MOLECULAR ECOLOGY

Molecular Ecology (2015) **24**, 5957–5975 doi: 10.1111/mec.13433

Speciation with gene flow in whiptail lizards from a Neotropical xeric biome

ELIANA F. OLIVEIRA,* MARCELO GEHARA,† VINÍCIUS A. SÃO-PEDRO,* XIN CHEN,‡§ EDWARD A. MYERS,‡§ FRANK T. BURBRINK,‡§¶ DANIEL O. MESQUITA,** ADRIAN A. GARDA,†† GUARINO R. COLLI,‡‡ MIGUEL T. RODRIGUES,§§ FEDERICO J. ARIAS,§§ HUSSAM ZAHER,¶¶ RODRIGO M. L. SANTOS§§ and GABRIEL C. COSTA***

**Pós-Graduação em Ecologia, Universidade Federal do Rio Grande do Norte, Natal, RN 59072-970, Brazil, †Pós-Graduação em Sistemática e Evolução, Universidade Federal do Rio Grande do Norte, Natal, RN 59072-970, Brazil, ‡Department of Biology, 6S-143, College of Staten Island, The City University of New York, 2800 Victory Boulevard, Staten Island, NY 10314, USA, §Department of Biology, The Graduate School, City University of New York, New York, NY 10016, USA, ¶Department of Herpetology, American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024-5192, USA, **Departamento de Sistemática e Ecologia, Universidade Federal da Paraíba, João Pessoa, PB 58000-00, Brazil, ††Departamento de Botânica e Zoologia, Centro de Biociências, Universidade Federal do Rio Grande do Norte, Natal, RN 59072-970, Brazil, ‡‡Departamento de Zoologia, Universidade de Brasília, Brasília, DF 70910-900, Brazil, §§Departamento de Zoologia, Universidade de São Paulo, São Paulo, SP 05422-970, Brazil, ¶¶Museu de Zoologia, Universidade de São Paulo, São Paulo, SP 04263-000, Brazil, ***Departamento de Ecologia, Universidade Federal do Rio Grande do Norte, Natal, RN 59072-970, Brazil*

Abstract

Two main hypotheses have been proposed to explain the diversification of the Caatinga biota. The *riverine barrier hypothesis* (RBH) claims that the São Francisco River (SFR) is a major biogeographic barrier to gene flow. The *Pleistocene climatic fluctuation hypothesis* (PCH) states that gene flow, geographic genetic structure and demographic signatures on endemic Caatinga taxa were influenced by Quaternary climate fluctuation cycles. Herein, we analyse genetic diversity and structure, phylogeographic history, and diversification of a widespread Caatinga lizard (*Cnemidophorus ocellifer*) based on large geographical sampling for multiple loci to test the predictions derived from the RBH and PCH. We inferred two well-delimited lineages (Northeast and Southwest) that have diverged along the Cerrado–Caatinga border during the Mid–Late Miocene (6–14 Ma) despite the presence of gene flow. We reject both major hypotheses proposed to explain diversification in the Caatinga. Surprisingly, our results revealed a striking complex diversification pattern where the Northeast lineage originated as a founder effect from a few individuals located along the edge of the Southwest lineage that eventually expanded throughout the Caatinga. The Southwest lineage is more diverse, older and associated with the Cerrado–Caatinga boundaries. Finally, we suggest that *C. ocellifer* from the Caatinga is composed of two distinct species. Our data support speciation in the presence of gene flow and highlight the role of environmental gradients in the diversification process.

Keywords: approximate Bayesian computation approach, Caatinga biome, *Cnemidophorus*, coalescent methods, Miocene diversification, phylogeography

Received 22 July 2015; revision received 15 October 2015; accepted 21 October 2015

Introduction

Knowing how many species exist and what processes generated current biodiversity are among the most

Correspondence: Eliana F. Oliveira,
E-mail: elianabio@gmail.com

fundamental questions in biology. Although 1.5 million eukaryotic species have been described, recent estimates suggest that over 85% of Earth's biota is unknown (Mora *et al.* 2011; Costello *et al.* 2013), and processes responsible for generating and maintaining such diversity are only beginning to be understood (Beheregaray 2008; Hickerson *et al.* 2010). Such paucity of information is even worse for the Neotropical region, where mechanisms generating the highest global biodiversity are poorly known. Diversification hypotheses based on geomorphological features and climatic fluctuations have been tested in the Neotropics (Turchetto-Zolet *et al.* 2013; Smith *et al.* 2014), but the relative roles of such processes are still debated (Rull 2008, 2011, 2013). Furthermore, data are scant or nonexistent for many taxonomic groups and entire biomes (Beheregaray 2008; Turchetto-Zolet *et al.* 2013). In particular, open vegetation biomes of South America have been far less studied than tropical rainforests (Werneck 2011; Turchetto-Zolet *et al.* 2013).

South America east of the Andes is crossed by a southwest-to-northeast trending diagonal of open vegetation biomes that separate the Amazon and Atlantic rainforests (see Werneck 2011). This 'dry diagonal' encompasses three biomes: Chaco, Cerrado and Caatinga in northeastern Brazil. The Caatinga represents the largest and most isolated Seasonally Dry Tropical Forest (hereafter SDTF), which is formed by numerous disjunct patches in the Neotropical region (Werneck 2011). The Caatinga harbours high levels of diversity and contains many endemic species for several groups (e.g. Rodrigues 2003; Zanella & Martins 2003; Queiroz 2006), which attest its importance as a Neotropical centre of endemism. The SDTFs are considered a relatively old biome (Middle Eocene; Pennington *et al.* 2009); however, estimates for at least Caatinga plant groups suggest that most diversification took place in the Mid–Late Miocene and Pliocene (Pennington *et al.* 2004, 2009), although comparisons among other groups are lacking.

It is premature to generalize processes of diversification and assembly that account for the high biodiversity in the Caatinga. However, from the few previous studies so far, there are two main hypotheses that seem to capture the biogeographic and evolutionary history of Caatinga and explain the origins of species diversity in this biome. The first hypothesis is analogous to the *riverine barrier hypothesis* (RBH) proposed for the Amazon basin (Ayres & Clutton-Brock 1992; Patton *et al.* 1994). This hypothesis claims that the São Francisco River (SFR) is a major biogeographic barrier to gene flow in the Caatinga (Rodrigues 1996, 2003). The geographic distribution of some populations and/or sister species and their phylogenetic and phylogeographic relationships indicate moderate structuring on opposing sides of the SFR in tropidurid lizards (Passoni *et al.* 2008; Werneck *et al.* 2015), eyelid-less lizards (Siedchlag *et al.* 2010) and rodents (Nascimento *et al.* 2011, 2013; Faria *et al.* 2013). However, some studies found no evidence for the SFR as an effective barrier. In fact, some widespread lineages of plants (Caetano *et al.* 2008), frogs (São-Pedro 2014) and gymnophthalmid lizards (Recoder *et al.* 2014) showed little genetic structure, and gene flow among populations seems unrestricted across the Caatinga.

The second hypothesis is also analogous to a general mechanism proposed for different ecosystems (e.g. Haffer 1969; Hewitt 2000; Anthony *et al.* 2007). The *Pleistocene climatic fluctuation hypothesis* (PCH) claims that gene flow, geographic genetic structure and demographic signatures on endemic Caatinga taxa were influenced by Quaternary climate fluctuation cycles (e.g. Machado *et al.* 2014; Magalhães *et al.* 2014; Werneck *et al.* 2015). Previous work suggests that the distribution of the Caatinga vegetation has changed during Quaternary climatic changes likely affecting the distribution and genetic diversity of local biota (Werneck *et al.* 2011). Wetter conditions have been inferred for distinct regions and periods during the last 210 000 years (Auler *et al.* 2004; Wang *et al.* 2004). Consequently, mesic climatic regimes favoured the expansion of humid vegetation and partial replacement of the vegetation represented in the present-day Caatinga, resulting in its fragmentation (De Oliveira *et al.* 1999; Behling *et al.* 2000). Current natural rainforest enclaves (so-called *brejos de altitude*) are likely remnants of these ancient forests (Sampaio 1995), although the precise timing of these expansions and their boundaries is still elusive (Werneck 2011). In addition, disjunct distributions of squamate species in isolated sandy soil patches suggest a past climate similar or even drier than current conditions, when these sandy areas were more extensive and continuous (Rodrigues 1996, 2003). Investigating the origin and historical demography of endemic species is crucial to clarify the interplay between Quaternary climatic fluctuations and geomorphological landscape features such as the SFR on the history of the Caatinga biota.

The whiptail *Cnemidophorus ocellifer* (Spix 1825) is one of the most common lizards in the Caatinga, and several studies have suggested that it comprises multiple cryptic species (e.g. Rocha *et al.* 1997; Rodrigues 2003), with new species of the *ocellifer* group described for the Caatinga recently (see Arias *et al.* 2011a,b; Silva & Ávila-Pires 2013). *Cnemidophorus ocellifer* is a heliophilic and active forager, highly abundant and being found mostly in open habitat types in the Caatinga. Herein, we analyse genetic diversity and structure,

phylogeographic history, and diversification of *C. ocellifer* based on range-wide sampling for multiple loci to test the RBH and PCH. If the SFR is a major vicariant barrier driving diversification in the Caatinga biome, we expect to see little gene flow across river margins and a clear pattern of lineage breaks and genetic structure coinciding with current or past river courses. We also expect to see a strong effect of the SFR in the historical biogeographic analyses of the lineages, where phylogeographic reconstructions would show the origin and expansion of lineages coinciding with the geographical position of the river as a major barrier. Alternatively, if PCH is corroborated, then we should see a signal of population structuring and demographic changes that mirror fluctuations in the habitats during Pleistocene cycles in the Caatinga. Based on our analyses, we reject both hypotheses. Our results revealed a strikingly complex diversification pattern where species diverged in the presence of gene flow along an environmental gradient.

Materials and methods

Sample collection and sequencing

We obtained 398 tissue samples of the *Cnemidophorus ocellifer* species complex (see 'Taxonomic background and implications' in Appendix S1, Supporting information) from 79 localities in the Caatinga biome and adjacent areas (Fig. 1 and Table S1, Supporting information). Samples were obtained through fieldwork led by the authors and through loans from different herpetological collections. *Cnemidophorus venetacaudus*, a member of *C. ocellifer* group (or *Cnemidophorus littoralis* subgroup; see 'Taxonomic background and implications' in Appendix S1, Supporting information), was used as outgroup when necessary.

We extracted DNA from liver or muscle tissue using Qiagen DNeasy kits. Five loci were amplified via polymerase chain reaction (PCR) using GoTaq Green MasterMix (Promega Corporation). Details about loci, primers and PCR protocols are listed in Table S2 (Supporting information). We cleaned PCR products with 2 μL of ExoSap (USB Corporation) and sequenced the products using 1 μL of each primer, 2 μL of DTCS (Beckman-Coulter) and 4 μL of ultrapure water. First, we sequenced all individuals for one mitochondrial DNA (mtDNA) gene (12S ribosomal RNA; 12S). For nuclear (nuDNA) genes, we sequenced a subset of individuals (137 samples), which were chosen to represent a wide geographic range within the Caatinga (i.e. the same 79 collecting localities). We sequenced four nuDNA genes: ATP synthase beta subunit (ATPSB), natural killer-tumour recognition sequence (NKTR), G protein-coupled receptor 149 (R35) and ribosomal protein 40 (RP40). All sequences were generated using Sanger sequencing at the College of Staten Island (City University of New York), aligned with the Clustal algorithm (Sievers *et al.* 2011) and checked by eye using GENEIOUS 6.1 (Biomatters). We found gaps in 12S and ATPSB genes and removed them using GBLOCKS 0.91b (Castresana 2000; Talavera & Castresana 2007), available as a web server (http://molevol.cmima.csic.es/castresana/Gblocks_server.html). This program reduces the

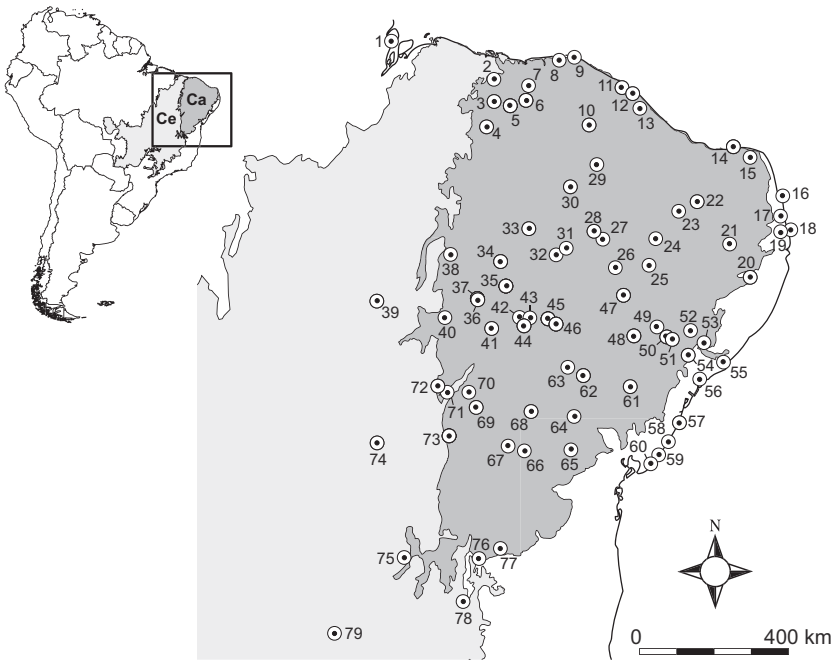

Fig. 1 Distribution of sampled localities for *Cnemidophorus ocellifer* species complex. Numbers correspond to 79 localities' names in Tables S1 and S6 (Supporting information). Grey shades represent biomes limits: Caatinga (Ca) in dark grey and Cerrado (Ce) in light grey.

need for manually editing multiple alignments, facilitating the reproduction of the alignments and subsequent phylogenetic analysis by other researchers. We used the default options and parameter values in GBLOCKS, which were: number of sequences for a conserved position [12S: 200 (for 398 samples) and 69 (for 137 samples); ATPSB: 64], minimum number of sequences for a flanking position [12S: 338 (for 398 samples) and 116 (for 137 samples); ATPSB: 107], maximum number of contiguous nonconserved positions (eight for both genes), minimum length of a block (10 for both genes) and allowed gap positions (none for both genes).

To determine the most probable pair of alleles for each nuDNA gene, we used the PHASE algorithm (Stephens *et al.* 2001) implemented in the DNASP 5.10 software (Librado & Rozas 2009) using default options. Only samples with probability of pairs of alleles in heterozygosity higher than 80% were considered in the following analyses. All sequences obtained in this study are available at GenBank (KT844957–KT845861 and KT886989–KT886992; see also Table S1, Supporting information).

Testing the RBH and PCH

If the SFR is a main vicariant barrier that explains lineages' origin and biogeographic history of Caatinga biota, we expect to see no gene flow, relatively similar population sizes and a clear pattern of main haplotypes shared within localities in each margin of the SFR. We expect larger genetic distances between the opposite margins of the SFR. We also expect a clear spatial structure pattern in the genetic lineages recovered by each gene, corresponding to the SFR geographical position. Therefore, the geographic origin of lineages should match the location of the river. We also expect to see a pattern of spatial expansion that reflects the position of the barrier. When considering the SFR as a barrier, it is important to note its geomorphological history. The SFR headwaters start in Minas Gerais state, and the river runs northwardly until its course turns abruptly east towards the Atlantic Ocean (Fig. 2). Geomorphological evidence indicates that the SFR's paleo-course differed from its current configuration by continuing north reaching the Atlantic Ocean (Fig. 2). This paleo-course could have persisted until the Middle–Late Miocene (Potter 1997). Therefore, all predictions described above could reflect isolation promoted by either current or paleo-course of the SFR.

If the PCH is corroborated, we expect that the time of divergence of lineages and origin of species in the Caatinga be within the Pleistocene. We expect a spatial structure pattern that reflects major habitat shifts in the Caatinga (e.g. Werneck *et al.* 2011). We also anticipate finding a recent geographical expansion following major vegetation shifts during the Pleistocene. Finally, we expect to see historical demographic responses of population expansion or contractions occurring during the Pleistocene.

To test the predictions described above and to determine the degree of genetic structure and the geographic position of genetic breaks, we first conducted two population assignment tests. The subsequent analysis followed population assignment results as described below.

Population assignment

We used a genotype matrix of the four nuDNA genes to investigate population structure with STRUCTURE 2.3.4

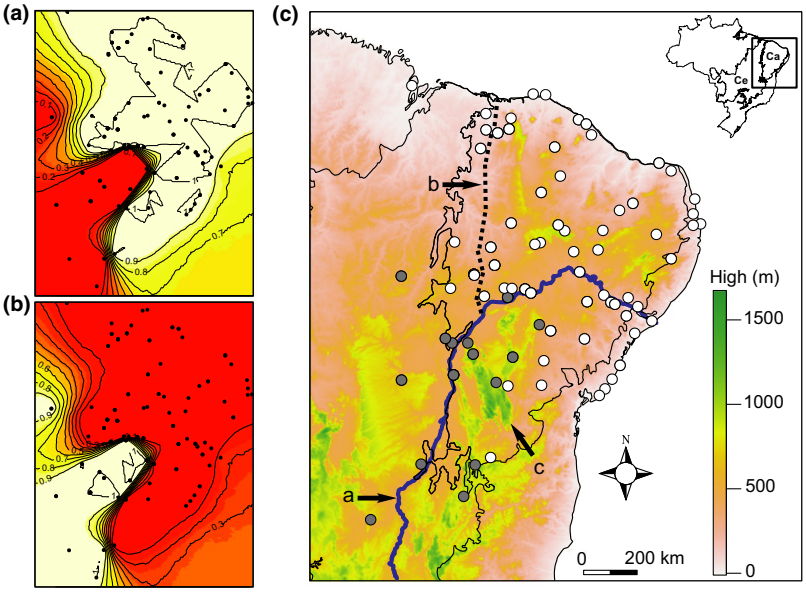

Fig. 2 GENELAND analysis (a, b) with posterior probability isoclines, which indicate extensions of the genetic populations found (black lines with inclusion probabilities). Light colour zones in each map indicate the groups of localities with greater probabilities of belonging to the same genetic unit. Black dots indicate locations of the 79 analysed localities. Map on the right (c) shows the distribution of Northeast (white circles) and Southwest (grey circles) lineages along Caatinga (Ca) and Cerrado (Ce) biomes, depicting São Francisco River (a), its hypothetical paleo-course (b) and Espinhaço Mountain Range (c). Green colour represents higher altitudes and white represents lower altitudes.

(Pritchard *et al.* 2000). To obtain this matrix, we converted aligned sequence data from different genomic regions (i.e. four nuDNA) into the Structure input file format using the program XMFAS2STRUCT (available from http://www.xavierdidelot.xtreemhost.com/clonalframe.htm). The genotype matrix is designed for use with linkage model and provides better resolution to study the historical process of admixture (see Falush *et al.* 2003). We explored a large range of values by running 10 replicate analyses over a range of number of populations (k) from 1 to 10. Each independent run implemented 5×10^4 generations following a burn-in of 5×10^4 generations, assuming a linkage model and uncorrelated allele frequencies. We chose the best value of k based on the rate of changes in the log-probability of data between successive k values, Δk (Evanno *et al.* 2005), using STRUCTURE HARVESTER (Earl & vonHoldt 2012). Second, we ran GENELAND 4.0.3 (Guillot *et al.* 2005a,b) implemented in R 3.1.1 (R Core Team 2015), which uses the clustering algorithm of STRUCTURE under a spatial model. This analysis evaluates the presence of population structure in a group of georeferenced genetic data by inferring and explicitly identifying genetic discontinuities along the landscape. Further, GENELAND can handle haploid and diploid data in the same model, allowing for combination of mtDNA and nuDNA. Thus, to extensively use all data available, we prepared a file containing spatial coordinates of the individuals sampled and combined it with two different haplotype data sets, one with only nuDNA data and another with both mtDNA and nuDNA data. The most probable number of population units (k) was determined by a Markov chain Monte Carlo (MCMC) method, with 10 repetitions (5×10^6 iterations in each) of k from 1 to 10. In addition, for each population identified by STRUCTURE and GENELAND (both programs generated identical results), we calculated the number of polymorphic sites (S), haplotype number (h), haplotype diversity (H_d), nucleotide diversity (π), Tajima's D and its P value for each locus using DNASP. We investigated genetic structure between and within populations and loci with analyses of molecular variance (AMOVAS) in ARLEQUIN 3.5 (Excoffier & Lischer 2010) using 10 000 permutations. We also estimated uncorrected pairwise genetic distances between and within identified populations for all genes in MEGA 6.06 (Tamura *et al.* 2013) using default options.

Gene tree estimation and haplotype genealogy

We estimated gene trees independently for mtDNA and nuDNA (unphased) genes using Bayesian inference in BEAST 1.8 (Drummond *et al.* 2012). Because the outgroup *C. venetacaudus* failed to amplify for NKTR, we used the closest sequence to *C. ocellifer* available in GenBank (i.e. HQ876282) to root this gene tree. We determined the most appropriate substitution model using Bayesian information criterion in JMODELTEST (Posada 2008; see Table S3, Supporting information). We ran 2×10^7 generations sampled every 2×10^3 generations, resulting in five gene trees. We visually assessed convergence of the MCMC runs and effective sample sizes (ESS values ≥ 200) using TRACER 1.6 (Drummond & Rambaut 2007). The first two thousand generations were discarded as burn-in, and the consensus tree for each locus was inferred with TREEANNOTATOR 1.8 (Drummond *et al.* 2012). Table S3 (Supporting information) shows other details of these analyses.

We estimated haplotype networks for mtDNA and nuDNA (phased) genes using the median-joining (MJ) method (Bandelt *et al.* 1999) in NETWORK 4.6.1.2 (www.fluxus-engineering.com). However, MJ networks recovered many unresolved loops in the genealogical connections between haplotypes (Fig. S1, Supporting information). Following Sequeira *et al.* (2011), we used phylogenetic algorithms to generate haplotype networks. We then used a maximum-likelihood (ML) approach with PHYML 3.1 (Guindon *et al.* 2010), using default options and the best fit model for each locus (Table S3, Supporting information). We used ML trees to estimate each network haplotype in HAPLOVIEWER (Salzburger *et al.* 2011). For both analyses, individuals were assigned to populations following STRUCTURE and GENELAND results.

Species tree estimation

We estimated a species tree in *BEAST 1.8 (Drummond *et al.* 2012). To calibrate the species tree, we used a 12S substitution rate derived from a calibrated gene tree using a Bayesian phylogenetic method. First, we downloaded teiid sequences from GenBank (one outgroup and other 59 sequences) and added two *C. ocellifer* sequences from this study (Table S4, Supporting information). We then ran a 12S gene tree applying four node constraints based on appropriate fossil evidence (see 'Fossil Record' in Appendix S1, Supporting information) as follows: (i) origin of Teiidae at 56 Ma; (ii) divergence of *Tupinambis* from other Tupinambinae at 21 Ma; (iii) origin of 'cnemidophorine' at 20.4 Ma; and (iv) divergence of *Dracaena* from other Tupinambinae at 13.8 Ma. Based on the most recent phylogeny of Squamata (i.e. Pyron *et al.* 2013), we enforced the monophyly of each clade used in the calibration scheme. Because 'cnemidophorine' is not monophyletic (see Pyron *et al.* 2013), we used the 'cnemidophorine' fossil to place a minimum constraint at the origin of Teiinae. We then enforced time constraints using lognormal distributions,

so that fossil ages would represent the youngest limit for the respective node divergence without defining a hard limit for older divergences. Accordingly, the resulting 5–95% prior distributions were as follows: 56–71.4 Ma (Teiidae), 21–40.8 Ma (*Tupinambis*), 20.4–40.25 Ma (Teiinae) and 13.8–35.1 Ma (*Dracaena*) (Fig. S2, Supporting information). We used an uncorrelated log-normal relaxed clock with a Yule speciation-process prior and ran BEAST for 1×10^8 generations, sampled every 1×10^4 generation. From this analysis, we obtained a 12S substitution rate of 5.11×10^{-3} (equivalent to 5.11×10^{-9} substitutions/site/year), which is similar to mtDNA rates found in other squamate groups (Eo & DeWoody 2010). We used this estimated rate as a fixed parameter for the 12S substitution rate in the species tree estimation. We then ran *BEAST using mtDNA and nuDNA (phased) genes. This analysis requires a priori assignment of individual alleles to a species before estimating the relationship, and we therefore made assignments based on STRUCTURE and GENELAND results (both generated identical results). All *BEAST analyses were run for 5×10^8 generations and sampled every 5×10^4 generation (more details in Table S3, Supporting information). We assessed convergence of the MCMC runs and ESS values (≥200) using TRACER. The first 20% of sampled genealogies were discarded as burn-in, and the maximum clade credibility tree was inferred with TREEANNOTATOR.

Migration

Our previous analyses showed that *C. ocellifer* from Caatinga is composed of two major lineages (i.e. Northeast and Southwest lineages). Because the two recovered lineages are not reciprocally monophyletic for all genes and moderate degrees of haplotype sharing were identified, it was therefore essential in our analysis to assess the levels of gene flow. To this end, we used the coalescent-based program IMA2 (Hey & Nielsen 2007; Hey 2010) to estimate gene flow, ancestral and current population sizes, and divergence time between Northeast and Southwest lineages. We used all loci (mtDNA and phased nuDNA genes) and all individuals available for each gene in this analysis. We provided estimates of mutation rates (substitutions/locus/year) based on estimates from the species tree for each gene (Table S3, Supporting information). We applied the HKY model (Hasegawa *et al.* 1985) for all genes and an inheritance scalar of 0.25 and 1.0 for mtDNA and the four nuclear loci, respectively. We used a generation time of 2 years, which was estimated for other teiid lizards (*Aspidoscelis tigris*, Dessauer *et al.* 2000; and *Ameiva chrysolaema*, Gifford & Larson 2008). Upper prior limits for population parameters were defined following the IMA2 manual ($q = 22.45$, $t = 8.98$, $m = 0.45$). First, we conducted a short preliminary run to check convergence of parameters with 20 chains and the geometric heating model as suggested by the manual. We then performed an M-mode run using 'IMburn' file to inspect the trend plots to ensure stationarity and control the length of the burn-in (>1 800 000 steps). The recording phase had 1×10^7 steps sampling genealogies every 100 step. Finally, we conducted an L-mode run using 100 000 sampled genealogies to test a total of 25 nested models. We used log-likelihood ratio tests to compare these nested models and AIC to discriminate between models not rejected by the tests.

Historical demography

We built Bayesian skyline plots (Drummond *et al.* 2005) for each recovered lineage in BEAST. We used a mtDNA substitution rate of 5.11×10^{-9} substitutions/site/year to calibrate the molecular clock. We ran three independent chains of 5×10^7 generations sampling every 5×10^3 generation (more details in Table S3, Supporting information). Parameter convergence, stationarity and ESS values (≥200) were visually assessed using TRACER, and the graphs of population dynamics through time were generated in the same program.

Phylogeographic reconstruction

We used one mtDNA and four nuDNA (phased) genes to generate the phylogeographic reconstructions from homogeneous Brownian model for the Northeast lineage (Lemey *et al.* 2010). Phylogeographic reconstructions were not conducted for the Southwest lineage (more details in Discussion section), because our sampling does not encompass sufficient geographical information (Lemey *et al.* 2010). Substitution rates of nuclear loci were obtained from the species tree estimation (see values in Table S3, Supporting information). We used a jitter option of 0.05, because some sample coordinates (used as a trait) were duplicated. In BEAST, MCMC and sampled steps were set differently according to each gene (see Table S3, Supporting information) to reach convergence checked by TRACER. Table S3 (Supporting information) also shows details of substitution model used for each gene and tree prior. The first 20% of sampled genealogies were discarded as burn-in, and the maximum clade credibility tree was computed with TREEANNOTATOR. These trees were used as input for the program SPREAD 1.0.4 (Bielejec *et al.* 2011) to generate a keyhole markup language file (.kml) containing the phylogeographic history. We visualized kml files in Google Earth. The

feasible centre of origin was considered by overlapping the centre of origin for all genes.

Species validation

To test the validity of the inferred lineages by STRUCTURE and GENELAND, we employed two complementary approaches that use different methods for species validation, following recommendations in Carstens *et al.* (2013). First, we used Bayesian Phylogenetics and Phylogeography (BPP 3.1) program that is a genealogical method that uses multiple independent loci in a coalescent framework and implements a reversible jump MCMC (rjMCMC) method to calculate the posterior probability for species validation models (Yang & Rannala 2010; Yang 2015). All four nuclear loci were included in the BPP analyses. We used two different data sets: one with all 137 samples and other with a subsample of 10 individuals per lineage (20 alleles) and per locus (some sequences failed to some genes). We decided to run both data sets, because mixing problems tend to be worse and occur more frequently for larger data sets (Yang 2015). In all runs, we implemented both algorithms (0 and 1) and adjusted the fine-tuning parameters to ensure swapping rates ranged between 0.30 and 0.70 for each parameter, as recommended by BPP tutorial. We implemented different combinations of priors for ancestral population size (θ) and the root age ($\tau 0$), according to Leaché & Fujita (2010). Both priors were parameterized using a gamma $G(\alpha, \beta)$ distribution. We tested four different models: large ancestral population sizes and deep divergences, $\theta \sim G(1, 10)$ and $\tau 0 \sim G(1, 10)$; large ancestral population sizes and recent divergences, $\theta \sim G(1, 10)$ and $\tau 0 \sim G(2, 2000)$; relatively small ancestral population sizes and recent divergences, $\theta \sim G(2, 2000)$ and $\tau 0 \sim G(2, 2000)$; and small ancestral population sizes and deep divergences, $\theta \sim G(2, 2000)$ and $\tau 0 \sim G(1, 10)$. We ran three independent analyses for each set of priors using different starting seeds 1×10^6 generations, with a burn-in of 1×10^5 and thinning every two generations.

In addition to BPP, we also used SPEDESTEM 2 (Ence & Carstens 2011). This method estimates the likelihood of a species tree given a collection of independent gene trees and uses information theory to generate metrics of comparison (Carstens & Dewey 2010). Nuclear gene trees (phased) were estimated using BEAST with models of DNA sequence substitution selected using JMODELTEST. We calculated average of nucleotide diversity from nuclear genes to estimate θ value (i.e. 0.05). Because θ changes the expectation of incomplete lineage sorting of the coalescent model, we also used other value of θ (i.e. $2 \times \theta = 0.1$). We then conducted SPEDESTEM analyses using $K = 1$ and $K = 2$ levels.

Tests of diversification scenarios

Our previous analyses showed that *C. ocellifer* is composed of two major lineages with significant migration between them. Our demographic analyses showed a significant population expansion for the Northeast lineage and a nonsignificant trend for the Southwest lineage. We used an approximate Bayesian computation (ABC) approach (Beaumont 2010) to test the relative importance of mechanisms pertaining to RBH and/or the PCH as drivers of Caatinga biodiversity while taking into account detailed information based on the results mentioned above. In addition, the ABC approach helped us to better determine the time of divergence between the lineages (see more details in Results section).

We built simulations to test four possible diversification scenarios (Fig. 3). (a) *Divergence with gene flow*. This is our simplest diversification scenario and it is equivalent to the IMA2 model, which does not incorporate population size changes. This scenario simulates diversification where lineages diverged in the presence of a barrier that has permitted some gene flow; this barrier could have been an ecological gradient (i.e. climate), a topographic gradient or a physical barrier that allowed for crossing in different periods. This scenario does not support any effect of the PCH, as it does not incorporate demographic responses. This is also the scenario with the lowest number of parameters, and its inclusion allows us to test whether a simpler diversification scenario could explain the observed genetic signature. Our remaining scenarios all consider demographical responses; because our previous results suggest population expansions, we did not consider scenarios that included population contractions. (b) *Divergence with gene flow and recent population expansion in both lineages*. In this scenario, lineages diverged similar to scenario (a), and subsequently both lineages were affected by climatic fluctuations and experienced recent population expansions. Scenarios (c) and (d) also simulate diversification where the initial lineage divergence was promoted by a barrier that has permitted some gene flow (i.e. climatic, topographic or physical). Both scenarios incorporate population expansion only in the Northeast lineage. In (c), *Divergence with gene flow and recent population expansion in Northeast lineage*, climatic fluctuations promote population expansion in the Northeast lineage while not affecting the Southeast lineage. This could occur, for example, if Southeast lineage distribution were located in areas more stable during climatic cycles. In (d), *Divergence with gene flow and founder effect for the Northeast lineage*, population expansion in the Northeast lineage is not necessarily related to climatic fluctuations. This scenario is based on the higher

genetic diversity of the Southwest lineage (longer branches) and the strong increase in population size in the Northeast lineage shown through the Bayesian skyline plot (BSP; see Results), which may suggest a founder effect with range expansion for the Northeast lineage. Within each of these four general diversification scenarios, we built four different specific models (Fig. 3). The models within the scenarios were built by varying divergence times and migration estimations. We used two different divergence times (T_1) based on *BEAST and the other based on IMA2 results (see Results). We also considered two possibilities of migration histories as follows: one with migration throughout the population history and another with recent migration (i.e. a possible secondary contact; T_2). All combinations resulted in 16 tested models [i.e. four models for each scenario; (i) *BEAST divergence times and migration throughout; (ii) IMA2 divergence times and migration throughout; (iii) *BEAST divergence times and recent migration; and (iv) IMA2 divergence times and recent migration]. Figure 3 shows a detailed illustration of the scenarios and models.

We performed 1 000 000 simulations for each model using MSABC (Pavlidis *et al.* 2010). We used an R script to sample parameters from prior distributions and call MSABC. Parameter prior distributions were based on results of *BEAST, BSP and IMA2 analyses (see Results and Table S5, Supporting information). We implemented priors in demographic quantities and transformed to *ms*-scaled parameters using equations from the *ms* manual (Hudson 2002). We called MSABC one time for each gene (i.e. five genes) using the same number of samples and length of loci. For the mitochondrial gene, we used one-fourth of the sampled population size. Five summary statistics for each gene were used: nucleotide diversity and Tajima's *D* for each population, and the F_{ST} between populations. We transformed observed sequence data into ms-like files using the *fas2ms* perl script (from MSABC package) and calculated the same summary statistics for each locus. To estimate posterior probabilities and model support ('postpr' function), we used R package 'abc' (Csilléry *et al.* 2012). We first compared the four possibilities within each scenario and kept the most probable one for a comparison among scenarios. We set tolerance to 0.0001 and implemented both the multinomial logistic regression and nonlinear neural network regression methods. Because ABC could not attain resolution in comparison among scenarios, we increased the tolerance rate to 0.001. We summarized simulated model fit to the observed data using a principal component analysis (PCA) calculated in R.

Divergence with gene flow is consistent with a scenario of speciation along gradients (e.g. Endler 1973; Nosil 2008). ABC simulation scenarios cannot distinguish among different sources of gradients, as they all would produce similar genetic signatures. To test whether the genetic variance observed can be explained by climatic gradients or topographic gradients, we conducted a redundancy analysis (RDA) using package 'vegan' (Oksanen *et al.* 2015) followed by an ANOVA to check for significance. RDA is a constrained ordination method, which combines multiple regression with PCA. We used the nuclear genotypes of individuals (i.e. GENELAND input) as dependent variables, and altitude, geographic coordinates and current climate as independent variables. We tested the influence of each variable and pairwise combination of variables while conditioning (*Condition* option within the *rda* function) to isolate the influence of each variable on the results. We used vegan's *varpart* function to disentangle and visualize the relative importance of each explanatory variable on the total genetic variance. Present climatic variables were downloaded from the WORLDCLIM database (see http://www.worldclim.org/ for variable descriptions) interpolated to 2.5 arc-min resolution (Hijmans *et al.* 2005). Among 19 climatic variables, we selected the five least correlated ones to represent the current climate (i.e. isothermality, precipitation seasonality, temperature annual range, annual precipitation and maximum temperature of the warmest month).

Results

DNA polymorphism and population structure

We sequenced all 398 individuals for mtDNA 12S gene (Table S1, Supporting information) and a subset of 137 individuals for nuDNA (Table S6, Supporting information). The nuDNA loci resulted in 115–134 sequences for each gene (Table 1). Only ~7% of the data was missing for all nuDNA loci. ATPSB and 12S sequences preserved ~89% (623 bp) and ~94% (370 bp) of their original size, respectively, after gap exclusion by GBLOCKS. The number of variable sites was highest in 12S (96 sites), with a maximum of 56 (ATPSB) and a minimum of 15 (RP40) for nuDNA loci. Table 1 shows additional information for all genes and population genetic statistics for each locus and each *Cnemidophorus ocellifer* lineage.

Using a genotype matrix, STRUCTURE detected two populations ($k = 2$; see Fig. S3, Supporting information). An identical result, with the same individuals being assigned to each population, was obtained using the haplotype data set in GENELAND (Fig. 2). One is distributed from north to southeastern Caatinga, occupying a large part of this biome (Northeast lineage), and the other occurs southwest of the Caatinga (i.e.

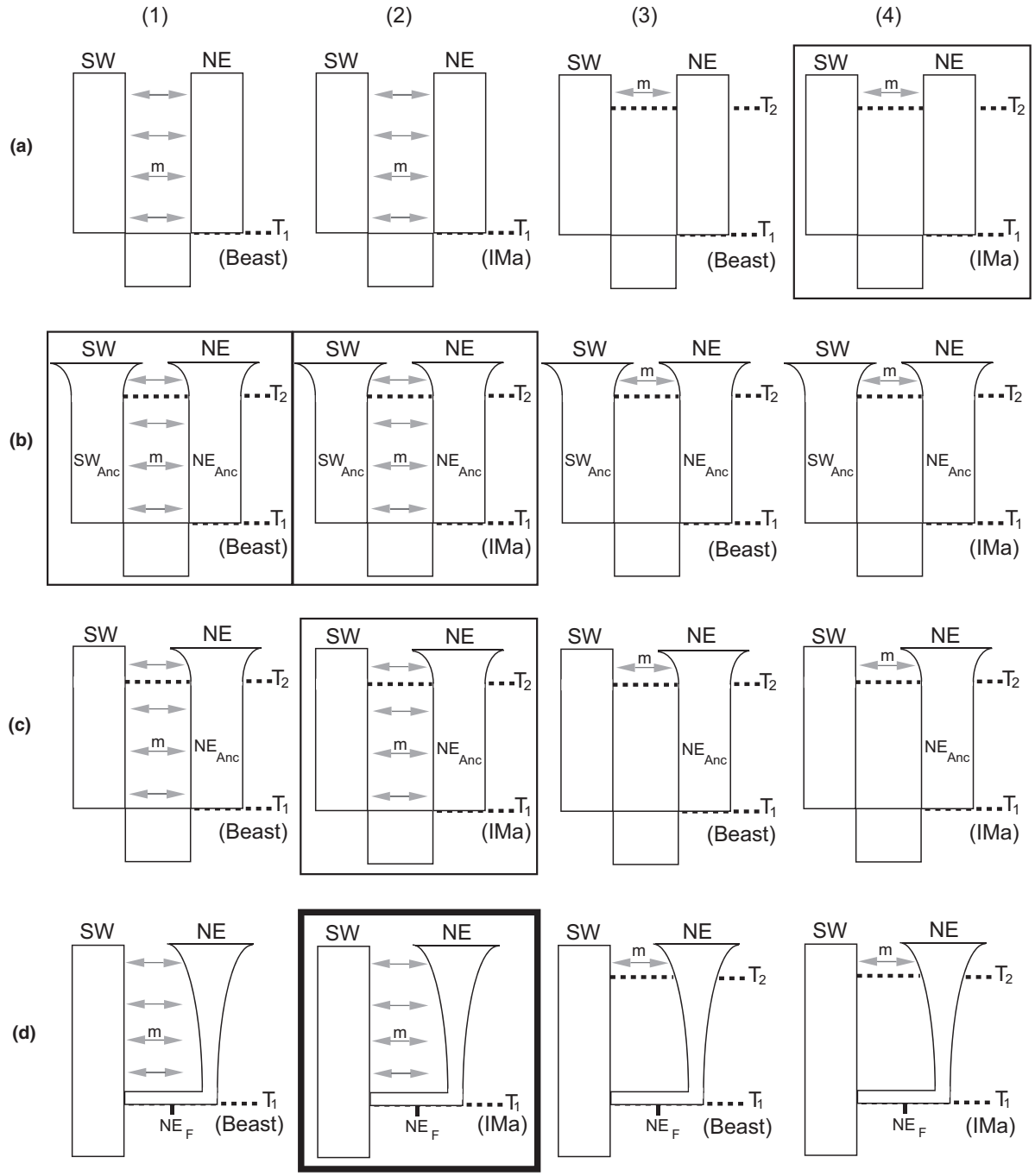

Fig. 3 Alternative scenarios for diversification of the Northeast (NE) and Southwest (SW) lineages tested with multilocus approximate Bayesian computation (ABC): (a) *divergence with gene flow*, (b) *divergence with gene flow and recent population expansion in both lineages*, (c) *divergence with gene flow and recent population expansion in Northeast lineage* and (d) *divergence with gene flow and founder effect for the Northeast lineage*. Within each of these four general diversification scenarios, we built four different specific models (1–4). The models within each scenario were built by varying divergence times and migration estimations. Divergence time is referred to as T_1 and was estimated from IMA2 (IMa) or *BEAST (Beast) analyses. Models include two migration rates (m) possibilities: migration throughout the population history (models 1 and 2; migration starting in T_1) and recent migration (models 3 and 4; migration starting in T_2). T_2 also represents the beginning of the population expansion (scenarios b and c). SW_{ANC} and NE_{ANC} represent Southwest and Northeast ancestral populations, respectively. Northeast founder population is referred to as NE_F. The best supported models under each scenario are reported within boxes, and the best among all models is highlighted within the thicker box. The posterior probabilities of all models are available in Table 4. Prior distributions for each parameter are available in Table S5 (Supporting information). See Materials and methods for more details.

Table 1 Genetic statistics for each locus for Northeast and Southwest lineages of *Cnemidophorus ocellifer*

Locus	Population	*L* (bp)	*N*	*S*	*H*	H_d	π	Tajima's *D*	*P*-value
12S	Northeast	370	330	66	103 (102)	0.947	0.01732	−1.29179	>0.10
	Southwest	370	68	47	34 (33)	0.967	0.02133	−1.6959	>0.10
ATPSB	Northeast	623	206*	32	35 (33)	0.830	0.00411	−1.49981	>0.10
	Southwest	623	46*	32	27 (25)	0.954	0.00984	−0.53478	>0.10
NKTR	Northeast	354	190*	24	41 (40)	0.647	0.00408	−1.79457	<0.05†
	Southwest	354	40*	17	22 (21)	0.937	0.01039	−0.42723	>0.10
R35	Northeast	396	220*	12	13 (11)	0.592	0.00178	−1.56108	0.10 > *P* > 0.05
	Southwest	396	48*	9	11 (9)	0.809	0.00421	−0.50002	>0.10
RP40	Northeast	354	216*	5	6 (4)	0.547	0.00185	−0.41487	>0.10
	Southwest	354	48*	11	11 (9)	0.793	0.00634	−0.27788	>0.10

L, length in base pairs; *N*, sample size; *S*, number of polymorphic sites; *H*, number of haplotypes (number of exclusive haplotypes in this lineage); H_d, haplotype diversity; π, nucleotide diversity.
Tajima's *D* tests and *P*-values.
*Phased sequences.
†Significant.

Espinhaço Mountain Range, hereafter EMR) and adjacent areas of Cerrado biome (Southwest lineage). Our data did not support the RBH because genetic breaks in *C. ocellifer* lineages did not match current or paleocourse of the SFR and both lineages are widespread on opposite riverbanks (Fig. 2).

According to Bayesian gene trees (Fig. S4, Supporting information), the Northeast lineage shows shallower subclades than those of the Southwest lineage. Most individuals are grouped on exclusive subclades for each lineage, although Northeast and Southwest lineages are not reciprocally monophyletic and show some admixture. Nevertheless, population structure can be easily visualized through mtDNA and nuDNA haplotype genealogies, and few haplotypes are shared between Northeast and Southwest lineages in all genes (Fig. 4 and Fig. S5, Supporting information). The Southwest lineage presents higher haplotype and nucleotide diversity than the Northeast lineage (Table 1). AMOVA showed high values of F_{ST} (28–66%) for all genes. In three genes (ATPSB, NKTR and RP40), the source of genetic variation was greater between lineages (Table 2). When we split both lineages in the opposite margins of the SFR (current and paleo), AMOVA showed low values of F_{ST} for all genes (Tables S7 and S8, Supporting information). This result did not support the RBH because the larger genetic distances of each lineage are within the same riverbanks. Uncorrected mtDNA *p*-distances exhibited substantial genetic differences (3.6%) between Northeast and Southwest lineages (Table 3).

Species tree estimation and migration

*BEAST analyses showed high support for Northeast and Southwest lineages ($P = 1$) and estimated the divergence around 2.07 Ma (95% HPD = 1.18–3.10 Ma), during the early Pleistocene.

In our IMA2 analyses, the likelihood ratio tests failed to reject three models in favour of the fully parameterized 'Isolation and Migration—IM' model (Table S9, Supporting information). AIC could not easily discriminate between two models given their low ΔAIC values. Both models have equal migration in both directions; one model suggests that extant population sizes are equal but both are different from the ancestral population size, while the other assumes distinct population sizes (Table S9, Supporting information). Equal migration in both directions was inferred at 0.46 migrants per generation per gene copy (95% HPD = 0.17–0.77). Effective population sizes of Northeast and Southwest lineages were estimated to be much larger (at least six times) than the ancestral population (Table S10, Supporting information). Divergence between Northeast and Southwest lineages was dated at 10.55 Ma (95% HPD = 6.79–14.76 Ma), during the Mid–Late Miocene, which is not consistent with estimates from the species tree. Because we detected significant migration rates between Northeast and Southwest lineages, which violates *BEAST assumptions, we should consider the divergence time estimated by IMA2. However, some violations of the IM model can also affect parameters estimated by IMA2 (see Strasburg & Rieseberg 2010). For instance, even moderate levels of gene flow from an unsampled species (i.e. third population) may overestimate divergence times (Strasburg & Rieseberg 2010). Considering that our sampling covers only the Caatinga and adjacent areas, and closely related species also reach the entire Cerrado biome, this violation is possible. To address divergence inconsistencies, we simulated different divergence times (i.e. *BEAST and IMA2

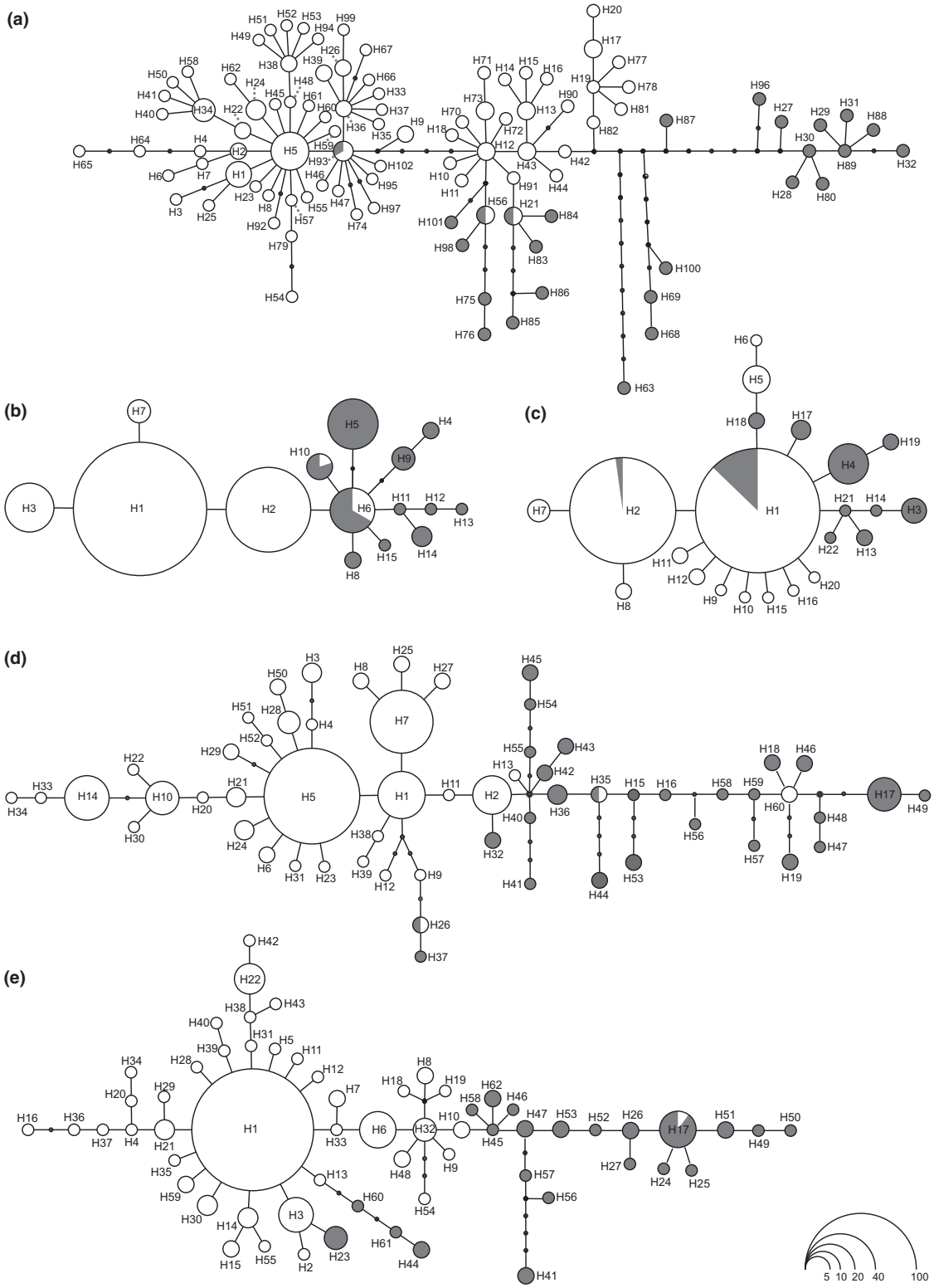

Fig. 4 Haplotype genealogies from maximum-likelihood analysis of 12S (a), RP40 (b), R35 (c), ATPSB (d) and NKTR (e) gene trees visualized in the software HAPLOVIEWER. Each haplotype is represented by a circle whose area is proportional to its frequency (indicated in legend). White and grey circles represent Northeast and Southwest lineages, respectively. Black dots represent inferred unsampled or extinct haplotypes. The localities where each haplotype (coded as a number) occurs are available in Table S6 (Supporting information).

Table 2 Variance percentages for components of analyses of molecular variance (AMOVAS) performed with different genes in Northeast and Southwest lineages of *Cnemidophorus ocellifer*

	Source of variation		
Locus	Between lineages	Within lineages	F_{ST}
12S	46.16	53.84	0.46158
ATPSB	58.89	41.11	0.58893
NKTR	55.47	44.53	0.55466
RP40	66.18	33.82	0.66179
R35	28.36	71.64	0.28359

All *P*-values < 0.0001.

Table 3 Genetic distance (uncorrected *p*-distance) between and within Northeast and Southwest lineages of *Cnemidophorus ocellifer* performed with different genes

		Within	
Locus	Between	Northeast	Southwest
12S	0.036	0.018	0.022
ATPSB	0.015	0.004	0.010
NKTR	0.014	0.004	0.011
R35	0.004	0.002	0.004
RP40	0.009	0.002	0.006

estimates) in the model-based ABC analysis (see Results below).

Historical demography and phylogeographic reconstruction

The Southwest lineage experienced no significant population size changes according to the BSP confidence interval, whereas the Northeast lineage, supporting the PCH, showed a significant increase in population size through time, with accelerated growth during the Late Pleistocene (Fig. 5).

Phylogeographic analyses suggested that the Northeast lineage originated in central-north Caatinga, encompassing a region in southeastern Piauí, southern Ceará, western Pernambuco and northern Bahia states (Fig. S6, Supporting information). All genes independently indicated this same region as a feasible area of origin, but some genes revealed wider areas. The ancestral population subsequently expanded towards the north, east and south in Caatinga, and colonized the entire biome during the Late Pleistocene (Fig. S6, Supporting information). This pattern of spatial expansion does not corroborate the RBH as it does not follow the spatial configuration of the SFR.

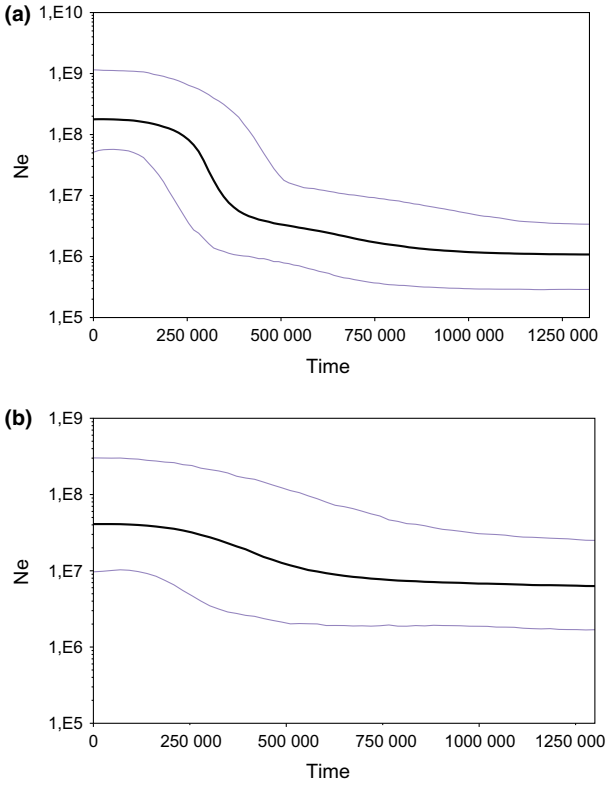

Fig. 5 Bayesian skyline plots illustrating effective population sizes (N_e) through time (in years) of Northeast (a) and Southwest (b) lineages. The black line represents the median population size, and the grey lines represent 95% higher posterior probability.

Species validation

BPP analyses presented mixing problems when we used the larger data set as input (137 individuals). In this case, the rjMCMC algorithm was trapped in the starting model, being it a one-species or a two-species model. Thus, independent runs yielded different results. Conversely, the chains mixed well when we used the small data set, yielding similar results between multiple runs using the two algorithms. BPP analyses delimited both lineages with high posterior probability (100%) under the four tested models. Likewise, SPEDESTEM analyses also recognized two lineages under two values of θ (Table S11, Supporting information). Therefore, our species validation approaches showed congruent results, which validate our population assignments and reinforce the validity of our lineages.

Diversification scenarios

Approximate Bayesian computation analysis showed higher support for the scenario of divergence with gene flow and founder effect for the Northeast lineage

(Fig. 3d). Within this scenario, we found highest support to the model with IMA2 divergence date estimates and constant migration ($P = 0.999$ or $P = 1$; see Table 4). The model with IMA2 divergence times was the best supported model in comparisons within scenarios a, c and b. In scenario b, the model with IMA2 divergence times was equally supported to the model with *BEAST divergence times. Constant migration was also supported in comparisons within scenarios b and c (Table 4). Summary statistics of our observed data were within the bounds of the summary statistics for founder effect simulations in the PCA predictive plots, confirming good model fitting (Fig. S7, Supporting information).

Redundancy analysis results showed that genetic variance between main lineages is significantly associated with spatial structure ($r = 0.04$; $P = 0.001$) and current climate ($r = 0.05$; $P = 0.001$), and it is not associated with altitude ($r = -0.001$; $P = 0.56$; Table S12, Supporting information). The fraction of explained genetic variance by the explanatory variables is 31%. The relative importance of current climate is around 16%, spatial structure 13% and 0 for altitude. However, most of the explained variance is shared among the three sets of explanatory variables (~32%) and between spatial structure and current climate (other ~32%).

Table 4 Results from approximate Bayesian computation analyses assuming four scenarios of diversification (a, b, c or d; see Fig. 3) between Northeast and Southwest lineages of *Cnemidophorus ocellifer*

		Regression method	
Scenario	Model	mnlogistic	Neuralnet
a	BeastCM	0	0
	BeastSC	0	0
	IMaCM	0.0002	0.0027
	IMaSC	**0.9998**	**0.9973**
b	BeastCM	0.1481	**0.4150**
	BeastSC	0.1166	0.3841
	IMaCM	**0.6912**	0.1566
	IMaSC	0.0441	0.0443
c	BeastCM	0.1709	0.0583
	BeastSC	0.2375	0.1923
	IMaCM	**0.3884**	**0.5969**
	IMaSC	0.2031	0.1525
d	BeastCM	0	0.0012
	BeastSC	0	0
	IMaCM	**0.9892**	**0.9006**
	IMaSC	0.0108	0.0982
a	IMaSC	†	0
b	BeastCM	0	0
b	IMaCM	0	0
c	IMaCM	0.0001	0
d	IMaCM	**0.9999**	**1**

Values represent posterior probabilities of comparisons within each scenario (tolerance of 0.0001) and among the best models selected for each scenario (tolerance of 0.001). Preferred model at each comparison is highlighted in bold.
Model: IMA2 divergence time (IMa), *BEAST divergence time (Beast), recent migration [secondary contact (SC)] and migration throughout the population history [constant migration (CM)].
†Zero simulations accepted in the rejection step, and therefore, the model was not included in the regression step.

Discussion

The RBH and the role of the SFR

Our results do not support the role of the SFR as a major barrier promoting diversification in the Caatinga. We found evidence for gene flow across the river, lacking the presence of clear lineages across river margins (Fig. 2). We also found that genetic distances were not higher across river margins (Table S8, Supporting information). In addition, the phylogeographic history of the Northeast lineage does not imply any major role of the SFR. Our results are consistent with a few recent studies on widespread Caatinga lineages that have showed no phylogeographic structure between SFR riverbanks (Caetano *et al.* 2008; Magalhães *et al.* 2014; Recoder *et al.* 2014; São-Pedro 2014). However, the role of the SFR is evident in many other studies. For example, several endemic genera and species pairs are isolated on opposite riverbanks (Rodrigues 1996, 2003). In addition, the phylogenetic relationships of some groups have supported the riverine hypothesis and the role of the SFR (Passoni *et al.* 2008; Siedchlag *et al.* 2010; Nascimento *et al.* 2011, 2013; Faria *et al.* 2013; Werneck *et al.* 2015). Therefore, our results together with previous studies suggest that the role of SFR may be idiosyncratic and/or may only be predictive given the specific biological characteristics of the organisms, such as dispersal capabilities and habitat specificity, as demonstrated for frogs in the Amazon region (Fouquet *et al.* 2015). The species studied so far in which the SFR was found to be a barrier are microendemics or associated with sandy soils near the riverbanks (e.g. Rodrigues 1996, 2003; Passoni *et al.* 2008; Siedchlag *et al.* 2010; Nascimento *et al.* 2013). However, more studies with a sampling strategy aiming to test the role of SFR (e.g. systematic sampling of different species with different characteristics along both river margins) are necessary to confirm this assertion.

The PCH and the role of Pleistocene climatic fluctuations

The PCH suggests that climatic fluctuations during the Pleistocene have promoted isolation and divergence

driving the origin and diversification of Caatinga's biota. Because we found gene flow between lineages and that the timing of divergence between lineages was older than the Pleistocene (6–14 Ma), our results do not support the PCH. We also found that the widespread Northeast lineage has low levels of genetic variation and a clear signal of population expansion. The timing of population expansion is consistent with an effect of Pleistocene climatic fluctuations. However, when different scenarios were simulated in our ABC approach, the model that best fit our data was one where demographic changes experienced by the lineage were caused by a founder effect. The genetic signature showed by the Northeast lineage is consistent with a scenario that does not require climatic fluctuations to explain the demographic responses. In this scenario, after the initial split between the Southwest and Northeast lineages, the Northeast lineage originated with a low population size and expanded continuously. This expansion may be just a consequence of the founder effect process or may have been expedited by climatic fluctuations. Ultimately though, our data do not show unequivocal evidence to support a major role of PCH in the diversification of the Caatinga biota.

Previous studies showed other endemic Caatinga groups with genetic signatures consistent with population expansions (e.g. São-Pedro 2014; Werneck *et al.* 2015) or at least short coalescent times/branch lengths (e.g. Werneck *et al.* 2012; Recoder *et al.* 2014). Some studies have reported population contractions during the Pleistocene (e.g. Magalhães *et al.* 2014), and other have found stable population sizes during climatic fluctuations (Nascimento *et al.* 2011; Faria *et al.* 2013). The Caatinga has experienced intense climatic fluctuations with wetter conditions favouring the expansion of humid forests in distinct regions during the last 210 000 years (De Oliveira *et al.* 1999; Behling *et al.* 2000; Auler *et al.* 2004; Wang *et al.* 2004). Many studies in different systems have interpreted these population fluctuation signatures as effects associated with range expansions/contractions in response to habitat shifts caused by Pleistocene climatic fluctuations (e.g. Carnaval & Bates 2007; Shepard & Burbrink 2008; Qu *et al.* 2011). Our results highlight that population expansion signatures driven by processes other than climatic cycles can be teased apart using simulation methods. As a consequence, the role of PCH in demographic responses in the Caatinga may be overestimated.

Diversification in Cnemidophorus ocellifer

We generated an extensive multilocus data set with large geographical coverage. Our species validation approaches recovered two well-delimited lineages that diverged despite the presence of gene flow. We reject both major hypotheses proposed for diversification in the Caatinga. Surprisingly, our results revealed a striking complex diversification pattern that is consistent with a model of speciation with gene flow along environmental gradients (Endler 1973; Nosil 2008). Our model-based analysis suggests that the Northeast lineage originated from individuals located along the edge of the Southwest lineage distribution, which eventually diverged into the Northeast lineage and expanded along the Caatinga. The Southwest lineage occurs at higher elevations and milder climates (i.e. Cerrado and EMR), and our RDA and variance-partitioning analysis showed that most of the genetic variance is explained by climatic variables with altitudinal differences having little importance. Therefore, our results support the role of climatic gradients as drivers of speciation. These same mechanisms have been suggested to explain diversification between Caatinga and Cerrado adjacent lineages of the gecko *Phyllopezus pollicaris* (Werneck *et al.* 2012).

Most of the analyses we implemented here gave us an overview on the phylogeography of *Cnemidophorus ocellifer* species complex, while the ABC analysis proved to be extremely valuable to circumvent limitations of descriptive and other model-based methods. One example of such limitations is in our IMA2 results. The likelihood or the information content regarding nested IM models was not significantly different. Thus, either different IM models are equally good at describing our data, or the IM model is not appropriate due to violation of essential assumptions. For instance, the IM model does not allow changes in population size through the history of a population. Therefore, one possible violation is the presence of population expansion evident in our data. For fitting population size change, *BEAST may represent a good alternative. However, the absence of a migration parameter in the *BEAST model poses another model violation. Thus, in our case, these methods were extremely useful for estimating confidence intervals of population parameters, narrowing down the possibilities of diversification scenarios. In this context, a simulation-based approach, such as ABC, emerges as a powerful way to choose among different more complex diversification scenarios, overcoming limitations of methods with fixed models containing a particular number of parameters. Although ABC was long used to test complex diversification scenarios in humans (Fagundes *et al.* 2007), the method is seldomly used for the investigation of population processes in Neotropical organisms (e.g. Carnaval *et al.* 2009; Batalha-Filho *et al.* 2012; Werneck *et al.* 2012; Thomé *et al.* 2014). Here, we highlight the importance of the method and suggest the use of ABC to test specific

hypothesis of species diversification and to test the generality of our results for the Caatinga.

The Cerrado may also be a source of diversity for the formation of Caatinga biota. Because the distribution of the *C. ocellifer* species complex extends further into the entire Cerrado, the Southwest lineage may be part of a more widespread lineage from central Brazil that reaches southwestern Caatinga. Therefore, additional samples from central Cerrado are necessary to better define genealogic relationships within the Southwest lineage and its geographic boundaries. Even considering that the Southwest lineage was undersampled, it showed higher genetic diversity than the Northeast lineage. The Southwest lineage is associated with the Cerrado, a biome characterized by high geomorphological complexity and landscape heterogeneity (Cole 1986; Nogueira *et al.* 2011). Recent studies have revealed deep cryptic diversity in many Cerrado lizards (Werneck *et al.* 2012; Domingos *et al.* 2014; Recoder *et al.* 2014; Santos *et al.* 2014). The Cerrado and Caatinga share a large contact zone, and many taxa (likely species complexes) are recognized as widely distributed across both biomes. Our results along with other studies suggest that Cerrado/Caatinga widespread taxa are in fact different species with Cerrado species consistently showing higher genetic diversity and deeper coalescence (Werneck *et al.* 2012; Recoder *et al.* 2014). In fact, many studies have found multiple lineages within Cerrado taxa (e.g. Werneck *et al.* 2012; Domingos *et al.* 2014; Santos *et al.* 2014), whereas widespread lineages from the Caatinga have shown weak genetic structure, including plants (Caetano *et al.* 2008), frogs (São-Pedro 2014) and lizards (our study; Werneck *et al.* 2012; Recoder *et al.* 2014).

Our species validation approaches supported the existence of two unrecognized lineages of *C. ocellifer*. We also show evidence for only one widespread lineage in the Caatinga. We sequenced specimens within the range and with morphological patterns attributed to *Cnemidophorus pyrrhogularis* (Silva & Ávila-Pires 2013), and they were genetically indistinguishable from our Northeast lineage. Hence, our findings suggest that Northeast lineage and *C. pyrrhogularis* correspond to the same species. Based on the distribution of species from the *C. ocellifer* group, our Northeast lineage is likely to be *C. ocellifer*. Therefore, we conclude that *C. pyrrhogularis* is a junior synonym of *C. ocellifer* (see details in 'Taxonomic background and implications' in Appendix S1, Supporting information). In addition, some individuals included in our analyses present morphological characteristics stated as diagnostic of *Cnemidophorus nigrigula* (D. O. Mesquita, personal communication; Garda *et al.* 2013). We also sequenced specimens that based on lepidosis were assigned to *Cnemidophorus confusionibus* (Arias *et al.* 2011a). Still, these samples were not genetically distinguishable from our Northeast lineage. Therefore, we recommend a more detailed investigation on the identity of *C. nigrigula* and *C. confusionibus*, by incorporating genetic samples from their type localities. Our Southwest lineage is the recent described species, *Cnemidophorus xacriaba* (Arias *et al.* 2014). Some specimens in our analyses were also used in the *C. xacriaba* description paper, including samples from the type locality. The description paper claims that *C. xacriaba* is endemic to the Planalto dos Gerais in the Cerrado region. Our results do not support this claim and suggest a much wider occurrence area for *C. xacriaba* reaching many localities along the southwest Caatinga. It is likely that *C. xacriaba* is also more widespread in the Cerrado biome. Hence, other populations formerly designated as *C. ocellifer* from central Brazil (Arias *et al.* 2014) are most likely our Southwest lineage (i.e. *C. xacriaba*) or other still undescribed taxa.

Conclusion

We conducted a detailed phylogeographic assessment of the *Cnemidophorus ocellifer* species complex from the Caatinga and adjacent areas through the coupling of multilocus data, coalescent methods, historical demographic reconstructions and model testing. The *C. ocellifer* species complex is structured into two distinct genetic lineages in the Caatinga, which have split during the Mid–Late Miocene and diverged despite continuous gene flow. The Northeast lineage occurs exclusively in the Caatinga and expanded its range rapidly during the Pleistocene. We found no evidence that the RBH or PCH influences its current geographic structure. The Southwest lineage is more diverse, older and associated with the Cerrado–Caatinga boundaries. Our results also suggest that the Northeast lineage originated from the Southwest lineage and expanded continuously, leaving a genetic signature concordant with a founder effect. We conclude that *C. ocellifer* from the Caatinga is composed of two distinct species associated with different geographic regions. Our data support a speciation with gene flow and highlight the role of the environmental gradients in the diversification process.

Acknowledgements

We thank researchers and curators of the following Brazilian herpetological collections for providing tissue samples and distribution records: Universidade Católica do Salvador (UCSAL), Universidade Federal de Alagoas (UFAL), Universidade Federal do Ceará (UFC), Universidade Federal do Rio Grande do Norte (UFRN), Universidade Federal da Paraíba (UFPB), Universidade Federal de Sergipe (UFS), Universidade Federal

de Viçosa (UFV), Universidade de Brasília (UnB) and Universidade de São Paulo (USP). We also thank L.O. Drummond and P.H. Medeiros for their assistance in the field expeditions; and A. McKelvy and S. Ruane from Burbrink laboratory for laboratorial assistance. E.F.O. thanks Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for providing a graduate fellowship and Instituto Chico Mendes de Conservação da Biodiversidade for collecting permit (no. 26255-1). This work was funded by two grants from Conselho Nacional de Desenvolvimento Científico e Tecnológico (SISBIOTA Grant CNPq no. 563352/2010-8 to G.C.C. and ICMBio/CNPq Grant no. 552031/2011-9 to A.A.G.), two from Conselho de Aprimoramento de Pessoal de Nível Superior (CAPES PVE no. 001/2012 to A.A.G. and BJT-A058/2013 to M.G.) and three from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP no. 03/8335-8, no. 11/50146-6, no. 12/02212-2 to M.T.R). D.O.M. thanks CAPES for a postdoctorate fellowship and CNPq for a research fellowship (303610/2014-0). G.R.C. thanks CAPES, CNPq and Fundação de Apoio à Pesquisa do Distrito Federal (FAPDF) for financial support. G.C.C. thanks CNPq grants 474392/2013-9, 201413/2014-0 and 302776/2012-5. This project was also supported in part by L. Clampitt, D. Rosenberg and S. Harris and a US National Science Foundation Grant (DEB 1257926) to F.T.B. We thank Felipe Grazziotin and Sérgio Lima for suggestions on earlier versions of the manuscript.

References

Anthony NM, Johnson-Bawe M, Jeffery K *et al*. (2007) The role of Pleistocene refugia and rivers in shaping gorilla genetic diversity in central Africa. *Proceedings of the National Academy of Sciences of the USA*, **104**, 20432–20436.

Arias F, Carvalho CM, Rodrigues MT, Zaher H (2011a) Two new species of *Cnemidophorus* (Squamata: Teiidae) from the Caatinga, Northwest Brazil. *Zootaxa*, **2787**, 37–54.

Arias F, Carvalho CM, Rodrigues MT, Zaher H (2011b) Two new species of *Cnemidophorus* (Squamata: Teiidae) of the *C. ocellifer* group, from Bahia, Brazil. *Zootaxa*, **3022**, 1–21.

Arias FJ, Teixeira M, Recoder R *et al*. (2014) Whiptail lizards in South America: a new *Ameivula* (Squamata, Teiidae) from Planalto dos Gerais, eastern Brazilian Cerrado. *Amphibia-Reptilia*, **35**, 227–242.

Auler AS, Wang X, Edwards RL *et al*. (2004) Quaternary ecological and geomorphic changes associated with rainfall events in presently semi-arid northeastern Brazil. *Journal of Quaternary Science*, **19**, 693–701.

Ayres J, Clutton-Brock T (1992) River boundaries and species range size in Amazonian primates. *The American Naturalist*, **140**, 531–537.

Bandelt HJ, Forster P, Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, **16**, 37–48.

Batalha-Filho H, Cabanne GS, Miyaki CY (2012) Phylogeography of an Atlantic forest passerine reveals demographic stability through the last glacial maximum. *Molecular Phylogenetics and Evolution*, **65**, 892–902.

Beaumont MA (2010) Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, **41**, 379–406.

Beheregaray LB (2008) Twenty years of phylogeography: the state of the field and the challenges for the southern hemisphere. *Molecular Ecology*, **17**, 3754–3774.

Behling HW, Arz H, Pätzold J, Wefer G (2000) Late quaternary vegetational and climate dynamics in northeastern Brazil, inferences from marine core GeoB 3104-1. *Quaternary Science Reviews*, **19**, 981–994.

Bielejec F, Rambaut A, Suchard MA, Lemey P (2011) SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics. *Bioinformatics*, **27**, 2910–2912.

Caetano S, Prado D, Pennington RT *et al*. (2008) The history of Seasonally Dry Tropical Forests in eastern South America: inferences from the genetic structure of the tree *Astronium urundeuva* (Anacardiaceae). *Molecular Ecology*, **17**, 3147–3159.

Carnaval AC, Bates JM (2007) Amphibian DNA shows marked genetic structure and tracks Pleistocene climate in Northeastern Brazil. *Evolution*, **61**, 2942–2957.

Carnaval AC, Hickerson MJ, Haddad CFB, Rodrigues MT, Moritz C (2009) Stability predicts genetic diversity in the Brazilian Atlantic Forest hotspot. *Science*, **323**, 785–789.

Carstens BC, Dewey TA (2010) Species delimitation using a combined coalescent and information-theoretic approach: an example from North American *Myotis* bats. *Systematic Biology*, **59**, 400–414.

Carstens BC, Pelletier TA, Reid NM, Satler JD (2013) How to fail at species delimitation. *Molecular Ecology*, **22**, 4369–4383.

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

Cole MM (1986) *The Savannas: Biogeography and Geobotany*. Academic Press, London.

Costello MJ, May RM, Stork NE (2013) Can we name Earth's species before they go extinct? *Science*, **339**, 413–416.

Csilléry K, François O, Blum MGB (2012) abc: an R package for approximate Bayesian computation (ABC). *Methods in ecology and evolution*, **3**, 475–479.

De Oliveira PE, Barreto AMF, Suguio K (1999) Late Pleistocene/Holocene climatic and vegetational history of the Brazilian caatinga: the fossil dunes of the middle São Francisco River. *Palaeogeography Palaeoclimatology Palaeoecology*, **152**, 319–337.

Dessauer HC, Cole CJ, Townsend CR (2000) Hybridization among western whiptail lizards (*Cnemidophorus tigris*) in southwestern New Mexico: population genetics, morphology, and ecology in three contact zones. *Bulletin of the American Museum of Natural History*, **246**, 1–148.

Domingos FM, Bosque RJ, Cassimiro J *et al*. (2014) Out of the deep: cryptic speciation in a Neotropical gecko (Squamata, Phyllodactylidae) revealed by species delimitation methods. *Molecular Phylogenetics and Evolution*, **80**, 113–124.

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

Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, **22**, 1185–1192.

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, **29**, 1969–1973.

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output

and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359–361.

Ence DD, Carstens BC (2011) SpedeSTEM: a rapid and accurate method for species delimitation. *Molecular Ecology Resources*, **11**, 473–480.

Endler JA (1973) Gene flow and population differentiation studies of clines suggest that differentiation along environmental gradients may be independent of gene flow. *Science*, **179**, 243–250.

Eo SH, DeWoody JA (2010) Evolutionary rates of mitochondrial genomes correspond to diversification rates and to contemporary species richness in birds and reptiles. *Proceedings of the Royal Society of London B: Biological Sciences*, **277**, 3587–3592.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

Fagundes NJ, Ray N, Beaumont M *et al.* (2007) Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences of the USA*, **104**, 17614–17619.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Faria MB, Nascimento FF, de Oliveira JA, Bonvicino CR (2013) Biogeographic determinants of genetic diversification in the mouse opossum *Gracilinanus agilis* (Didelphimorphia: Didelphidae). *Journal of Heredity*, **104**, 613–626.

Fouquet A, Courtois EA, Baudain D *et al.* (2015) The trans-riverine genetic structure of 28 Amazonian frog species is dependent on life history. *Journal of Tropical Ecology*, **31**, 361–373.

Garda AA, Costa TB, Santos-Silva CRd *et al.* (2013) Herpetofauna of protected areas in the Caatinga I: Raso da Catarina Ecological Station (Bahia, Brazil). *Check List*, **9**, 405–414.

Gifford ME, Larson A (2008) In situ genetic differentiation in a Hispaniolan lizard *Ameiva chrysolaema*: a multilocus perspective. *Molecular Phylogenetics and Evolution*, **49**, 277–291.

Guillot G, Estoup A, Mortier F, Cosson JF (2005a) A spatial statistical model for landscape genetics. *Genetics*, **170**, 1261–1280.

Guillot G, Mortier F, Estoup A (2005b) Geneland: a computer package for landscape genetics. *Molecular Ecology Notes*, **5**, 712–715.

Guindon S, Dufayard J-F, Lefort V *et al.* (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, **59**, 307–321.

Haffer J (1969) Speciation in Amazonian forest birds. *Science*, **165**, 131–137.

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

Hewitt G (2000) The genetic legacy of the Quaternary ice ages. *Nature*, **405**, 907–913.

Hey J (2010) Isolation with migration models for more than two populations. *Molecular Biology and Evolution*, **27**, 905–920.

Hey J, Nielsen R (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences of the USA*, **104**, 2785–2790.

Hickerson MJ, Carstens BC, Cavender-Bares J *et al.* (2010) Phylogeography's past, present, and future: 10 years after. *Molecular Phylogenetics and Evolution*, **54**, 291–301.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Leaché AD, Fujita MK (2010) Bayesian species delimitation in West African forest geckos (*Hemidactylus fasciatus*). *Proceedings of the Royal Society of London B: Biological Sciences*, **277**, 3071–3077.

Lemey P, Rambaut A, Welch JJ, Suchard MA (2010) Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology and Evolution*, **27**, 1877–1885.

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**, 1451–1452.

Machado T, Silva VX, Silva MJdJ (2014) Phylogenetic relationships within *Bothrops neuwiedi* group (Serpentes, Squamata): geographically highly-structured lineages, evidence of introgressive hybridization and Neogene/Quaternary diversification. *Molecular Phylogenetics and Evolution*, **71**, 1–14.

Magalhães IL, Oliveira U, Santos FR *et al.* (2014) Strong spatial structure, Pliocene diversification and cryptic diversity in the Neotropical dry forest spider *Sicarius cariri*. *Molecular Ecology*, **23**, 5323–5336.

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How many species are there on Earth and in the Ocean? *PLoS Biology*, **9**, e1001127.

Nascimento FF, Pereira LG, Geise L *et al.* (2011) Colonization process of the Brazilian common vesper mouse, *Calomys expulsus* (Cricetidae, Sigmodontinae): a biogeographic hypothesis. *Journal of Heredity*, **102**, 260–268.

Nascimento FF, Lazar A, Menezes AN *et al.* (2013) The role of historical barriers in the diversification processes in open vegetation formations during the Miocene/Pliocene using an ancient rodent lineage as a model. *PLoS ONE*, **8**, e61924.

Nogueira CC, Ribeiro S, Costa GC, Colli GR (2011) Vicariance and endemism in a Neotropical savanna hotspot: distribution patterns of Cerrado squamate reptiles. *Journal of Biogeography*, **38**, 1907–1922.

Nosil P (2008) Speciation with gene flow could be common. *Molecular Ecology*, **17**, 2103–2106.

Oksanen J, Blanchet FG, Kindt R *et al.* (2015) *vegan: Community ecology package*. R package version 2.2-1. Available from http://CRAN.R-project.org/package=vegan.

Passoni J, Benozzati M, Rodrigues M (2008) Phylogeny, species limits, and biogeography of the Brazilian lizards of the genus *Eurolophosaurus* (Squamata: Tropiduridae) as inferred from mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution*, **46**, 403–414.

Patton JL, Da Silva MNF, Malcolm JR (1994) Gene genealogy and differentiation among arboreal spiny rats (Rodentia: Echimyidae) of the Amazon basin: a test of the riverine barrier hypothesis. *Evolution*, **48**, 1314–1323.

Pavlidis P, Laurent S, Stephan W (2010) msABC: a modification of Hudson's ms to facilitate multi-locus ABC analysis. *Molecular Ecology Resources*, **10**, 723–727.

Pennington RT, Lavin M, Prado DE *et al.* (2004) Historical climate change and speciation: Neotropical Seasonally Dry Forest plants show patterns of both Tertiary and Quaternary diversification. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **359**, 515–537.

Pennington RT, Lavin M, Oliveira-Filho A (2009) Woody plant diversity, evolution, and ecology in the tropics: perspectives from Seasonally Dry Tropical Forests. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 437–457.

Posada D (2008) jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution*, **25**, 1253–1256.

Potter PE (1997) The Mesozoic and Cenozoic paleodrainage of South America: a natural history. *Journal of South American Earth Sciences*, **10**, 331–344.

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

Pyron R, Burbrink F, Wiens J (2013) A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. *BMC Evolutionary Biology*, **13**, 93.

Qu Y, Luo X, Zhang R *et al.* (2011) Lineage diversification and historical demography of a montane bird *Garrulax elliotii*-implications for the Pleistocene evolutionary history of the eastern Himalayas. *BMC Evolutionary Biology*, **11**, 174.

Queiroz LP (2006) The Brazilian Caatinga: phytogeographical patterns inferred from distribution data of the Leguminosae. In: *Neotropical Savannas and Seasonally Dry Forests: Plant Diversity, Biogeography and Conservation* (eds Pennington RT, Lewis GP, Ratter JA), pp. 121–157. CRC Press, Boca Raton, Florida.

R Core Team (2015) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org.

Recoder RS, Werneck FP, Teixeira M *et al.* (2014) Geographic variation and systematic review of the lizard genus *Vanzosaura* (Squamata, Gymnophthalmidae), with the description of a new species. *Zoological Journal of the Linnean Society*, **171**, 206–225.

Rocha CFD, Bergallo HG, Peccinini-Seale D (1997) Evidence of an unisexual population of the Brazilian whiptail lizard genus *Cnemidophorus* (Teiidae), with description of a new species. *Herpetologica*, **53**, 374–382.

Rodrigues MT (1996) Lizards, snakes, and amphisbaenians from the Quaternary sand dunes of the middle Rio São Francisco, Bahia, Brazil. *Journal of Herpetology*, **30**, 513–523.

Rodrigues MT (2003) Herpetofauna da Caatinga. In: *Ecologia e Conservação da Caatinga* (eds Leal IR, Tabarelli M, Silva JMC), pp. 181–236. Editora Universitária da UFPE, Recife.

Rull V (2008) Speciation timing and Neotropical biodiversity: the Tertiary-Quaternary debate in the light of molecular phylogenetic evidence. *Molecular Ecology*, **17**, 2722–2729.

Rull V (2011) Neotropical biodiversity: timing and potential drivers. *Trends in Ecology & Evolution*, **26**, 508–513.

Rull V (2013) Some problems in the study of the origin of Neotropical biodiversity using palaeoecological and molecular phylogenetic evidence. *Systematics and Biodiversity*, **11**, 415–423.

Salzburger W, Ewing GB, Von Haeseler A (2011) The performance of phylogenetic algorithms in estimating haplotype genealogies with migration. *Molecular Ecology*, **20**, 1952–1963.

Sampaio EVSB (1995) Overview of the Brazilian Caatinga. In: *Seasonally Dry Tropical Forests* (eds Bullock SH, Mooney HA, Medina E), pp. 35–63. Cambridge University Press, Cambridge, UK.

Santos MG, Nogueira C, Giugliano LG, Colli GR (2014) Landscape evolution and phylogeography of *Micrablepharus atticolus* (Squamata, Gymnophthalmidae), an endemic lizard of the Brazilian Cerrado. *Journal of Biogeography*, **41**, 1506–1519.

São-Pedro VA (2014) *Filogeografia de anfíbios da diagonal de áreas abertas da América do Sul*. PhD Thesis, Universidade Federal do Rio Grande do Norte.

Sequeira F, Sodre D, Ferrand N *et al.* (2011) Hybridization and massive mtDNA unidirectional introgression between the closely related Neotropical toads *Rhinella marina* and *R. schneideri* inferred from mtDNA and nuclear markers. *BMC Evolutionary Biology*, **11**, 264.

Shepard DB, Burbrink FT (2008) Lineage diversification and historical demography of a sky island salamander, *Plethodon ouachitae*, from the Interior Highlands. *Molecular Ecology*, **17**, 5315–5335.

Siedchlag AC, Benozzati ML, Passoni JC, Rodrigues MT (2010) Genetic structure, phylogeny, and biogeography of Brazilian eyelid-less lizards of genera *Calyptommatus* and *Nothobachia* (Squamata, Gymnophthalmidae) as inferred from mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution*, **56**, 622–630.

Sievers F, Wilm A, Dineen D *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, **7**, 539.

Silva MB, Ávila-Pires TCS (2013) The genus *Cnemidophorus* (Squamata: Teiidae) in state of Piauí, northeastern Brazil, with description of a new species. *Zootaxa*, **3681**, 455–477.

Smith BT, McCormack JE, Cuervo AM *et al.* (2014) The drivers of tropical speciation. *Nature*, **515**, 406–409.

Spix JBRv (1825) *Animalia Nova sive species novae Lacertarum, quas in itinere per Brasiliam annis MDCCCXVII-MDCCCXX jussu et auspiciis Maximiliani Josephi I. Bavariae regis*. Typis Franc. Seraph Hubschmanni, Munich.

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, **68**, 978–989.

Strasburg JL, Rieseberg LH (2010) How robust are "isolation with migration" analyses to violations of the IM model? A simulation study. *Molecular Biology and Evolution*, **27**, 297–310.

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, **56**, 564–577.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, **30**, 2725–2729.

Thomé MTC, Zamudio KR, Haddad CF, Alexandrino J (2014) Barriers, rather than refugia, underlie the origin of diversity in toads endemic to the Brazilian Atlantic Forest. *Molecular Ecology*, **23**, 6152–6164.

Turchetto-Zolet AC, Pinheiro F, Salgueiro F, Palma-Silva C (2013) Phylogeographical patterns shed light on evolutionary process in South America. *Molecular Ecology*, **22**, 1193–1213.

Wang X, Auler AS, Edwards RL *et al.* (2004) Wet periods in northeastern Brazil over the past 210 kyr linked to distant climate anomalies. *Nature*, **432**, 740–743.

Werneck FP (2011) The diversification of eastern South American open vegetation biomes: historical biogeography and perspectives. *Quaternary Science Reviews*, **30**, 1630–1648.

Werneck FP, Costa GC, Colli GR, Prado DE, Sites JW Jr (2011) Revisiting the historical distribution of Seasonally Dry

Tropical Forests: new insights based on palaeodistribution modelling and palynological evidence. *Global Ecology and Biogeography*, **20**, 272–288.

Werneck FP, Gamble T, Colli GR, Rodrigues MT, Sites JJW (2012) Deep diversification and long-term persistence in the South American 'dry diagonal': integrating continent-wide phylogeography and distribution modeling of geckos. *Evolution*, **66**, 3014–3034.

Werneck FP, Leite RN, Geurgas SR, Rodrigues MT (2015) Biogeographic history and cryptic diversity of saxicolous Tropiduridae lizards endemic to the semiarid Caatinga. *BMC Evolutionary Biology*, **15**, 94.

Yang Z (2015) The BPP program for species tree estimation and species delimitation. *Current Zoology*, **61**, 854–865.

Yang Z, Rannala B (2010) Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences of the USA*, **107**, 9264–9269.

Zanella FC, Martins CF (2003) Abelhas da Caatinga: biogeografia, ecologia e conservação. In: *Ecologia e Conservação da Caatinga* (eds Leal IR, Tabarelli M, Silva JMC), pp. 75–134. Editora Universitária da UFPE, Receife, Pernambuco.

E.F.O. and G.C.C. conceived the initial idea of the study. E.F.O., V.A.S.-P., D.O.M., A.A.G., G.R.C., M.T.R., F.J.A., H.Z. and R.M.L.S. collected the samples. E.F.O., V.A.S.-P., X.C. and E.A.M. generated the sequence data. E.F.O., M.G., V.A.S.-P., X.C. and E.A.M. performed the analyses. E.F.O., M.G., V.A.S.-P., F.T.B., D.O.M., A.A.G., G.R.C. and G.C.C. contributed to writing the manuscript. All authors commented and improved the final version of the manuscript.

Data accessibility

DNA sequences deposited in GenBank (Accession nos KT844957–KT845861 and KT886989–KT886992; see Table S1, Supporting information). Voucher number, geographic and haplotypic data of individuals (Tables S1 and S6, Supporting information). Primers (Table S2, Supporting information). Aligned sequences for all individuals and loci, tree files resulting from this study and input files (STRUCTURE, GENELAND and IMA2) are archived on Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.44372).

Supporting information

Additional supporting information may be found in the online version of this article.

Appendix S1 Taxonomic background and implications, and Fossil Record information.

Table S1 Samples of *Cnemidophorus* used in present study (398 samples + 1 outgroup).

Table S2 Primers used for amplification and sequencing of the five loci used in this study.

Table S3 Analyses performed in BEAST program.

Table S4 Species used in estimating the 12S substitution rate for Teiidae.

Table S5 Parameters and prior distributions used to test alternative scenarios for diversification of the Northeast and Southwest lineages under ABC approach.

Table S6 Samples of *Cnemidophorus* used in this study (subset of 137 samples).

Table S7 Variance percentages for components of analyses of molecular variance (AMOVA) performed with different genes in Southwest lineage considering current and paleo-course course of SFR.

Table S8 Variance percentages for components of analyses of molecular variance (AMOVA) performed with different genes in Northeast lineage considering current and paleo-course of SFR.

Table S9 Tests of nested models in IMA2 for Northeast and Southwest lineages.

Table S10 Population parameter estimates in IMA2 for Northeast and Southwest lineages.

Table S11 SPEDESTEM species delimitation results.

Table S12 Models tested in the redundancy analysis (RDA).

Fig. S1 Haplotype network for 12S (a), RP40 (b), NKTR (c), R35 (d) and ATPSB (e) using median-joining method.

Fig. S2 Gene tree used in estimating of the 12S substitution rate for Teiidae and inferred by Bayesian inference in the program BEAST.

Fig. S3 STRUCTURE results showing (a) plot of the log-likelihood value (LnPr($X \mid K$)) vs. the number of potential populations (K), (b) plot of Evanno ΔK method to evaluate the most supported K based on rate of change of the likelihood distribution as a function of K, and (c) plot of ancestry estimates, which represent the estimated membership for K-inferred clusters.

Fig. S4 Gene trees for 12S (a), RP40 (b), ATPSB (c), R35 (d) and NKTR (e) inferred using Bayesian inference in the program BEAST.

Fig. S5 Distribution of Northeast (white circles) and Southwest (gray circles) lineages along Caatinga (Ca) and Cerrado (Ce) biomes, depicting São Francisco River (blue line).

Fig. S6 Phylogeographic reconstructions of Northeast lineage.

Fig. S7 Principal components analysis predictive plots, PC1 × PC2 (a) and PC1 × PC3 (b), for the prior predictive distributions of summary statistics for the five best models (see Table 4) compared using the Approximate Bayesian Computation approach.