the Royal Society B: Biological Sciences 277: 3587-3592, 2010.

EXCOFFIER, L., FOLL, M., PETIT, R. J. Genetic consequences of range expansions. Annual Review of Ecology, Evolution, and Systematics, 40: 481–501. 2009.

FAIRCLOTH, B. C., et al. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Systematic Biology, 61: 717–726. 2012.

FAIRCLOTH, B. C. PHYLUCE is a software package for the analysis of conserved genomic loci. Bioinformatics, 32: 786–788. 2016.

FAN, X. K., WU, et al. Phylogenomic, morphological, and niche differentiation analyses unveil species delimitation and evolutionary history of endangered maples in Acer series Campestria (Sapindaceae). Journal of Systematics and Evolution, 61, 284–298, 2023. https://doi.org/10.1111/jse.12919.

FARAHANI, A., et al. A brief review of domain adaptation. Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020, 877–894, 2021.

FEDOROV, V. B., et al. Conflicting nuclear and mitogenome phylogenies reveal ancient mitochondrial replacement between two North American species of collared lemmings (Dicrostonyx groenlandicus, D. hudsonius). Molecular Phylogenetics and Evolution, 168. 2022.

FENG S., et al. Dense Sampling of Bird Diversity Increases Power of Comparative Genomics. Nature 587: 252–257, 2020.

FIRNENO, T. J., et al. Delimitation despite discordance: evaluating the species limits of a confounding species complex in the face of mitonuclear discordance. Ecology and Evolution, 11: 12739–12753. https://doi.org/10.1002/ece3.8018. 2021.

FLAGEL, L., et al. The unreasonable effectiveness of convolutional neural networks in population genetic inference. Molecular Biology and Evolution, 36, 220–238, 2019.

FLOURI, T., et al. Species Tree Inference with BPP using Genomic Sequences and the Multispecies Coalescent. Molecular Biology and Evolution, 35, 2585–2593, 2018. Doi:10.1093/molbev/msy147.

FLOURI, T., JIAO, X., RANNALA, B., YANG, Z. Species tree inference with BPP using genomic sequences and the multispecies coalescent. Molecular Biology and Evolution, 35: 2585–2593. https://doi.org/10.1093/molbev/msy147. 2018.

FLOURI, T., et al. A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. Molecular Biology and Evolution, 37, 1211–1223, 2020.

FLOURI, T., JIAO, X., HUANG, J., RANNALA, B., YANG, Z. Efficient Bayesian inference under the multispecies coalescent with migration. Proceedings of the National Academy of Sciences, 120: e2310708120. 2023.

FONSECA, E. M., et al. Phylogeographic model selection using convolutional neural networks. Molecular Ecology Resources, 21, 2661–2675, 2021. https://doi.org/10.1111/1755-0998.13427.

FONSECA, E. M., & CARSTENS, B. C. Artificial intelligence enables unified analysis of historical and landscape influences on genetic diversity. Molecular Phylogenetics and Evolution, 108116, 2024.

FONSECA, R. R., et al. Next-generation biology: sequencing and data analysis approaches for non-model organisms. Marine Genomics, 30, 3–13, 2016.

FONTANELLA F. M., et al. Molecular dating and diversification of the South American lizard genus Liolaemus (subgenus Eulaemus) based on nuclear and mitochondrial DNA sequences. Zoological Journal of the Linnean Society 164: 825-835, 2012.

FOUNTAIN-JONES, N. M., et al. Machine learning in molecular ecology. Molecular Ecology Resources, 21, 2589–2597, 2021. https://doi.org/10.1111/1755-0998.13532.

FROST, D. R., CRAFTS, H. M., FITZGERALD, L. A., TITUS, T. A. Geographic variation, species recognition, and molecular evolution of cytochrome oxidase I in the Tropidurus spinulosus complex (Iguania: Tropiduridae). Copeia, 839–851. 1998.

FUJITA, M. K., et al. Coalescent-based species delimitation in an integrative taxonomy. Trends in Ecology & Evolution, 27, 480–488, 2012.

FUNK D. J., OMLAND K. E. Species-Level Paraphyly and Polyphyly: Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. Annual Review of Ecology, Evolution, and Systematics 34: 397–423, 2003. Doi:10.1146/annurev.ecolsys.34.01

GABLE S. M., et al. The State of Squamate Genomics: Past, Present, and Future of Genome Research in the Most Speciose Terrestrial Vertebrate Order. In Genes (Vol. 14, Issue 7). Multidisciplinary Digital Publishing Institute (MDPI), 2023. https://doi.org/10.3390/genes14071387

GANIN, Y., & LEMPITSKY, V. Unsupervised domain adaptation by backpropagation. International Conference on Machine Learning, 1180–1189, 2015.

GEHARA, M., et al. PipeMaster: inferring population divergence and demographic history with approximate Bayesian computation and supervised machine-learning in R. bioRxiv, 2020-12, 2020. https://doi.org/10.1101/2020.12.04.410670

GENEREUX D. P., et al. A Comparative Genomics Multitool for Scientific Discovery and Conservation. Nature 587: 240–245, 2020.

GHIFARY, M., et al. Deep Reconstruction Classification Networks for Unsupervised Domain Adaptation. In: LEIBE, B., MATAS, J., SEBE, N., WELLING, M., eds. Computer Vision ECCV 2016. Lecture Notes in Computer Science. Cham: Springer International Publishing, p. 597, 2016.

GHIROTTO, S., et al. Distinguishing among complex evolutionary models using unphased whole-genome data through random forest approximate Bayesian computation. Molecular Ecology Resources, 21, 2614–2628, 2021. https://doi.org/10.1111/1755-0998.13263.

GOMPERT, Z., et al. Analysis of population genomic data from hybrid zones. Annual Review of Ecology, Evolution, and Systematics, 48, 207–229, 2017.

GOOD, J. M., VANDERPOOL, D., KEEBLE, S., BI, K. Negligible nuclear introgression despite complete mitochondrial capture between two species of chipmunks. Evolution, 69: 1961–1972. https://doi.org/10.1111/evo.12712. 2015.

GOODFELLOW, I., et al. Generative adversarial nets. Advances in Neural Information Processing Systems, 2672–2680, 2014.

GRAY M. W., BURGER G., LANG B. F. Mitochondrial evolution. Science 283: 1476-1481, 1999.

GREENER, J. G., et al. A guide to machine learning for biologists. Molecular Cell Biology, 23, 40–55, 2021. https://doi.org/10.1038/s41580-021-00407-0.

GUILLORY, W. X., BROWN, J. L. A new method for integrating ecological niche modeling with phylogenetics to estimate ancestral distributions. Systematic Biology, 70: 1033–1045. 2021.

GUILLORY, W. X., et al. Geoclimatic drivers of diversification in the largest arid and semi-arid environment of the Neotropics: perspectives from phylogeography. Molecular Ecology, e17431. 2024.

GUPTA, M. K.; VADDE, R. Next-generation development and application of codon model in evolution. Frontiers in Genetics, v. 14, p. 1091575, 2023.

GUTENKUNST, R. N., HERNANDEZ, R. D., WILLIAMSON, S. H., BUSTAMANTE, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP data. PloS Genetics, 5: e1000695. 2009.

GUYON, I., & ELISSEEFF, A. An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157–1182, 2003.

HARRIS, R. S. Improved pairwise alignment of genomic DNA. The Pennsylvania State University. 2007.

HARRISON, R. G., & LARSON, E. L. Heterogeneous genome divergence, differential introgression, and the origin and structure of hybrid zones. Molecular Ecology, 25, 2454–2466, 2016.

HARVEY, M. B., GUTBERLET JR, R. L. Lizards of the genus *Tropidurus* (Iguania: Tropiduridae) from the Serranía de Huanchaca, Bolivia: new species, natural history, and a key to the genus. Herpetologica, 493–520. 1998.

HARVEY, M. G., et al. Sequence capture versus restriction site associated DNA sequencing for shallow systematics. Systematic Biology, 65: 910–924. 2016.

HASSANIN A. Phylogeny of Arthropoda inferred from mitochondrial sequences: strategies for limiting the misleading effects of multiple changes in pattern and rates of substitution. Molecular Phylogenetics and Evolution 38: 100-116, 2006.

HASTIE, T., et al. Unsupervised learning. In: The Elements of Statistical Learning, 485–585, Springer, New York, NY, 2009.

HAUSDORF, B. Progress toward a general species concept. Evolution, 65, 923–931, 2011.

HEBERT P. D., et al. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proceedings of the Royal Society of London. Series B: Biological Sciences 270: S96-S99, 2003.

HEIL, B. J., et al. Reproducibility standards for machine learning in the life sciences. Nature Methods, 18, 1132–1135, 2021.

HELED, J., DRUMMOND, A. J. Bayesian inference of population size history from multiple loci. BMC Evolutionary Biology, 8: 1–15. 2008.

HELLER, R., et al. The confounding effect of population structure on Bayesian skyline plot inferences of demographic history. PloS ONE, 8: e62992. https://doi.org/10.1371/journal.pone.0062992. 2013.

HIBBINS, M. S., HAHN, M. W. Phylogenomic approaches to detecting and characterizing introgression. Genetics, 220. 2022.

HINOJOSA, J. C., et al. A mirage of cryptic species: genomics uncover striking mitonuclear discordance in the butterfly Thymelicus sylvestris. Molecular Ecology, 28: 3857–3868. https://doi.org/10.1111/mec.15153. 2019.

HIPSLEY C. A., MÜLLER J. Beyond fossil calibrations: realities of molecular clock practices in evolutionary biology. Frontiers in genetics 5: 138, 2014.

HO S. Y. Calibrating molecular estimates of substitution rates and divergence times in birds. Journal of Avian Biology 38: 409-414, 2007.

HODEL, R. G., et al. A phylogenomic approach, combined with morphological characters gleaned via machine learning, uncovers the hybrid origin and biogeographic diversification of the plum genus. bioRxiv, 2023-09, 2023. https://doi.org/10.1101/2023.09.13.557598.

HOORN, C.; et al. Amazonia Through Time: Andean Uplift, Climate Change, Landscape Evolution, and Biodiversity. Science, v. 330, p. 927–931, 2010.

HUA, X., & MORITZ, C. A phylogenetic approach to delimitate species in a probabilistic way. Systematic Biology, syaf004, 2025.

HUANG, J. P. Is population subdivision different from speciation? From phylogeography to species delimitation. Ecology and Evolution, 10, 6890–6896, 2020.

HUDSON, M. E. Sequencing breakthroughs for genomic ecology and evolutionary biology. Molecular ecology resources, v. 8, n. 1, p. 3-17, 2008.

HUELSENBECK, J. P., et al. Structurama: Bayesian inference of population structure. Evolutionary Bioinformatics, 7, 55–59, 2011.

HÜLLERMEIER, E., et al. Inductive bias. Encyclopedia of Systems Biology, 1018–1019, 2013.

JACKSON, N. D., et al. Species delimitation with gene flow. Systematic Biology, 66, 799–812, 2017.

JACOBS, S. J., et al. Incongruence in molecular species delimitation schemes: What to do when adding more data is difficult. Molecular Ecology, 27, 2397–2413, 2018.

JAMDADE, R., et al. Multilocus marker-based delimitation of Salicornia persica and its population discrimination assisted by supervised machine learning approach. PloS ONE, 17, 2022. https://doi.org/10.1371/journal.pone.0270463.

JARVIS, E. D., et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science, 346, 1320–1331, 2014.

JING Y., et al. Influence of life-history traits on mitochondrial DNA substitution rates exceeds that of metabolic rates in teleost fishes. Current Zoology: zoae045, 2024.

JOHNS G. C., AVISE J. C. A comparative summary of genetic distances in the vertebrates from the mitochondrial cytochrome b gene. Molecular Biology and Evolution 15: 1481-1490, 1998.

JÖRGER, K. M., & SCHRÖDL, M. How to describe a cryptic species? Practical challenges of molecular taxonomy. Frontiers in Zoology, 10, 1–27, 2013.

JORNA, J., et al. Species boundaries in the messy middle—A genome-scale validation of species delimitation in a recently diverged lineage of coastal fog desert lichen fungi. Ecology and Evolution, 11, 18615–18632, 2021.

JUKES, T. H.; CANTOR, C. R. Evolution of protein molecules. In: MUNRO, H. N. (ed.). Mammalian Protein Metabolism. New York: Academic Press, 1969. P. 21-132.

KALYAANAMOORTHY, S., MINH, B. Q., WONG, T. K., VON HAESELER, A., JERMIIN, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods, 14: 587–589. 2017.

KARBSTEIN, K., et al. Species delimitation 4.0: integrative taxonomy meets artificial intelligence. Trends in Ecology & Evolution, 2023.

KATOH, K., STANDLEY, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution, 30: 772–780. 2013.

KHALIGHIFAR, A., et al. Deep learning improves acoustic biodiversity monitoring and new candidate forest frog species identification (genus Platymantis) in the Philippines. Biodiversity and Conservation, 30, 643–657, 2021.

KIMURA M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of molecular evolution 16: 111-120, 1980.

KNAUS, B. J., GRÜNWALD, N. J. vcfr: a package to manipulate and visualize variant call format data in R. Molecular Ecology Resources, 17: 44–53. 2017.

KNOWLES, L. L., & CARSTENS, B. C. Delimiting species without monophyletic gene trees. Systematic Biology, 56, 887–895, 2007.

KORFMANN, K., et al. Deep learning in population genetics. Genome Biology and Evolution, 2023. https://doi.org/10.1093/gbe/evad008.

KUZENKOV, O., et al. Exploring evolutionary fitness in biological systems using machine learning methods. Entropy, 23, 1–35, 2020.

LANFEAR, R.; CALCOTT, B.; HO, S. Y.; GUINDON, S. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Molecular Biology and Evolution, v. 29, n. 6, p. 1695-1701, 2012.

LEACHÉ, A. D., & LINKEM, C. W. Phylogenomics of horned lizards (Genus: Phrynosoma) using targeted sequence capture data. Copeia 103: 586-594, 2015.

LEACHÉ, A. D., et al. The influence of gene flow on species tree estimation: a simulation study. Systematic Biology, 63, 17–30, 2014.

LEACHÉ, A. D., et al. The spectre of too many species. Systematic Biology, 68, 168–181, 2019.

LEAL, B. S. S.; PALMA DA SILVA, C.; PINHEIRO, F. Phylogeographic studies depict the role of space and time scales of plant speciation in a highly diverse Neotropical region. Critical Reviews in Plant Sciences, v. 35, n. 4, p. 215-230, 2016.

LECUN, Y., et al. Deep learning. Nature, 521, 436–444, 2015.

LEIGH, J. W., et al.. POPART: full-feature software for haplotype network construction. Methods in Ecology & Evolution, 6. 2015.

LEMMON A. R., BROWN J. M., STANGER-HALL K., LEMMON E. M. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Systematic Biology 58: 130-145, 2009.

LI, H., et al. The sequence alignment/map format and SAMtools. Bioinformatics, 25: 2078–2079. 2009.

LI, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv, 1303.3997v2 [q-bio.GN]. 2013.

LIBBRECHT, M.W. & NOBLE, W.S. 2015. Machine learning applications in genetics and genomics. Nature Reviews Genetics 16, 32–332.

LIMA, A.P. et al. Not as widespread as thought: Integrative taxonomy reveals cryptic diversity in the Amazonian nurse frog Allobates tinae Melo-Sampaio, Oliveira and Prates, 2018 and description of a new species. Journal of Zoological Systematics and Evolutionary Research, 58(4), 1173–1194. 2020a.

LIMA, L. R. et al. Below the waterline: cryptic diversity of aquatic pipid frogs (Pipa carvalhoi) unveiled through an integrative taxonomy approach. Systematics and Biodiversity, 18(8), 771–783. 2020b.

LIU, B. BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. Briefings in bioinformatics 20, 1280–1294. 2019.

LIU, M.Y. & TUZEL, O. Coupled Generative Adversarial Networks. In: Advances in Neural Information Processing Systems 29. Curran Associates, Inc. https://papers.nips.cc/paper/2016/hash/502e4a16930e414107ee22b6198c578f-Abstract.html. 2016.

LIU, X., FU, Y. X. Stairway Plot 2: demographic history inference with folded SNP frequency spectra. Genome Biology, 21: 280. 2020.

LOSOS J. B. Lizards in an evolutionary tree: ecology and adaptive radiation of anoles (Vol. 10). Univ of California Press, 2011.

LUKHTANOV, V. A. Species Delimitation and Analysis of Cryptic Species Diversity in the XXI Century. Entmol. Rev. 99, 463–472. https://doi.org/10.1134/S0013873819040055. 2019.

LÜRIG, M.D., et al. Computer vision, machine learning, and the promise of phenomics in ecology and evolutionary biology. Frontiers in Ecology and Evolution 9. 2021.

LUTTEROPP, S., et al.. NetRAX: accurate and fast maximum likelihood phylogenetic network inference. Bioinformatics, 38: 3725–3733. 2022.

MACKIEWICZ P., et al. Phylogeny and evolution of the genus Cervus (Cervidae, Mammalia) as revealed by complete mitochondrial genomes. Scientific Reports 12: 16381, 2022.

MAGALHÃES, F. D. M., et al. Taxonomic review of South American Butter Frogs: Phylogeny, geographic patterns, and species delimitation in the Leptodactylus latrans species group (Anura: Leptodactylidae). Herpetological Monographs, 34(1), 131–177. 2020.

MALLET, J., et al. How reticulated are species? BioEssays, 38, 140–149. 2016.

MALINSKY, M., et al. Dsuite—fast D-statistics and related admixture evidence from VCF files. Molecular Ecology Resources, 21: 584–595. 2021.

MANGUL, S. et al. Systematic benchmarking of omics computational tools. Nature communications 10. 2019.

MAO, X., ROSSITER, S. J. Genome-wide data reveal discordant mitonuclear introgression in the intermediate horseshoe bat (Rhinolophus affinis). Molecular Phylogenetics and Evolution, 150: 106886. 2020.

MARTIN, B. T., et al. The choices we make and the impacts they have: Machine learning and species delimitation in North American box turtles (Terrapene spp.). Molecular Ecology Resources 21, 2801–2817. 2021.

MAYR, E. M. The biological meaning of species. Biological Journal of the Linnean society, 1, 311–320. 1969.

MAYR, E. M. What is a species, and what is not? Philosophy of science, 63, 262–277. 1996.

MAYR, E. M. The biological species concept. Species concepts and phylogenetic theory: a debate, 17–29. 2000.

MCCLURE, E. C., et al. Artificial intelligence meets citizen science to supercharge ecological monitoring. Patterns 1. 2020.

MCKENNA, A., et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research, 20: 1297–1303. 2010.

MELLO B., SCHRAGO C. G. Assignment of calibration information to deeper phylogenetic nodes is more effective in obtaining precise and accurate divergence time estimates. Evolutionary Bioinformatics 10: EBO-S13908, 2014.

MELO, B. F.; et al. Cryptic species in the Neotropical fish genus Curimatopsis (Teleostei, Characiformes). Zoologica Scripta, v. 45, n. 6, p. 650-658, 2016.

MEYER, S.; VON HAESELER, A. Identifying site-specific substitution rates. Molecular Biology and Evolution, v. 20, n. 2, p. 182-189, 2003.

MINH, B. Q., et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Molecular Biology and Evolution, 37: 1530–1534. 2020.

MIRARAB, S., et al. Multispecies coalescent: theory and applications in phylogenetics. Annual Review of Ecology, Evolution, and Systematics, 52, 247-268. 2021.

MITCHELL, T. M. Machine Learning. McGraw-Hill, New York. 1997.

MO, Z., & SIEPEL, A. Domain-adaptive neural networks improve supervised machine learning based on simulated population genetic data. PLOS Genetics, 19. 2023.

MO, Y. K., et al. Applications of machine learning in phylogenetics. Molecular Phylogenetics and Evolution, 196, 108066. 2024.

MOMIGLIANO, P.; FLORIN, A. B.; MERILÄ, J. Biases in demographic modeling affect our understanding of recent divergence. Molecular Biology and Evolution, v. 38, n. 7, p. 2967-2985, 2021.

MOORE, W. S. Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees. Evolution, 49: 718–726. 1995.

MORIMOTO, J., et al. Editorial: Applications of Machine Learning to Evolutionary Ecology Data. Frontiers in Ecology and Evolution. 2021.

MOWER J. P., et al. Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants. BMC Evolutionary Biology 7, 2007. https://doi.org/10.1186/1471-2148-7-135

MUELLER R. L. Evolutionary rates, divergence dates, and the performance of mitochondrial genes in Bayesian phylogenetic analysis. Systematic biology 55: 289-300, 2006.

MURRELL, B., et al. Detecting individual sites subject to episodic diversifying selection. PloS Genetics, 8: e1002764. 2012.

MURRELL, B., et al. FUBAR: a fast, unconstrained Bayesian approximation for inferring selection. Molecular Biology and Evolution, 30: 1196–1205. 2013.

MURRELL, B., et al. Gene-wide identification of episodic selection. Molecular Biology and Evolution, 32: 1365–1371. 2015.

MYERS, E. A., et al. Interspecific gene flow and mitochondrial genome capture during the radiation of Jamaican Anolis lizards (Squamata: Iguanidae). Systematic Biology, 71: 501–511. 2022.

NABHOLZ B., et al. Strong variations of mitochondrial mutation rate across mammals-the longevity hypothesis. Molecular Biology and Evolution 25: 120-130, 2008. 10.1093/molbev/msm248.

NABHOLZ B., et al. Body mass-corrected molecular rate for bird mitochondrial DNA. Molecular ecology 25: 4438-4449, 2016.

NASER-KHDOUR, S. et al. The prevalence and impact of model violations in phylogenetic analysis. Genome Biology and Evolution, v. 11, n. 12, p. 3341-3352, 2019.

NESTERENKO, L., et al. Phyloformer: towards fast and accurate phylogeny estimation with self-attention networks. bioRxiv, 2022-06. https://doi.org/10.1101/2022.06.24.496975. 2022.

NEWTON, L. G., et al. Integrative species delimitation reveals cryptic diversity in the southern Appalachian Antrodiaetus unicolor (Araneae: Antrodiaetidae) species complex. Molecular Ecology 29, 2269–2287. 2020.

NON A. L., et al. Identification of the most informative regions of the mitochondrial genome for phylogenetic and coalescent analyses. Molecular Phylogenetics and Evolution 44: 1164-1171, 2007.

O'MEARA B. C. New heuristic methods for joint species delimitation and species tree inference. Systematic Biology 59, 59–73. 2010.

O'MEARA B. C. Evolutionary inferences from phylogenies: a review of methods. Annual Review of Ecology, Evolution, and Systematics 43, 267–285. 2012.

OLAVE M., et al. Model-based approach to test hard polytomies in the Eulaemus clade of the most diverse South American lizard genus Liolaemus (Liolaemini, Squamata). Zoological Journal of the Linnean Society 174: 169-184, 2015.

OLSON, D. M., et al. Terrestrial ecoregions of the world: a new map of life on Earth. BioScience, 51: 933–938. 2001.

PÄCKERT M., et al. Calibration of a molecular clock in tits (Paridae)—Do nucleotide substitution rates of mitochondrial genes deviate from the 2% rule? Molecular Phylogenetics and Evolution 44: 1-14, 2007.

PADIAL, J. M., et al. The integrative future of taxonomy. Frontiers in zoology, 7, 1–14. 2010.

PALACIOS, C., CAMPAGNA, L., PARRA, J. L., CADENA, C. D. Mito-nuclear discordance in the phenotypically variable Andean hummingbirds Coeligena bonapartei and Coeligena helianthea (Trochilidae). Biological Journal of the Linnean Society, 139: 145–157. 2023.

PAN, S. J. & YANG, Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22, 1345–1359. 2010.

PANTE, E., et al. Species are hypotheses: avoid connectivity assessments based on pillars of sand. Molecular Ecology 24, 525–544. 2015.

PARADIS, E. pegas: an R package for population genetics with an integrated–modular approach. Bioinformatics, 26: 419–420. 2010.

PARK E., et al. Estimation of divergence times in cnidarian evolution based on mitochondrial protein-coding genes and the fossil record. Molecular Phylogenetics and Evolution 62: 329-345, 2012.

PARKINSON C. L., et al. Multiple major increases and decreases in mitochondrial substitution rates in the plant family Geraniaceae. BMC Evolutionary Biology 5: 1–12, 2005. https://doi.org/10.1186/1471-2148-5-73

PATWARDHAN A., et al. Molecular markers in phylogenetic studies-a review. Journal of Phylogenetics & Evolutionary Biology 2: 131, 2014.

PEI, J., et al. CLADES: A classification-based machine learning method for species delimitation from population genetic data. Molecular Ecology Resources 18, 1144–1156. https://doi.org/10.1111/1755-0998.12887. 2018.

PEREZ, M. F., et al. Coalescent-based species delimitation meets deep learning: Insights from a highly fragmented cactus system. Molecular Ecology Resources. 2021.

PHUONG, M. A., BI, K., MORITZ, C. Range instability leads to cytonuclear discordance in a morphologically cryptic ground squirrel species complex. Molecular Ecology, 26: 4743–4755. https://doi.org/10.1111/mec.14238. 2016.

PIANKA E. P., VITT L. J. Lizards: windows to the evolution of diversity (Vol. 5). Univ of California Press, 2003.

PICHLER, M., et al. Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. Methods in Ecology and Evolution 11, 281–293. 2020.

PIGLIUCCI, M. Species as family resemblance concepts: the (dis-)solution of the species problem? BioEssays, 25, 596–602. 2003.

PLAZZI F., PUCCIO G., PASSAMONTI M. Comparative large-scale mitogenomics evidences clade specific evolutionary trends in mitochondrial DNAs of Bivalvia. Genome Biology and Evolution 8: 2544-2564, 2016.

PONS J., et al. Nucleotide substitution rates for the full set of mitochondrial protein-coding genes in Coleoptera. Molecular Phylogenetics and Evolution 56: 796–807, 2010. https://doi.org/10.1016/j.ympev.2010.02.007.

PONS, J., et al. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Systematic Biology 55, 595–609. 2006.

PORTELLI S. N., et al. Historical biogeographic reconstruction of the South American Liolaemus boulengeri group (Iguania: Liolaemidae). South American Journal of Herpetology 25: 41-56, 2022.

POSADA, D.; CRANDALL, K. A. Selecting the best-fit model of nucleotide substitution. Systematic Biology, v. 50, n. 4, p. 580-601, 2001.

PRATES, I.; XUE, A.T.; BROWN, J.L.; ALVARADO-SERRANO, D.F.; RODRIGUES, M.T.; HICKERSON, M.J.; CARNAVAL, A.C. Inferring responses to climate dynamics from historical demography in neotropical forest lizards. Proceedings of the National Academy of Sciences, v. 113, n. 7978-7985, 2016.

PRATES I., et al. Phylogenetic relationships of Amazonian anole lizards (Dactyloa): taxonomic implications, new insights about phenotypic evolution and the timing of diversification. Molecular phylogenetics and evolution 82: 258-268, 2015.

PRICE, T. D., et al. The role of phenotypic plasticity in driving genetic evolution. Proceedings of the Royal Society of London. Series B: Biological Sciences 270, 1433–1440. 2003.

PRITCHARD, J. K., et al. Inference of population structure using multilocus genotype data. Genetics 155, 945–959. 2000.

PUDLO, P., et al. Reliable ABC model choice via random forests. Bioinformatics 32, 859–866. https://doi.org/10.1093/bioinformatics/btv684. 2016.

PULLER, V.; SAGULENKO, P.; NEHER, R. A. Efficient inference, potential, and limitations of site-specific substitution models. Virus Evolution, v. 6, n. 2, veaa066, 2020.

PURCELL, S., et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics, 81: 559–575. 2007.

PYRON, R. A. Unsupervised machine learning for species delimitation, integrative taxonomy, and biodiversity conservation. Molecular Phylogenetics and Evolution, v. 189, p. 107939, 2023.

PYRON, R. A., et al. Speciation hypotheses from phylogeographic delimitation yield an integrative taxonomy for Seal Salamanders (Desmognathus monticola). Systematic Biology, 72, 179–197. 2023.

PYRON, R. A., et al. Discerning structure versus speciation in phylogeographic analysis of Seepage Salamanders (Desmognathus aeneus) using demography, environment, geography, and phenotype. Molecular Ecology, 33, e17219. 2024.

QI, Y., et al. Genetic evidence for male-biased dispersal in the Qinghai toad-headed agamid Phrynocephalus vlangalii and its potential link to individual social interactions. Ecology & Evolution, 3: 1219–1230. 2013.

QU, K., et al. Application of machine learning in microbiology. Frontiers in Microbiology 10. 2019.

DE QUEIROZ, K. The general lineage concept of species, species criteria, and the process of speciation. Endless forms: species and speciation. 1998.

DE QUEIROZ, K. The General Lineage Concept of Species and the Defining Properties of the Species Category. In book: Species: New Interdisciplinary Essays, Chapter: 3, Publisher: MIT Press, Editors: Robert A. Wilson. 1999.

DE QUEIROZ, K. Ernst Mayr and the modern concept of species. Proceedings of the National Academy of Sciences, 102, 6600–6607. 2005a.

DE QUEIROZ, K. Different species problems and their resolution. BioEssays 27, 1263–1269. 2005b.

DE QUEIROZ, K. Species concepts and species delimitation. Syst. Biol., v. 56, p. 879–886, 2007.

DE QUEIROZ, K. Branches in the lines of descent: Charles Darwin and the evolution of the species concept. Biol. J. Linn. Soc. 103, 19–35. 2011.

DE QUEIROZ, K. An updated concept of subspecies resolves a dispute about the taxonomy of incompletely separated lineages. Herpetological Review. 2020.

RAMBAUT A., DRUMMOND A. J., SUCHARD M. Tracer v1. 6 http://beast.bio.ed.ac.uk, 2007.

RAMBAUT A., DRUMMOND A. J. LogCombiner v2. 1.3. Institute of Evolutionary Biology,University of Edinburgh, 2014.

RAMBAUT, A. FigTree v.1.4.2: tree drawing tool. Available at: http://tree.bio.ed.ac.uk/software/figtree/. 2014.

RAMBAUT, A., DRUMMOND, A. J., XI, D., BAELE, G., SUCHARD, M. A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Systematic Biology, 67: 901–904. 2018.

RANCILHAC, L., et al. Phylotranscriptomic evidence for pervasive ancient hybridization among Old World salamanders. Molecular Phylogenetics and Evolution, 155. https://doi.org/10.1016/j.ympev.2020.106967. 2021.

RANNALA, B. The art and science of species delimitation. Current Zoology, v. 61, n. 5, p. 846-853, 2015.

RANNALA, B., & YANG, Z. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164, 1645–1656. 2003.

RANNALA, B., & YANG, Z. Bayesian species delimitation using multilocus sequence data. Proceedings of the National Academy of Sciences 107, 9264–9269. 2010.

RANNALA, B., & YANG, Z. Species Delimitation. In: Phylogenetics in the genomic era. 2020.

RANNALA, B; YANG, Z. Species delimitation. In: Scornavacca, C and Delsuc, F and Galtier, N, (eds.) Phylogenetics in the Genomic Era. (5.5:1-5.5:18). Self published, 2020.

RANNALA, B., et al. The Multi-species Coalescent Model and Species Tree Inference. SCORNAVACCA, CELINE; DELSUC, FRÉDÉRIC; GALTIER, NICOLAS. Phylogenetics in the Genomic Era, No commercial publisher | Authors open access book. 2020.

RAPOSO DO AMARAL, F., et al. Recent chapters of Neotropical history overlooked in phylogeography: Shallow divergence explains phenotype and genotype uncoupling in Antilophia manakins. Molecular Ecology, 27: 4108–4120. 2018.

RASMUSSEN M. D., KELLIS M. Accurate gene-tree reconstruction by learning gene- and species specific substitution rates across multiple complete genomes. Genome research 17: 1932-1942, 2007.

RAYNAL, L., et al. ABC random forests for Bayesian parameter inference. Bioinformatics 35, 1720–1728. 2019.

RIPPLINGER, J.; SULLIVAN, J. Assessment of substitution model adequacy using frequentist and Bayesian methods. Molecular Biology and Evolution, v. 27, n. 12, p. 2790-2803, 2010.

RITCHIE A. M., et al. The impact of the tree prior on molecular dating of data sets containing a mixture of inter-and intraspecies sampling. Systematic Biology 66: 413-425, 2017.

ROE A. D., SPERLING F. A. Patterns of evolution of mitochondrial cytochrome c oxidase I and II DNA and implications for DNA barcoding. Molecular Phylogenetics and evolution 44: 325-345, 2007.

ROGERS T. F., et al. Using recent genetic history to inform conservation options of two Lesser Caymans iguana (Cyclura otali caymanensis) populations. Conservation Genetics 25: 711-724, 2024.

ROMÁN-PALACIOS C., et al. When did anoles diverge? An analysis of multiple dating strategies. Molecular Phylogenetics and Evolution 127: 655-668, 2018.

ROUX, C., et al. Shedding light on the grey zone of speciation along a continuum of genomic divergence. PloS Biology, 14. 2016.

ROZANTSEV, A., SALZMANN, M. & FUA, P. Beyond sharing weights for deep domain adaptation. IEEE transactions on pattern analysis and machine intelligence 41, 801–814. 2018.

RUBINOFF D., HOLLAND B. Between Two Extremes: Mitochondrial DNA is neither the Panacea nor the Nemesis of Phylogenetic and Taxonomic Inference. Systematic Biology 54: 952–961, 2005. Doi:10.1080/10635150500234674

RULL, V.; CARNAVAL, A. C. (Ed.). Neotropical diversification: patterns and processes. Berlin: Springer, 2020.

SACCONE C., et al. Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. Gene 238: 195-209, 1999.

SALLES, M. M. A., et al. Ancient introgression explains mitochondrial genome capture and mitonuclear discordance among South American collared *Tropidurus* lizards. bioRxiv, 2025-04, 2025. https://doi.org/10.1101/2025.04.25.650633.

SALLES, M. M. A. S.; DOMINGOS, F. M. C. B. Towards the next generation of species delimitation methods: An overview of machine learning applications. Molecular Phylogenetics and Evolution, v. 108368, 2025.

SANCHEZ, T., et al. Deep learning for population size history inference: Design, comparison and combination with approximate Bayesian computation. Molecular Ecology Resources 21, 2645–2660. 2020.

SARYAN, P., et al. Species complex delimitations in the genus Hedychium: A machine learning approach for cluster discovery. Applications in Plant Sciences 8. https://doi.org/10.1002/aps3.11377. 2020.

SCALON, M.C., et al. Diversity of functional trade-offs enhances survival after fire in Neotropical savanna species. Journal of Vegetation Science, 31, 139-150. 2020.

SCARPETTA S. G. The first known fossil Uma: ecological evolution and the origins of North American fringe-toed lizards. BMC Evolutionary Biology 19: 178, 2019.

SCHENK J. J. Consequences of secondary calibrations on divergence time estimates. PloS one 11: e0148228, 2016.

SCHRIDER, D. R. & KERN, A. D. Supervised Machine Learning for Population Genetics: A New Paradigm. Trends in Genetics 34, 301–312. https://doi.org/10.1016/j.tig.2017.12.005. 2018.

SEARLS, D. B. The Roots of Bioinformatics. PloS Comput Biol 6. https://doi.org/10.1371/journal.pcbi.1000809. 2010.

SEIXAS, F. A., et al. The genomic impact of historical hybridization with massive mitochondrial DNA introgression. Genome Biology, 19. https://doi.org/10.1186/s13059-018-1471-8. 2018.

SHEEHAN, S., & SONG, Y.S. Deep learning for population genetic inference. PloS computational biology 12. 2016.

SHEN, C. C., et al. Exploring Mitonuclear Discordance: Ghost Introgression From an Ancient Extinction Lineage in the Odorrana swinhoana Complex. Molecular Ecology, e17763. 2025.

SHERRATT E., et al. Amber fossils demonstrate deep-time stability of Caribbean lizard communities. Proceedings of the National Academy of Sciences 112: 9961-9966, 2015.

SHURTLIFF, Q. R. Mammalian hybrid zones: a review. Mammal Review, 43, 1–21. 2013.

SIDEY-GIBBONS, J. A., & SIDEY-GIBBONS, C. J. Machine learning in medicine: a practical introduction. BMC medical research methodology 19, 1–18. 2019.

SILVA, D. C., et al. Cerrado bat community assembly is determined by both present-day and historical factors. Journal of Biogeography. 2024.

SIMÕES T. R., PYRON R. A. The squamate tree of life. Bulletin of the Museum of Comparative Zoology 163: 47-95, 2021.

SIMONSEN, K. L., et al. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 1411, 413–429. 1995.

SITES, JR J. W. & MARSHALL, J. C. Operational criteria for delimiting species. Annual Review of Ecology, Evolution, and Systematics, 199-227. 2004.

SLATKO, B. E., et al. Overview of next-generation sequencing technologies. Current protocols in molecular biology 122. 2018.

SLOAN D. B., OXELMAN B., RAUTENBERG A., TAYLOR D. R. Phylogenetic analysis of mitochondrial substitution rate variation in the angiosperm tribe Sileneae. BMC Evolutionary Biology 9, 2009. https://doi.org/10.1186/1471-2148-9-260

SMITH, B. T., et al. Target Capture and Massively Parallel Sequencing of Ultraconserved Elements (UCEs) for Comparative Studies at Shallow Evolutionary Time Scales. Systematic Biology, 64: 83–95. Doi:10.1093/sysbio/syt061. 2014.

SMITH, M. D., et al. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Molecular Biology and Evolution, 32: 1342–1353. 2015.

SMITH, M. L., & CARSTENS, B. C. Process-based species delimitation leads to identification of more biologically relevant species. Evolution, 74, 216–229. https://doi.org/10.1111/evo.13878. 2020.

SMITH, M. L.; CARSTENS, B. C. Process-based species delimitation leads to identification of more biologically relevant species. Evolution, v. 74, n. 2, p. 216-229, 2020.

SMITH, M. L., et al. Demographic Model Selection using Random Forests and the Site Frequency Spectrum. Molecular Ecology. 2017.

SMITH, M. L., & HAHN, M. W. Phylogenetic inference using generative adversarial networks. Bioinformatics, 39. 2023.

SMITH, S. A., BROWN, J. W., WALKER, J. F. So many genes, so little time: a practical approach to divergence-time estimation in the genomic era. PloS one 13: e0197433, 2018.

SMITH, S. A., et al. So many genes, so little time: a practical approach to divergence-time estimation in the genomic era. PloS One, 13: e0197433. 2018.

SOLIS-LEMUS, C., et al. Accurate phylogenetic inference with a symmetry-preserving neural network model. arXiv preprint arXiv:2201.04663. 2022.

SOLÍS-LEMUS, C., ANÉ, C. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PloS Genetics, 12: e1005896. 2016.

SOLÍS-LEMUS, C., YANG, M., ANÉ, C. Inconsistency of species tree methods under gene flow. Systematic Biology, 65: 843–851. 2016.

SOLÍS-LEMUS, C., BASTIDE, P., ANÉ, C. PhyloNetworks: a package for phylogenetic networks. Molecular Biology and Evolution, 34: 3292–3298. 2017.

STANKOWSKI, S., RAVINET, M. Defining the speciation continuum. Evolution, 75: 1256–1273. 2021.

SU Z., et al. The impact of incorporating molecular evolutionary model into predictions of phylogenetic signal and noise. Frontiers in Ecology and Evolution 2: 11, 2014.

SUKUMARAN, J. & KNOWLES, L.L. Multispecies coalescent delimits structure, not species. Proceedings of the National Academy of Sciences 114, 1607–1612. 2017.

SUKUMARAN, J.; HOLDER, M. T.; KNOWLES, L. L. Incorporating the speciation process into species delimitation. PloS computational biology, v. 17, n. 5 (e1008924), 2021.

SUKUMARAN, J., et al. Incorporating the speciation process into species delimitation. PloS Computational Biology 17. 2021.

SULLIVAN, J.; JOYCE, P. Model selection in phylogenetics. Annual Review of Ecology, Evolution, and Systematics, v. 36, n. 1, p. 445-466, 2005.

SUVOROV, A., et al. Accurate inference of tree topologies from multiple sequence alignments using deep learning. Systematic biology 69, 221–233. 2020.

TAGU, D., et al. Genomic data integration for ecological and evolutionary traits in non-model organisms. BMC genomics 15, 1–16. 2014.

TAMURA, K., QIQING, T., KUMAR, S. Theoretical Foundation of the RelTime Method for Estimating Divergence Times from Variable Evolutionary Rates. Molecular Biology and Evolution, 35: 1770–1782. 2018.

TAMURA, K., STECHER, G., KUMAR, S. MEGA 11: Molecular Evolutionary Genetics Analysis Version 11. Molecular Biology and Evolution. https://doi.org/10.1093/molbev/msab120. 2021.

TANG, B.; PAN, Z.; YIN, K.; KHATEEB, A. Recent advances of deep learning in bioinformatics and computational biology. Frontiers in genetics, v. 10, n. 214, 2019.

TAO, Q., TAMURA, K., MELLO, B., KUMAR, S. Reliable Confidence Intervals for RelTime Estimates of Evolutionary Divergence Times. Molecular Biology and Evolution, 37: 280–290. 2020.

TAUTZ, D., et al. A plea for DNA taxonomy. Trends Ecol. Evol. 18, 70–74. 2003.

TILEY, G. P. et al. Estimation of species divergence times in presence of cross-species gene flow. Systematic Biology, v. 72, n. 4, p. 820-836, 2023.

TOEWS, D. P., BRELSFORD, A. The biogeography of mitochondrial and nuclear discordance in animals. Molecular Ecology, 21: 3907–3930. 2012.

TOWNSEND T. M., et al. Eastward from Africa: palaeocurrent-mediated chameleon dispersal to the Seychelles islands. Biology Letters 7: 225-228, 2011.

TURCHETTO-ZOLET, A. C.; PINHEIRO, F.; SALGUEIRO, F.; PALMA-SILVA, C. Phylogeographical patterns shed light on evolutionary process in South America. Mol Ecol, v. 22, p. 1193–1213, 2013.

UETZ P., et al. A quarter century of reptile and amphibian databases. Herpetol. Rev. 52: 246-255, 2021.

UETZ P., FREED P., AGUILAR R., REYES F., KUDERA J., HOŠEK J. (eds.). The Reptile Database, http://www.reptile-database.org, accessed January 31, 2025.

VALLETTA, J. J., et al. Applications of machine learning in animal otaling studies. Animal Behaviour 124, 203–220. 2017.

VINK, C. J., et al. Taxonomy and irreproducible biological science. BioScience 62, 451–452. 2012.

VOGLER, A. P., MONAGHAN, M. T. Recent advances in DNA taxonomy. J. Zool. Syst. Evol. Res. 45, 1–10. 2007.

WAKE, D. B., et al. Homoplasy: from detecting patterns to determining process and mechanism of evolution. Science 331, 1032–1035. 2011.

WÄLDCHEN, J. & MÄDER, P. Machine learning for image-based species identification. Methods in Ecology and Evolution 9, 2216–2225. 2018.

WANG, G. Machine learning for inferring animal behavior from location and movement data. Ecological informatics 49, 69–76. 2019.

WANG, Z., et al. Automatic inference of demographic parameters using generative adversarial networks. Molecular ecology resources 21, 2689–2705. 2021.

WARNOCK R. C., et al. Calibration uncertainty in molecular dating analyses: there is no substitute for the prior evaluation of time priors. Proceedings of the Royal Society B: Biological Sciences 282: 20141013, 2015.

WELCH J. J., BININDA-EMONDS O. R., BROMHAM L. Correlates of substitution rate variation in mammalian protein-coding sequences. BMC Evolutionary Biology 8: 1-12, 2008.

WEN, D., YU, Y., ZHU, J., NAKHLEH, L. Inferring phylogenetic networks using PhyloNet. Systematic Biology, 67: 735–740. 2018.

WERNECK F. P., et al. Biogeographic history and cryptic diversity of saxicolous Tropiduridae lizards endemic to the semiarid Caatinga. BMC Evolutionary Biology 15: 1–24, 2015. https://doi.org/10.1186/s12862-015-0368-3

WIENS, J. J., & PENKROT, T. A. Delimiting species using DNA and morphological variation and discordant species limits in spiny lizards (Sceloporus). Syst. Biol., 51, 69–91. 2002.

WIENS, J. J. Species delimitation: new approaches for discovering diversity. Syst. Biol. 56, 875–8. 2007.

WILKINS, J. S., ZACHOS, F. E., & PAVLINOV, I. Y. (Eds.). Species Problems and Beyond: Contemporary Issues in Philosophy and Practice. CRC Press. 2022.

WILLIAMS E. J. B., HURST L. D. Is the synonymous substitution rate in mammals gene-specific? Molecular Biology and Evolution 19: 1395-1398, 2002.

XIA, X. Phylogenetic relationship among horseshoe crab species: effect of substitution models on phylogenetic analyses. Systematic Biology, v. 49, n. 1, p. 87-100, 2000.

XU W., et al. The relationship between the rate of molecular evolution and the rate of genome rearrangement in animal mitochondrial genomes. Journal of molecular evolution 63: 375-392, 2006.

YAN H., HU Z., et al. PhyloAcc-GT: A Bayesian method for inferring patterns of substitution rate shifts on targeted lineages accounting for gene tree discordance. Molecular Biology and Evolution 40: msad195, 2023.

YAN L., XU W., ZHANG D., LI J. Comparative analysis of the mitochondrial genomes of flesh flies and their evolutionary implication. International Journal of Biological Macromolecules 174: 385–391, 2021. https://doi.org/10.1016/j.ijbiomac.2021.01.188

YANG, B., et al. Identification of species by combining molecular and morphological data using convolutional neural networks. Systematic Biology, 71, 690–705. 2022.

YANG Z. Among-site rate variation and its impact on phylogenetic analyses. Trends in Ecology & Evolution, v. 11, n. 9, p. 367-372, 1996.

YANG Z., NIELSEN, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Molecular Biology and Evolution 17: 32-43, 2000.

YANG, Z. PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution 24, 1586–1591, 2007.

YANG, Z.; RANNALA, B. Molecular phylogenetics: principles and practice. Nature Reviews Genetics, v. 13, n. 5, p. 303-314, 2012.

YANG H., LI T., DANG K., BU W. Compositional and mutational rate heterogeneity in mitochondrial genomes and its effect on the phylogenetic inferences of Cimicomorpha (Hemiptera: Heteroptera). BMC Genomics 19, 2018. https://doi.org/10.1186/s12864-018-4650-9

YELMEN, B. & JAY, F. An Overview of Deep Generative Models in Functional and Evolutionary Genomics. Annual Reviews of Biomedical Data Science. https://doi.org/10.1146/annurev-biodatasci-020722. 2023.

ZACHOS, F. E. Species concepts in biology (Vol. 801). Cham: Springer. 2016.

ZACHOS, F. E. (New) Species concepts, species delimitation and the inherent limitations of taxonomy. Journal of genetics, 97, 811–815. 2018.

ZAHARIAS, P., et al. Re-evaluating Deep Neural Networks for Phylogeny Estimation: The Issue of Taxon Sampling. Journal of Computational Biology 29, 74–89. https://doi.org/10.1089/cmb.2021.0383. 2022.

ZARDOYA R., MEYER A. Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. Molecular biology and evolution 13: 933-942, 1996.

ZARZA, E., et al. Hidden histories of gene flow in highland birds revealed with genomic markers. Molecular Ecology, 25: 5144–5157. 2016.

ZARZA E., REYNOSO V. H., EMERSON B. C. Diversification in the northern neotropics: mitochondrial and nuclear DNA phylogeography of the iguana Ctenosaura otaling and related species. Molecular Ecology 17: 3259-3275, 2008.

ZHANG, C., et al. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics, 19: 15–30. 2018.

ZHENG Y., WIENS J. J. Do missing data influence the accuracy of divergence-time estimation with BEAST? Molecular Phylogenetics and Evolution 85: 41-49, 2015.

ZHENG, Y., WIENS, J. J. Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. Molecular Phylogenetics and Evolution, 94: 537–547. 2016.

ZHU A., GUO W., JAIN K., MOWER J. P. Unprecedented heterogeneity in the synonymous substitution rate within a plant genome. Molecular Biology and Evolution 31: 1228–1236, 2014. https://doi.org/10.1093/molbev/msu079

ZIELIŃSKI, P., et al. No evidence for nuclear introgression despite complete mtDNA replacement in the Carpathian newt (Lissotriton montandoni). Molecular Ecology, 22: 1884–1903. 2013.

ZOZAYA, S. M., MACOR, S. A., SCHEMBRI, R., HIGGIE, M., HOSKIN, C. J., O'HARA, K., … MORITZ, C. Contact zones reveal restricted introgression despite frequent hybridization across a recent lizard radiation. Evolution, 79: 411–422. 2025.