BIOLOGY LETTERS

rsbl.royalsocietypublishing.org

Research

Cite this article: Pelletier TA, Carstens BC. 2018 Geographical range size and latitude predict population genetic structure in a global survey. *Biol. Lett.* **14**: 20170566.
http://dx.doi.org/10.1098/rsbl.2017.0566

Received: 12 September 2017
Accepted: 11 December 2017

Subject Areas:
evolution

Keywords:
GBIF, genetic structure, machine learning, latitude, elevation

Authors for correspondence:
Tara A. Pelletier
e-mail: tarapell@gmail.com
Bryan C. Carstens
e-mail: carstens.12@osu.edu

Electronic supplementary material is available online at http://dx.doi.org/10.6084/m9.figshare.c.3967920.

THE ROYAL SOCIETY PUBLISHING

Population genetics

Geographical range size and latitude predict population genetic structure in a global survey

Tara A. Pelletier and Bryan C. Carstens

Department of Evolution, Ecology, and Organismal Biology, Ohio State University, 318 W. 12th Avenue, Columbus, OH 43210-1293, USA

TAP, 0000-0003-3190-3053; BCC, 0000-0002-1552-227X

While genetic diversity within species is influenced by both geographical distance and environmental gradients, it is unclear what other factors are likely to promote population genetic structure. Using a machine learning framework and georeferenced DNA sequences from more than 8000 species, we demonstrate that geographical attributes of the species range, including total size, latitude and elevation, are the most important predictors of which species are likely to contain structured genetic variation. While latitude is well known as an important predictor of biodiversity, our work suggests that it also plays a key role in shaping diversity within species.

1. Background

Intraspecific genetic variation is a key component of evolution. Population genetic theory predicts that the physical separation of individuals limits the exchange of alleles, producing genetic variation that is geographically structured [1]. Within a species, genetic distance should be positively correlated with geographical distance under an isolation-by-distance (IBD) model, and might enable local adaptation along environmental gradients [2]. For example, a meta-analysis of 70 studies by [3], found that isolation-by-environment (IBE) plays a strong role in structuring populations. Whether correlated with geographical or environmental distance, genetic structure has been detected in a variety of species with vastly different distribution patterns [4–6].

While thousands of phylogeographic investigations have been published [7], the discipline has not addressed questions on the broadest scales. Several meta-analyses have examined IBD, IBE or both [3,8,9], and while informative, are limited in scope due to the nature of meta-analyses and often contain conflicting results [3,9,10], which stem from differences in study design, search criteria and publication bias [8] that are difficult to circumvent. Rather than attempt such a meta-analysis, we repurpose existing georeferenced genetic data from online repositories: GenBank and Global Biodiversity Information Facility (GBIF). Because the collection of these data was motivated by a variety of reasons, repurposing enabled us to assess IBD and IBE in an unbiased manner on a larger scale. We compare both geographical and environmental distance matrices to a matrix of genetic distance for over 8000 species and apply a machine learning approach to identify intrinsic and extrinsic characteristics that best explain variation in population genetic structure among species.

2. Material and methods

We downloaded all occurrence data from GBIF and identified records that included GenBank accessions, retrieved these sequences from GenBank and conducted multiple sequence alignment on a gene-by-gene basis for each species. All statistical

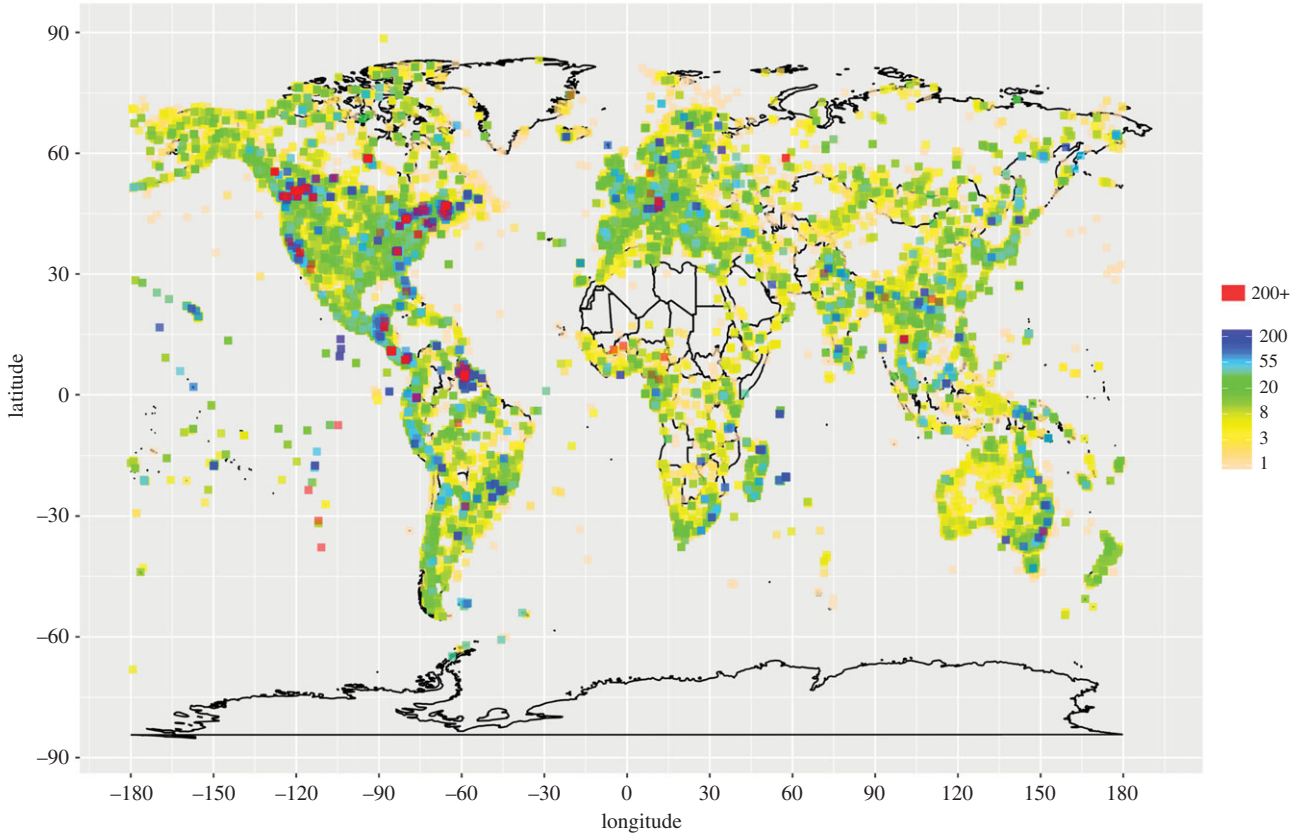

Figure 1. Global sampling. Collection localities from 561,534 georeferenced sequences. Colours correspond to the numbers of individuals sampled from that locality.

Table 1. The proportion of datasets significant for geography (Geo) and the environment (Env). *p*-value of a binomial test ($p = 0.05$) of whether the proportion of significant datasets was greater than expected by chance.

Group	*n* datasets	prop.sig Geo.	*p*-value Geo.	prop.sig Env.	*p*-value Env.
fungi	23	0.04	0.69	0.04	0.69
mosses	10	0	1	0	1
ferns	7	0	1	0	1
Gymnosperms	111	0.07	0.19	0.06	0.32
angiosperms	870	0.1	<0.01	0.1	<0.01
arthropods	6014	0.15	<0.01	0.13	<0.01
vertebrates	2577	0.29	<0.01	0.21	<0.01
Annelida	33	0.21	0	0.15	0.02
Cnidaria	6	0.5	0	0	1
Echinodermata	14	0.21	0.03	0.21	0.03
Mollusca	44	0.18	0.01	0.16	0.01
Nematoda	6	0.33	0.03	0.33	0.03
Platyhelminthes	15	0	1	0.2	0.04
total	9730	0.19	<0.01	0.15	<0.01

analyses were conducted using R v. 3.2.3 [11]. See electronic supplementary material for more details. The distribution of georeferenced data was mapped by calculating the frequency of localities associated with each GPS coordinate (figure 1; electronic supplementary material, table S1). We calculated genetic, geographical and environmental distance matrices for each dataset. In order to characterize the environmental conditions experienced by each species, we followed [12]. Given that geography and environment are often correlated, as we observe in our data (mean $r = 0.77$), we conducted a multiple matrix regression with randomization (MMRR) [12] to examine the effects of two different distance matrices (geographical (IBD) and environmental (IBE)) on the response variable (genetic distance), while controlling for the other matrix.

A data table was developed to identify the strongest predictors of population genetic structure: habit (terrestrial, aquatic, volant, parasitic), metabolism (ectotherm, endotherm, photosynthetic), gene type (nDNA, mtDNA, cpDNA), number of individuals (n) in the dataset, total area of species' range, minimum distance from the equator, mid-point of latitude, the

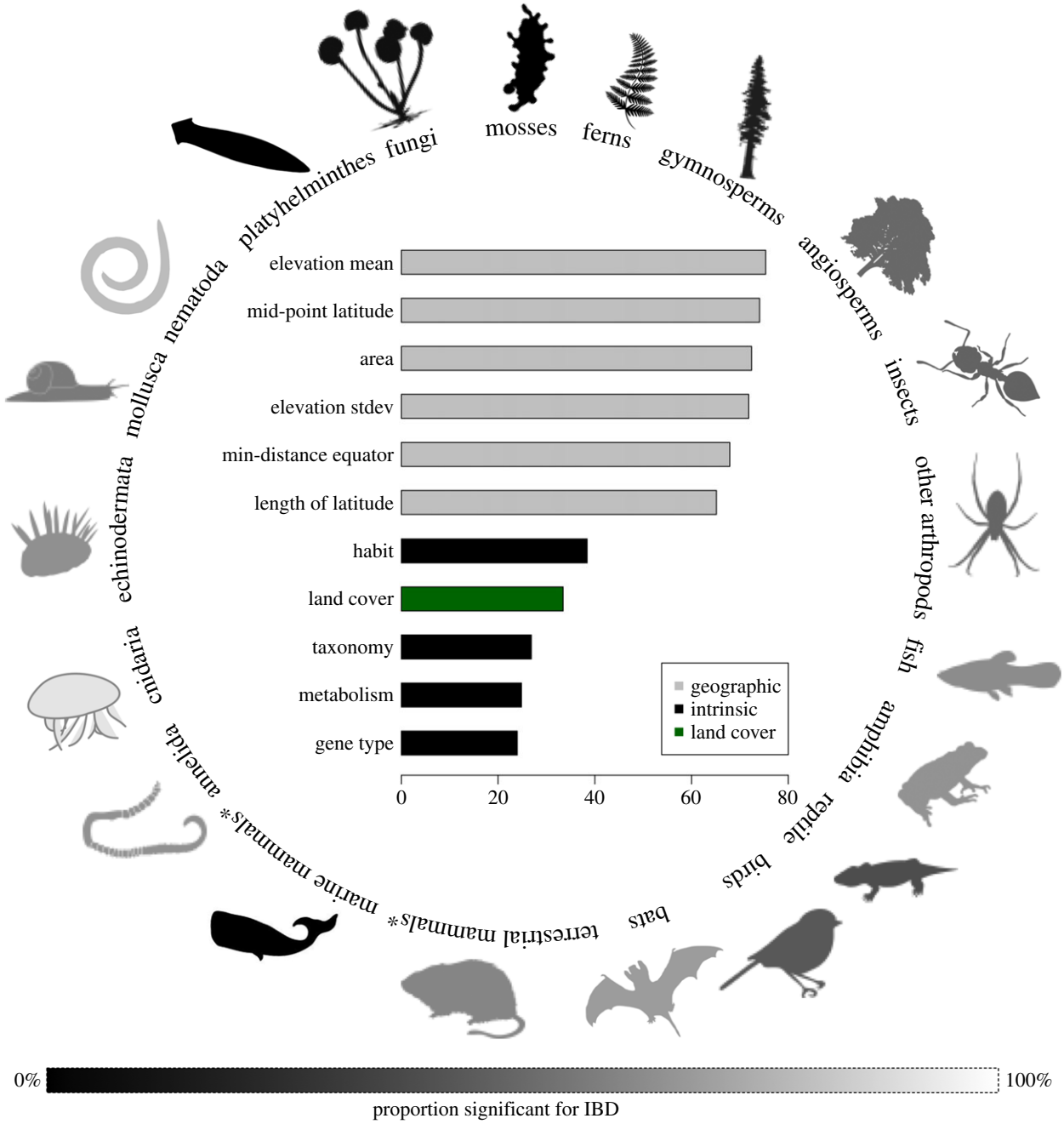

Figure 2. Predictor variables and the proportion of IBD by group. (Inset) Mean decrease in accuracy of predictor variables. Taxonomic rank and land cover variables were averaged for clarity of presentation. (Outer) Proportion of species that exhibit significant population genetic structure (see scale on the lower portion of the figure). Groups marked with an asterisk (*) are not monophyletic clades, but are grouped together due to small sample sizes and similarity of life-history traits. (Online version in colour.)

extent of latitude and elevation mean and standard deviation. Taxonomy was included to assess the role that phylogenetic relatedness plays in structuring populations, and served as a proxy for organismal traits common to particular clades. Finally, the proportion of GPS coordinates within each of the 23 land cover classes described by the European Space Agency GlobCover Portal [13] was included to evaluate environmentally dependent organismal traits [14].

Random forest analysis was used to determine which of the above variables were the most important predictors of IBD or IBE [15,16]. This is a machine learning approach that uses multiple decision trees (a forest) to predict the response based on many potential predictor variables, and is designed to deal with large correlated datasets. The importance of each variable is determined by measuring the mean decrease in accuracy (MDA) of the prediction after the removal of each variable from the predictive function. We categorized datasets as either being significant for IBD or IBE (p-value < 0.05), or not. We conducted a series of random forest analyses with different cut-offs for n, and used several downsampling schemes to assess biases in the data, such as uneven response variables, and uneven geographical sampling. Classification error rates were calculated to assess the accuracy of the models.

3. Results

After filtering data that did not contain sufficient sample sizes, we analysed 9730 datasets from 8955 species. A total of 19% of the datasets were significant for IBD and 15% were significant for IBE (table 1). In most taxonomic groups, there were more datasets with population genetic structure than expected by chance ($p < 0.05$) (table 1; figure 2; electronic supplementary material, figure S1). Out of the datasets that were significant for either IBD or IDE,

Table 2. Comparison of geographical range characteristics for species with and without IBD.

variable	mean with IBD	mean without IBD	*t*-test *p*-value
area (km^2)	6.24×10^6	3.50×10^6	1.93×10^{-7}
minimum distance from equator	29.6712	32.1637	7.12×10^{-10}
mid-point latitude of range	31.4947	34.1037	6.92×10^{-8}
length of latitude°	14.5924	10.4653	2.20×10^{-16}
mean elevation	113.9991	112.9753	0.4563
standard deviation elevation	63.5219	59.7870	3.92×10^{-7}

57% were significant for both, 27% were significant for IBD only, and 15% were significant for IBE only (electronic supplementary material, figure S2).

The variable with the most predictive ability was sample size (n; electronic supplementary material, table S4), which is indicative of a bias introduced by low sample sizes. To address this, we plotted a rarefaction curve to see where the proportion of datasets that are significant for IBD levels off as a function of sample size (electronic supplementary material, figure S3). There was a large jump from $n > 3$ to $n > 10$, and at $n > 20$ the proportion of datasets that are significant levels off. We therefore repeated the analysis with $n > 10$ and $n > 20$ (electronic supplementary material, tables S5–S10 and S13–S14; $n > 10$ datasets = 4,304). The accuracy of the random forest model improved when the response variable was even, and slightly improved when geographical sampling was even (electronic supplementary material, tables S15–S17). The top predictor variables in all analyses were related to the geographical range: latitude, area and elevation (figure 2; electronic supplementary material, tables S4–S14). This suggests that regardless of potential sampling size and/or biases, the importance of the geographical variables in predicting IBD and IBE is a strong signal in the data. Results were similar when using $p < 0.01$ as significant for IBD or IBE (electronic supplementary material, tables S18–S19).

Whether the most important variables were significantly different in species with or without IBD was examined using *t*-tests (table 2). The mean size of the geographical range of a species with IBD was almost twice that of a species without population genetic structure, while the total latitudinal length is 1.5× longer, and datasets with IBD were significantly further from the equator for both mid-point latitude and minimum distance. The standard deviation of elevation for those with IBD was significantly larger than those without, while the mean elevation was higher, but not significantly different.

4. Discussion

There is a considerable amount of population genetic structure within species that can be explained by geographical and environmental differences. Geographical distance had a slightly stronger signal, although neither is substantially more responsible for genetic structure across all taxonomic groups (table 1; electronic supplementary material, figures S2–S5). Our random forest analyses identified several predictor variables related to the geographical range of species, such as area and measurements related to latitude (figure 2; electronic supplementary material, tables S4–S12), as important in predicting population genetic structure, similar to findings by [17]. These variables are likely important because they are related to both organismal dispersal ability and physiological adaptations to conditions in the abiotic environment [18,19], and hold true even after controlling for latitudinal sampling bias. Attributes of the geographical range were significantly different in species with and without IBD (table 2). As observed in other studies (e.g. [20]), IBD and IBE are identified along elevation gradients. We suspect that the reason why elevation mean is an important predictor in the random forest, but not significantly different between species with and without IBD, is because while important, the elevation at which it influences IBD depends on geographical location. The complex relationship between these variables should be considered in future studies.

While mapping genetic diversity on a global scale provides important information [21], identifying factors that influence genetic diversity within species will improve our ability to protect biodiversity [22]. This structure is important as species adapt across their geographical ranges and their life-history traits evolve in response to environmental pressures. Furthermore, we are likely underestimating global genetic structure given limitations of available data. This supposition is supported by the difference in rate of IBD estimated from the full dataset (approx. 15%) as opposed to that estimated from species where more than 100 samples are available (approx. 40%; electronic supplementary material, figure S3). While our analyses suggest that we have detected IBD and IBE in a greater number of species than expected by chance (electronic supplementary material, figure S1), it is very likely that we lack sufficient genetic data for most species and thus are underestimating the proportion of species that are structured by geography, the environment, or both.

Geographical variation in intraspecific genetic structure likely results from variation in speciation, migration and extinction rates. Lower rates of speciation in temperate regions of the world [23–25] might explain the difference in IBD due to latitude because as species remain intact, there is more time for genetic differentiation to accumulate across geographical space. We suspect that area is an important predictor of IBD and IBE due to both the intrinsic dispersal ability of species and the larger amounts of landscape variability that are likely to be found in large ranges.

Our findings were made possible by repurposing existing georeferenced genetic data that contain immense potential for insight [7,26,27]. Unfortunately, most available sequence data are not linked to geographical coordinates [28]. This disassociation of genetic and geographical accessions limits the

utility of open source databases and must be addressed if biodiversity scientists are to leverage the information contained within existing data to meet the challenges associated with conservation of species on a global scale.

Data accessibility. Scripts, data table (electronic supplementary material, appendix S1) and GenBank accessions (electronic supplementary material, appendix S2) are available on Dryad: DOI: http://dx.doi.org/10.5061/dryad.q1j20 [29].

Authors' contributions. All authors conceived the idea, wrote and approved the final manuscript. T.A.P. collected and analysed the data.

Competing interests. We have no competing interests.

Funding. This study was funded by Division of Environmental Biology (NSF DEB-60046038).

Acknowledgements. We thank A. Morales, S. Hird, A. Espindola, G. Wheeler for comments on earlier drafts of this manuscript. Computational resources provided by Ohio Supercomputer Center.

References

1. Wright S. 1943 Isolation by distance. *Genetics* **28**, 114–138.
2. Felsenstein J. 1976 The theoretical population genetics of variable selection and migration. *Annu. Rev. Genet.* **10**, 253–280. (doi:10.1146/annurev.ge.10.120176.001345)
3. Sexton JP, Hangartner SB, Hoffmann AA. 2013 Genetic isolation by environment or distance: which pattern of gene flow is most common? *Evolution* **68**, 1–15. (doi:10.1111/evo.12258)
4. Kuchta SR, Tan A-M. 2005 Isolation by distance and post-glacial range expansion in the rough-skinned newt, *Taricha granulosa*. *Mol. Ecol.* **14**, 225–244. (doi:10.1111/j.1365-294X.2004.02388.x)
5. Frantz AC, Pope LC, Etherington TR, Wilson GJ, Burke T. 2010 Using isolation-by-distance-based approaches to assess the barrier effect of linear landscape elements on badger (*Meles meles*) dispersal. *Mol. Ecol.* **19**, 1663–1674. (doi:10.1111/j.1365-294X.2010.04605.x)
6. Relethford J. 2004 Global patterns of isolation by distance based on genetic and morphological data. *Hum. Biol.* **76**, 499–513. (doi:10.1353/hub.2004.0060)
7. Garrick RC *et al.* 2015 The evolution of phylogeographic datasets. *Mol. Ecol.* **24**, 1164–1171. (doi:10.1111/mec.13108)
8. Jenkins DG *et al.* 2010 A meta-analysis of isolation by distance: relic or reference standard for landscape genetics? *Ecography* **33**, 215–320. (doi:10.1111/j.1600-0587.2010.06285.x)
9. Shafer ABA, Wolf JBW. 2013 Widespread evidence for incipient ecological speciation: a meta-analysis of isolation-by-ecology. *Ecol. Lett.* **16**, 940–950. (doi:10.1111/ele.12120)
10. Wang IJ, Glor RE, Losos JB. 2013 Quantifying the roles of ecology and geography in spatial genetic divergence. *Ecol. Lett.* **16**, 175–182. (doi:10.1111/ele.12025)
11. R Development Core Team 2016 *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
12. Wang IJ. 2013 Examining the full effects of landscape heterogeneity on spatial genetic variation: a multiple matrix regression approach for quantifying geographic and ecological isolation. *Evolution* **67**, 3403–3411. (doi:10.1111/evo.12134)
13. Louvain & ESA. GLOBCOVER. 2009 Université Catholique de Louvain & European Space Agency. http://due.esrin.esa.int/page_globcover.php.
14. Van Oudtshoorn KVR, Van Rooyen MW. 2013 *Dispersal biology of desert plants*. Berlin, Germany: Springer Science & Business Media.
15. Breiman L. 2001 Random forests. *Mach. Learn.* **45**, 5. (doi:10.1023/A:1010933404324)
16. Biau G. 2012 Analysis of a random forests model. *J. Mach. Learn. Res.* **13**, 1063–1095.
17. Martin PR, McKay JK. 2003 Latitudinal variation in genetic divergence of populations and the potential for future speciation. *Evolution* **58**, 938–945. (doi:10.1111/j.0014-3820.2004.tb00428.x)
18. Mori GM, Zucchi MI, Souza AP. 2015 Multiple-geographic-scale genetic structure of two mangrove tree species: the roles of mating system, hybridization, limited dispersal and extrinsic factors. *PLoS ONE* **10**, e0118710. (doi:10.1371/journal.pone.0118710)
19. McDevitt AD, Oliver MK, Piertney SB, Szafrańska PA, Konarzewski M, Zub K. 2013 Individual variation in dispersal associated with phenotype influences fine-scale genetic structure in weasels. *Conserv. Genet.* **14**, 499–509. (doi:10.1007/s10592-012-0376-4)
20. Guarnizo CE, Amézqutia A, Bermingham E. 2009 The relative roles of vicariance versus elevational gradients in the genetic differentiation of the high Andean tree frog, *Dendroposophus labialis*. *Mol. Phylogenet. Evol.* **50**, 84–92. (doi:10.1016/j.ympev.2008.10.005)
21. Miraldo A *et al.* 2016 An anthropocene map of genetic diversity. *Science* **353**, 1532–1535. (doi:10.1126/science.aaf4381)
22. Chen IC, Hill JK, Ohlemüller R, Roy DB, Thomas CD. 2011 Rapid range shifts of species associated with high levels of climate warming. *Science* **333**, 1024–1026. (doi:10.1126/science.1206432)
23. Willig MR, Kaufman DM, Stevens RD. 2003 Latitudinal gradients of biodiversity: pattern, process, scale, and synthesis. *Annu. Rev. Ecol. Evol. Syst.* **34**, 273–309. (doi:10.1146/annurev.ecolsys.34.012103.144032)
24. Jablonski D, Roy K, Valentine JW. 2006 Out of the tropics: evolutionary dynamics of the latitudinal diversity gradient. *Science* **314**, 102–106. (doi:10.1126/science.1130880)
25. Leffer EM, Bullaughey K, Matute DR, Meyer WK, Ségurel L, Venkat A, Andolfatto P, Przeworski M. 2012 Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol.* **10**, e1001388. (doi:10.1371/journal.pbio.1001388)
26. Dawson MN. 2014 Natural experiments and meta-analyses in comparative phylogeography. *J. Biogeogr.* **41**, 52–65. (doi:10.1111/jbi.12190)
27. Soltis DE, Soltis PS. 2016 Mobilizing and integrating big data in studies of spatial and phylogenetic patterns of biodiversity. *Plant Divers.* **38**, 264–270. (doi:10.1016/j.pld.2016.12.001)
28. Marques AC, Maronna MM, Collins AG. 2013 Putting GenBank data on the map. *Science* **341**, 1341. (doi:10.1126/science.341.6152.1341-a)
29. Pelletier TA, Carstens BC. 2017 Data from: Geographic range size and latitude predict population genetic structure in a global survey. Dryad Digital Repository. (http://dx.doi.org/10.5061/dryad.q1j20)