Supplementary Methods

Study area and sample processing

The study area includes the Atlantic rainforest of the Brazilian states of Espírito Santo (ES) and Rio de Janeiro (RJ) and the east coast of the states of São Paulo (SP), Paraná (PR), Santa Catarina (SC) and Rio Grande do Sul (RS) (Figure 1 and Table S1A). Sampling was performed in 13 conservation areas within and outside the Serra do Mar biological corridor, which ranges from Rio de Janeiro to Paraná States. The altitudes of the sampling localities ranged from sea level up to 1,000 m. From all of the species of land planarians sampled, we selected *C. bergi* (213 specimens) because it exhibits the widest distribution across the SAF and a relatively high abundance and due to its level of genetic variability [1]. For phylogenetic analysis, we used *Cephaloflexa* sp., *Choeradoplana iheringi* von Graff, 1899 [2] and *Ch. banga* Carbayo and E. M. Froehlich, 2012 [3] as outgroup because *Choeradoplana* is the sister genus of *Cephaloflexa* [4]. The animals were collected directly from the ground in natural conservation areas and surrounding localities. Each animal was photographed and cut into two pieces. One part was fixed in absolute ethanol for DNA extraction and the other in 10% formalin for identification after histological processing.

Morphological analysis

Thirty-seven specimens were analysed morphologically for identification in order to check that individuals throughout the whole geographic range sampled presented the diagnostic characters for the species. Tissue blocks from the cephalic region, the

pharynx and the copulatory organs were embedded in Paraplast, sectioned at a thickness of 7 µm and stained with Mallory ⁄ Cason trichrome stain [5]. As the copulatory apparatus is the main structure used for unequivocal identification, we reconstructed the copulatory apparatuses with a camera lucida attached to a light microscope. Voucher specimens have been deposited in the *Museu de Zoologia da Universidade de São Paulo* (MZUSP) (Table S1B).

Cephaloflexa bergi **distribution modelling**

We modelled the distribution of *Cephaloflexa bergi* for current and LGM (21 kya) climatic periods using the localities in this study (23 points) and the coordinates of the few localities (9 points) cited in the literature [6]. The models were generated by Maxent 3.3.3 [7,8] through 1000 replicates. We used as input data for the distribution models the same seven climatic variables as in other studies in the same region [9,10]; that is annual mean temperature, temperature seasonality, mean temperature of warmest quarter, mean temperature of coldest quarter, annual precipitation, precipitation of wettest quarter and precipitation of driest quarter; this information was downloaded from the WORLDCLIM 1.4 database [11].

For modelling we used the default values for factor of regularization parameter, and the convergence threshold (1 and 0.00001 respectively), setting the maximum number of iterations to 1000 using the bootstrap option. We evaluated the models using the area under the receiver operating characteristic curve (AUC; [12]) to ensure that model performance was satisfactory. Climatic retrojections were performed using the Community Climate System Model (CCSM) and the Model for Interdisciplinary

Research On Climate (MIROC), available from PMIP2, at 2.5 arc-minutes resolution, downloaded from http: //www.worldclim.org/past. We used 75% of the localities in each replicate to train the model and 25% to test it, using the default convergence threshold and regularization values; the maximum number of iterations was set at 1000. Finally, binary maps of both models were superimposed with the program ArcMap v10 under the GIS environment (ESRI 2011. ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute)

DNA extraction, gene amplification and sequencing

We amplified a section of almost 0.8kb of the mitochondrial gene cytochrome oxidase I (COI) and ~500bp of the nuclear ribosomal internal transcribed spacer (ITS-1) intron as described in [1]. For COI we used the PCR products as template in sequencing reactions using Big-Dye (3.1, Applied Biosystems, Foster City, CA, USA) and the reaction products were separated on the ABI Prism 3730 automated sequencer (Unitat de Genòmica dels Serveis Científico-Tècnics de la UB). ITS-1 sequences were performed at Macrogen Inc. (Korea). After revising the chromatograms, we aligned sequences using MAFFT version 6 [13] and then checked them by eye with Bioedit v.7.0.9.0 software [14]. COI sequences were translated into amino acids and used as a guide for the nucleotide alignment. For ITS-1 sequences, those positions that could not be unambiguously aligned were subsequently excluded from the analyses using GBlocks 0.91b [15].

Phylogenetic analysis

We determined the nucleotide substitution model that best fit the data using jModelTest 0.1.1 [16] and applying the Akaike information criterion (AIC); the model obtained was GTR+I+G for both genes. We used Maximum Likelihood (ML) and Bayesian Inference (BI) methods to estimate phylogenetic relationships independently for COI and ITS-1 datasets. ML analysis was run in RAxML 7.0.0 software [17] and bootstrap support (BS) values [18] were calculated from 10,000 replicates. BI trees were inferred with MrBayes v. 3.1.2 [19]. Two independent runs were performed for 3 million generations, sampling every 100 generations. The convergence of the runs was checked through the standard deviation of split frequencies. To test whether population 01-AR, showing a long branch in the trees, could be wrongly situated due to a Long Branch Attraction artefact (LBA), we performed two extra ML analyses to remove the fastest evolving sites from the alignments, a strategy that has been demonstrated to improve the accuracy of the reconstruction methods in these circumstances [20,21]. We applied the GTR model in MEGA 5 [22] to estimate the substitution rates for each position in the COI dataset. In one case we removed the most variable positions (5^{th}category,) estimated from the entire alignment. In the second case we removed (in all sequences) the most variable positions (5^{th}category) from the 01-AR population.

We estimated the divergence time of the sampled populations with BEAST v. 1.6 [23] using COI sequence information. We used the uncorrelated lognormal relaxed clock model with a mean substitution rate of 0.0173 nucleotide substitutions per site and per million years [24]. We ran 10 million iterations of the Markov Chain Monte

Carlo (MCMC), from which we sampled 10,000 trees and discarded 2500 (burn-in period) to obtain the posterior estimates of the node ages. We determined the convergence of the MCMC sampler using TRACER v1.5 [24].

Population genetic analysis and neutrality tests

The analyses were conducted using all COI sequences as well as ITS-1 sequences without ambiguous positions (using the complete deletion option). To perform both intra- and inter-population genetic analyses, we used the program DnaSP v5.10.1 [25]. We estimated the intra-population genetic diversity based on the number of haplotypes (h), haplotype diversity (H), nucleotide diversity (π; [26]), and the Watterson parameter (θ; [27]). To test whether all of the populations (sampling localities within the same conservation area) are under the neutral hypothesis, we conducted three neutrality tests (by calculating Tajima's D [28], Fu's F_S [29] and R_2 [30]) for both the nuclear and mitochondrial markers. The levels of linkage disequilibrium (LD) were assessed using the Z_{nS} [31] statistic and the association among segregating sites with Wall's Q statistic [32]. The statistical significance of these tests was assessed through 10,000 replicates of computer simulations based on the coalescent process [33]. The inter-population genetic diversity levels were measured with the D_{XY} and D_a parameters (the average number and the net number of nucleotide substitutions per site between populations, respectively), applying the Jukes and Cantor correction [34]. Moreover, a Neighbor-Joining tree was estimated basing on D_{XY} values both for COI and ITS-1 data independently using MEGA 5 [22]. We determined whether these populations are genetically differentiated using the S_{nn} statistic [35] and estimated its

statistical significance through a permutation test (10,000 replicates). We also estimated the levels of gene flow among the populations from the N_{st} statistic [36] assuming the infinite island model [37]. We assessed the regression significance between pairwise genetic distances (D_{XY}) and the natural logarithm-transformed geographical distances by the Mantel Test [38] using the Isolation by Distance Web Service, 3.2 [39] (30,000 randomizations).

ABC-based analyses: regional models

We used the ABC-GLM method [40] implemented in ABCtoolbox [41] to compare different coalescent-based evolutionary models that might explain the land planarians diversification across the Brazilian AF. We tested four scenarios (based on a structural serial founder model) to infer general patterns of the distribution of genetic diversity along the SAF (Figure S1). In the first model (model 1), we assume a range expansion of *C. bergi* populations, attended with several serial founder events, from the most northerly population (01-AR) to the higher latitudes (31-ST). In the second model (model 2), these events occur in the opposite direction, from higher latitudes (31-ST) to lower ones (01-AR). The other two models assume that populations are funded from the ends toward the centre (model 3) and from the centre toward the ends of the current species distribution (model 4). To avoid spurious results using populations at similar latitude (mainly in the C-SAF region), we restricted the analysis to only 8 populations (all populations except 13-EC, 19-PC and 22-PI). Furthermore, we also assume that there is no migration among the populations due to the low mobility of

the studied individuals. We assume that the time required for a new founder event from an existing population (t_C), the relative population size of founder populations (x), and the duration of bottlenecks (t_1) are the same for all consecutive founder events. Priors of the parameters shared by all populations were set as uniform distributions, ranging from 0 to 5 (in units of $4N$ generations) for coalescence times (t_0, t_C and t_1) and from 0 to 1 for the parameter x. We estimated the effective population size (N) of the populations relative to population 1 (which was set to 1) using a truncated normal distributed hyper prior with a mean of 1 and a standard error uniformly distributed from 0.25 to 2. Our statistical inferences were based on a number of summary statistics describing the intraspecific variation (the nucleotide diversity, π; the Watterson parameter, θ; the haplotype diversity, H; the number of haplotypes weighed by the sample size, K_W; plus j statistics, one for each of the j frequency classes -i.e. the unfolded frequency spectrum). These statistics were calculated separately for each population and for each gene. Additionally we used one statistic capturing the information of the genetic variation between populations (the nucleotide diversity between populations, D_{XY}, calculated for all pairs). The final vector includes 239 statistics. To reduce the putative random noise introduced by the use of too many summary statistics in the estimation procedure, we performed a partial least square (PLS) transformation (as proposed in [42]), which finally resulted in 5 linear combinations of the initial vector of statistics. Using the program mlcoalsim [43], we simulated 2,000,000 data sets, each one corresponding to a vector of 198 summary statistics, under each of the four competing evolutionary models. The observed values of these statistics in empirical data were computed in mstatpop [44]. From the

simulated data, we retained the 10,000 replicates with the smallest Euclidean distances δ (between the simulated and the observed data) to perform the post-sampling adjustment and to obtain the marginal densities and the *P*-value of the model, the fraction of the retained simulations with a smaller or equal likelihood than the observed data under the estimated general linear model (GLM). The model choice (among the four competitive models) was performed with Bayes factors (BF) as in Leuenberger & Wegmann [40].

References

1. Alvarez-Presas M, Carbayo F, Rozas J, Riutort M. 2011 Land planarians (Platyhelminthes) as a model organism for fine-scale phylogeographic studies: Understanding patterns of biodiversity in the Brazilian Atlantic Forest hotspot. *J. Evol. Biol.* **24**, 887-896

2. Graff, L. v. 1899 Monographie der Turbellarien II. Tricladida terricola (landplanarien). *Engelmann, Leipzig.*

3. Carbayo F & Froehlich EM. 2012 Three new Brazilian species of the land planarian *Choeradoplana* (Platyhelminthes: Tricladida: Geoplaninae), and an emendation of the genus. *J. Nat. Hist.* **46**, 1153-1177

4. Carbayo F, Álvarez-Presas M, Olivares CT, Marques FPL, Froelich EM, Riutort M 2013 Molecular phylogeny of Geoplaninae (Platyhelminthes) challenges current classification: proposal of taxonomic actions. *Zoologica Scripta* **42** (5), 508–528.

5. Romeis B. 1989 *Mikroskopische Technik*. München.

6. Carbayo F & Froehlich EM. 2008 Estado do conhecimento dos macroturbelários (Platyhelminthes) do Brasil. *Biota Neotrop.* **8**, 177-197.

7. Phillips SJ, Anderson RP, Schapire RE. 2006 Maximum entropy modeling of species geographic distributions. *Ecol. Model.* **190**, 231-259.

8. Phillips SJ & Dudík M. 2008 Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*. **31**, 161-175.

9. 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.

10. Amaro RC, Rodrigues MT, Yonenaga-Yassuda Y, Carnaval AC. 2012 Demographic processes in the montane Atlantic rainforest: Molecular and cytogenetic evidence from the endemic frog *Proceratophrys boiei. Mol. Phylogenet. Evol.* **62**, 880-888.

11. Hijmans R J, Cameron SE, Parra JL, Jones P, Jarvis A. 2005 Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology,* **25**, 1965– 1978.

12. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology,* **143**, 29-36.

13. Katoh K, Asimenos G, Toh H. 2009 Multiple alignment of DNA sequences with MAFFT. *Methods Mol. Biol.* **537**, 39-64.

14. Hall TA. 1999 BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41**, 95-98.

15. Castresana J & Talavera G. 2007 Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, 564–577.

16. Posada D. 2008 jModelTest: Phylogenetic Model Averaging. *Mol Biol Evol.* **25**, 1253-1256.

17. Stamatakis A. 2006 RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics.* **22**, 2688-2690.

18. Felsenstein J. 1985 Confidence limits on phylogenies: An approach using the bootstrap. *Evolution.* **39**, 783-791.

19. Ronquist F & Huelsenbeck aJP. 2003 MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics.* **19**, 1572-1574.

20. Philippe H, Lopez P, Brinkmann H, Budin K, Germot A, Laurent J, Moreira D, Müller M, Le Guyader H. 2000 Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. *Proc. R. Soc. B Biol. Sci.* **267**, 1213-1221.

21. Lartillot N, Brinkmann H, Philippe H. 2007 Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.* **7**.

22. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011 MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol. Biol. Evol.* **28**, 2731-2739.

23. Drummond AJ & Rambaut A. 2007 BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7.**

24. Rambaut A, D. A. 2007. *Tracer v1.4 , Available from* http://beast.bio.ed.ac.uk/Tracer.

25. Librado P & Rozas J. 2009 DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics.* **25**, 1451-1452.

26. Nei, M. 1987 Molecular evolutionary genetics. New York: Columbia University Press.

27. Watterson GA. 1975 On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256-276 .

28. Tajima F. 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics.* **123**, 585-595.

29. Fu YX. 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics.* **147**, 915-925.

30. Ramos-Onsins SE & Rozas J. 2002 Statistical properties of new neutrality tests against population growth. *Mol. Biol. Evol.* **19**, 2092-2100.

31. Kelly JK. 1997 A test of neutrality based on interlocus associations. *Genetics.* **146**, 1197-1206.

32. Wall JD. 1999 Recombination and the power of statistical tests of neutrality. *Genet. Res.* **74**, 65-79 .

33. Hudson RR. 1990 Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.* **7**, 1–44.

34. Jukes TH & Cantor CR. 1969 Evolution of protein molecules. *Mammalian Protein Metabolism Academic Press, New York*, 21-132.

35. Hudson RR. 2000 A New Statistic for Detecting Genetic Differentiation. *Genetics.* **155**, 2011-2014.

36. Lynch M & Crease TJ. 1990 The analysis of population survey data on DNA sequence variation. *Mol. Biol. Evol.* **7**, 377-394.

37. Wright S. 1931 Evolution in mendelian populations. *Genetics.* **16**, 97-159.

38. Mantel N. 1967 The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27**, 209-220.

39. Jensen J, Bohonak A, Kelley S. 2005 Isolation by distance, web service. *BMC Genetics.* **6**, 13 .

40. Leuenberger C & Wegmann D. 2010 Bayesian computation and model selection without likelihoods. *Genetics.* **184**, 243-252.

41. Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L. 2010 ABCtoolbox: A versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics.* **11**, 116-123.

42. Wegmann D, Leuenberger C, Excoffier L (2009) Efficient Approximate Bayesian Computation coupled with Markov chain Montecarlo without likelihood. Genetics, 182, 1207-1218.

43. Ramos-Onsins SE & Mitchell-Olds T. 2007 Mlcoalsim: Multilocus Coalescent Simulations. *Evol. Bioinform.* **3**, 41–44.

44. Ramos-Onsins SE, Ferretti L, Marmorini G. statispop: Statistical Analysis using Multiple Populations to pipeline with mlcoalsim. http://bioinformatics.cragenomica.es/numgenomics/people/sebas/software/software.html