biology letters

A coalescent framework for comparing alternative models of population structure with genetic data: evolution of Celebes toads

Ben J Evans, Jimmy A McGuire, Rafe M Brown, Noviar Andayani and Jatna Supriatna

Biol. Lett. 2008 **4**, 430-433
doi: 10.1098/rsbl.2008.0166

biology letters

Biol. Lett. (2008) **4**, 430–433
doi:10.1098/rsbl.2008.0166
Published online 29 May 2008

Population genetics

A coalescent framework for comparing alternative models of population structure with genetic data: evolution of Celebes toads

Ben J. Evans[1,*], Jimmy A. McGuire[2], Rafe M. Brown[3], Noviar Andayani[4] and Jatna Supriatna[4]

[1]*Department of Biology, Life Sciences Building Room 328, McMaster University, 1280 Main Street West, Hamilton, Ont., Canada L8S 4K1*
[2]*Department of Integrative Biology, 3101 Valley Life Sciences Building, University of California, Berkeley, CA 94720-3160, USA*
[3]*Department of Ecology and Evolutionary Biology, Dyche Hall, University of Kansas, 1345 Jayhawk Boulevard, Lawrence, KS 66045-7561, USA*
[4]*Departmen Biologi, FMIPA, University of Indonesia, Depok, Java 16424, Indonesia*
**Author for correspondence (evansb@mcmaster.ca).*

Isolation of populations eventually leads to divergence by genetic drift, but if connectivity varies over time, its impact on diversification may be difficult to discern. Even when the habitat patches of multiple species overlap, differences in their demographic parameters, molecular evolution and stochastic events contribute to differences in the magnitude and distribution of their genetic variation. The Indonesian island of Sulawesi, for example, harbours a suite of endemic species whose intraspecific differentiation or interspecific divergence may have been catalysed by habitat fragmentation. To further test this hypothesis, we have performed phylogenetic and coalescent-based analyses on molecular variation in mitochondrial and nuclear DNA of the Celebes toad (*Bufo celebensis*). Results support a role for habitat fragmentation that led to a population structure in these toads that closely matches distributions of Sulawesi macaque monkeys. Habitat fragmentation, therefore, may also have affected other groups on this island.

Keywords: coalescence; rejection sampling; demography

1. INTRODUCTION

Fragmentation of populations can cause divergence by genetic drift depending on their effective population sizes, mutation rate, the duration of isolation, and the level of gene flow between them. However, the role of fragmentation in causing population structure may be subtle after what we call 'cryptic fragmentation', in which a barrier to gene flow is either not apparent or no longer present. Cryptic fragmentation may have affected, for example, fauna of the island of Cuba (Glor *et al.* 2004) and the Baja California Peninsula (Leaché *et al.* 2007). In contrast to diversification caused by enduring barriers to dispersal, the genetic signature of cryptic fragmentation might fade over time if the margins of the subdivided populations move, or if recent gene flow causes populations to amalgamate.

Cryptic fragmentation has been proposed on the Indonesian island of Sulawesi based on similar patterns of diversity in multiple groups, including the Celebes toad (Evans *et al.* 2003*c*), monkeys (Evans *et al.* 2003*b*), fanged frogs (Evans *et al.* 2003*a*) and flying lizards (McGuire *et al.* 2007). However, the validity of a demographic model of isolation by distance (IBD) plus fragmentation as opposed to a model of exclusively IBD has been questioned (Bridle *et al.* 2004). To further explore this, we used a coalescent-based approach to compare the fit of models that approximate each of these demographic scenarios for the Celebes toad, *Bufo celebensis*.

2. MATERIAL AND METHODS

(a) *Molecular data, genealogies, networks and population subdivision*

New data were collected from mitochondrial DNA (mtDNA) and two nuclear loci (nDNA) from throughout the range of the Celebes toad, including sequences from up to 166 individual toads from up to 56 localities (figures 1*a* and 2*a*; see the electronic supplementary material). MtDNA sequences are from the 12S rDNA gene and nuclear sequences are from the recombination activation gene 1 (*RAG*) and intron 3 of the rhodopsin gene (*RHO*). A phylogeny was estimated from the mtDNA data under a doublet model for ribosomal genes using MrBayes v. 3.1.2 (Huelsenbeck & Ronquist 2001) with secondary structure inferred from a model for *Xenopus laevis* (Cannone *et al.* 2002). Networks were estimated from inferred alleles for the nuclear loci (see the electronic supplementary material).

(b) *Coalescent comparison of alternative demographic models*

For computational efficiency (Wilkins 2004), we used a lattice model to approximate an IBD model (IBD_L), and compared it with an alternative model that also has simultaneous fragmentation (IBD_L+F) at the sites of each macaque contact zone, except a displaced location of the margin in toads between the NW and WC areas of endemism (AOEs; figure 2*a*). The locations of macaque contact zones are well characterized (Evans *et al.* 2003*b* and references therein). Both models assume mutation–drift equilibrium, constant population size over time and symmetrical migration between connected neighbouring demes.

The IBD_L model has three parameters: the effective population size of the locus in each deme ($N_{e\text{-nDNA-deme}}$), the mutation rate per sequence (μ) and the fraction of subpopulation i in each generation that are migrants from subpopulation j (m_{ij}). The IBD_L+F model has an additional parameter (τ), which is the time in $4N_{e\text{-nDNA-deme}}$ generations from the present that fragmentation started simultaneously at all boundaries between AOEs (figure 2*a*).

Model likelihood was estimated using rejection sampling of coalescent simulations (Weiss & von Haeseler 1998) based on three summary statistics: the average pairwise nucleotide divergence per sequence (π), the number of segregating sites (S) and F_{ST} (table 1). π and S were calculated for simulated data using sample_stat (Hudson 2002) and for the observed data using DNAsp v. 4.10.9 (Rozas *et al.* 2003). F_{ST} was calculated using Perl scripts according to: $F_{ST}=(\pi_{between}-\pi_{within})/\pi_{between}$ where $\pi_{between}$ and π_{within} are the average number of pairwise differences between and within AOEs, respectively (Hudson *et al.* 1992). To avoid bias due to differences in sample size, the average π_{within} of each AOE was used in this calculation. F_{ST} was transformed according to $F_{ST}/(1-F_{ST})$ (Rousset 1997) before rejection sampling. Model likelihood for each locus was estimated as the proportion of 100 000 simulations whose summary statistics were $\pm 10\%$ of the observed values for all three statistics; multilocus likelihoods are the product of the likelihood of each locus. Simulations were performed with the program ms (Hudson 2002) under an approximation of the finite sites model by using the total mutation rate for each sequence instead of the mutation rate per site, which is appropriate when $4N_e\mu$/site is small. Scaling factors were applied to $N_{e\text{-nDNA-deme}}$ and μ as a coarse measure to accommodate

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsbl.2008.0166 or via http://journals.royalsociety.org.

Received 23 March 2008
Accepted 14 May 2008

Table 1. Polymorphism statistics for sequence data from the Celebes toad for mtDNA, *RAG* and *RHO*. (The number of unique haplotypes (no.) on Sulawesi (all) and each area of endemism (AOE) are indicated, with labelling of AOEs following figure 1. Other statistics (π, S and F_{ST}) are discussed in the text.)

	mtDNA	*RAG*	*RHO*
base pairs	727	961	338
S	80	21	10
π per site	0.03144	0.00254	0.00379
π per sequence	22.85688	2.44094	1.28102
F_{ST}	0.91460	0.46068	0.46469
no. all	32	22	15
no. NE AOE	5	0	0
no. NC AOE	0	2	1
no. NW AOE	1	2	1
no. CW AOE	11	3	2
no. CE AOE	1	1	0
no. SW AOE	9	4	2
no. SE AOE	7	6	3

Table 2. Scaling factors used in coalescent simulations for mtDNA, *RAG* and *RHO*. ($\pi_{between}$ is the average number of substitutions per site between the Sulawesi toads and *Bufo divergens*. The effective population size of mtDNA in each deme is scaled to 0.25 of the size of autosomal DNA (N_e scaling). The mutation rate scaling (μ scaling) is calculated by dividing the $\pi_{between}$ of each locus by the $\pi_{between}$ of *RAG1*. The finite sites scaling factor (FS) is obtained by dividing the number of base pairs of data at the locus by the number of base pairs of data collected for *RAG1*. The product of these scaling factors is the composite θ scaling factor. The time scaling factor is the reciprocal of the N_e scaling factor.)

	mtDNA	*RAG*	*RHO*
base pairs	727	961	338
$\pi_{between}$	0.087	0.015	0.041
N_e scaling	0.25	1.00	1.00
μ scaling	5.929	1.000	2.797
FS	0.757	1.000	0.352
θ scaling	1.1	1.0	1.0
time scaling	4	1	1

differences in uniparental or biparental inheritance, ploidy, mutation rate per nucleotide, and the number of nucleotides sequenced per locus (table 2).

Nested evolutionary models can be compared by assuming that twice the difference of the natural logarithm of their likelihoods of each model (2δ) follows a χ^2 distribution with degrees of freedom equal to the number of free parameters (Goldman 1993). However, because the IBD_L model is equivalent to the IBD_L+F model with τ equal to a boundary of zero, in our case this distribution can be expressed as a 50 : 50 mixture of χ^2_0 and χ^2_1 distributions (Self & Liang 1987). This is equal to half of the probability of a χ^2 distribution with degrees of freedom equal to 1 (Goldman & Whelan 2000).

3. RESULTS

The likelihood of the IBD_L+F model is significantly higher than the likelihood of the IBD_L model when considering all loci ($p=0.0416$), only nuclear loci ($p=0.0269$) or only mtDNA ($p=0.0195$; figure 2; see the electronic supplementary material). When all loci are considered, the 95% confidence interval (CI) of the IBD_L+F model dips below significance. However, this likelihood is compromised by divergent mtDNA lineages in three demes that could be derived from recent gene flow across contact zones (figure 2; see the electronic supplementary material). Support for fragmentation is also apparent in the phylogeography of mtDNA (figure 1) and nuclear DNA (see the electronic supplementary material). Population structure between AOEs is significant at each locus (see the electronic supplementary material). Each AOE has an endemic clade of toad mtDNA, and most have private nuclear alleles in both nuclear loci (table 1).

These new findings support and extend previous work. Notably, the statistical framework reported here incorporates stochasticity of genealogical coalescence. New samples in this study (166 versus 29 samples in Evans *et al.* 2003*c*) demonstrate that contact zones between toad mtDNA clades closely match the locations of macaque hybrid zones (figure 1). In units of $4N_{e\text{-nDNA-deme}}$ generations, the maximum-likelihood time of simultaneous fragmentation is 1.0 when all loci are considered or 2.0 when only nDNA is considered. Using an independent estimate of μ, this time is estimated to be Late Pleistocene (see the electronic supplementary material).

4. DISCUSSION

Using data from up to three loci, rejection sampling of coalescent simulations based on three summary statistics rejects the IBD_L model in favour of the IBD_L+F model. While extensions of the rejection sampling approach, such as approximate Bayesian computation, may improve the accuracy of parameter estimates (Beaumont *et al.* 2002), overall this illustrates, under the assumptions of each model, that population structure in Celebes toads cannot be attributed exclusively to IBD.

Because models simplify real demographic histories, caveats exist in their interpretation. These results do not demonstrate, for example, that the IBD_L+F model is better than other models that we did not consider. Significant improvement over the IBD_L model could also be recovered if not all of these AOEs arose by habitat fragmentation, or if toad AOEs do not precisely match macaque AOEs on a fine geographical scale. However, it is also plausible that more complex scenarios involving multi-taxon fragmentation at the sites of monkey contact zones are significantly better than the ones tested here. These models could include non-simultaneous divergence at different contact zones, fragmentation at some contact zones followed by recent gene flow, and non-simultaneous fragmentation of different sympatric taxa.

Habitat fragmentation of Celebes toads in the same or similar locations as multiple macaque hybrid zones could be a consequence of physical barriers, such as marine inundation in low-lying areas between the SW and WC AOE and between the NW and NC AOE (figure 1), and/or multi-taxon adaptation to substantial ecological transitions between substrate, vegetation and climatic zones (Evans *et al.* 2003*c* and

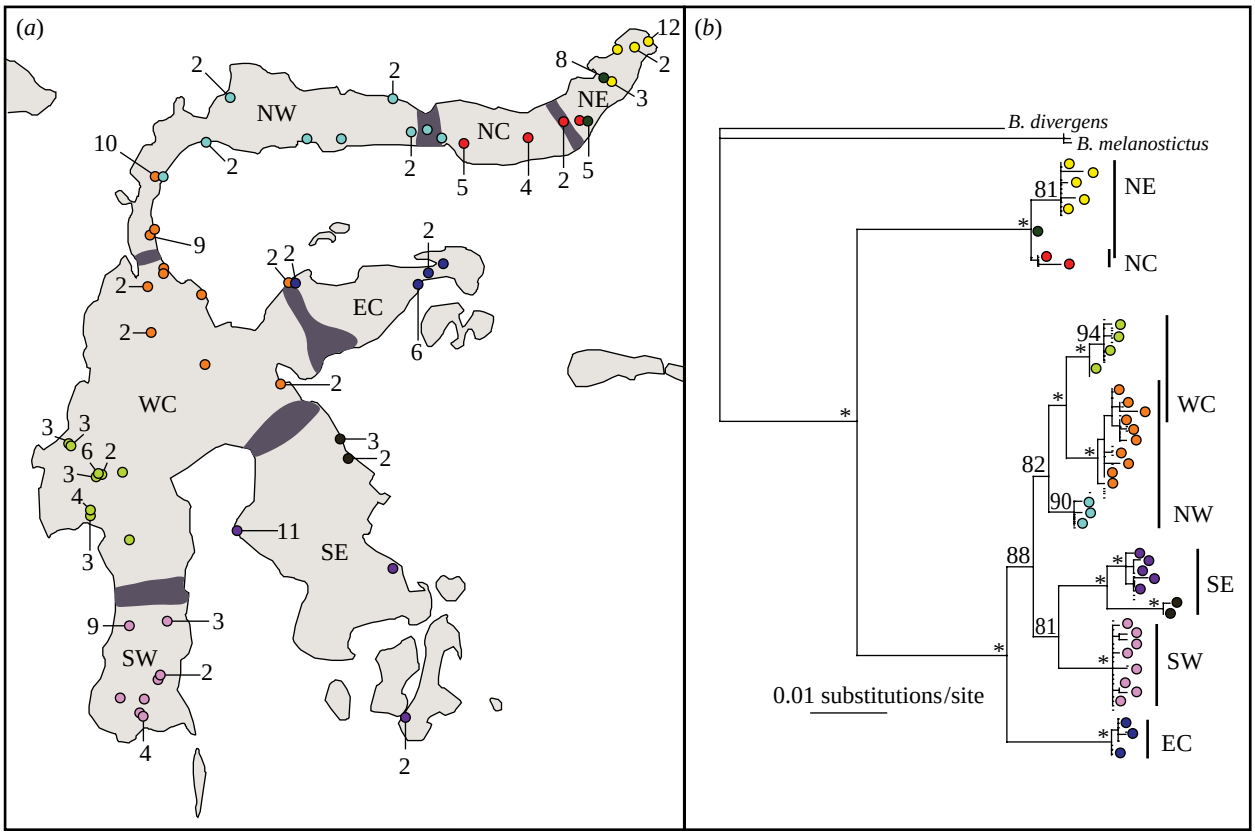

Figure 1. Samples and mtDNA phylogeny of the Celebes toad. (*a*) Sampling localities; the number of individuals sampled at each locality is labelled if more than one. Shaded areas indicate the locations of macaque contact zones. AOEs are labelled: northeast (NE), north-central (NC), northwest (NW), west-central (WC), east-central (EC), southwest (SW) and southeast (SE). (*b*) MtDNA phylogeny of the Celebes toad. Posterior probabilities are above branches; asterisks indicate those above 95%.

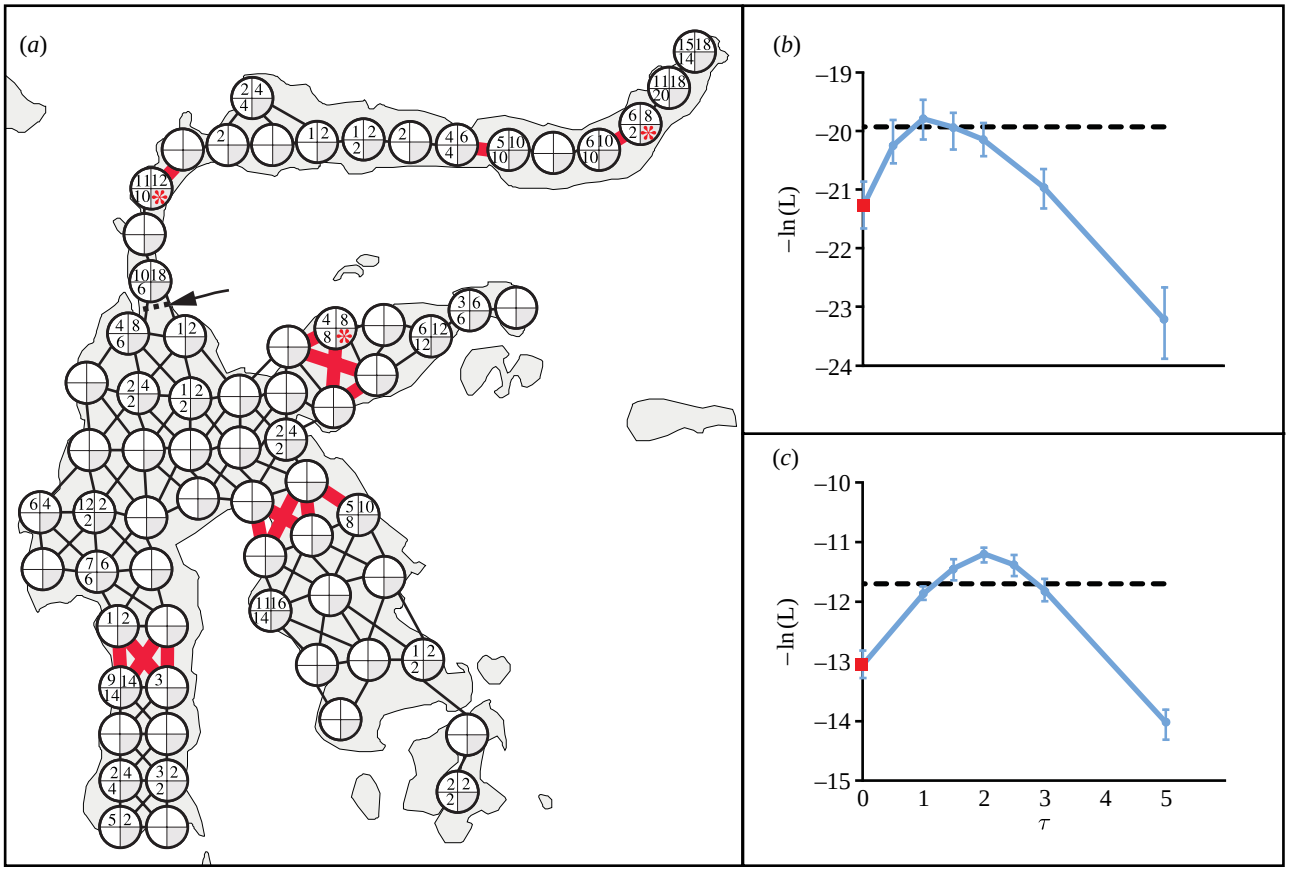

Figure 2. (*Caption opposite.*)

references therein). These processes probably affected many other species and therefore provide scientific rationale for geographically dispersed conservation areas on Sulawesi.

We thank P. Andolfatto, J. Bridle, R. Butlin, B. Golding, N. Goldman, R. Hudson, J. Wilkins and S. Wright and members of the McGuire lab for their advice and comments. This research was supported by the National Science Foundation, Canadian Foundation for Innovation, National Science and Engineering Research Council and McMaster University.

Beaumont, M. A., Zhang, W. & Balding, D. J. 2002 Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035.

Bridle, J. R., Pedro, P. M. & Butlin, R. K. 2004 Habitat fragmentation and biodiversity: testing for the evolutionary effects of refugia. *Evolution* **58**, 1394–1396. (doi:10.1111/j.0014-3820.2004.tb01718.x)

Cannone, J. J. *et al.* 2002 The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinform.* **3**, 2. (doi:10.1186/1471-2105-3-2)

Evans, B. J., Brown, R. M., McGuire, J. A., Supriatna, J., Andayani, N., Diesmos, A., Iskandar, D. T., Melnick, D. J. & Cannatella, D. C. 2003*a* Phylogenetics of fanged frogs (Anura; Ranidae; *Limnonectes*): testing biogeographical hypotheses at the Asian–Australian faunal zone interface. *Syst. Biol.* **52**, 794–819. (doi:10.1080/10635150390251063)

Evans, B. J., Supriatna, J., Andayani, N. & Melnick, D. J. 2003*b* Diversification of Sulawesi macaque monkeys: decoupled evolution of mitochondrial and autosomal DNA. *Evolution* **57**, 1931–1946. (doi:10.1111/j.0014-3820.2003.tb00599.x)

Evans, B. J., Supriatna, J., Andayani, N., Setiadi, M. I., Cannatella, D. C. & Melnick, D. J. 2003*c* Monkeys and toads define areas of endemism on Sulawesi. *Evolution* **57**, 1436–1443. (doi:10.1111/j.0014-3820.2003.tb00350.x)

Glor, R. E., Gifford, M. E., Larson, A., Losos, J. B., Schettino, L. R., Lara, A. R. C. & Jackman, T. R. 2004 Partial island submergence and speciation in an adaptive radiation: a multilocus analysis of the Cuban green anoles. *Proc. R. Soc. B* **271**, 2257–2265. (doi:10.1098/rspb.2004.2819)

Goldman, N. 1993 Statistical tests of models of DNA substitution. *J. Mol. Evol.* **36**, 182–198. (doi:10.1007/BF00166252)

Goldman, N. & Whelan, S. 2000 Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. *Mol. Biol. Evol.* **17**, 975–978.

Hudson, R. R. 2002 Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338. (doi:10.1093/bioinformatics/18.2.337)

Hudson, R. R., Slatkin, M. & Maddison, W. P. 1992 Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589.

Huelsenbeck, J. P. & Ronquist, F. 2001 MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755. (doi:10.1093/bioinformatics/17.8.754)

Leaché, A., Crews, S. C. & Hickerson, M. J. 2007 Two waves of diversification in mammals and reptiles of Baja California revealed by hierarchical Bayesian analysis. *Biol. Lett.* **3**, 646–650. (doi:10.1098/rsbl.2007.0368)

McGuire, J. A., Brown, R. M., Mumpuni, Riyanto, A. & Andayani, N. 2007 The flying lizards of the *Draco lineatus* group (Squamata: Iguania: Agamidae): a taxonomic revision with descriptions of two new species. *Herpetol. Monogr.* **21**, 179–212. (doi:10.1655/07-012.1)

Rousset, F. 1997 Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. *Genetics* **145**, 1219–1228.

Rozas, J., Sanchez-DelBarrio, J. C., Messegyer, X. & Rozas, R. 2003 DNAsp, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**, 2496–2497. (doi:10.1093/bioinformatics/btg359)

Self, S. G. & Liang, K.-Y. 1987 Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.* **82**, 605–610. (doi:10.2307/2289471)

Weiss, G. & von Haeseler, A. 1998 Inference of population history using a likelihood approach. *Genetics* **149**, 1539–1546.

Wilkins, J. F. 2004 A separation-of-timescales approach to the coalescent in a continuous population. *Genetics* **168**, 2227–2244. (doi:10.1534/genetics.103.022830)

Figure 2. (*Opposite.*) Demographic models and likelihoods. (*a*) The IBD_L and IBD_L+F models consist of 68 demes, represented by circles, some of which exchange migrants, represented by lines between circles. In the IBD_L+F model, no migration occurs between demes connected by red lines from time τ until the present. The number of haplotypes or alleles sampled is indicated inside each deme for mitochondria (top left), *RAG* (top right) and *RHO* (bottom left). An arrow and dashed line indicates the location of a monkey contact zone that is displaced from the margin of toad mtDNA clades. (*b*) Likelihood of the IBD_L+F model as a function of duration of fragmentation (τ) in units of $4N_{e\text{-nDNA-deme}}$ generations, based on mtDNA and nDNA loci, and (*c*) likelihood of the IBD_L+F model based on nDNA loci only. In (*b*,*c*), the likelihood of the IBD_L model indicated by red squares is equal to that at $\tau=0$. Likelihoods above the dashed line indicate significant improvement of the IBD_L+F model (i.e. that $2\delta>2.7065$). Parameters were calculated to the nearest 0.01 units for Θ ($=4N_e\mu$), the nearest 0.2 for $M_{ij}(=4N_{e\text{-nDNA-deme}}m_{ij})$ and the nearest $2N_{e\text{-nDNA-deme}}$ generations for τ. 95% CIs, indicated as vertical bars, were obtained with 40 replicate simulations. Demes discussed in the electronic supplementary material contain asterisks.