Ecography

ECOG-04937
Marquitti, F. M. D., Fernandes, L. D. and de Aguiar, M. A. M. 2020. Allopatry increases the balance of phylogenetic trees during radiation under neutral speciation. – Ecography doi: 10.1111/ecog.04937

Supplementary material

In this *Supplementary Material*, we present the Individual Based Model of speciation we used, as initially proposed by Costa et al. (2019), with similar analyses of phylogenetic tree metrics. We included explanation about how the phylogenetic trees were built from the Individual based model and also details about phylogenetic tree metrics in order to make the present study easier to understand on its own.

Phylogenetic trees

In this manuscript we opted to use the MRCAT algorithm (most recent common ancestor time), which focuses on genealogies (Costa et al., 2018). It consists in maintaining the parenthood of each individual, registering the time to the most recent common ancestor in the MRCAT matrix (T), which is

$$T_{t+1}(i,j) = \min_{\{k,l\}}\{T_t(P_k(i), P_l(j))\} + 1 \tag{1}$$

with $k, l = \{1, 2\}$, $T_0(i,j) = 1$ and $T_t(i,i) = 0$; $P_1(i)$ and $P_2(i)$ represent the parents of individual i. In words, we evaluate the most recent common ancestor of individuals i and j by considering the ancestry of all parental combinations, taking the smallest time and adding one for the present generation. The MRCAT matrix is symmetric, i.e., $T_{ij} = T_{ji}$. The phylogeny is obtained from the full genealogy by selecting one individual per species at each moment in time. We used the UPGMA method to recalculate the times every time a group was formed (Felsenstein, 2004). The choice of the individual for constructing the phylogenetic tree with MRCAT can matter, but previous results show that it is not relevant for the parameters used in our simulations. Although MRCAT algorithm is only an approximation of the speciation time, since speciation events might happen several generations later than the common ancestor indicates, it is still a good proxy when comparing it to the phylogenetic tree draw based on all Sequential Speciation and Extinction Events (SSEE method) (Costa et al., 2018).

The Sackin index

Tree imbalance is one of the most common phylogenetic structural patterns and measures asymmetries between the numbers of species in each side of the tree's branches. A widely used metric for tree imbalance is the Sackin index I (Sackin, 1972; Blum and François, 2005; Frost and Volz, 2013; Dearlove and Frost, 2015), defined as:

$$I(R) = \sum_{j=1}^{R} d_j \tag{2}$$

where d_j represents the number of nodes that separate the root from each leaf (species) j (with the root included), and R is the number of species. The maximum value for tree imbalance is given for completely asymmetric trees (branching happens always on the same side for each branching event), which is related to the maximum Sackin index of $\max(I) = (R-1)(R+2)/2$. Average and variance of the Sackin index as a function of the number of leaves R can be obtained for the Yule model, a stochastic model with constant branching probability per branch. These quantities are given by (Cardona et al., 2013),

$$E[I(R)] = 2R(h_R - 1) \approx 2R \log R \tag{3}$$

$$\sigma_R^2 = 7R^2 - 4R^2 h_R^{(2)} - 2Rh_R - R \approx \left(7 - \frac{2\pi^2}{3}\right) R^2 \approx 0.42R^2 \tag{4}$$

where

$$h_R = \sum_{k=1}^{R} \frac{1}{k} \tag{5}$$

$$h_R^{(2)} = \sum_{k=1}^{R} \frac{1}{k^2} \tag{6}$$

are the harmonic numbers of first and second kind.

The Sackin index has a known dependence on the number of leaves, making it unsuitable for comparing trees with different number of species. To make this comparison possible, we work with the normalized Sackin index,

$$I_n(R) = \frac{I(R) - E[I(R)]}{\sqrt{\sigma_R^2}}. \tag{7}$$

Although $I_n(R)$ would be close zero for trees generated with the Yule model, independent of the species richness R, different geographic modes of speciation may introduce important deviations from the behaviour of the Yule model. For the case of parapatric speciation, sequential branching of species from the spatially distributed population generates highly asymmetric trees, with the Sackin index $I(R)$ approximated by $\max(I) \sim R^2/2$. In this scenario, we have for the normalized Sackin index:

$$I_n(R) \approx \frac{R^2/2 - 2R\log R}{0.65R} \approx 0.74R. \tag{8}$$

Thus, for spatially distributed populations under parapatric speciation, I_n itself grows linearly with R, producing unrealistic values of tree imbalance for large species richness.

Gamma statistic (γ) and α-value

Another important structural pattern in phylogenies is related to the distribution of branch lengths, or times between two consecutive branching (speciation) events. A common metric used to evaluate this distribution is the γ-statistic, which is defined as (Pybus and Harvey, 2000):

$$\gamma = \frac{1}{D}\left[\frac{1}{R-2}\sum_{k=2}^{R-1}\Theta(k) - \Theta(R)/2\right] \tag{9}$$

where

$$\Theta(k) = \sum_{j=2}^{k} j g_j; \tag{10}$$

$$D = \Theta(R)/\sqrt{12(R-2)} \tag{11}$$

where g_k is the time interval between speciation events k and $k-1$, and $k=1$ corresponds to the root of the tree. The gamma-statistic is constructed in reference to a continuous time process in which all species bifurcate with fixed rate b, for which $\langle g_k\rangle = 1/bk$ and, therefore, $\langle\gamma\rangle = 0$ with $\langle\gamma^2\rangle = 1$ (see Costa et al. (2019)).

One of the problems of using the γ-statistic is that, like the Sackin index, it depends on the number of species in a given phylogeny (McPeek, 2008; Phillimore and Price, 2008), making it unsuitable for comparison of trees with different sizes. To tackle this problem, we used the α-value, defined in (Costa et al., 2019) by the formal relation

$$g_k(\alpha) = \frac{1}{bk^{\alpha}}. \tag{12}$$

The gamma-statistic corresponding to this sequence of speciation events can be computed with Eq. (9) to provide an index $\gamma = \gamma(R, \alpha)$, which can be numerically inverted to give $\alpha = \alpha(\gamma, R)$. The α-value measures the average acceleration of speciation along the tree. Therefore, positive values of α correspond to tippy trees, where speciation events accumulating close to the leaves, whereas negative values indicate steammy trees, with most speciation occurring close to the root.

Periodic and Non periodic boundary conditions

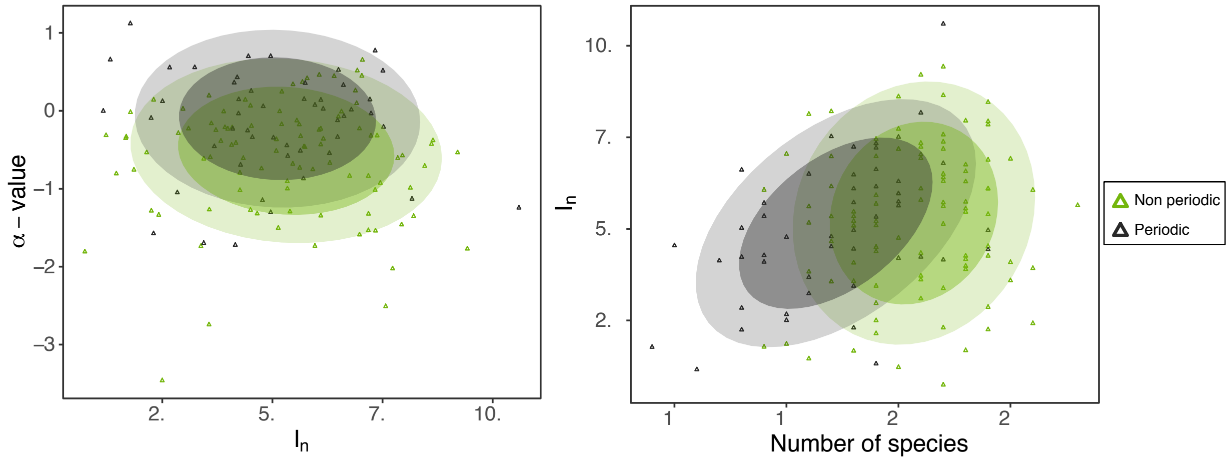

Figure A1: Populations of M=1,000 individuals under different boundary conditions. Green dots represent populations that evolved under non-periodic boundary conditions. Black dots represent populations that evolved under a periodic boundary conditions. The ellipses represent 0.63 (inner) and 0.87 (outer) of the variation in both axis.

Connectivity between demes

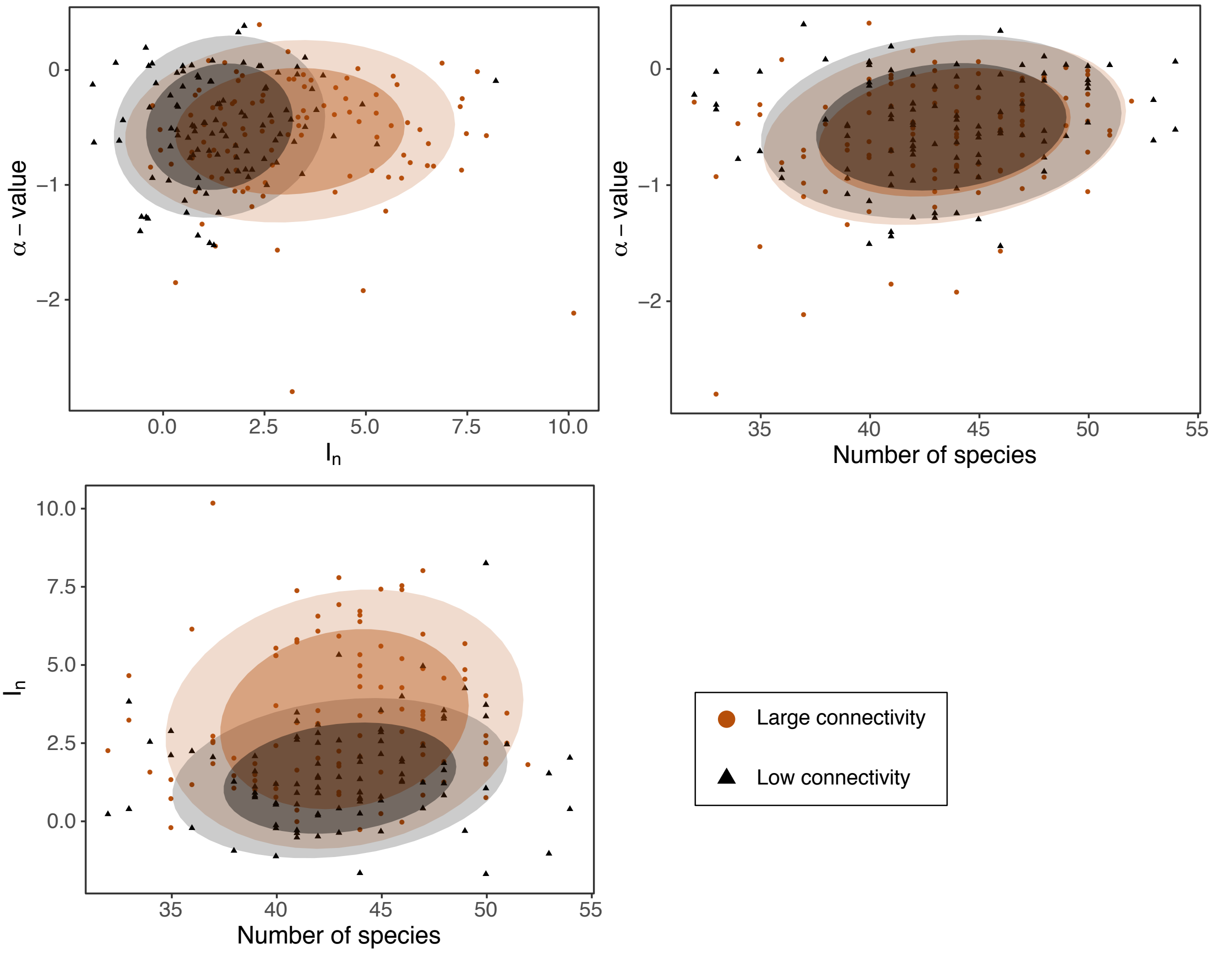

Figure A2: Populations of $M = 2,000$ individuals in a four-demes configuration with different degrees of connectivity between demes. Orange dots represent populations that evolved under large connectivity between demes (the gap is 1.9 of diameter size of each individual neighborhood). Black dots represent populations that evolved under low connectivity between demes (the gap is 0.9 of diameter size of each individual neighborhood). The ellipses represent 0.63 (inner) and 0.87 (outer) of the variation in both axis.

Equilibration time

The equilibration time used in our simulations were estimated based on 50 runs of each population size (M) and number of demes for a long time. The table A1 below are the values used on for the 100 runs simulated for each scenario:

Table A1: Equilibration time for each scenario of population size (M)), geographic structure (number of demes), and average of total number of species ($\langle N \rangle$) formed in the simulations.

Number of demes	Number of individuals	Equilbration time (generations)	$\langle N \rangle$
1	250	550	5.54
1	500	725	10.66
1	1,000	900	20.67
1	2,000	800	40.44
1	3,000	700	60.32
1	4,000	700	79.76
2	500	700	11.47
2	1,000	1,000	21.12
2	2,000	900	41.28
2	3,000	900	61.31
2	4,000	900	82.62
3	1,000	1,050	21.21
3	2,000	1,000	41.80
3	3,000	900	61.41
3	4,000	800	83.43
4	1,000	1,100	21.80
4	2,000	1,050	43.18
4	3,000	950	62.07
4	4,000	900	84.24

Number of species x Number of individuals

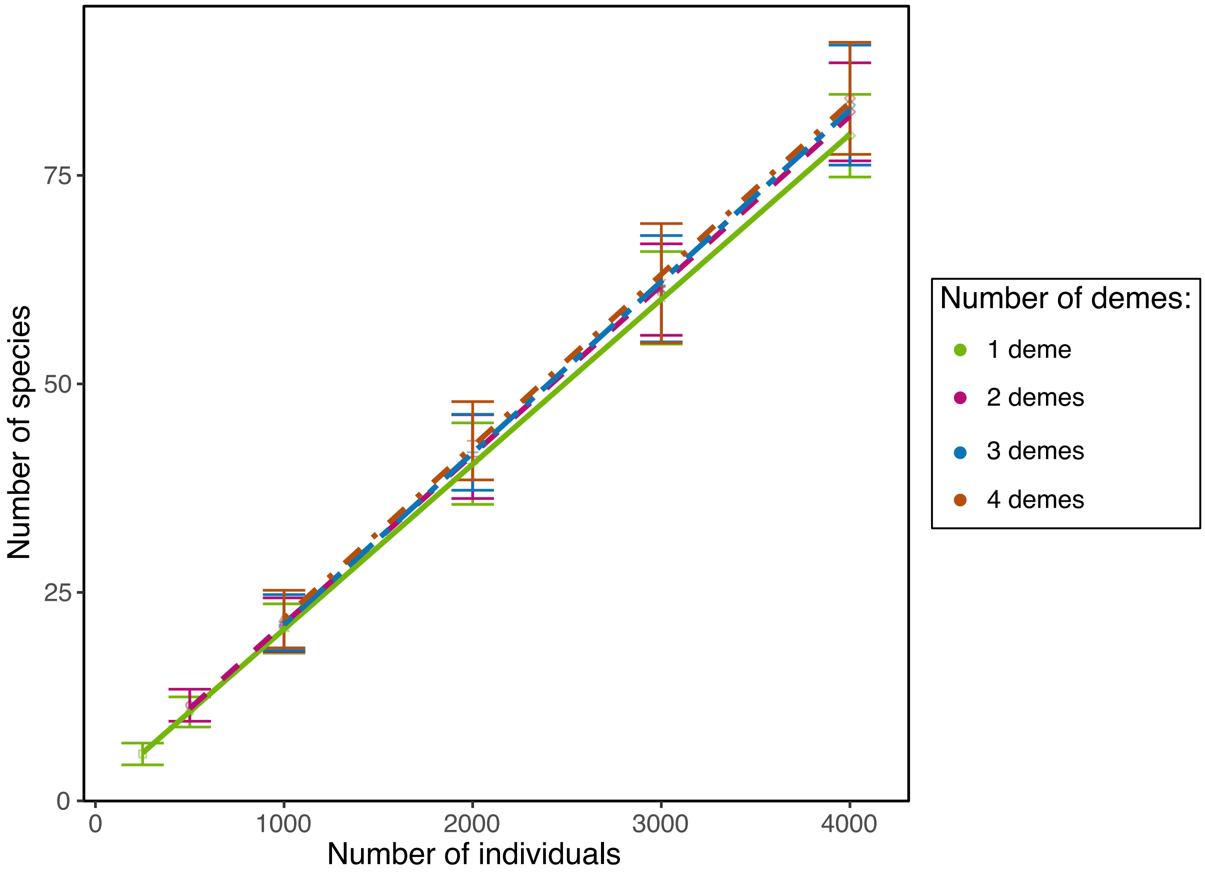

Figure A3: Number of species formed under differing total number of individuals. Species richness depends linearly on the number of individuals and it is not affected by the number of demes. The number of species here was measured at equilibration time. Check Table A1 for the equilibration time of each case.

Statistical tests for I_n and α-value

Table A2: Anova test for I_n and α-value with number of demes as treatment, according to number of individuals (only for $M = \{1000, 2000, 3000, 4000\}$).

Number of individuals	$\mathbf{I_n}$	**α-value**

1,000	F= 145.5
	df=3
	p¡0.0001***
		F= 0.515
		df=3
		p=0.672

2,000	F=220.4
	df=3
	p¡0.0001***
		F= 5.326
		df=3
		p¡0.01**

3,000	F= 287.3
	df=3
	p¡0.0001***
		F= 8.739
		df=3
		p¡0.0001***

4,000	F=277.1
	df=3
	p¡0.0001***
		F= 12.6
		df=3
		p¡0.0001***

Table A3: Post hoc Tukey's test for I_n and α-value. For α-value, Tukey's test was computed only for $M = \{2000, 3000, 4000\}$ because of the Anova test results. Here, $diff$ is the difference in the observed mean values and p is the p-value after adjustment for the multiple comparisons.

Number of individuals	**Demes comparisons**	$\mathbf{I_n}$		α**-value**	
		diff	*p*	*diff*	*p*
1,000	1-2	3.11	¡0.0001		
	1-3	3.99	¡0.0001		
	1-4	4.95	¡0.0001		
	2-3	0.88	¡0.001		
	2-4	1.84	¡0.0001		
	3-4	0.96	¡0.0001		
2,000	1-2	3.90	¡0.0001	0.14	0.14
	1-3	5.98	¡0.0001	0.25	¡0.0001
	1-4	8.26	¡0.0001	0.19	¡0.05
	2-3	2.08	¡0.0001	0.11	0.34
	2-4	4.36	¡0.0001	0.05	0.86
	3-4	2.29	¡0.0001	0.06	0.81
3,000	1-2	6.37	¡0.0001	0.16	¡0.01
	1-3	8.67	¡0.0001	0.18	¡0.01
	1-4	11.05	¡0.0001	0.25	¡0.0001
	2-3	2.30	¡0.0001	0.02	0.98
	2-4	4.68	¡0.0001	0.09	0.31
	3-4	2.38	¡0.0001	0.07	0.53
4,000	1-2	5.78	¡0.0001	0.063	0.41
	1-3	8.77	¡0.0001	0.19	¡0.0001
	1-4	11.61	¡0.0001	0.21	¡0.0001
	2-3	3.00	¡0.0001	0.13	¡0.01
	2-4	5.84	¡0.0001	0.15	¡0.01
	3-4	2.84	¡0.0001	0.02	0.95

Table A4: Linear regressions of the normalized Sackin index (I_n) and α-value by number of species for all the four geographic configurations.

Number of demes	$\mathbf{I_n}$			$\alpha -$ **value**		
	slope	R^2	p	slope	R^2	p
1	0.21	0.79	¡0.0001***	0.009	0.06	¡0.0001***
2	0.13	0.67	¡0.0001***	0.008	0.10	¡0.0001***
3	0.09	0.42	¡0.0001***	0.006	0.07	¡0.0001***
4	0.07	0.46	¡0.0001***	0.004	0.03	¡0.001**

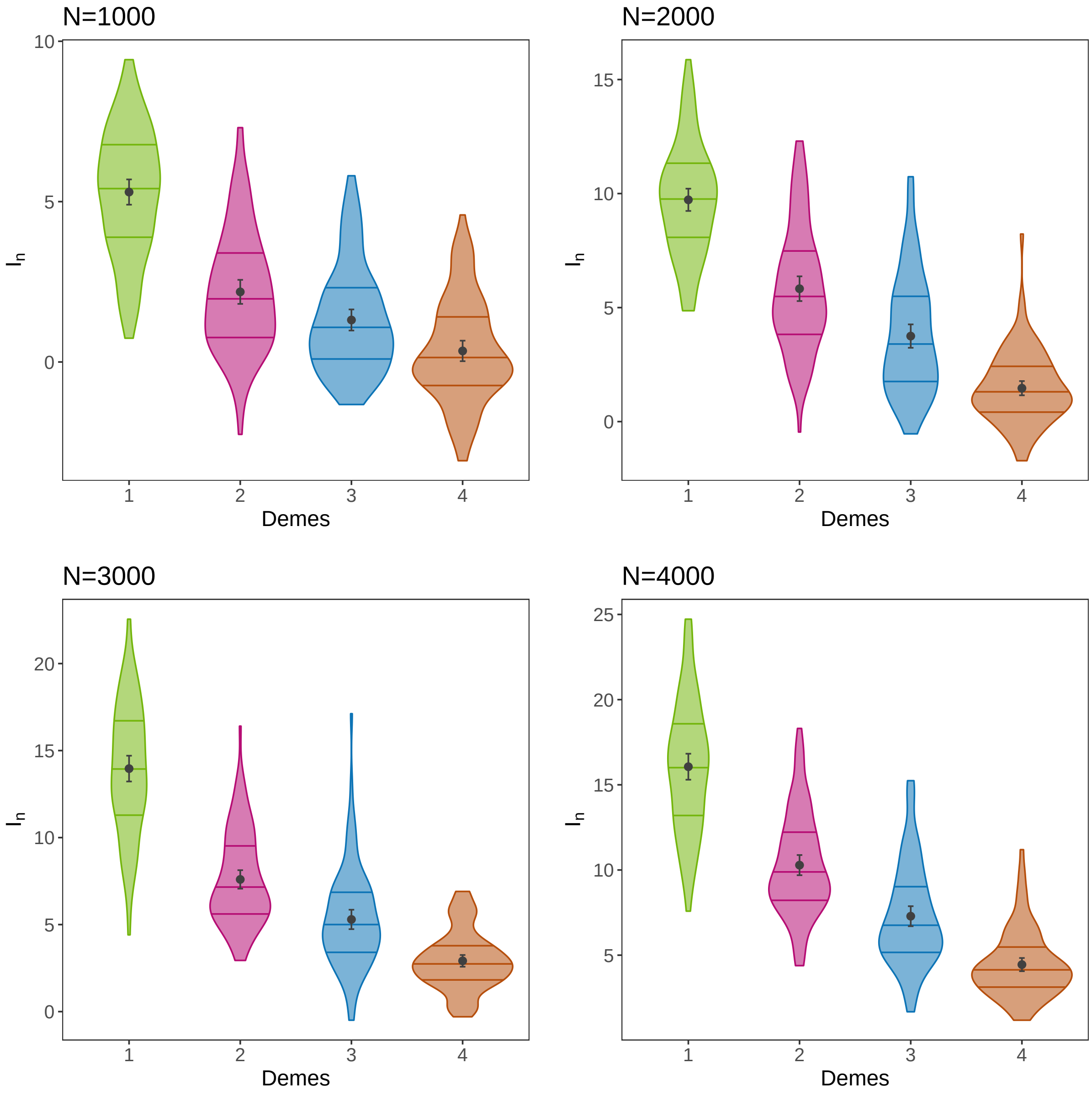

Figure A4: Plots of Normalized Sackin index (I_n) by the number of demes. The violin plots represent the distribution of runs according to I_n. The three colored lines inside the violin represent 25%, 50% and 75% quartiles. The dark point represents the average value of I_n with the respective confidence interval (error bar of 95%). As analysed by the Anova test, for the I_n the mean values are significantly different.

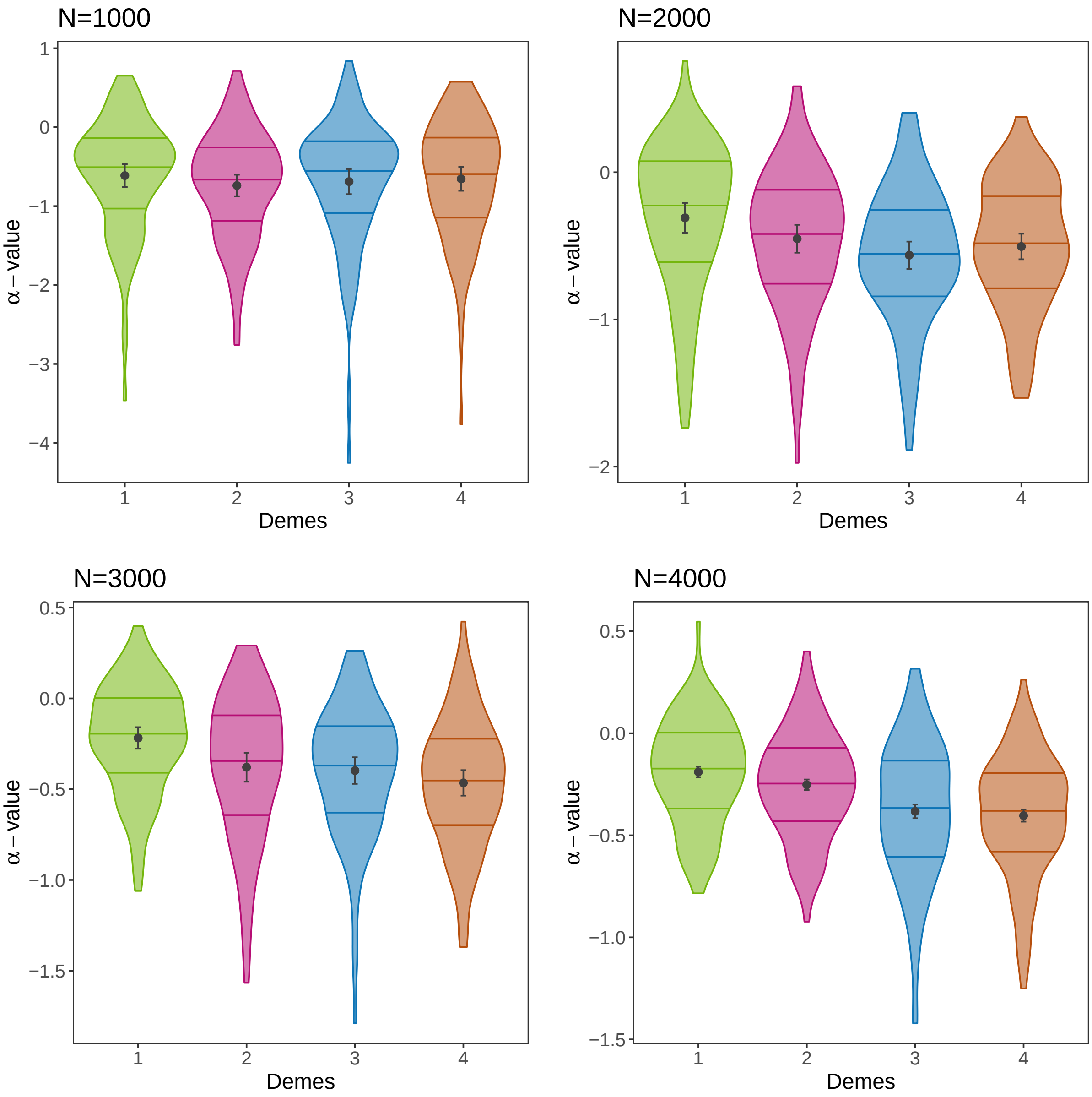

Figure A5: Plots of α-value by the number of demes. The violin plots represent the distribution of runs according to α-value. The three colored lines inside the violin represent 25%, 50% and 75% quartiles. The dark point represents the average value of α-value with the respective confidence interval (error bar of 95%). As analysed by the Anova test, for the α-value the mean values are significantly different for $M = \{2000, 3000, 4000\}$. The post-hoc Tukey's test shows: for M=2,000 only the 1 deme configuration differs from the 3 and 4 demes configurations; for M=3,000 the 1 deme configuration differs from the 2, 3 and 4 demes configurations; the 1 and 2 deme configurations differ from the 3 and 4 demes configurations.

Tree imbalance measured with β splitting method

Table A5: Anova test for β value measured at the equilibration time with number of demes as treatment. We performed 50 runs for each deme configuration only for $M = 2,000$.

Number of individuals	β **value**
2,000	F= 20.43.5
	df=196
	p¡0.0001***

Table A6: Post hoc Tukey's test for I_n and α-value. For α-value, Tukey's test was computed only for $M = 2,000$ individuals. Here, $diff$ is the difference in the observed mean values and p is adjusted p-value for the multiple comparisons.

Number of individuals	**Demes comparisons**	β **value**	
		$diff$	p
2,000	1-2	0.05	0.63
	1-3	0.19	¡0.001
	1-4	0.31	¡0.0001
	2-3	0.13	¡0.05
	2-4	0.26	¡0.0001
	3-4	0.12	¡0.05

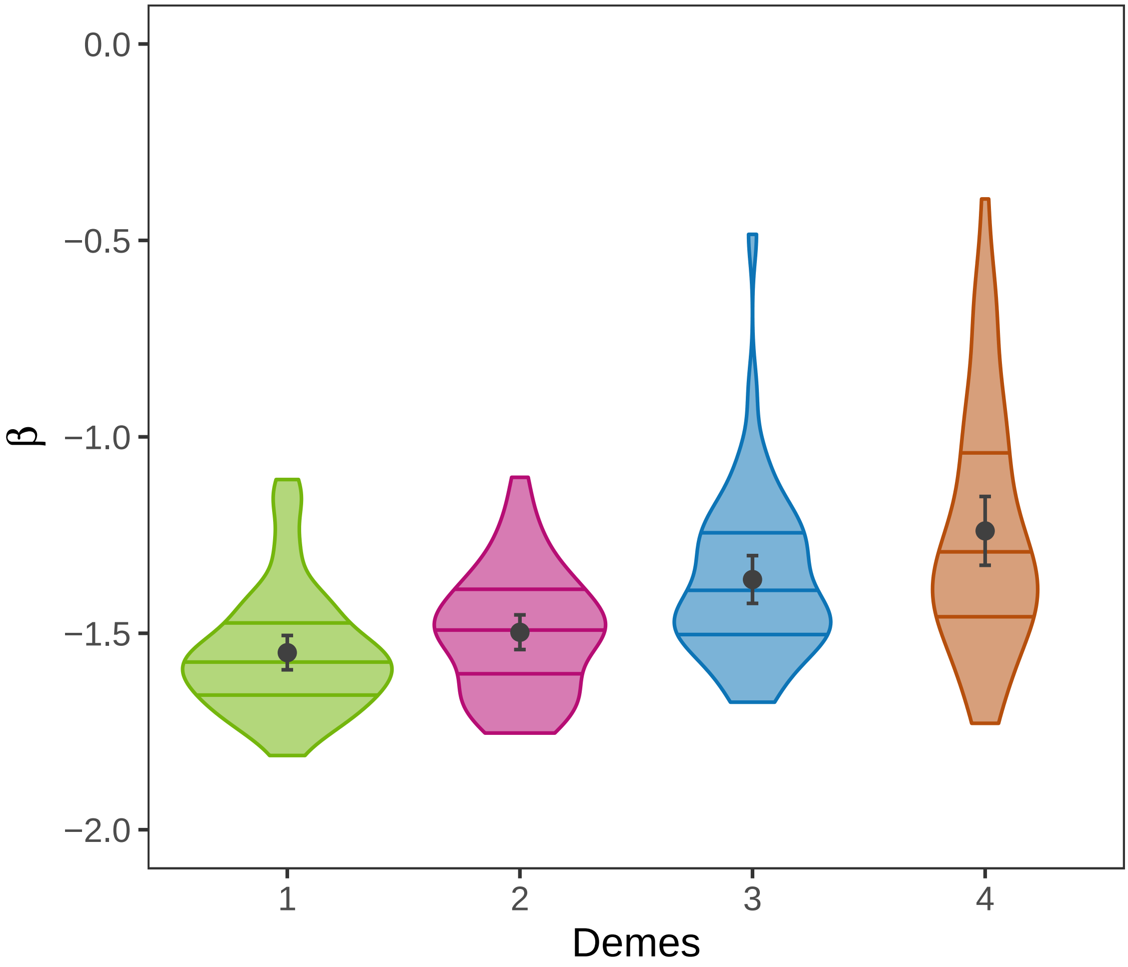

Figure A6: Balance of phylogenetic trees measured by β value. In total, we performed 50 runs for each deme configuration for $M = 2,000$ individuals. $\beta = 0$ is a tree with the expected balance by the Yule model. Negative values of β represent unbalanced trees compared to Yule. Negative values of β represent more balanced trees than expected by the Yule model. The three colored lines inside the violin represent 25%, 50% and 75% quartiles. The dark point represents the average value of α-value with the respective confidence interval (error bar of 95%). The average values of β are significantly different, and the post hoc Tukey's test reveals that only the average β values for 1 and 2 demes configurations are not different.

Statistical tests for I_n and α-value for long term

We performed analysis on the last generation of the long term simulations to detect differences in number of species formed, normalized Sackin index and α-value metrics of the correspondent phylogenetic trees under the different geographic scenarios.

Table A7: Anova test for I_n, α-value and number of species (N_{spp}) at time=10,000 generations with number of demes as treatment (only for $M = 2000$).

Number of individuals	$\mathbf{I_n}$	α-value	$\mathbf{N_{spp}}$

2,000	F= 14.11
	df=3
	p¡0.001***
		F= 1.41
		df=3
		p¿0.2***
			F= 6.38
			df=3
			p¡0.005**

Table A8: Post hoc Tukey's test for I_n and number of species (N_{spp}) computed only for $M = 2000$. Here, $diff$ is the difference in the observed mean values and p is the p-value after adjustment for the multiple comparisons.

Number of individuals	**Demes comparisons**	$\mathbf{I_n}$		$\mathbf{N_{spp}}$	
		$diff$	p	$diff$	p
2,000	1-2	1.79	¡0.005	5.4	¡0.05
	1-3	1.01	¡0.0005	6.6	¡0.05
	1-4	2.64	¡0.0005	7.4	¡0.01
	2-3	0.48	0.70	1.2	0.91
	2-4	0.84	0.26	2.0	0.71
	3-4	0.37	0.84	0.8	0.97

References

Blum, M. G. and François, O. 2005. On statistical tests of phylogenetic tree imbalance: the sackin and other indices revisited. – Mathematical Biosciences 195: 141–153.

Cardona, G. et al. 2013. Exact formulas for the variance of several balance indices under the yule model. – Journal of mathematical biology 67: 1833–1846.

Costa, C. L. et al. 2019. Signatures of microevolutionary processes in phylogenetic patterns. – Systematic Biology 68: 131–144.

Costa, C. L. N. et al. 2018. Registering the evolutionary history in individual-based models of speciation. – Physica A: Statistical Mechanics and its Applications 510: 1–14.

Dearlove, B. L. and Frost, S. D. 2015. Measuring asymmetry in time-stamped phylogenies. – PLoS Computational Biology 11: e1004312.

Felsenstein, J. 2004. Inferring phylogenies, vol. 2. – Sinauer associates Sunderland.

Frost, S. D. and Volz, E. M. 2013. Modelling tree shape and structure in viral phylodynamics. – Philosophical Transactions of the Royal Society B 368: 20120208.

McPeek, M. A. 2008. The ecological dynamics of clade diversification and community assembly. – The American Naturalist 172: E270–E284.

Phillimore, A. B. and Price, T. D. 2008. Density-dependent cladogenesis in birds. – PLoS Biology 6: e71.

Pybus, O. G. and Harvey, P. H. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. – Proceedings of the Royal Society B: Biological Sciences 267: 2267–2272.

Sackin, M. 1972. “good” and “bad” phenograms. – Systematic Biology 21: 225–226.