Quantifying the scale of genetic diversity extinction in the Anthropocene

Moises Exposito-Alonso[1,2,3*], Tom R. Booker[4,5,&], Lucas Czech[1,&], Tadashi Fukami[2,&], Lauren Gillespie[1,7,&], Shannon Hateley[1,&], Christopher C. Kyriazis[8,&], Patricia Lang[2,&], Laura Leventhal[1,2,&], David Nogues-Bravo[9,&], Veronica Pagowski[2,&], Megan Ruffley[1,&], Jeffrey P. Spence[6,&], Sebastian E. Toro Arana[1,2,&], Clemens L. Weiß[6,&], Erin Zess[1,&].

[1]Department of Plant Biology, Carnegie Institution for Science, Stanford, CA 94305, USA.
[2]Department of Biology, Stanford University, Stanford, CA 94305, USA.
[3]Department of Global Ecology, Carnegie Institution for Science, Stanford, CA 94305, USA.
[4]Department of Zoology, University of British Columbia, Vancouver, Canada.
[5]Biodiversity Research Centre, University of British Columbia, Vancouver, Canada.
[6]Department of Genetics, Stanford University, Stanford, CA 94305, USA.
[7]Department of Computer Science, Stanford University, Stanford, CA 94305, USA.
[8]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095, USA.
[9]Center for Macroecology, Evolution and Climate, GLOBE Inst., Univ. of Copenhagen, Copenhagen, Denmark.
[*]To whom correspondence should be addressed: moisesexpositoalonso@gmail.com
[&]Authors are listed alphabetically

Draft last updated: 13 Oct 2021

Keywords: extinction, genetic diversity, climate change, habitat loss, Anthropocene

More species than ever before are at risk of extinction due to anthropogenic habitat loss and climate change. But even species that are not threatened have seen reductions in their populations and geographic ranges, likely impacting their genetic diversity. Although preserving genetic diversity is a key conservation target to maintain the adaptability of species, we lack predictive tools and global estimates of genetic diversity loss across ecosystems. By bridging biodiversity and population genetics theories, we introduce the first mathematical framework to understand the loss of naturally occurring DNA mutations within a species—what we call *genetic diversity extinction*. Analyzing genome-wide variation data of 10,126 geo-tagged individuals from 19 plant and animal species, we show that genome-wide diversity follows a power law with geographic area, which can predict genetic diversity decay in simulated spatial extinctions. Given pre-21st century values of ecosystem transformations, we estimate that over 10% of genetic diversity may be extinct, already surpassing the United Nations targets for genetic preservation. These estimated losses could rapidly increase with advancing climate change and habitat destruction, highlighting the need for new forecasting tools that assist in the rapid implementation of policies to protect genetic resources.

Anthropogenic habitat loss and climate change (*1*, *2*) are putting approximately one million species (25% of all species) at risk of extinction (*3*). One thousand species are already documented as extinct (*1*, *2*). However, it has been estimated that an even larger fraction—47%—of plant and animal species have lost part of their geographic range (*4*, *5*). Though this might seem inconsequential compared to losing an entire species, this range loss detrimentally impacts genetic diversity. Genetic diversity dictates a species' ability to adapt to new environments (*6–8*), so its loss can spiral into a feedback loop where diversity loss further increases the risk of complete species extinction (*9*, *10*). Because of these damaging consequences, and as population declines outpace mutation replenishment making these diversity losses effectively irreversible, we term this "*genetic diversity extinction"*.

Although genetic diversity is recognized by scientists as a key dimension of biodiversity (*11*), it has been overlooked in conservation plans of international policy groups. Only in 2021 did the United Nations' Convention of Biological Diversity propose to preserve at least 90% of all species' genetic diversity (*12*, *13*). Although meta-analyses of empirical genetic markers in animals through time are emerging (*14–16*), scalable approaches to estimate genetic diversity extinction across species do not yet exist, impairing prioritization and evaluation of conservation targets. Here, we introduce a framework to estimate global genetic diversity extinction by bridging biodiversity theory with population genetics, and by combining global ecosystem transformations with new genomic datasets.

The first studies that predicted biodiversity reductions in the 1990s and 2000s projected species extinctions due to habitat loss and climate change using the relationship of biodiversity with geographic area— termed the Species-Area Relationship (SAR) (*17*) (see **Supplementary Materials [SM] I** for a review of mathematical models). In this framework, ecosystems with a larger area (A) harbor a larger number of species (S), and the more a study area is extended, the more species are found. The SAR has been empirically shown to follow a power law, $S = A^z$. It scales consistently across continents and ecosystems (*18*), with a higher z characterizing speciose and highly spatially structured ecosystems. Conversely, given estimates of decreasing ecosystem areas over time ($A_{t-1} > A_t$), Thomas et al. (*19*) proposed rough estimates of the percentage of species extinctions in the 21st century ranging from 15 to 37%—though this may be an oversimplification it has become a common tool for policy groups including the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) (*3*). However, despite the similarities between species' distributions within an ecosystem and the distribution of individual DNA mutations within the same species, there have been no attempts to describe the extent of genetic diversity extinction using an analogous "*Mutations-Area Relationship"*.

Defining the Mutations-Area Relationship

Genetic mutations, defined here as DNA nucleotide variants appearing in individuals of a species (e.g. ACGGGTA vs ACGG**A**TA), typically remain low frequency in a population, though a few become prevalent through stochastic genetic drift and natural selection (27, 28). As the "commonness of rarity" principle for species abundances is a key statistical condition that led to the power law relationship of the SAR, we thus hypothesized that if mutations follow the same rarity principle, the genetic equivalent should exist, namely the Mutations-Area Relationship (MAR). Therefore, we examined the rarity of mutations using 11,769,920 biallelic genetic variants of the *Arabidopsis thaliana* 1001 genomes dataset (**Fig. 1A**) (*20*) by fitting several common models of species abundances (*21*) to the distribution of mutation frequencies (q), termed the Site Frequency Spectrum in population genetics (**Fig. 1A** inset, **SM II.1**). The canonical L-shaped probability distribution ($1/q$) of this spectrum fit this data well (R^2 = 0.998), though Preston's species abundance log-normal model achieved the best AIC value with R^2 = 0.999 (**SM III.1, Table SIII.1, Table IV.1**). This showcases

the similarities of abundance distributions of mutations within species and species within ecosystems, suggesting they may behave similarly in their relationship to geographic area (*21*, *22*).

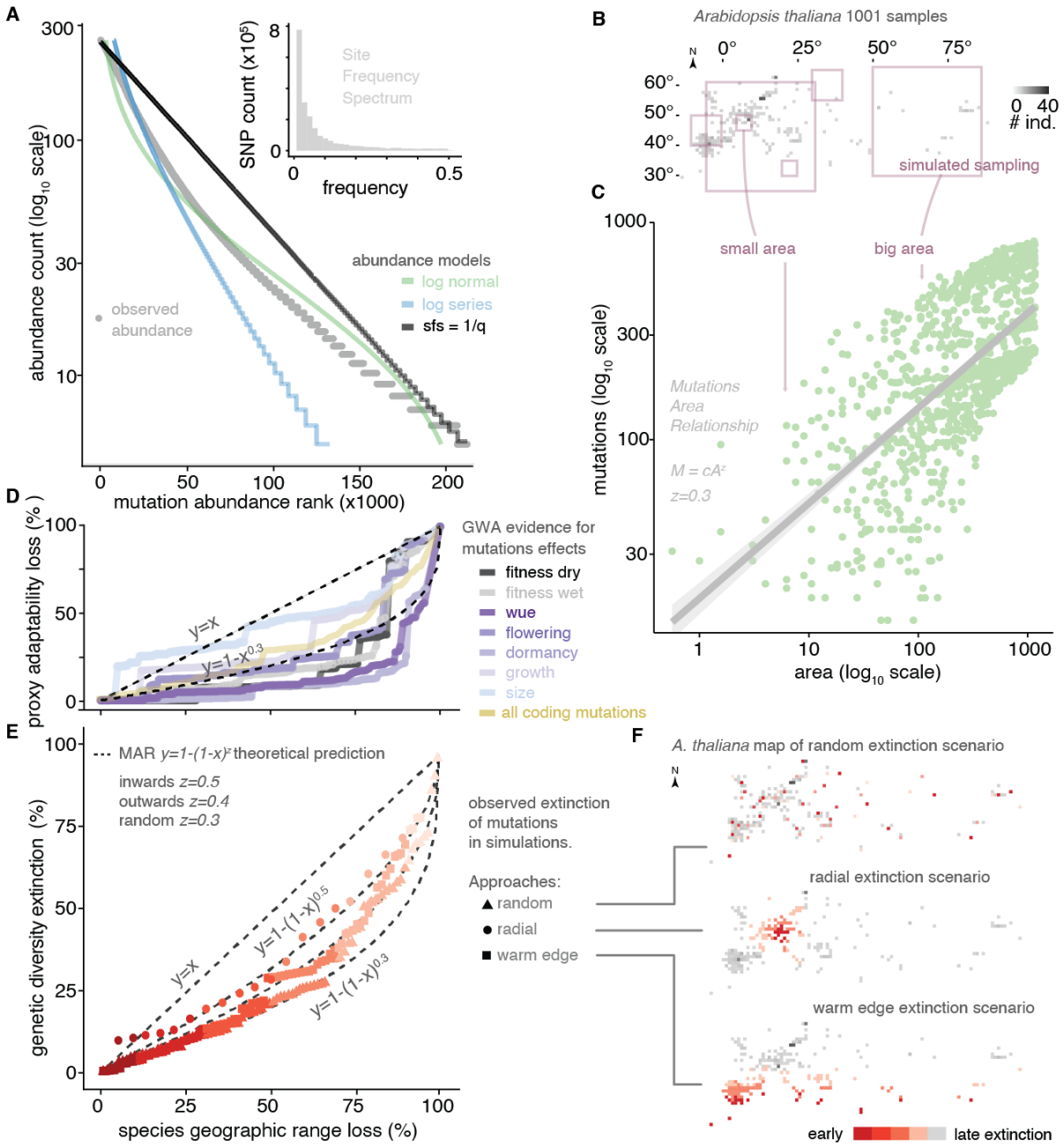


Fig. 1 | Mutations across a population follow a log-normal abundance distribution and a power law with range area. (A) *Distribution of mutation frequencies in 1,001* Arabidopsis thaliana *plants using a Site Frequency Spectrum histogram (grey inset) and a Whittaker's rank abundance curve plot, and the fitted models of common species abundance functions.* ***(B)*** *Density of individuals projected in a 1x1 degree latitude/longitude map of Europe and exemplary subsample areas of different sizes.* ***(C)*** *The Mutations-Area Relationship (MAR) in log-log space built from subsamples of 10,000 mutations in* A. thaliana *in (B) using 10 replicates of every possible size of square areas randomly placed in space.* ***(D)*** *A metric of adaptive capacity loss during extinction in (E-F). Using Genome Wide Associations (GWA) to estimate effects of mutation on fitness in different rainfall conditions, water use efficiency [wue], flowering time, seed dormancy, plant growth rate, and plant size. Plotted are the fraction loss of the summed squared effects ($\sum a^2$) of 10,000 mutations from the top 1% tails of effects. We also plot the fraction of gene-coding alleles (yellow).* ***(E)*** *Percentage of extinction of total genetic diversity (grey) from stochastic simulations of extinction in (F), and theoretical model projections (red) of genetic diversity extinction using the MAR.* ***(F)*** *Cartoon of several possible range contractions simulated by progressively removing grid cells following different hypothesized spatial extinction patterns.*

To define a metric for measuring how genetic diversity within a species increases with geographic area, we constructed the MAR by subsampling different regions of the native range of *Arabidopsis thaliana* using over one thousand geo-tagged genomes (**Fig. 1B**). As a metric of genetic diversity, we modelled the number of mutations (M) in space (number of segregating sites) consistent with the species-centric approach of SAR, which uses species richness as the metric of biodiversity (**SM II.2**). The MAR followed the power law relationship $M = cA^z$ with a scaling value $z_{MAR} = 0.324$ (CI95% = 0.238–0.41) (**Fig. 1C**). This scaling is robust to different methods of subsampling areas, the effects of non-random spatial patterns, random area sampling, fully nested outward or inward sampling (*18*), raster area calculations, raster grid resolution (~10–1,000 km side cell size), and is adjusted for limited sample sizes (**SM III.3, II.3.2, Fig. SIII.3-6, Tables SIII.2-7**).

To test the generality of the MAR, we leveraged genomic datasets of hundreds to thousands of individuals of the same species across large geographic areas (**Table 1, SM IV**). We assembled datasets of 10,126 individuals of 19 plant and animal species, ranging from 1,522 to 88,332,015 naturally occurring mutations per species, covering a geographic area ranging from 0.03 to 115 million km^2. Similar values to *A. thaliana's* z_{MAR} were recovered for these diverse species (mean z_{MAR} scaled = 0.24, median = 0.23, IQR = 0.18, **Table 1, SM IV, Fig. SIV.1, Table SIV.1-2).** Theoretical derivations show that z_{MAR} is a consequence of fundamental evolutionary and ecological forces (mutation rate, dispersal, selection) and should range from 0–1, depending on the strength of population structure (**SM II.3,** see **Fig SII.7** for its relationship with isolation-by-distance). These predictions were confirmed by spatial population genetics coalescent and individual-based simulations in 2D and continuous space (**SM II.3**), as well as with mainland-island community assembly simulations according to the Universal Neutral Theory of Biodiversity (UNTB) (**SM V)**. Naturally, each species has a different total genetic diversity—which can vary substantially due to species-specific traits such as census size, mating system, genome size and structure, etc. (*23*)—but the relationship of genetic diversity with sample area within a species' geographic range appears relatively consistent across species.

Finally, to predict the effect of species' range contractions on genetic diversity, we applied the MAR to estimate genetic diversity extinction. We set up several scenarios of range contraction in *A. thaliana* by removing map grid cells *in silico* (**Fig. 1F**). Our simulations included local extinction in the warmest regions within a species range, which may rapidly become the least hospitable with climate change (*4*, *24*). Other simulations represent additional scenarios, such as random local population extinction representing deforestation scattered across large continents (**Fig. 1F-E**). The MAR-based predictions of genetic extinction, using $1-(1-A_t/A_{t-1})^z$, conservatively followed the simulated local extinctions in *A. thaliana* ($R^2 = 0.87$) (**SM III.4)**. Our model can predict a simulated local extinction using the z_{MAR} value of another species or the average z_{MAR} across species given that the z_{MAR} estimates were robust and relatively consistent across species, and no association was found between z_{MAR} and different ecology, mating systems, home continents, etc. (**Table 1**, **Table SIV.3-4)**. These results are encouraging both because a typical $z_{MAR} = 0.2\text{-}0.3$ may be predictive of genetic diversity extinction in range-reduced species that lack genomic information and because any scaling $z < 1$ given our predictive equation implies that the genetic diversity loss is slower than area loss.

Since genetic diversity is ultimately created by spontaneous DNA errors passed onto offspring every generation, the loss of genetic diversity may seem reversible. However, the recovery of genetic diversity through natural mutagenesis is extremely slow (57), especially when restricting to mutations affecting adaptation. Simulating a species undergoing "only" a 5–10% area extinction, it would take at

least ≈140–520 generations to recover its original genetic diversity (2,100–7,800 years for a fast-growing tree or medium-lifespan mammal), although for most simulations, recovery virtually never happened over millennia (see **SM II.4-5, Fig. SII.8, SM III.6**).

Table 1 |The Mutations-Area Relationship across diverse species

Summary statistics of individuals sampled broadly across species distributions, sequencing method and mutations studied, and convex hull area extent of all samples within a species. Mutations Area Relationship (MAR) parameters, which capture how spatially restricted mutations area, including a scaled correction for low sampling genomic effort. Area that needs to be kept for a species to maintain 90% of its genetic diversity, using the per-species MAR value estimates. Area predictions are not provided for threatened species, as these have likely already lost substantial genetic diversity and require protection of their full geographic range (Fig. 2).

Species	N	M_{tot} Method	A_{tot} Km^2x10^6	MAR z [CI95%]	MAR scaled z* [CI95%]	Min area $_{90\%}$ %
Arabidopsis thaliana	1,135 (1,001)[#]	11,769,920 W	27.34	0.324 (0.238–0.41)	0.312 (0.305 - 0.32)	71–78
Arabidopsis lyrata	108	17,813,817 W	2.79	0.236 (0.218–0.254)	0.151 (0.137–0.165)	50–66
Amaranthus tuberculatus	162 (155)	1,033,443 W	0.80	0.109 (0.081–0.136)	0.142 (0.136–0.149)	48–65
Eucalyptus melliodora[VU]	275 (36)[*]	9,378 GBS	0.95	0.466 (0.394–0.538)	0.403 (0.398–0.407)	77–82
Yucca brevifolia[CA]	290	10,695 GBS	1.21	0.128 (0.109–0.147)	0.049 (0.037–0.062)	-
Mimulus guttatus	521 (286)[#*]	1,522 GBS	25.14	0.274 (0.259–0.29)	0.231 (0.221–0.241)	63–73
Panicum virgatum	732 (576)[†]	33,905,044 W	6.29	0.232 (0.211–0.252)	0.126 (0.116–0.136)	43–63
Panicum hallii	591	45,589 W	2.19	0.68 (0.546–0.813)	0.652 (0.559–0.746)	85–88
Pinus contorta	929	32,449 GC	0.89	0.015 (0.014–0.016)	-	-
Pinus torreyana[CR]	242	478,238 GBS	0.03	0.236 (0.19–0.282)	0.105 (0.099–0.11)	-
Populus trichocarpa	882	28,342,826 W	1.12	0.275 (0.218–0.332)	0.165 (0.155–0.176)	53–67
Anopheles gambiae	1142 (29)[*]	52,525,957 W	19.96	0.214 (0.164–0.264)	0.122 (0.111–0.132)	42–62
Acropora millepora	253 (12)[*]	17,931,448 W	0.03	0.246 (0.209–0.283)	0.287 (0.28–0.294)	69–77
Drosophila melanogaster	271[%]	5,019 W	115.21	0.437 (0.397–0.477)	0.325 (0.314–0.336)	72–79
Setophaga petechia	219 (199)[&]	349,014 GBS/GC	7.03	0.214 (0.174–0.254)	0.074 (0.047–0.102)	24–54
Peromyscus maniculatus	80 (78)[&]	14,076 GBS	22.61	0.488 (0.264–0.713)	0.683 (0.615–0.751)	86–88
Dicerorhinus sumatrensis[CR]	16	8,870,513 W	3.33	0.412 (0.369–0.456)	0.127 (0.11–0.144)	-
Canis lupus	349 (230)[†]	1,517,226 W	19.10	0.256 (0.232–0.28)	0.184 (0.175–0.193)	56–70
Homo sapiens	2504 (24)[*]	88,332,015 W	80.76	0.431 (0.347–0.514)	0.281 (0.23–0.332)	-

[#]*Only individuals in the native range were used for the analyses.*
[&]*Only individuals with available coordinates or matching IDs were used for analyses.*
[%]*Numbers indicate pools of flies used for Pool-Sequencing.*
[*]*Number of geographically separated populations, as multiple individuals were collected per population.*
[$]*GPS locations unknown, MAR calculated with area equal number of individuals.*
[†]*Only natural populations were used, excluding breeds, landraces, and cultivars.*
[^ns.]*Reported z_{MAR} calculated assuming number of individuals equals area. Computations using raster areas provided noisy estimates, see Table SV.1. As these estimates were not reliable or different from zero, they were not used for extinction calculations.*
Acronyms: W = whole-genome re-sequencing or discovery SNP calling. GBS = genotyping by sequencing of biallelic SNP markers. GC = Genotyping Chip; logN = log Normal distribution. logS = log Series distribution. Wei = Weibull distribution. CR = Red List Critically Endangered. VU= Red List Vulnerable. CA= Included in the California Endangered Species Act.

Estimating the global magnitude of genetic diversity extinction

Using the MAR, we estimated the average global genetic diversity extinction caused by pre-21st century land transformations. Although accurate species-specific geographic area reduction data in the last centuries are scarce, we leveraged global land cover transformations from primary ecosystems to urban or cropland systems (*3*, *25*) (**Table SV.1**). Using the average scaled z_{MAR} (**Table SIV.2**), several global averages of Earth's land transformation for present day (38% from (*25*), 34% from (*26*), and 43-50% from (*27*)), we estimated a global genetic diversity extinction across per species between

10-16%. While these estimates may correctly approximate central values across species in an ecosystem, we expect variation in the extent of loss across species ranging from 0 to 100% (**Fig. SV.4**). One cause of this variation is the heterogeneity in land cover transformations across ecosystems; for example, more pristine high altitude systems have only lost 0.3% of their area, while highly managed temperate forests and woodlands have lost 67% (**Fig. 2**, **Table SV.1-5**).

Another cause for the variability in genetic extinction among species (even within the same ecosystem) may be their differential geographic ranges and abundances, life histories, or conservation risk factors. We gathered data from species red-listed by the International Union for Conservation of Nature (IUCN) (*1*), which evaluates recent population or area reduction in ±10 years / ±3 generations to place assessed species in different threat categories using several thresholds. Assuming an average z_{MAR} can capture general patterns, we translate these category thresholds into genetic diversity extinction (see **SM V, Table SV.4)**. *Vulnerable* species that have lost at least 30% of their populations may have experienced >9% of genetic diversity extinction, *endangered* species that have lost over 50% of their populations should have incurred >16% of genetic diversity extinction, and *critically endangered* species with over 80% area reduction likely suffered >33% of genetic diversity extinction (**Fig. 2B**).

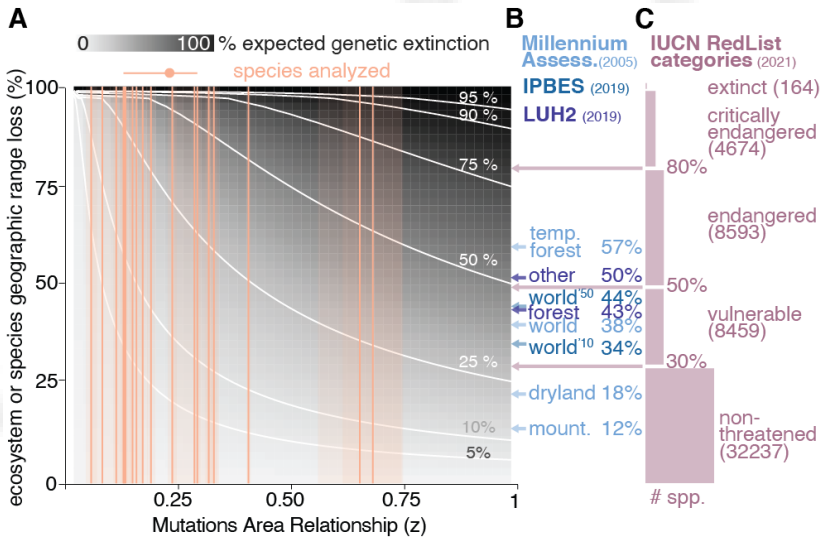


Fig. 2 | The parameter space of genetic diversity extinction compared to pre-21st century ecosystem transformations and endangered species categories. (A) *Possible values of two key parameters, z scaling of the Mutations-Area Relationship (MAR) and % of area reduction of a species geographic range (as a proxy of entire ecosystem transformation). The theoretical % of genetic diversity extinction (grey gradient) is represented as filled color, with isolines in white. Estimates of* z_{MAR} *from Table 1 are vertical orange lines with semi-transparent CI95% boxes.* ***(B)*** *Percentage of transformed ecosystem area from the Millennium Ecosystem Assessment (25) are represented by light blue arrows, from the Intergovernmental Science-Policy Panel for Biodiversity and Ecosystem Services (IPBES) (26) are dark blue arrows, and from the Land Use Harmonization 2 dataset (27) are in dark purple.* ***(C)*** *The minimum criterion value of population or geographic distribution loss to be classified in each category of the IUCN's red list are indicated with pink arrows. The number of plant species included in each category is shown as box sizes (1).*

How does genetic diversity extinction impact species' adaptability?

To quantitatively understand how MAR relates to the loss of adaptive capacity, we leveraged the extensive knowledge of the effect of mutations in ecologically relevant traits in *A. thaliana* from Genome-Wide Associations (GWA) (**Fig. 1E, SM III**). Again conducting spatial warm edge extinction simulations, we tracked metrics of adaptive capacity, including the total sum of effects of

remaining mutations ($\sum a^2$), the additive genetic variance [$V_a = \sum p(1-p)a^2$], and the loss of nonsynonymous mutations. Although determining the effect of mutations is technically challenging (*28*, *29*) to impossible with environmental or genomic context change (*28*, *30*), our simplistic analyses suggest these potentially adaptive alleles may be lost more slowly than neutral genetic diversity (**Fig. 1C, SM II.3.4.**). Therefore, the loss of adaptive mutations may lag behind the extinction of neutral mutations—and additive variance may even temporarily increase due to bottlenecks (*31*) (**Fig SIII.10**)— then sharply collapse in late stages of the whole-species extinction event (*32*) (**Fig. 1C**, **Fig SII.6**).

To achieve the recently-published target of the United Nations to protect "at least 90% of genetic diversity within all species" (*13*), it will be necessary to aggressively plan for early protection of populations from different geographic enclaves. Here, we developed the Mutations-Area Relationship (MAR)—the first mathematical framework to forecast genetic diversity extinction with shrinking geographic species ranges. The MAR contrasts with existing studies on the risk of losing entire species by focusing on quantifying the magnitude and dynamics of genetic diversity extinction likely ongoing in the majority of species. The MAR demonstrates that even with conservative estimates, substantial area protection will be needed to meet the UN's Sustainable Goal A. For vulnerable or critically endangered species we have likely already failed.

Supplementary Materials are available in this ***Google Drive link***

Data availability. The analyzed datasets are publicly available or were shared by authors upon request (see Supplementary Materials). Code and intermediate data are available at https://github.com/moiexpositoalonsolab/mar <tobepushed>.

Acknowledgements. We are grateful to authors that shared genome variant files: Sujan Mamidi, John Lovell, Dan Jacobson, Manesh Shah, Julia Kreiner, Kay Lucek, Yvonne Willi, Juan D Palacio Mejia, Justin Borevitz, Megan Supple, Mario Vallejo-Marin, Nicolas Dussex, Lionel Di Santo, and Jill Hamilton. We thank John Wiens, Sue Rhee, Detlef Weigel, Ellie Armstrong, Marty Kardos, Dmitri Petrov, Rob Colwell, Rasmus Nielsen, Anna Michalak, Jonathan Pritchard, and members of the Moi and Mordecai Labs for comments, discussion, or references. M.E.-A. is supported by the Office of the Director of the National Institutes of Health's Early Investigator Award with award number: 1DP5OD029506-01, and by the U.S. Department of Energy, Office of Biological and Environmental Research, grant number: DE-SC0021286; and the Carnegie Institution for Science. J.P.S. was supported by an NIH training grant (5T32HG000044-23), and S.T.A by the NIGMS Center of the NIH under award number T32GM007276. S.H. and C.L.W. are supported by Stanford's Center for Computational, Evolutionary, and Human Genomics. M. R. is supported by the NSF's Plant Genome Postdoctoral Research Fellowship in Biology. L. L. and L.G. are supported by NSF's GRFP. Computational analyses were done on the High-Performance Computing clusters *Memex* and *Calc* supported by the Carnegie Institution for Science.

Author contribution M.E.-A. conceived and led the project. M.E.-A., J.P.S., M.R., S.H., L.G., L.C., L.L., S.T.A., V.P., E.Z., P.L., C.K., T.B., C.W. conducted research, all authors interpreted the results and wrote the manuscript.

REFERENCES

1. IUCN, The IUCN Red List of Threatened Species. *https://www.iucnredlist.org*.

2. S. Díaz, J. Settele, E. S. Brondízio, H. T. Ngo, J. Agard, A. Arneth, P. Balvanera, K. A. Brauman, S. H. M. Butchart, K. M. A. Chan, L. A. Garibaldi, K. Ichii, J. Liu, S. M. Subramanian, G. F. Midgley, P. Miloslavich, Z. Molnár, D. Obura, A. Pfaff, S. Polasky, A. Purvis, J. Razzaque, B. Reyers, R. R. Chowdhury, Y.-J. Shin, I. Visseren-Hamakers, K. J. Willis, C. N. Zayas, Pervasive human-driven decline of life on Earth points to the need for transformative change. *Science*. **366**, eaax3100 (2019).

3. IPBES, *Global Assessment Report on Biodiversity and Ecosystem Services* (IPBES Secretariat, Bonn, Germany, 2019; https://www.ipbes.net/global-assessment-report-biodiversity-ecosystem-services).

4. J. J. Wiens, Climate-Related Local Extinctions Are Already Widespread among Plant and Animal Species. *PLoS Biol.* **14**, e2001104 (2016).

5. W. Thuiller, S. Lavorel, M. B. Araújo, M. T. Sykes, I. C. Prentice, Climate change threats to plant diversity in Europe. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 8245–8250 (2005).

6. M. Exposito-Alonso, F. Vasseur, W. Ding, G. Wang, H. A. Burbano, D. Weigel, Genomic basis and evolutionary potential for extreme drought adaptation in Arabidopsis thaliana. *Nat Ecol Evol*. **2**, 352–358 (2018).

7. C. Parmesan, Ecological and Evolutionary Responses to Recent Climate Change. *Annu. Rev. Ecol. Evol. Syst.* **37**, 637–669 (2006).

8. T. Capblancq, M. C. Fitzpatrick, R. A. Bay, M. Exposito-Alonso, S. R. Keller, Genomic Prediction of (Mal)Adaptation Across Current and Future Climatic Landscapes. *Annu. Rev. Ecol. Evol. Syst.* **51**, 245–269 (2020).

9. M. Lynch, J. Conery, R. Burger, Mutation Accumulation and the Extinction of Small Populations. *Am. Nat.* **146**, 489–518 (1995).

10. D. Spielman, B. W. Brook, R. Frankham, Most species are not driven to extinction before genetic factors impact them. *Proceedings of the National Academy of Sciences*. **101** (2004), pp. 15261–15264.

11. W. Steffen, K. Richardson, J. Rockström, S. E. Cornell, I. Fetzer, E. M. Bennett, R. Biggs, S. R. Carpenter, W. de Vries, C. A. de Wit, C. Folke, D. Gerten, J. Heinke, G. M. Mace, L. M. Persson, V. Ramanathan, B. Reyers, S. Sörlin, Planetary boundaries: Guiding human development on a changing planet. *Science*. **347** (2015), doi:10.1126/science.1259855.

12. S. Díaz, N. Zafra-Calvo, A. Purvis, P. H. Verburg, D. Obura, P. Leadley, R. Chaplin-Kramer, L. De Meester, E. Dulloo, B. Martín-López, M. R. Shaw, P. Visconti, W. Broadgate, M. W. Bruford, N. D. Burgess, J. Cavender-Bares, F. DeClerck, J. M. Fernández-Palacios, L. A. Garibaldi, S. L. L. Hill, F. Isbell, C. K. Khoury, C. B. Krug, J. Liu, M. Maron, P. J. K. McGowan, H. M. Pereira, V. Reyes-García, J. Rocha, C. Rondinini, L. Shannon, Y.-J. Shin, P. V. R. Snelgrove, E. M. Spehn, B. Strassburg, S. M. Subramanian, J. J. Tewksbury, J. E. M. Watson, A. E. Zanne, Set ambitious goals for biodiversity and sustainability. *Science*. **370**, 411–413 (2020).

13. CBD, 1st Draft of The Post-2020 Global Biodiversity Framework (2021), (available at https://www.cbd.int/doc/c/abb5/591f/2e46096d3f0330b08ce87a45/wg2020-03-03-en.pdf).

14. D. M. Leigh, A. P. Hendry, E. Vázquez-Domínguez, V. L. Friesen, Estimated six per cent loss of genetic variation in wild populations since the industrial revolution. *Evol. Appl.* **12**, 1505–1512 (2019).

15. K. L. Millette, V. Fugère, C. Debyser, A. Greiner, F. J. J. Chain, A. Gonzalez, No consistent

effects of humans on animal genetic diversity worldwide. *Ecol. Lett.* **23**, 55–67 (2020).

16. S. Theodoridis, C. Rahbek, D. Nogues‐Bravo, Exposure of mammal genetic diversity to mid‐21st century global change. *Ecography (Cop.)* (2021), doi:10.1111/ecog.05588.

17. O. Arrhenius, Species and Area. *J. Ecol.* **9**, 95–99 (1921).

18. D. Storch, P. Keil, W. Jetz, Universal species-area and endemics-area relationships at continental scales. *Nature*. **488**, 78–81 (2012).

19. C. D. Thomas, A. Cameron, R. E. Green, M. Bakkenes, L. J. Beaumont, Y. C. Collingham, B. F. N. Erasmus, M. F. de Siqueira, A. Grainger, L. Hannah, L. Hughes, B. Huntley, A. S. van Jaarsveld, G. F. Midgley, L. Miles, M. A. Ortega-Huerta, A. Townsend Peterson, O. L. Phillips, S. E. Williams, Extinction risk from climate change. *Nature*. **427**, 145–148 (2004).

20. 1001 Genomes Consortium, 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell*. **166**, 481–491 (2016).

21. F. W. Preston, The canonical distribution of commonness and rarity: Part I. *Ecology*. **43**, 185 (1962).

22. R. A. Fisher, XVII.—The Distribution of Gene Ratios for Rare Mutations. *Proceedings of the Royal Society of Edinburgh*. **50**, 204–219 (1931).

23. V. Buffalo, Why do species get a thin slice of π? Revisiting Lewontin's Paradox of Variation. *bioRxiv* (2021), p. 2021.02.03.429633.

24. A. Hampe, R. J. Petit, Conserving biodiversity under climate change: the rear edge matters. *Ecol. Lett.* **8**, 461–467 (2005).

25. Millennium Ecosystem Assessment, *Millennium ecosystem assessment* (Millennium Ecosystem Assessment, 2005; http://chapter.ser.org/europe/files/2012/08/Harris.pdf).

26. Ipbes, *Global assessment report of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services* (IPBES Secretariat, Bonn, Germany, 2019; https://www.ipbes.net/news/ipbes-global-assessment-summary-policymakers-pdf).

27. G. C. Hurtt, L. Chini, R. Sahajpal, S. Frolking, B. L. Bodirsky, K. Calvin, J. C. Doelman, J. Fisk, S. Fujimori, K. Klein Goldewijk, T. Hasegawa, P. Havlik, A. Heinimann, F. Humpenöder, J. Jungclaus, J. O. Kaplan, J. Kennedy, T. Krisztin, D. Lawrence, P. Lawrence, L. Ma, O. Mertz, J. Pongratz, A. Popp, B. Poulter, K. Riahi, E. Shevliakova, E. Stehfest, P. Thornton, F. N. Tubiello, D. P. van Vuuren, X. Zhang, Harmonization of global land use change and management for the period 850–2100 (LUH2) for CMIP6. *Geosci. Model Dev.* **13**, 5425–5464 (2020).

28. M. Exposito-Alonso, P. Wilton, R. Nielsen, Non-additive polygenic models improve predictions of fitness traits in three eukaryote model species. *bioRxiv* (2020), p. 2020.07.14.194407.

29. M. Kardos, E. Armstrong, S. W. Fitzpatrick, S. Hauser, P. Hedrick, J. Miller, D. Tallmon, W. Chris Funk, The crucial role of genome-wide genetic variation in conservation. *bioRxiv* (2021), p. 2021.07.05.451163.

30. M. J. Harms, J. W. Thornton, Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat. Rev. Genet.* **14**, 559–571 (2013).

31. H. R. Taft, D. A. Roff, Do bottlenecks increase additive genetic variance? *Conserv. Genet.* **13**, 333–342 (2012).

32. P. Ehrlich, B. Walker, Rivets and redundancy. *Bioscience*. **48**, 387 (1998).

DRAFT

Supplementary Materials for

Exposito-Alonso et al.: Quantifying the scale of genetic diversity extinction in the Anthropocene

---

Table of content

I. Background on species biodiversity and biogeography **3**

I.1 Theoretical models of biodiversity 3

Fig. SI.1 | Example of typical plots used for species abundance curve studies 3

I.1.2. Niche apportionment approaches 3

I.1.2. Niche statistical approaches of species sampling 4

Fig. SI.2 | Summary of theoretical models of Species Abundance Curves. 5

I.2 Metric of species diversity 5

I.3 Biogeography of species and extinction. 5

Fig. SI.3 | Example of a Species-Area Relationship in Galapagos Islands 6

I.4 Estimating extinction of species from the species area relationship 6

II. Population genetics models and the site frequency spectrum. **6**

II.1 The Wright-Fisher model and the Site Frequency Spectrum 6

Fig. SII.1 | Similarity between the Species Abundance Distribution and the Site Frequency Spectrum 7

II.2 Metrics of genetic diversity 7

II.3 Spatial genetics and the mutations-area relationship (MAR) 8

II.3.1 Panmictic population 8

Fig. SII.2 | Expected ranges of z MAR given sample sizes. 10

II.3.2 Scaling zMAR for low sampling and low census size 10

II.3.3 Meta-populations in space 11

Fig. SII.3 | msprime 2D deme simulations and the mutations-area relationship 11

Table SII.1 | msprime population genetic simulations in 2D. 12

II.3.4 Meta-populations in space with local adaptation 12

Fig SII.4 | SLiM population genetic simulations in 2D with selection and local adaptation 12

Table SII.2 | Linear model explaining zMAR by migration rate and strength of spatially-varying selection 13

II.3.5. Meta-populations in space with purifying selection 13

Fig SII.5 | SLiM population genetic simulations in 2D with purifying selection. 13

II.3.6 Continuous-space non-Wright-Fisher models 13

Fig. SII.6 | Continuous space SLiM population genetic simulations 14

II.3.7 Connection with isolation-by-distance 14

Fig SII.7 | SLiM population genetic simulations in 2D comparing Fst and zMAR. 15

II.4 The extinction of mutations in space 15

II.5 Recovery of genetic diversity after a bottleneck 16

Fig. SII.8 | 2D stepping-stone msprime simulations with extinction of a fraction of the population and recovery 16

II.6 Similar attempts to describe a genetic analog of the Species-Area Relationship and related work 16

II.7 Notes on conservation genetics state-of-the-art 17

III. Testing the mutation area relationship theory with the 1001 Arabidopsis Genomes. **18**

III.1 The Site Frequency Spectrum of the 1001 Arabidopsis Genomes. 18

Fig. SIII.1 | Mutation abundance study in A. thaliana. 18

Fig. SIII.2 | Fit of mutation abundance study in A. thaliana with different SAD models 19

Table SIII.1 | AIC values for model fit of common species distribution curves. 19

III.2 Building the Mutations-Area Relationship 19

Table SIII.2 | Different SAR curves fit to mutations. 19

Table SIII.3 | Mutations-Area Relationship (MAR). 20

Table SIII.4 | Endemic-Mutation Area Relationship (EMAR). 20
Fig. SIII.3 | Mutation Area and Endemic-Mutation Area Relationships in A. thaliana. 20
III.3 Testing numerical artifacts 21
Table SIII.5 | MAR built with different area calculations and grid sizes 21
Fig. SIII.4 | Cartoon of raster sampling to build the MAR 22
Fig. SIII.5 | MAR comparison with different area calculations. 22
Fig. SIII.6 | MAR and EMAR in Arabidopsis thaliana using outward and inward sampling. 23
Table SIII.6 | Outward and inward MAR and EMAR 23
Table SIII.7 | MAR for putatively neutral, deleterious, and locally adaptive alleles in Arabidopsis thaliana 24
III.4 Local extinction in Arabidopsis 24
Fig. SIII.7 | Extinction of mutations with habitat loss in A. thaliana. 25
III.5 Potential impacts of genetic extinction in adaptability 25
Fig. SIII.8 | Bias of low frequency mutations and effect size for fitness traits in A. thaliana. 26
Fig. SIII.9 | Simulations illustrating the potential extinction of locally-adaptive mutations in A. thaliana. 27
Fig. SIII.10 | Extinction simulations as in Fig. SIII.7 showing proxies of adaptive capacity of A. thaliana. 27
III.6 Case study of a massive natural bottleneck 28

IV. The mutations-area parameters for diverse species 29
Fig. SIV.1 | MAR summaries across species. 33
Table SIV.1 | The mutations-area relationship across species. Extended Table 1 34
Table SIV.2 | Mean zMAR across species. 34
Table SIV.3 | Traits, life history, and other characteristics of the analyzed species. 35
Table SIV.4 | Association of traits, life history, and other characteristics with zMAR. 35

V. An estimate of global genetic diversity extinction 36
V.1 Estimates of ecosystem area losses 36
Table SV.1 | Millennium Ecosystem Assessment land cover transformation. 36
Table SV.2 | IPBES land cover transformation, 37
Table SV.3 | Land Use Harmonization 2 from 850 to 2015 37
Table SV.4 | IUCN Red List categories of extinction risk and number of species. 38
V.2 A global estimate of genetic extinction 38
Table SV.5 | Estimates of average expected genetic extinction for different ecosystems. 38
Fig. SV.1 | The parameter space of genetic diversity extinction, extended 40
V.3 Community ecology simulations and MAR 40
Fig. SV.2 | z MAR calculated from MESS eco-evolutionary simulations 41
V.4 The nested species and genetic extinction process 41
Fig. SV.3 | Cartoon of nested extinction of species and genetic diversity. 41
Fig. SV.4 | The distribution of genetic extinction in simulated ecosystems with 1000 species 42
Fig. SV.5 | Numeric simulation of nested genetic extinction. 42

VI. Limitations and outlook 44
VI.1 Reasons for overestimations and ideas for improvement 44
VI.2 Reasons for underestimations and ideas for improvement 44
VI.3 Final notes 44

VII. Supplemental References 46

I. Background on species biodiversity and biogeography

I.1 Theoretical models of biodiversity

Studies in biogeography have modeled the species-area relationship with several functions. Below we summarize the different approaches using an example of richness of S = 100 species, with variable abundance or area, A.

We may visualize the different area or abundance of species as a frequency histogram (Fig. SI. 1, Preston plot), with x-axis: logarithm of abundance bins (historically log2 as a rough approximation to the natural logarithm), and y-axis: number of species at given abundance. Alternatively, as a rank-abundance diagram (Fig. SI. 1, Whittaker plot): x-axis: species list, ranked in order of descending abundance (i.e. from common to rare), and y-axis: logarithm of % relative abundance.

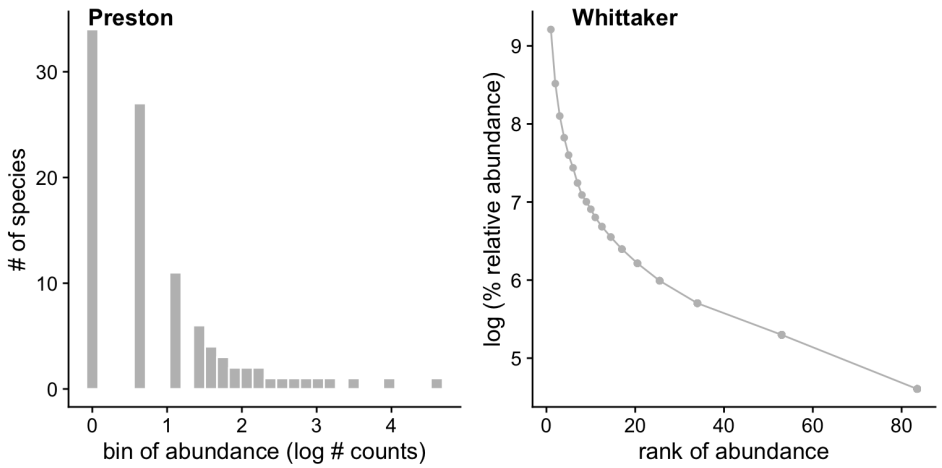


Fig. SI.1 | Example of typical plots used for species abundance curve studies
Due to their strong skew, Species Abundance Curves are often plotted using the Preston plot (left) where the x axis represents bins of log2 abundances (also referred to as octaves), or using the Whittaker plot (right) where the x axis is the rank of each species in a dataset and y axis the species' relative abundance.

Over the years

I.1.2. Niche apportionment approaches

A series of theoretical deterministic and stochastic "niche apportionment models" have been put forward (summarized in (Hubbell, 2001) or (Tokeshi, 1990, 1993)).

The Motomura (1932) geometric series suggests that each species that arrives takes half the area. The first would take 50%, the second 50% of 50%, and so forth, which can be expressed as:

$$P_i = 0.5^i.$$

Similarly, one can imagine that as a species colonizes a habitat, it takes up a fraction different than 50%. This gives a geometric series with parameters k which can be written as

$$P_i = k(1-k)^{i-1}.$$

Other geometric series-related models include stochasticity, where k instead of being a fixed parameter it is a random uniform variable and there is a k_i each time i a new species arrives to the

ecosystem. The "dominance preemption" model draws from 50-100% at any new arrival of a species, the random fraction model draws from 0-100%. Then the abundance of a species depends on the stochastic process of previous *f = 1...i-1* species arriving first:

$$E[P_i|P_1,...,P_{i-1},k_i] = k_i \times \left(1 - \sum_{f=1}^{i-1} P_f\right)$$

Another approach is the broken stick by MacArthur (MacArthur, 1957), which theorized a habitat is broken into *S-1* places at random, which creates S fractions of an area. Then the relative area of a species is:

$$E[P_i] = (\tfrac{1}{S}) \sum_{w=i}^{S} \tfrac{1}{w}.$$

I.1.2. Niche statistical approaches of species sampling

Differently from niche partitioning functions, statistical approaches such as the log-series from Fisher (Fisher, Corbet and Williams, 1943) and log-normal from Preston (Preston, 1962) are probability distributions, and approach modeling in a conceptually different way: they model the sampling process of species collections given an underlying relative abundance (see below).

Statistical-based derivations began with Fisher (Fisher, Corbet and Williams, 1943), with the log-series distribution. It assumes that species abundances in the community are independent identically distributed variables, sampling is a Poisson process, sampling is done with replacement, or the fraction sampled is small enough to approximate a sample with replacement. Here,

$$S_n = \frac{\alpha x^n}{n},$$

where x is a constant $x \in [0,1]$ related to the sample dataset (typically close to 1), $x = \frac{N}{\alpha+N}$, and α is a new constant term (ecosystem-specific) that is used as a measure of biodiversity. Fisher proposed the number of species could be estimated as:

$$S = \alpha \times log\left(1 + \frac{N}{\alpha}\right).$$

Finally, Preston (Preston, 1948) posed that the skewness of previous proposals is due to lack of sampling. With little data, common species are collected sooner, but with more abundant sampling, the rarest species are also well-sampled and have abundances well above 0. Preston then proposed that the octaves (bins of doubling abundance) follow a normal distribution, making the raw abundance log-normal distributed. Given S_0 is the number of species in the model octave of abundance and a variance composite of the log-Normal σ^2, the number of species per abundance (octave) bin R (*=log(n)*) is:

$$S_R = S_0 e^{-R^2/2\sigma^2}.$$

The Unified Neutral Theory of Biodiversity (UNTB) by Hubbell (Hubbell, 2001) takes a stochastic approach of a community with immigrants, extinctions, and speciation in continuous dynamics. Interestingly, the UNTB's key parameter, θ, coincides with Fisher's α, as the log-series is a limiting case of UNTB. Hubbell's discovery was that $\alpha=2J_m v$, where J_m is the size of the external metacommunity that provides migrants of species to the focal community, and v is the speciation rate. Alonso and McKane (Alonso and McKane, 2004) derived the so-called Metacommunity Zero-Sum Multinomial (MZSM) distribution from the UNTB. In practice, both distributions have almost-identical fits (lines completely overlapping in Fig. SI. 2 below).

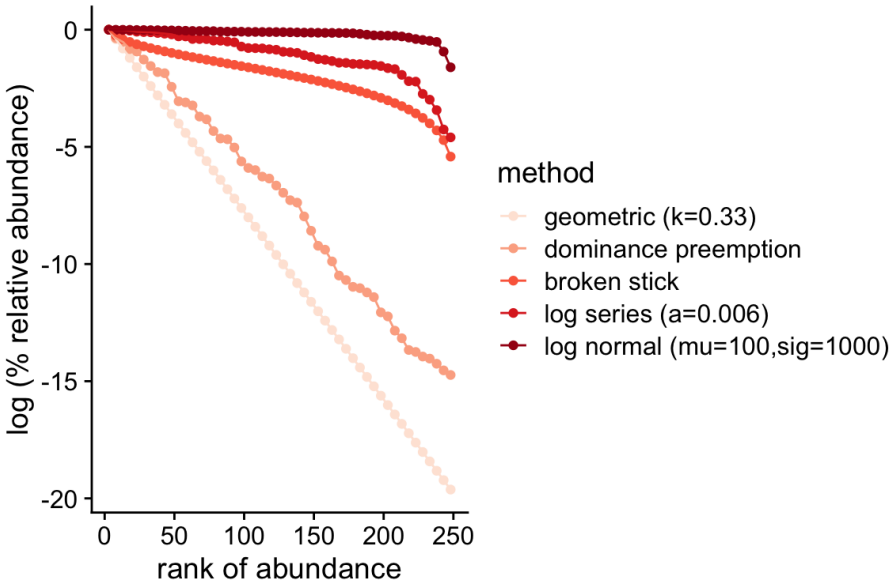


Fig. SI.2 | Summary of theoretical models of Species Abundance Curves.
Five niche partitioning or statistical models shown in a Whittaker plot. The different models expect different levels of evenness in abundance across the species in the community, from the lowest (geometric series) to the highest (log-normal).

I.2 Metric of species diversity

Although a number of metrics exist to measure species diversity, such as the Shannon index, $H' = -\sum_{i=1}^{S} P_i log P_i$ (with Pi the relative proportions of species abundances) or Fisher's non-dimensional α parameter, the study of species abundances and area relationships has focused on species richness *S*, that is, the total number of species in a given location or area.

I.3 Biogeography of species and extinction.

SAD and SAR connection

Due to many species being rare, it is expected that as researchers sample an area, the most common species will be sampled first, and as the area studied increases, more and more species will be discovered. This is thought to happen following a power law relationship, where the number of species in that area S_A increases with the sampled area A, with scaling *z* (slope in a log-log plot), and with a constant *c*:

$$SAR(A) = S_A = cA^z.$$

Preston (1962) derived theoretically that from a log-normal series, one would expect *z*=0.27, under a number of assumptions (Fig. SI. 3). This has been empirically shown to be close to reality (1962; Storch, Keil and Jetz, 2012), although there is some variation across ecosystems and spatial scales.

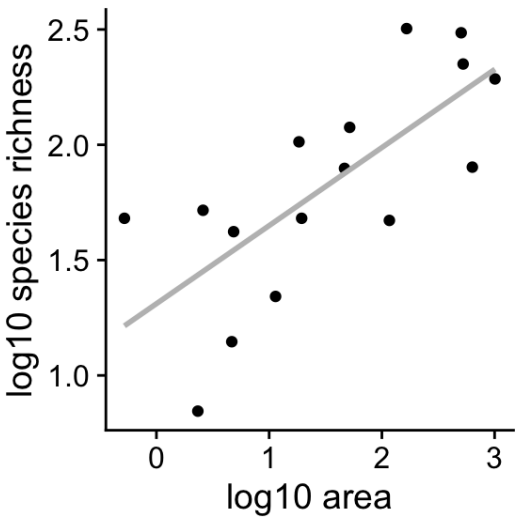


Fig. SI.3 | Example of a Species-Area Relationship in Galapagos Islands
Classic species richness dataset from the Galapagos Islands (Preston, 1962). It depicts species richness as a function of island area in a log-log plot

I.4 Estimating extinction of species from the species area relationship

The first estimates of species extinction used the SAR relationship. Given a reduction of ecosystem area, A, by an area of a (Pimm *et al.*, 1995; Thomas *et al.*, 2004). If these areas, as well as the SAR scaling, z, are known, then one can predict the number of species in the future as:

$$S_{\text{now}} - S_{\text{fut}} = cA_{\text{now}}^{z} - cA_{\text{fut}}^{z},$$

although we are normally interested in the fraction of species that will go extinct.

$$X_s = \frac{S_{\text{now}} - S_{\text{fut}}}{S_{\text{now}}} = 1 - \frac{cA_{\text{fut}}^{z}}{cA_{\text{now}}^{z}} = 1 - \left(\frac{A_{\text{fut}}}{A_{\text{now}}}\right)^{z}.$$

II. Population genetics models and the site frequency spectrum.
II.1 The Wright-Fisher model and the site frequency spectrum

The Site Frequency Spectrum (SFS) in population genetics theory is remarkably similar to the Species Abundance Relationship. In fact, Fisher himself (Fisher, 1931) also proposed that mutation abundances should follow a logarithmic series, with the number of mutations of a given abundance, n, being inverse to their frequency category q:

$$M_n = c\frac{1}{q}.$$

Rearranging terms, one can see this is a constrained version of the log-series Probability Mass Function (PMF), which Fisher also proposed for the distribution of species abundances (Fisher, Corbet and Williams, 1943). Below, one can graphically see the similarities (Fig. SII. 1):

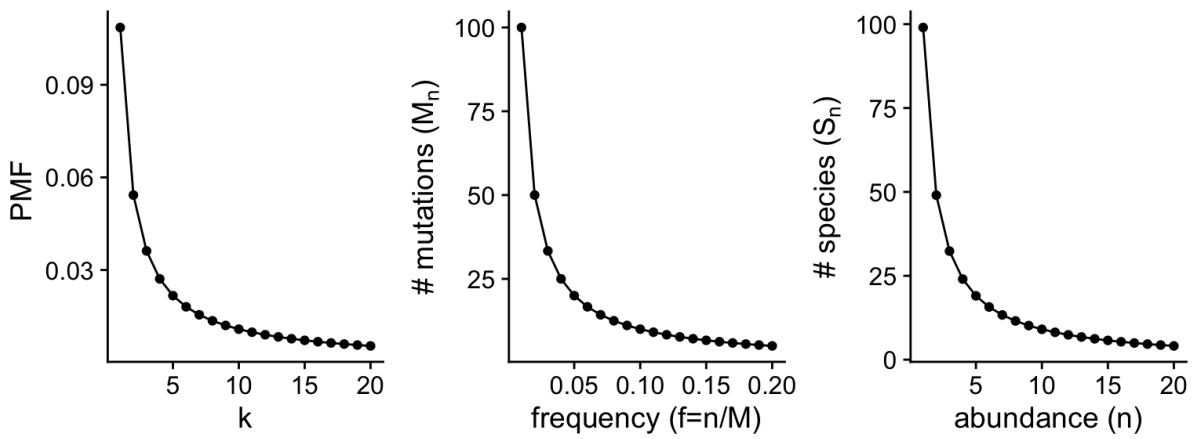


Fig. SII.1 | Similarity between the Species Abundance Distribution and the Site Frequency Spectrum
Left is the Probability Mass Function of the log-series (p=0.999), middle is the SFS (N=100, c=1), and right is the log-series-based abundance of species (alpha=100, N=10000).

Keeping the abundance, n, constant (and low), when the number of individuals $N \to \infty$, we know that the constant x from Fisher's SAD approaches 1, $x = \frac{N}{N+\alpha} \to 1$. Then, we can rewrite the number of species at any given abundance (S_n) as:

$$S_n = \alpha \frac{(\frac{N}{\alpha+N})^n}{n} = \alpha \frac{1^n}{n} = c\frac{1}{n} = M_n$$

So both have the same form as the log series PMF: $f(k) = \frac{-1}{ln(1-p)} \frac{p^k}{k}$ when $p \to 1$. In the next section we will see that the constants of the SAD and the SFS are proportional to species and mutation diversity, although the Site Frequency Spectrum (SFS) is a specific case of SAD. One can also see that because the constant in the SFS is the population scaled mutation rate, $c = \theta = Ne\mu$, Fisher's $\alpha \approx \theta$ for large N.

II.2 Metrics of genetic diversity

In population genetics, multiple measurements of genetic diversity have been put forward. The most straightforward is the allelic richness, also number of mutations, or also called the number of segregating sites. Segregating sites, *M*, is the direct equivalent of the species richness, *S*, and it depends on the number of samples used and length of DNA sequence explored (Note we use the non-standard notation, *M*, as the standard in population genetics is *S* but this is already in use for species richness. We then use *M* for mutations and *S* for species). This metric can also be thought of as the area under the curve of the SFS. Two other metrics that describe the SFS but that aim to be sequence-length- and individual independent are Watterson's Theta, θ_W, and Nucleotide diversity, π, (also called θ_π). These two metrics of diversity are identical at population equilibrium and are estimates of *4Neμ* (when the SFS follows a 1/q relationship), with effective population size *Ne* and per-generation mutation rate *μ*, whereas they differ in non-equilibrium demographics, under natural selection, or under other behaviors not considered in the Wright-Fisher neutral model, such as different mating systems (Hahn, 2018).

First, π is described as:

$$\pi = \frac{\sum_{i=1}^{n-1} i(n-i)M_i}{n(n-1)/2},$$

and θ_W as:

$$\theta_W = \frac{\sum_{i=1}^{n-1} M_i}{\sum_{i=1}^{n-1} 1/i},$$

where $\sum_{i=1}^{n-1} 1/i$ is the *n-1*th Harmonic number, which serves to scale the segregating sites based on the assumption that the abundance of mutations follows a 1/q SFS. The diversity metrics π and θ_W are both functions of the SFS, as opposed to Fisher's α from the Species Abundance Distribution, which is a parameter that changes the shape of the distribution.

Although often π is reported as a typical measure of genetic diversity of a species, since it can be calculated for a single genome and it captures the process of inbreeding of a population (Buffalo, 2021), classic literature relating germplasm management for conservation and breeding has advocated for allelic richness (Marshal and Brown, 1975).

II.3 Spatial genetics and the mutations-area relationship (MAR)

Since its inception, a number of concepts in population genetics have dealt with genetic variation structuring in space. Already in 1943, Sewall Wright proposed that populations sampled further apart geographically must differ more in allele frequency due to more independent drift (Wright, 1943), leading to the commonly used correlation between geographic distance and the metric of differentiation *Fst*. Most prominently, the use of correlation in the accumulation of mutations of populations that are geographically close or share evolutionary history has been uncovered using dimensionality reduction approaches such as PCA (Novembre and Stephens, 2008). Others have discretized space into populations and used process-based models with explicit individuals of a population and their genomes. The latter, so-called Wright-Fisher population process and its reverse the Coalescent, have allowed inferences of population sizes across time and migration rates across space (Li and Durbin, 2011; Petkova, Novembre and Stephens, 2016). Despite these enormous advances in understanding spatial genetic structures, surprisingly little quantitative work has been done to parametrize the loss of genetic diversity by direct loss of habitat.

Because of the abundance of rare mutations in populations, it is straightforward to think that the more area and individuals sampled, the more segregating sites will be found. Analogous to the Species Area Relationship (SAR), $S=cA^z$, we should thus be able to estimate the equivalent scaling for a mutations-area relationship (MAR):

$$M=cA^z,$$

with a scaling $z = z_{MAR}$, which corresponds to the slope of best fit in a log-log-plot of A and M for a given species. (Other functions are often fit empirically for SAR datasets, which we explore later in section II.2. We work with the power law because of its historical use, mathematical convenience, and because other more complicated functions only improved fitting marginally, see Table SIII.2).

This differs from other efforts to understand the number of segregating sites or heterozygosity differences across species that differ in their total census size or geographic distribution (see section II.6). The MAR instead is built within a species, as its ultimate aim is to relate the number of mutations left in a species as it loses spatial populations.

Below we derive what are the expectations of MAR taking two opposite scenarios of neutral population evolution, and study how many segregating sites or mutations M are discovered with increasing area in the simulations. We further test the scenario of meta-populations in space with varying migration rates and neutral or natural selection processes.

II.3.1 Panmictic population

The expected number of mutations, M, is a constant that depends on the mutation rate, μ, and the expected total branch length of the population genealogy, L, with $M=\mu L$. Under the coalescent, the total branch length is equal to the number of lineages or individuals sampled from the population, n, times the time of the genealogy during which there are such lineages, T_n, plus $n-1$ times the time in the genealogy with such number of lineages, and so forth:

$$L = nT_n + (n-1)T_{n-1} + ... + 2T_2.$$

Under the coalescent,

$$E[T_n] = \frac{2Ne}{n(n-1)},$$

and thus:

$$E[L] = n\frac{2N_e}{n(n-1)} + (n-1)\frac{2N_e}{(n-1)(n-2)} + ...,$$

which simplifies to

$$E[L] = 2N_e(\frac{1}{n-1} + \frac{1}{n-2} + ... + 1) = 2N_e H_{n-1},$$

where H_{n-1} is the (n-1)th harmonic number. Finally, following the Taylor expansion approximation of the harmonic number:

$$H_n = \gamma + \log(n) + \frac{1}{2n} + O(\frac{1}{n^2}) \simeq \gamma + \log(n) + \frac{1}{2n},$$

which we can further approximate as:

$$E[L] \approx 2N_e \log(n-1) + c.$$

Therefore, assuming a constant mutation rate and effective population size (N_e) under panmixia, M grows like $log(n)$. In such a case, a log-log plot (typical power law plot) does not display a linear relationship, and the slope is asymptotic to $z \rightarrow 0$ for $N \rightarrow \infty$. On the other hand, with low values of x (area or individuals sampled close to 0), the slope z_{MAR} will be incorrectly high. We can show this effect trivially by studying the local derivative of the function $log_{10}(M) = log_{10}(log(N))$. The local slope of that function is an approximation of our z_{MAR} parameter. This can be locally estimated at any given point N by taking the derivative:

$$\frac{d\log_{10}(\log(N))}{d(\log_{10}(N))} = \frac{1}{\log_{10}(N)\log(10)}.$$

The implication of this nonlinear function is that if we sampled only few individuals or areas of a species (e.g., n=100), even if this species was completely panmictic we would expect a non-zero z_{MAR}. We can roughly approximate z_{MAR} by the local slope of the number in the midpoint of the graph, e.g., for n=100 we look at the slope at $n=50$, and obtain $1/(log_{10}(50) \times log(10)) \cong 0.256$. Therefore, with small sample sizes, this parameter will not be helpful to understand whether a species behaves panmictically or is limited by migration, which may be problematic for estimates of extinction later. We can visualize our expectation of the z_{MAR} under panmixia plotting the first derivative above (Fig. SII.2). Because—as we will show below—we do expect a power law relationship under a migration-limited scenario, z_{MAR} should theoretically not change with sample size. The graphical study of the (non-)linearity of the log-log plots between the number of mutations and area sampled should be diagnostic to this problem (We see for instance that *Pinus contorta* has a highly non-linear relationship, likely due to the use of ascertained intermediate frequency markers instead of genome-wide data, Fig. SIV.1).

Finally, we used msprime (Kelleher, Etheridge and McVean, 2016) to corroborate this finding (z_{MAR} being constant with respect to sample size) with simulations, simulating 1600 demes in a 40x40 grid of demes or populations of $N=N_e=1000$ that are completely panmictic (universal gene flow or dispersal, so this is equivalent to a single panmictic deme). We observed the z_{MAR} for $t=100...10,000$ generations in log_{10} increments. After this time, we sample $n=1...100$ individuals in increasingly large groups of adjacent demes. The range of estimates of z_{MAR} in these simulations was 0.07-0.15.

Fig. SII.2 indicates that the minimum average z_{MAR} even under panmixia would continuously increase with lower numbers of individuals of a species sampled. This is due to the fact that the site frequency spectrum is not fully sampled with small numbers of individuals. Therefore, we devised an approach to rescale z_{MAR}.

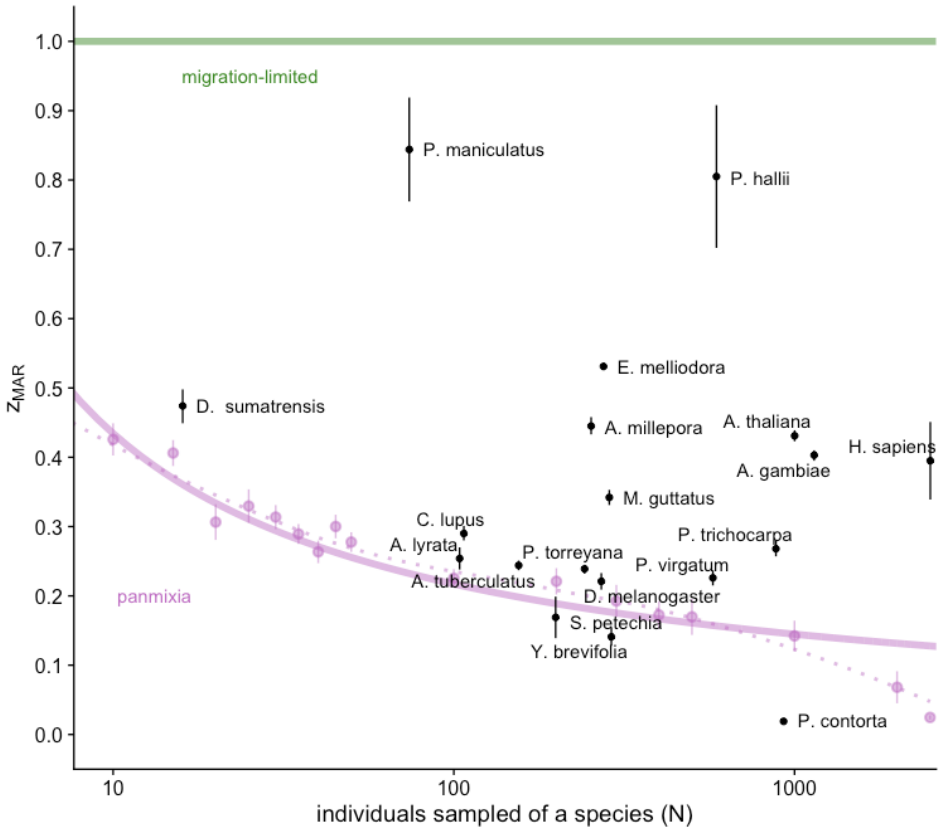


Fig. SII.2 | Expected ranges of z_{MAR} given sample sizes.
For increasing numbers of individuals sampled, we plot the expected mean z_{MAR} under two theoretical trends of a migration-limited (green) and a panmictic (purple) species (Purple dots indicate averages from SLiM simulations under panmixia to confirm the theoretical trend based on the derivative approach above). In black, z_{MAR} and 95% Confidence Interval of species analyzed in section IV are plotted (see section for details).

II.3.2 Scaling z_{MAR} for low sampling and low census size

Let $z_{pan\text{-}n} = E[z_{MAR} \mid n, panmixia]$, be the expected value of z_{MAR} of a panmictic species given that we only have small sampling of n. Although theoretically z_{MAR} should approach 0, with small samples it can be upwardly biased. In order to force the possible values of z_{MAR} to range 0-1 despite small sample sizes, we can scale it as:

$$z_{naive\ scaled} = (z_{MAR} - z_{pan\text{-}n}) / (1-z_{pan\text{-}n}).$$

In words, this moves the purple line in Fig. SII.2 to zero, stretching the space above it accordingly.

Most species have census sizes that large that z_{MAR} should indeed approach 0 under panmixia, so we should correct the sample estimate z_{MAR} to range 0-1. However, some species have such low census size N that even if we sample all individuals of a species, the sample size will still be small. In those cases, we should not scale z_{MAR} to range 0-1, but rather scale it from $z_{pan\text{-}N}$ - *1,* where $z_{pan\text{-}N} = E[z_{MAR} \mid N, panmixia]$ is the expected value of z_{MAR} given a census size N (plants or animals living in the wild). The updated scaling approach for both census and sample size would then be:

$$z_{scaled} = (1\text{-}z_{pan\text{-}N})\ (z_{MAR} - z_{pan\text{-}n}) \,/\, (1\text{-}z_{pan\text{-}n}) + z_{pan\text{-}N}.$$

Note that this scaled estimate must be conservative because while we adjust the minimum z for the average value expected for low sample sizes, we do not adjust for the maximum possible z, which only under very extraordinary theoretical conditions can be z=1, namely under an unrealistic complete disconnection of populations by gene flow (see below). Because deriving the maximum z would require more biological knowledge of the species' demography, landscape connectivity, genome structure, etc., and because we rather create conservatie estimates, we do not create further scaling approaches.

II.3.3 Meta-populations in space

A more realistic simulation than a panmictic population is that of the same 40x40 deme grid where migration can happen between adjacent demes. This migration rate can be changed to understand the effect of population structure and migration on z_{MAR}. Under no migration (or very low migration), we expect the mutations in two distinct populations (and thus their SFS) to be (almost) completely independent. Hence, when explored demes are doubled (N_e doubles), we discover twice as many mutations. In this case, the number of mutations should scale linearly with the area, so we expect the following to be true: $M=A$, $log(M) = log(A)$, and $z_{MAR}=1$. Our analyses under different sampling schemes, and with different numbers of "burn-in generations" (generations since a single deme colonized the full 40x40 space) confirm that z_{MAR} approaches 1 in the limit of high migration (see Table SII.1 and Fig. SII.3). Different from the panmictic situation, as we increase the sampled area, we not only increase n but also N_e.

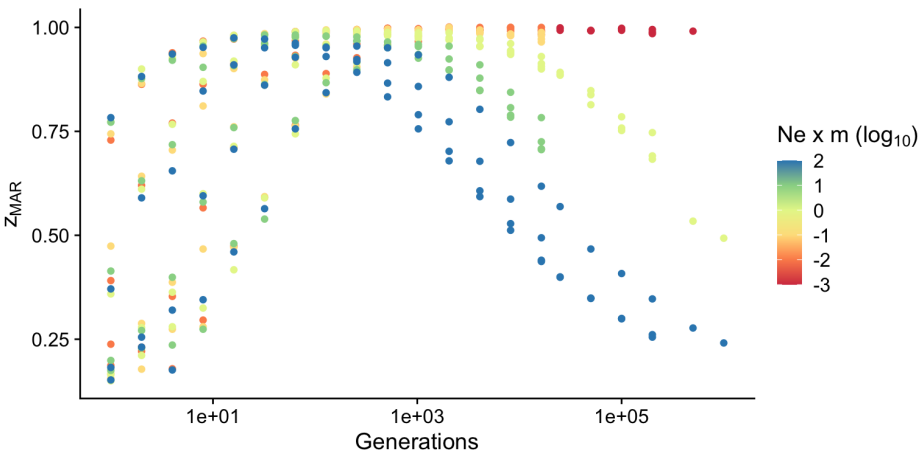


Fig. SII.3 | msprime 2D deme simulations and the mutations-area relationship
Simulations with different burn-in and migration rates under neutrality, and their corresponding zmar.

Table SII.1 | msprime population genetic simulations in 2D.
Simulations summarized by grouping ranges of the resulting z_{MAR} parameters. The average parameters of the simulations with similar z_{MAR} EW provided. (Acronyms: Nemt = product of effective population size, migration rate, and simulated generations).

z_{MAR}	Samples/deme	Generations	Migration rate	$N_e mt$
0.2 +/- 0.05	2.4	50001.7	0.0271675	5000044.23
0.3 +/- 0.05	20.25	70003	0.0561655	7000075.77
0.4 +/- 0.05	26.5714286	13057.4286	0.04450857	1305497.96
0.5 +/- 0.05	12.9230769	121759.462	0.04017769	752221.743
0.6 +/- 0.05	15.6111111	3218.77778	0.045735	321174.768
0.7 +/- 0.05	35.6842105	35034.8421	0.03395895	143791.614
0.8 +/- 0.05	35.030303	15655.1212	0.03055818	58023.5539
0.9 +/- 0.05	36.5806452	3057.12903	0.0253029	15290.4081
1 +/- 0.05	42.0140845	13625.4085	0.00861178	1798.36141

These simulations corroborated that we can recover z_{MAR} values ranging between 0-1 just varying migration and burn-in generation parameters. We found that it was both the time of the system to reach an equilibrium as well as the migration rate that determined z_{MAR}. In the future, it will be interesting to study different non-equilibrium scenarios to better understand how genetic drift, gene flow, and different landscape structures may shape the z_{MAR} .

II.3.4 Meta-populations in space with local adaptation

In order to simulate local adaptation, we use the individual-based simulation software SLiM (Haller and Messer, 2019) following the approach of (Booker, Yeaman and Whitlock, 2021). These simulations were set up for 196 demes arranged in a 14 x 14 grid. Each grid cell contains a population of $N=1000$ and has an environment attribute, e, which varied spatially from the lower-left to the upper-right corners (approx. $-7 < e < 7$). 12 locations in the genome were allowed to be under directional natural selection. The selection coefficient was fixed for a simulation, and grid runs were conducted with $0<s<0.05$, but this selection would vary based on the environmental selection value of a grid cell, according to $e \times s$. Therefore, these alleles are antagonistic pleiotropic. Selected mutations across the 12 loci in the genome behaved additively (e.g. if an individual in grid cell i had two of the selected mutations, fitness would be $w=1+2s \times e_i$). The migration rate varied from one individual in a billion (1×10^{-9}), to one individual every ten (1×10^{-1}). Finally, the mutation rate was set to 10^{-8} mutations/bp/generation and the recombination rate to 10^{-7} crossovers/bp/generation.

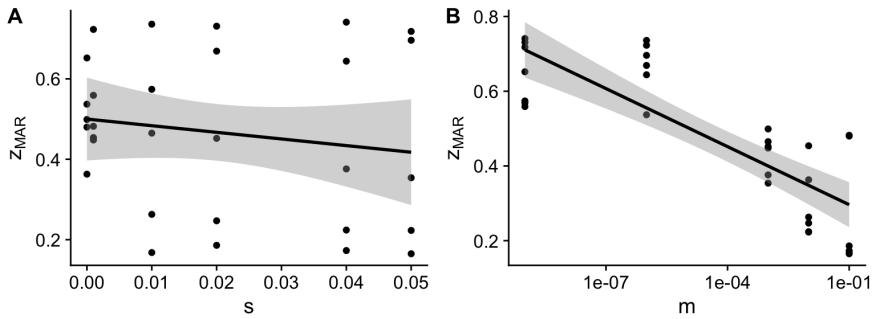


Fig SII.4 | SLiM population genetic simulations in 2D with selection and local adaptation
Simulations were carried out with different combinations of migration rates and strength of antagonistic pleiotropic selection at 12 QTLs. (A) Marginal relationship between z_{MAR} with the strength of spatially-varying selection s. (B) Marginal relationship between z_{MAR} with the migration rate m.

These results, together with individual-based simulations, corroborate what we had observed with Coalescent simulations, i.e. that z_{MAR} is lowest with a high migration rate. The simulations also

appear to show a negative effect of selection on z_{MAR} . Generating a linear model fitting migration rate and selection and their interaction to understand what factors explain the scaling coefficient: $z_{MAR} \sim log_{10}(m) + s + log_{10}(m)xs$ we confirm that both had a significant effect, and that selection significantly reduces z_{MAR} (Fig. SII.4, see below summary table SII.2). This may seem counterintuitive, as one may expect that locally-adaptive mutations are rare and will be localized only to where they are adaptive. More work is necessary to understand the signatures that spatially-varying natural selection (and its different types) create on z_{MAR}, but we can think that under migration limited scenarios (where z approaches 1) adaptive alleles and their linked mutations permeate faster to similar neighbor environments than neutral alleles.

Table SII.2 | Linear model explaining z_{MAR} by migration rate and strength of spatially-varying selection
Linear Model summary table

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3385022	0.0469174	7.214859	0.0000001
log10(mig)	-0.0419733	0.0085804	-4.891792	0.0000407
s	-4.6934926	1.6290184	-2.881178	0.0076725
log10(mig):s	-0.4998393	0.2426463	-2.059950	0.0491621

II.3.5. Meta-populations in space with purifying selection

To understand the effect of purifying selection on z_{MAR} we also ran 2D simulations with a fraction of the genome allowed to be globally-deleterious (i.e. independent of the spatially-varying environment). We simulated an increasingly strong purifying selection ($|s|$ range from 0.0 to 0.1), simulating roughly that 29% of the genome of Arabidopsis is coding (arabidopsis.org) and mutations can be deleterious. We also varied the degree of recombination. Following our expectation, with stronger purifying selection deleterious mutations are pushed to lower allele frequencies, stopping their geographic spread, which increases z_{MAR}. Recombination rate appears to have a minor role on z_{MAR} (Fig SII.5).

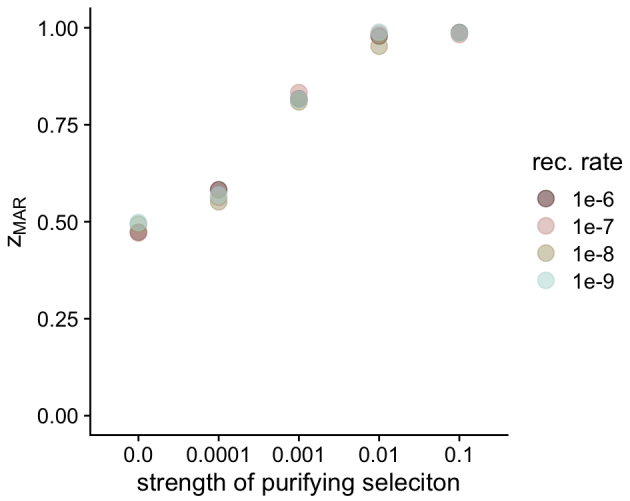


Fig SII.5 | SLiM population genetic simulations in 2D with purifying selection.
Simulations were carried out with varying strengths of purifying selection ($|s|$ range from 0.0 to 0.1) at coding positions, representing about 29% of the genome. Different values of recombination rate were also used in all pairwise combinations with $|s|$.

II.3.6 Continuous-space non-Wright-Fisher models

In order to confirm z_{MAR} generality in highly realistic conditions and its behaviour through the extinction process (II.4), we set up SLiM simulations using continuous space and non Wright-Fisher

dynamics. These are highly customizable and we chose parameters that involved the least assumptions and lead to realistic F_{st} across populations: dispersal kernel, local mate choice, spatial competition, age structure of population, local carrying capacity (github repo with scripts *in preparation*). Fitness, and thus selection, was modeled as a polygenic trait under stabilizing selection, and we kept track of variants that affected fitness during simulations. We ran simulations forward-in-time for 2,000 generations and distributed neutral mutations on the tree-sequence encoding of the data (Kelleher *et al.*, 2018). After that, the geographic distribution of the species experienced impacts as expected during global change: every generation a 0.001 of one edge of the species distribution got its carrying capacity reduced to 0. This meant that over 1,000 generations the whole species would disappear (note that this is a reasonable fraction of area reduction given the estimates of yearly deforestation and habitat change in section V).

Throughout these realistic simulations, at different timepoints before extinction we tracked important parameters such as *Va* (which corroborated the temporary inflation observed in section III.5) or z_{MAR} itself, which appeared to decrease as (Fig SII.2)

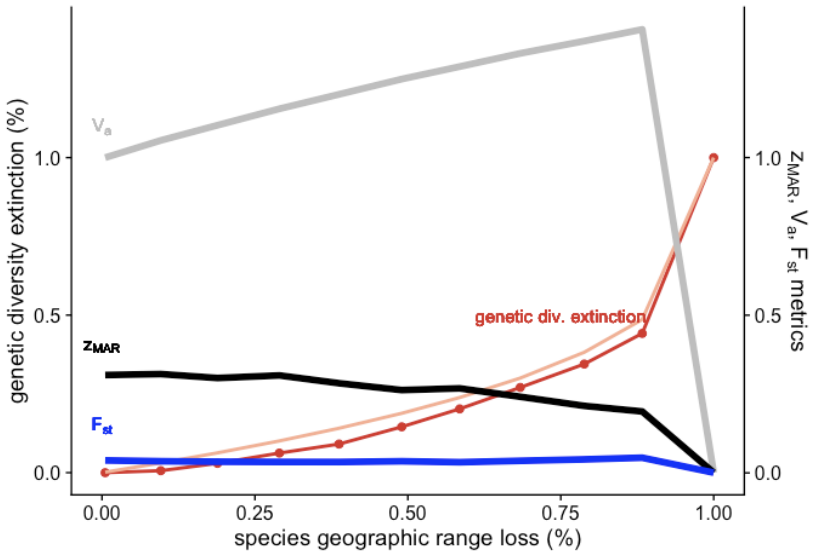


Fig. SII.6 | Continuous space SLiM population genetic simulations
At ten timepoints prior to final extinction, 1,000 individuals were sampled randomly in continuous space to quantify diversity extinction (dark red). The prediction of MAR (light red) using the starting z_{MAR} *seemed to follow the real trend accurately (dark red), indicating that even if* z_{MAR} *varies during the extinction process, it is relevant to understand genetic extinction by area reduction. We also tracked metrics of population structure (*z_{MAR}, F_{st}*) and a proxy of adaptive capacity (Va), which showed qualitatively similar patterns as the GWA-based trends (Fig III.10).*

II.3.7 Connection with isolation-by-distance

Ultimately, z_{MAR} is a complex integrator of evolutionary forces acting in space (mutation, migration, drift, selection) and captures how structured the distribution of a species' mutations is. Although the isolation-by-distance pattern conceptually resembles z_{MAR}, we have found no obvious analytical expression that relates both. Note that F_{st} is defined based on heterozygosity or π, instead of the number of segregating sites (i.e., mutations M). For instance, using Hudson's estimator (Hudson, Slatkin and Maddison, 1992) to compute Fst across a set of populations we calculate $F_{st} = 1 - (\pi_w / \pi_b)$, where π_w is the diversity or heterozygosity within a population and π_b is the same parameter calculated for the meta-population. Plotting F_{st} of a metapopulation by the distance of the farthest demes shows the typical non-linear trend of isolation-by-distance, which shows that very close populations have similar allele frequencies whereas populations further away drift apart. A challenge of Fst is that it requires pre-defining discrete populations, which is straightforward in stepping-stone simulations but

hard in real data. Comparing average F_{st} of our 14x14 spatial demes and z_{MAR}, we see that the two parameters correlate (Fig. SII.7C). However, it appears that for low values of F_{st}, z_{MAR} captures more variation across the simulations (Fig. SII.7). These patterns were also confirmed in continuous space simulations (not shown).

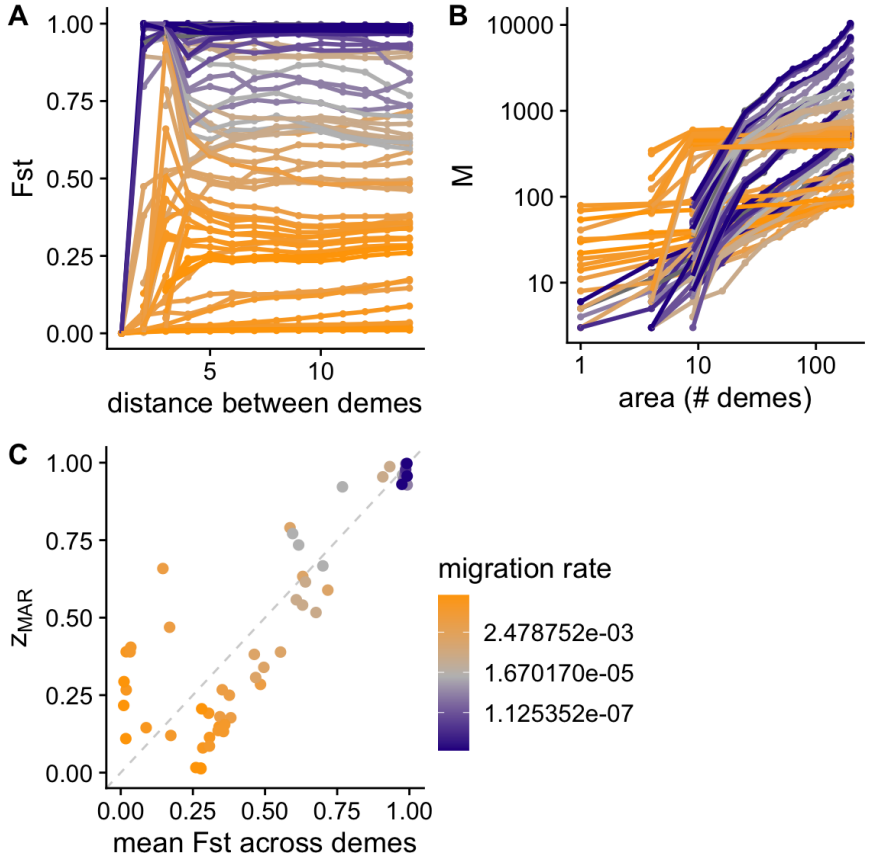


Fig SII.7 | SLiM population genetic simulations in 2D comparing Fst and z_{MAR}.
Neutral SLiM simulations with different degrees of migration. (A) Hudson's Fst across populations with different area subsamples. Following the expectation of the isolation-by-distance pattern, as the distance between the farthest demes in the subsample increases, Fst becomes larger and saturates at large distances. (B) The Mutations-Area-Relationship. (C) Comparison between Fst and z_{MAR}

II.4 The extinction of mutations in space

Having fit the MAR relationship as a function of area, and having estimated z_{MAR}, we can use this relationship to project the number of mutations (genetic diversity) lost as the geographic distribution (A) of a species is reduced by a due to habitat loss or climate change following equation:

$$X_M = 1 - \frac{MAR(A-a)}{MAR(A)} = 1 - \left(\frac{A_t}{A_{t-1}}\right)^z.$$

In the case of a single panmictic population, and again assuming $A \sim N$, the loss of area A by a (and fraction of area extinct $x=a/A$), would cause a minor damage, from: $M_{t-1} \approx log(A)$; to: $M_t \approx log(1-x) + log(A)$; which is also expected from the z_{MAR} framework as $z_{MAR} \approx 0$ for panmixia. A substantial loss of genetic diversity in this case only happens when population extinction is almost complete. This is of course not the case in highly structured populations, which is the other extreme scenario with $z_{MAR} \approx 1$, where the fraction loss of geographic area directly translates to the same fraction loss of genetic diversity.

Reality should be in between the panmictic and fully-migration-limited cases. With combinations of environmental selection, non-equilibrium demography, and long-range dispersal, we may get intermediate z_{MAR} values, and it will be empirical estimates that can inform us how much may be lost (Section III.3).

II.5 Recovery of genetic diversity after a bottleneck

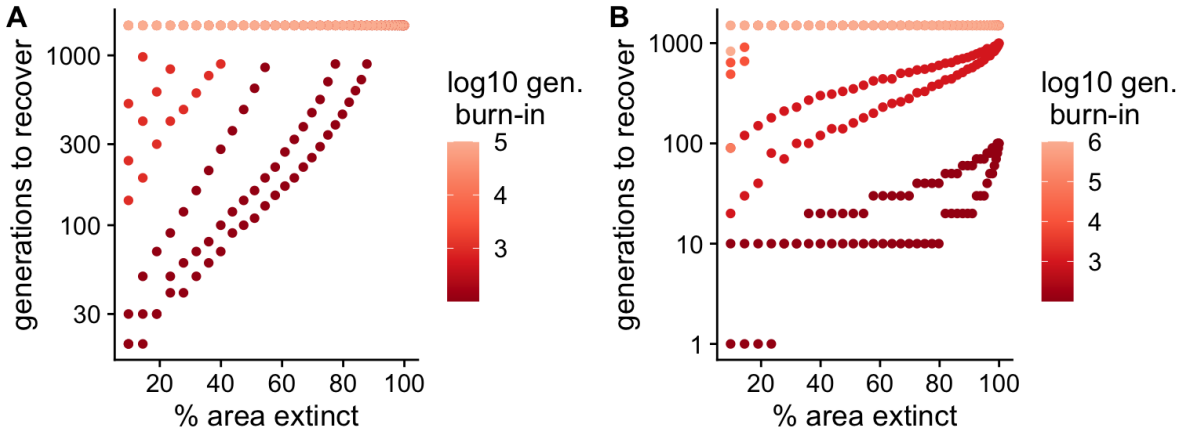


Fig. SII.8 | 2D stepping-stone msprime simulations with extinction of a fraction of the population and recovery
(A) Recovery of genetic diversity (number mutations) after extinction of a fraction of the population. (B) Recovery of genetic diversity after instantaneous extinction of a fraction of the population and consecutive repopulation.

The intuition that rapid recovery of genetic diversity may be possible is likely flawed. While genetic recovery may be faster than speciation rates, which are on the order of millions of years, the time for a set of populations that went through a simulation burn-in of 1,000 generations (not yet in diversity equilibrium), and that suffer an instantaneous 5% extinction of area and an instantaneous recovery (e.g., through reforestation) would range from 20-90 generations. This number of generations for long-lived species would translate into centuries or millennia of recovery without further impacts. About 49% of simulations – including every simulation that reached equilibrium (burnin generations >10,000) – have a recovery time of more than a thousand generations.

II.6 Similar attempts to describe a genetic analog of the Species-Area Relationship and related work

Although the Mutations-Area Relationship, to our knowledge, has never been empirically tested to follow a power law, nor formally derived through population genetics, here we want to highlight some similar concepts and their differences:

Campos, Oliveira, and Rosa (2010) created a customized simulator of species adapting in a heterogeneous landscape, with mutations with fitness effects depending on the environment epistatically interacting with other mutations. To study genetic differentiation with area, they used average pairwise differences (unscaled heterozygosity or θ_π) as a metric of speciation. Although there are no barriers to cross-species hybridization, the authors indicate that the parameter space of habitat heterogeneity and their reproduction scheme make genetically-different individuals effectively isolated. Then, the authors built a power law Species-Area Relationship with their genetic proxy of species diversity that recovered typical z_{SAR} values.

Fan and colleagues (2019) attempted to describe the differences in genetic diversity across bird species with different geographic distribution ranges. They focused on heterozygosity, although in their paper they also tested allelic richness (the equivalent of our metric of number of mutations or segregating sites used here too). The data analyzed are microsatellites from various studies, known to be ascertained to be variable and thus would not be expected to follow the 1/q probability distribution

of the SFS. Then the authors fit the relationship between phylogenetically-corrected geographic distribution range area (x) and body-size corrected heterozygosity or allelic richness. They found that the best-fitting relationship was a Monod equation. Although they tested a power law, their scaling was z=0.01 for heterozygosity and z=0.05 for allelic richness, as expected due to the ascertained SFS characteristic of microsatellite markers. The authors called this the Genetic Diversity Area Relationship (GAR). We note this is a different relationship to what we describe here for two reasons: (1) it focuses on heterozygosity or highly-variable sites rather than the number of segregating sites genome-wide, (2) it quantifies a spatial relationship across all species and their different geographic ranges rather than a relationship for each species. Therefore, it is hard (or not impossible) to derive the empirical GAR from Fan et al. from population genetics models, nor does the GAR informs about species-specific genetic diversity-area relationships.

Finally, while not explicitly aiming at describing a genetic diversity-area relationship, Buffalo (2021) revisited a paradox in population genetics that levels of heterozygosity, π, are fairly consistent across species relative to their effective (N_e) or census (N_c) population sizes. The aim of their paper is to study whether linked selection can explain the lack of a relationship between $\pi \approx 4N_c\mu$ (where μ is the mutation rate). By using a species range-based estimate to approximate population size, and correcting both diversity and population size by phylogenetic signal, the paper built a relationship that resembles the genetic diversity-area relationship from Fan et al. The 95% confidence intervals of the scaling parameter in a log-log plot relationship between π and approximate population size (z equivalent) from a phylogenetic mixed-effect model was: 0.03-0.11.

II.7 Notes on conservation genetics state-of-the-art

The literature of population genetic applications to conservation is extremely rich, since a series of foundational papers of Lande and Lynch was published in the 1990s. These focused on the processes that occur when populations become too small: Small populations can accumulate more deleterious alleles due to elevated genetic drift (Lande, 1993, 1995; Lynch and Lande, 1998). Declining or fragmented populations increase inbreeding and thus demographic depression (Franklin and Frankham, 1998; Kyriazis, Wayne and Lohmueller, 2020; Kardos *et al.*, 2021), due to strong recessive deleterious mutations (Kyriazis, Wayne and Lohmueller, 2020) and excessive mutational load (Willi, 2013). We consider this work as extremely illuminating and inspiring, and an important downstream process in the extinction trajectory. The purpose of our work, however, is understanding the process of loss of genetic variation, whether it is adaptive, neutral, or deleterious, which happens when species lose area (both endangered or non-endangered). Further, MAR is a scaling relationship that is phenomenological, and forward-in-time predictions do not incorporate the feedbacks that occur during population size reduction such as the mentioned above. Our work, however, fills a technical gap that could have important consequences to current United Nations' Convention on Biological Diversity agreement and its Sustainable Development Goals, which recently proposed to preserve 90% of genetic diversity within each and all species (Díaz *et al.*, 2020).

III. Testing the mutation area relationship theory with the 1001 Arabidopsis Genomes.

We begin testing the idea of a general mutations-area relationship using the extensive sampling of the model plant species *Arabidopsis thaliana* and the 1001 Arabidopsis Genomes Project (1001 Genomes Consortium, 2016). This section will serve as a case study to explore different approaches and biases when building MAR to then apply the learned lessons across species (section IV).

III.1 The Site Frequency Spectrum of the 1001 Arabidopsis Genomes.

We began analyzing the frequency distribution of 11,769,920 biallelic genetic variants (i.e., mutations), which is typically called the Site Frequency Spectrum (SFS) in population genetics.

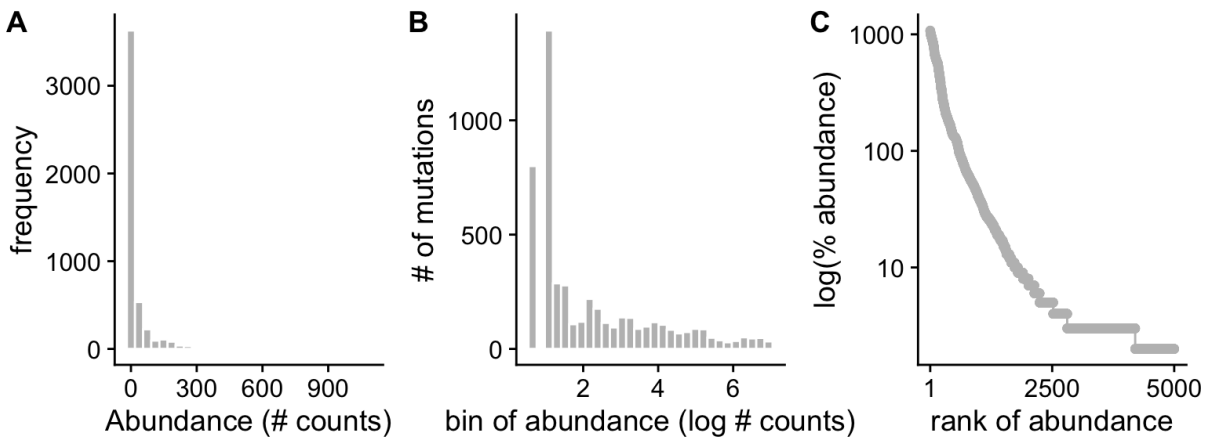


Fig. SIII.1 | Mutation abundance study in A. thaliana.
(A) Site Frequency Spectrum (SFS). (B) Preston plot of mutation abundances. (C) Whittaker plot of mutation rank abundances.

To showcase the similarities to the Species Abundance Distributions (SAD), we use the Whittaker plot of mutation rank abundance (Fig. SIII.2) suggests a log-normal of S-shape may be the best fitting model (Table SIII.1). For a review listing many popular models, see (McGill *et al.*, 2006), and for implementation details of 13 SAD models see the thorough manual of R package SADS (Prado, Miranda and Chalom, 2018). As we shall see later, the log-normal distribution seems to be the best fit across species.

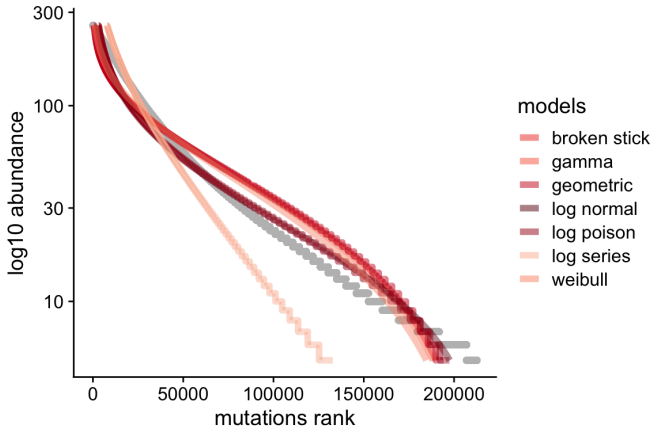


Fig. SIII.2 | Fit of mutation abundance study in A. thaliana with different SAD models
Representative models from Table SIII.1 are plotted along with the observed frequency of 11,769,920 mutations

Table SIII.1 | AIC values for model fit of common species distribution curves.
For each SAD model, the degrees of freedom and the delta AIC compared to the top model are reported.

Model	dAIC	df
log-Normal	0	2
Poisson	7204.37509	2
Geometric	44267.5475	1
Weibull	45872.3678	2
Gamma	48805.6065	2
Borken Stick	49076.4368	0
UNTB (MTZSM)	168434.181	1
log-Series	168434.726	1

III.2 Building the Mutations-Area Relationship

In the following, we explain how the area was estimated that was used to compute z_{MAR} on real world data. In short, we used a grid on the world map, with samples placed on the map based on their geo-coordinates of origin (Fig. 1). We first create square spatial subsamples of the *Arabidopsis thaliana* geographic distribution (Fig. 1, Fig. SIII.4) and quantify diversity *M* as the total segregating sites. Excluding zeros, these two variables are fed to the sars_power function from the R SARS package (Matthews *et al.*, 2019).

Although the power law Mutations-Area Relationship was already theoretically motivated (II.3), here we also fit different types of functions typically applied to the Species-Area Relationship. Doing this, we reach the conclusion that multiple models perform very similarly, and the classic power law is among the top models, see Table SIII.2. Although small marginal fitting accuracy could be achieved with other models, for mathematical convenience and historical continuity, we use the power law for later sections and the study of MAR across species (Sections IV and V).

Table SIII.2 | Different SAR curves fit to mutations.
We fit 20 different functions and calculated the variance explained (R2), Pearson's r, and Spearman's rho

Model	R2	r	rho
Asymptotic regression	0.21825683	0.46717965	0.53510077
Beta-P cumulative	0.22012799	0.46917799	0.53374757
Chapman Richards	0	NA	NA
Cumulative Weibull 3 par.	0.21929646	0.468291	0.53374757
Cumulative Weibull 4 par.	0.21930145	0.46829633	0.53374757
Extended Power model 1	0.21833611	0.46726449	0.53026812
Extended Power model 2	0.21682584	0.46564561	0.53462775
Gompertz	0.16393078	0.40488366	0.45964364
Heleg(Logistic)	0.21929721	0.4682918	0.53531975
Kobayashi	0.22228406	0.47147011	0.53526975
Linear model	0.19579007	0.44248171	0.53510077
Logarithmic	0.20280401	0.45033767	0.53430311
Logistic(Standard)	0.22536996	0.47473146	0.53549765
Monod	0.22500999	0.47435217	0.53579276
Negative exponential	0.22801633	0.47751055	0.53447179
Persistence function 1	0.21929612	0.46829063	0.53501182
Persistence function 2	0.21760028	0.46647645	0.53409266

Power	0.21929556	0.46829004	0.53543785
PowerR	0.21753225	0.46640353	0.53493321
Rational function	0.22072491	0.46981369	0.53451874

Because in the species literature it is recommended to only quantify richness of endemic species (He and Hubbell, 2011), we also count segregating sites that are private to the area subsample, creating the equivalent Endemic-Mutation Area relationship (EMAR) (He and Hubbell, 2011). The MAR slope and 95% Confidence Interval was z = 0.324 (0.238 - 0.41), while the EMAR was z = 1.241 (1.208 - 1.274). Interestingly, the endemics-area relationship of $z \approx 1$ resembles that of endemic species, whereas the total mutation relationship with area is above that of species relationships, which typically follows the canonical $z \approx 0.2 - 0.4$.

We must note that EMAR, the genetic analogy of the Endemic-(species)-Area Relationship (EAR) may not be that meaningful when analyzing genomic data (we did not find a way to theoretically motivate it in section II), and later we see it overestimates extinction in our simulations (Fig SIII.7)

Table SIII.3 | Mutations-Area Relationship (MAR).
Fit values in a log-log power function between area sampled and mutations discovered.

	Estimate	Std. Error	t value	P	2.5%	97.5%	nls.Est.	nls.2.5%	nls.97.5%
c	494.565432	135.6314588	3.646392	0.0003138	223.3025141	765.8283493	494.5531270	278.1107276	822.829918
z	0.323727	0.0430277	7.523681	0.0000000	0.2376715	0.4097824	0.3237367	0.2430303	0.413162

Table SIII.4 | Endemic-Mutation Area Relationship (EMAR).
Fitted values in a log-log power function between area sampled and endemic mutations discovered.

	Estimate	Std. Error	t value	P	2.5%	97.5%	nls.Est.	nls.2.5%	nls.97.5%
c	0.0001001	0.0000231	4.337758	1.98e-05	0.0000539	0.0001463	0.0001001	0.0000635	0.0001555
z	1.2411831	0.0165268	75.101442	0.00e+00	1.2081296	1.2742366	1.2412125	1.2096087	1.2737927

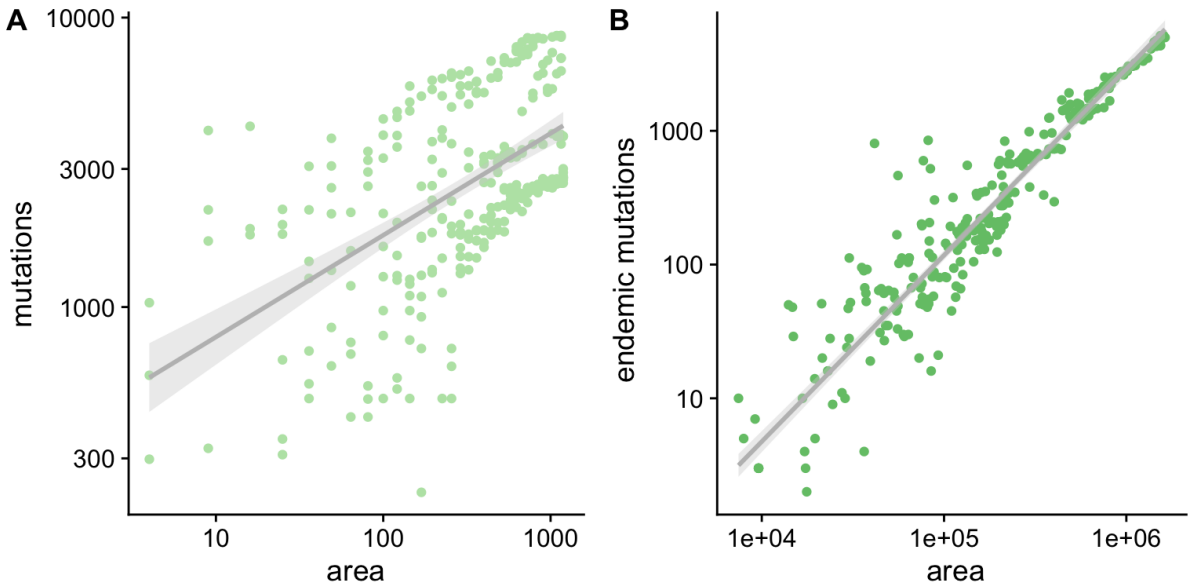


Fig. SIII.3 | Mutation Area and Endemic-Mutation Area Relationships in A. thaliana.
Dividing A. thaliana native distribution into a 1 degree lat/long grid, square areas with 1 degree side-length to 36 degrees side-length were randomly placed (n=100 for each size) across the distribution, and genetic diversity metrics were computed to produce the (A) Mutations-Area Relationship and (B) Endemic-Mutations Area relationship.

III.3 Testing numerical artifacts

We wondered whether MAR estimates may be affected by some numerical artifacts in our software pipeline (available at https://github.com/moiexpositoalonsolab/mar). For instance, real world data may have uneven sampling in space, the spatial resolution of GPS-tagged samples may vary, projection of samples into gridded maps may have limited resolution, software pipelines may produce biased estimates, etc. To test this, we conducted several experiments:

Lower bound of the method for z_{MAR} Our first experiment when building the MAR aimed to make sure that bias in the spatial sampling or genome sequencing would produce an artificially large z_{MAR} when the relationship should approach zero. We then took a permutation and simulation approach, where the SNPs in the *A. thaliana* dataset were shuffled across individuals. The same number of mutations, samples, and geographic locations were kept but the MAR relationship should be null. This exercise confirmed we get a value approaching zero: z=0.033, (-0.095 - 0.162) (Table 1, Table SIII.5*)*.

Table SIII.5 | MAR built with different area calculations and grid sizes

Grid resolution (deg.)	z_{MAR} [CI95%] (cell area)	z_{MAR}[CI95%] (total area)
A=N	0.431 (0.423 - 0.439)	NA
0.1	0.435 (0.424 - 0.446)	0.367 (0.281 - 0.454)
0.25	0.454 (0.449 - 0.459)	0.422 (0.376 - 0.467)
0.5	0.488 (0.465 - 0.511)	0.352 (0.152 - 0.551)
1	0.543 (0.529 - 0.558)	0.389 (0.295 - 0.483)
2.5	0.644 (0.6 - 0.688)	0.388 (0.251 - 0.526)
5	0.617 (0.205 - 1.029)	0.403 (-0.204 - 1.011)

Grid sizes, area calculations, and non-random spatial sampling. In order to streamline geospatial operations, we implemented the MAR relationship calculations in this project using R raster objects (van Etten, 2012). This required projecting the collected samples of a species and the observations of any given mutation into a world map (i.e., each mutation's geographic distribution). Necessarily, in order to be able to assign areas to sets of samples or mutations on the map, the projection requires the choice of a grid size. The larger the grid size (e.g., lower spatial resolution), the faster the spatial operations can be performed computationally. Further, for larger grid sizes, we expect the slope of MAR to be more influenced by larger-scale patterns, while for smaller grid sizes, the MAR will be influenced by smaller-scale patterns. To test this, we repeated the subsampling of *A. thaliana* distribution with grid sizes ranging 0.1 degrees latitude/longitude (roughly 10km side-length in temperate regions) to 10 degrees (roughly 1,000 km side-length). The estimates were roughly consistent between 0.4-0.6, which resembles that of species in ecosystems (Storch, Keil and Jetz, 2012), including that at larger geographic scales (row in Table SIII.5 for large grid size values), the z appears to be steeper (e.g. 0.6).

Because we often have sparse samples of individuals in space, we devised two strategies to calculate areas during the subsampling of MAR (see cartoon in Fig. SIII.4): (A) the total square area of the minimum and maximum latitude/longitude values of all the samples analyzed. That is, simply the area of the red box in the figure. (B) the sum of areas of grid cells that contain at least one sample. That is, the sum of the grey squares within the red box in the figure. In addition, we also calculated the MAR relationship assuming the total area is equal to the number of individuals ($A=N$), which should be equivalent to very high grid resolution.

Table SIII.5 values suggest there is a dependency of z_{MAR} with the grid size when areas are calculated as the sum of grid cells with at least one sample. Our intuition for this pattern is that lower resolution grids (e.g., 5 degrees side) lead to some grid cells having many samples, which would increase the number of mutations discovered when discovering the area. On the other hand, the calculation of z_{MAR} using the total area does not seem to affect the z_{MAR} estimate; however, because large areas often do not have samples (limiting the potential to find new mutations), it creates a higher variance in the estimate of z_{MAR} (see confidence intervals in Table SIII.5 and Fig. SIII.5). Here, we favored consistency of z at the expense of broader, more conservative confidence intervals. All the estimates reported below and in the main text therefore use the total area approach.

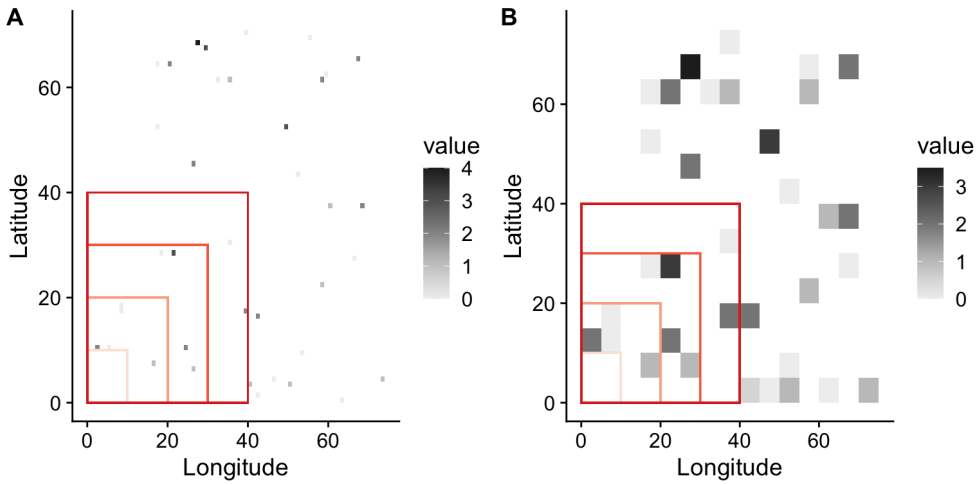


Fig. SIII.4 | Cartoon of raster sampling to build the MAR
Map of mock samples of a species projected into a raster. Grey scale indicates the number of samples per grid cell. Red boxes exemplify the process of spatial subsampling of increasing area to build the MAR relationship. Two example grid sizes were created for illustrative purposes: (A) Small grid size or high spatial resolution. (B) Large grid size or low spatial resolution.

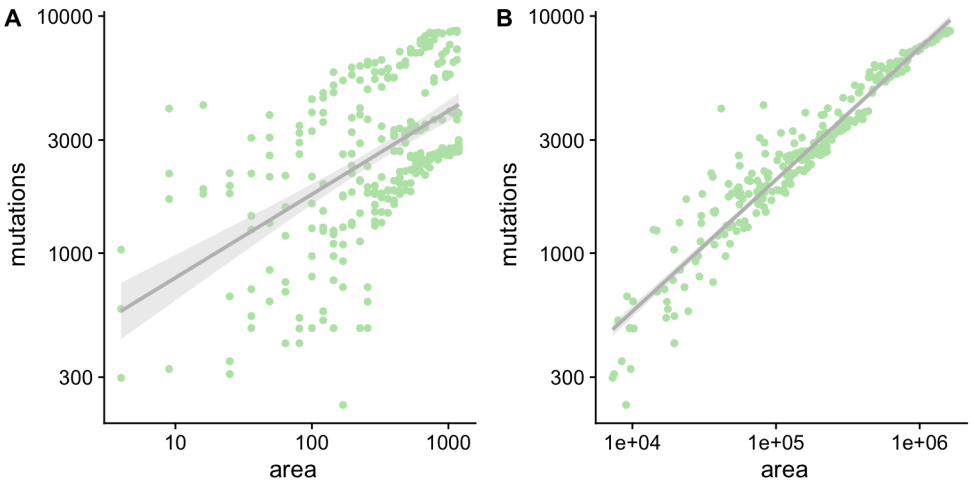


Fig. SIII.5 | MAR comparison with different area calculations.
(A) Using total area, (B) using grid cell sum with at least one sample.

Geographic subsampling strategy (inwards, outwards, random). It has been indicated that the way the Species-Area Relationship (SAR) and Endemics-Area Relationship (EAR) are created may create differences in the scaling parameter z. The plots and estimates above were produced by

randomly placing boxes of different size or area across the distribution of the species. Often, however, either discovery of species or extinction happen in certain patterns. For instance, we often imagine sampling an ecosystem concentrically outwards from a focal point, whereas we may think of the extinction process of species area reductions being concentrically inwards (He and Hubbell, 2011). Because these patterns seem of importance, we also calculated the MAR and EMAR outwards from the latitude and longitude median of all the samples in the map, moving outwardly until the map is filled. Likewise, the inward pattern is conducted in an inverse manner. See Fig. 1F of the main article for examples of further patterns.

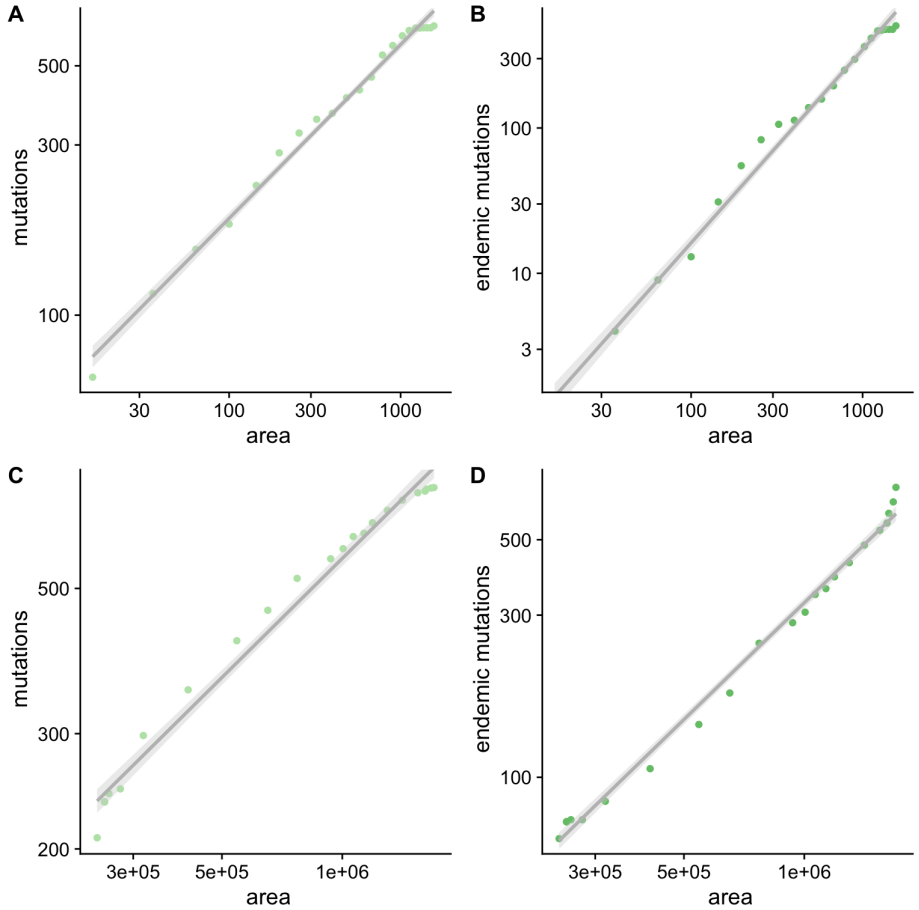


Fig. SIII.6 | MAR and EMAR in Arabidopsis thaliana using outward and inward sampling.
Dividing A. thaliana native distribution in 1 degree lat/long grid, a square area of 1 degree was placed at the median of the sampling range and was expanded iteratively by 1 degree lat/long until all the area of the distribution was covered. (A-B) MAR and EMAR using a typical outward sampling. (C-D) MAR and EMAR using an inward sampling. The latter may not be a common process of sample collection, but it is common for extinction progress.

Table SIII.6 | Outward and inward MAR and EMAR
The MAR and EMAR relationship computed with inward or outward nested subsampling, calculating area only as those cells with samples.

Relationship	*z*
MAR outwards	0.444 (0.412 - 0.476)
EMAR outwards	1.086 (0.982 - 1.189)
MAR inwards	0.561 (0.524 - 0.597)
EMAR inwards	1.295 (1.192 - 1.399)

Incomplete sampling of the species. To check whether the relationship holds with few individuals of a species or limited geographic distributions, we compared the species-wide MAR with that of subset populations. Downsampling the native distribution of *A. thaliana* to a region within North-East Spain (-2.00–4.25 degrees East, 36.52–42.97 degrees North), or to a region within Germany (2.69–13.73 degrees East, 50.0–52.0 degrees North), and using only 1,000 SNPs, we recovered z_{MAR}= 0.423(0.233-0.614) for Spain and 0.525(0.242-0.807) for Germany, which were close to the estimate based on the whole distribution (Table 1). This result is reassuring in that if migratory patterns are relatively homogeneous, one may be able to estimate this parameter from a subset of the species.

Number of genome-wide SNPs used. To check whether different numbers of SNPs used for the analyses would lead to different z_{MAR}, we conducted analyses with random subsets consisting of 100, 1,000, 10,000, and 1,000,000 SNPs, replicated 3 times. Estimates were tightly around the mean 0.5301 with a coefficient of variation of 0.0372348 in z_{MAR}.

Locally-adaptive variants. We then aimed to understand the effect of utilizing SNPs that appear to be related to adaptation. To study this, we utilized an outdoor climate-manipulated experiment that recorded fitness data (survivorship and reproduction output of seeds) for 517 *Arabidopsis thaliana* ecotypes part of the 1001 Genomes set in 8 environments (Exposito-Alonso, 2019). We devised two sets of alleles: 10,000 that were negatively correlated with fitness in a Genome-Wide Association across 8 different environments, and 10,000 alleles that were associated positively with fitness in one environment but negatively in another (antagonistic pleiotropic). The MAR relationship was computed as before and compared to the original random (putatively neutral) set of alleles from the previous sections (Table SIII.7). Although we see a trend that locally-adaptive alleles have a slightly higher *z*, estimates overlap. The effects seen here of having smaller *z* for adaptive alleles than neutral variation could, however, be due to top GWA SNPs often being ascertained to higher frequency than background SNPS.

Table SIII.7 | MAR for putatively neutral, deleterious, and locally adaptive alleles in Arabidopsis thaliana

SNP set	*z*
neutral	0.324 (0.238 - 0.41)
globally-deleterious	0.209 (0.13 - 0.288)
locally-adaptive	0.291 (0.217 - 0.365)
globally-positive	0.234 (0.137 - 0.332)

III.4 Local extinction in Arabidopsis

Using the MAR framework, we can make projections of extinction of mutations (or its inverse, the remaining genetic diversity. By doing this, the known intuition is that with $z > 1$ (as from EMAR) the decrease of diversity is much faster than the decrease of habitat, but with $z < 1$ (as from MAR), there is a (desirable) slower dynamics of genetic extinction. In the latter, despite habitats disappearing, reservoirs of mutations distributed across different locations enable conservation of certain variation. To study which one is more likely and to observe the stochastic nature of extinction, we simulated in silico extinction of map cells from the Arabidopsis map (Fig. 1) and directly estimated from the genome matrix of remaining individuals the remaining genetic diversity. These simulations were implemented to capture different hypothesized patterns of extinction (see main text). All, however, agree with the more hopeful estimate of $z_{MAR} \approx 0.25$.

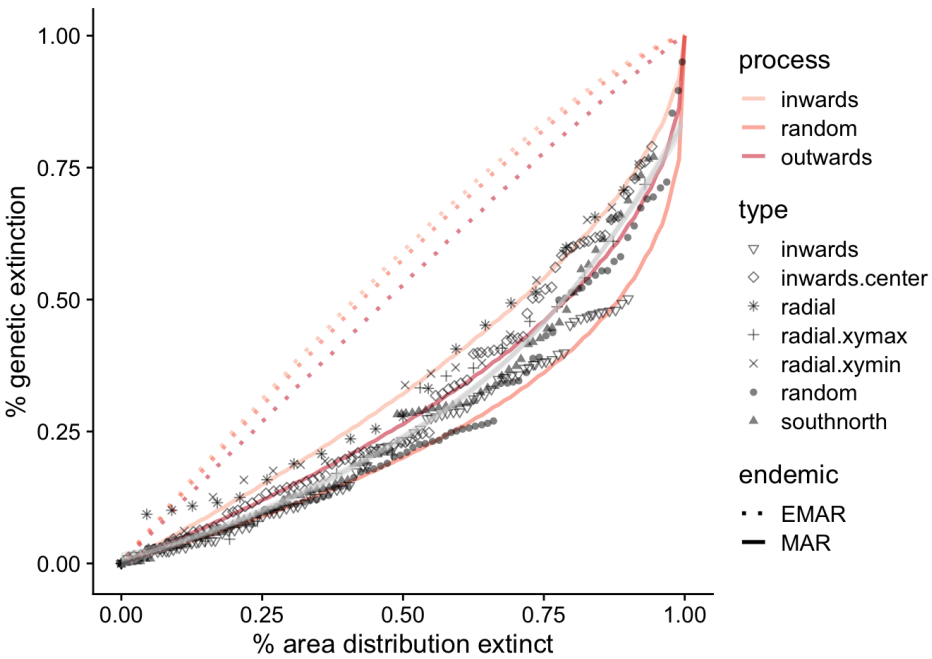


Fig. SIII.7 | Extinction of mutations with habitat loss in A. thaliana.
Predictions based on MAR and EMAR functions and in silico extinction stochastic simulations in A. thaliana.

To study the fit of the extinction predictions based on MAR relationships and the results from extinction simulations, we calculated a pseudo-R^2 based on the squared differences between the predicted line and the "observed" extinction as: $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$. This results in a high fit R^2=0.872 of the MAR, built from random samples of distribution areas, while the EMAR had a poor fit due to overestimation of extinction: R^2=-0.710 (negative values indicate predictions are worse than the mean of the data).

III.5 Potential impacts of genetic extinction in adaptability

Mutations may differ in evolutionary relevance (Rockman, 2012), but the methods described here to estimate genomic diversity loss are agnostic to these. To use the exact analog of species richness, here each mutation counts to the total genetic diversity, whether or not they create a novel morphological or physiological trait, are advantageous in an environment, are neutral, or even semi-lethal. Indeed, analyses expect the majority of mutations to be neutral, i.e. they do not have any cellular, morphological or physiological effect, but while they may be neutral in the present, they could be the key of evolutionary novelty in future environments, as expected in evolutionary rescue theory (Orr and Unckless, 2014). Alternatively, deleterious mutations may be found at lower frequency (Simons *et al.*, 2018). Many factors, including the architecture of traits under selection, the natural selective forces applied by spatially-varying environments, and others, may play a role in dictating what genetic variation may be most relevant to strategies of assisted migration and rescue of threatened populations, and since often this will be unknown (but see (Kyriazis, Wayne and Lohmueller, 2020)), it would be crucial to use genome-wide variation in conservation (Kyriazis, Wayne and Lohmueller, 2020; Kardos *et al.*, 2021).

Although likely imperfect, Genome-Wide Associations could help to understand the relevance of mutations in different frequency classes in model organisms such as Arabidopsis thaliana. Fig. SIII.7 shows the site frequency spectrum and a metric of the "total accumulated effect in fitness" of the alleles in every bin. Effect sizes were retrieved from GWA on lifetime fitness of 515 ecotypes in outdoor experiments (Exposito-Alonso *et al.*, 2019). The average effect size across 8 fitness GWA

from 8 experimental combinations were used: high/low precipitation, high/low latitude of outdoor stations, and high/low plant density. This exercise showcases the phenomenon that low frequency variants often have strong effect sizes, which is expected under a stabilizing selection quantitative model (Simons *et al.*, 2018). Because low frequency alleles will be the first to be lost during a bottleneck (as would happen with the rapid extinction of populations of a species), we may expect to lose variants that are related to fitness and thus potentially lose diversity that could be advantageous in some environments.

To further build intuition on the progress of extinction in relation to genetic diversity that is not neutral, we repeated warm edge extinction simulations with several subsets of alleles: randomly selected SNPs, SNPs that were associated positively in 2 environments (low precipitation Spain and high precipitation Germany) (labeled globally positive), and SNPs that were associated positively in one environment and negatively in the other (labeled antagonistic pleiotropic or putatively locally-adaptive). This Fig. SIII.8 supports our intuition that although these alleles may have slower extinction dynamics than neutral variants due to a high frequency and z_{MAR} , certain extinction patterns can actually lead to rapid extinction of potentially-adaptive genetic diversity. The complexity of these patterns, together with the evolutionary feedback created by lowering genetic standing variation that affects fitness, make the inference of adaptive capacity loss even more difficult than just inferring the extinction of genetic diversity itself.

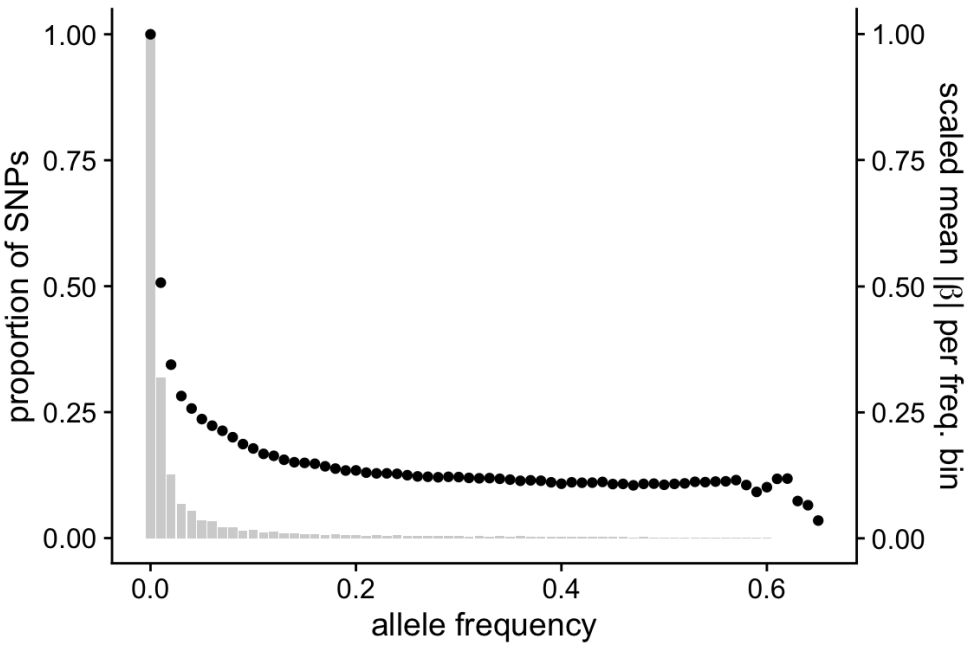


Fig. SIII.8 | Bias of low frequency mutations and effect size for fitness traits in A. thaliana.
Grey bars represent the site frequency spectrum (scaled for visualization purposes). The black dots represent the mean absolute effects of alleles as estimated from GWAs with 515 accessions scored for fitness traits in 8 outdoor experiments.

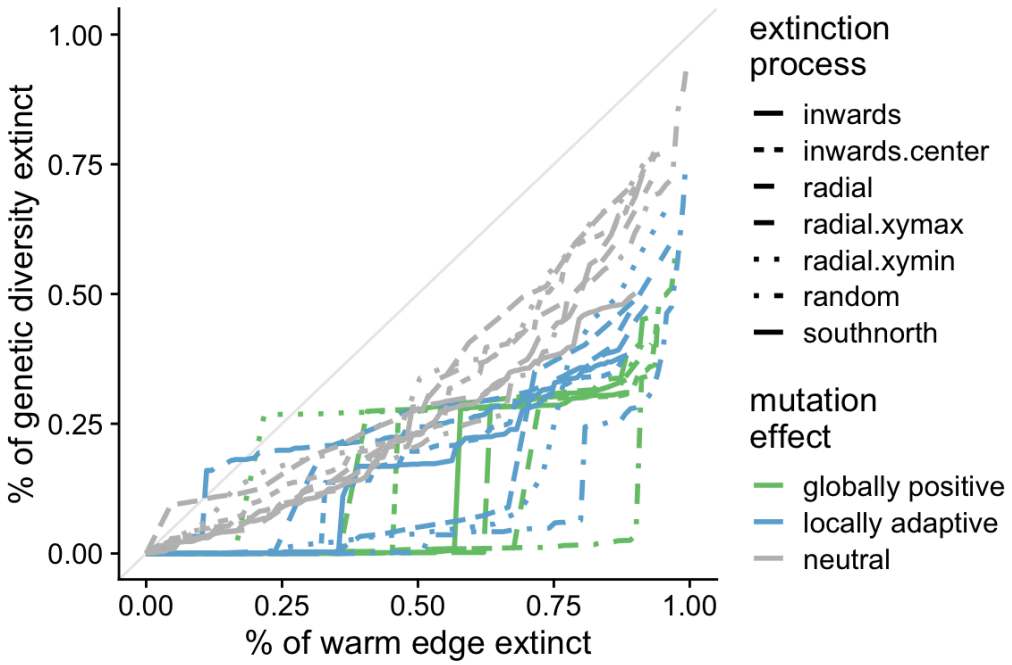


Fig. SIII.9 | Simulations illustrating the potential extinction of locally-adaptive mutations in A. thaliana.
Simulations of extinction using the warm edge pattern with different subsets of alleles ascertained to show positive associations in fitness GWA at least in two outdoor experiments (red), to show negative associations in fitness GWA at least in two outdoor experiments (dark red), to show positive associations in fitness GWA in one environment (low precipitation) but negative in a second environment (high precipitation) or vice versa (green). These were compared to a random set (light blue).

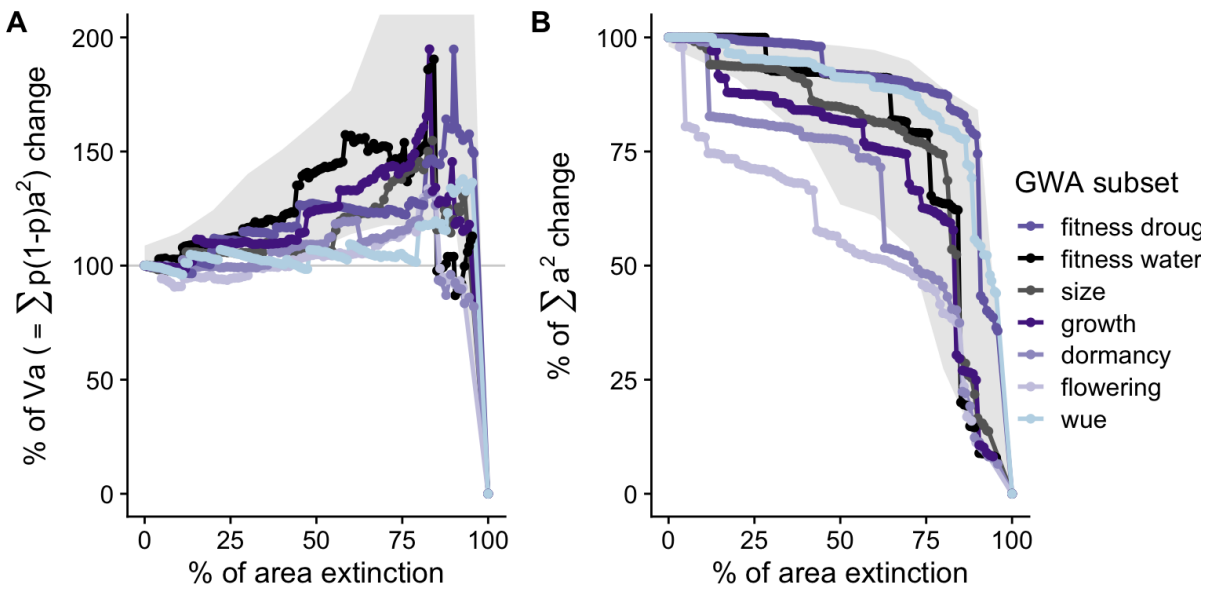


Fig. SIII.10 | Extinction simulations as in Fig. SIII.7 showing proxies of adaptive capacity of A. thaliana.
Using estimated allele effect sizes from 10,000 SNPs in the 1% P-value tails of several Genome-Wide Associations, we show (A) Percentage of change of Va as a proxy of adaptive potential ($= \sum p(1-p)a^2$) and (B) raw square sum of allele effects to showcase the inflating effect of intermediate frequency alleles. Grey background indicates the minimum and maximum values created by a frequency-matched non-effect set of SNPs as the effect SNPs for every GWA.

III.6 Case study of a massive natural bottleneck

A recent colonization of North America by *Arabidopsis thaliana* can help us understand the recovery of genetic variation. Whole-genome sequencing of 100 specimens of North American *A. thaliana* indicates that it migrated from its native range of Europe to North America in the 17th century, and began spreading across the continent from a genetically-homogeneous population (Exposito-Alonso *et al.*, 2018). Despite ideal conditions to re-gain genetic diversity—a continental population expansion aided by human travel (Seebens *et al.*, 2015, 2017)—only ~8,000 new mutations were detected through spontaneous accumulation, equivalent to only ~0.067% of the species-wide native genetic diversity. Because most of these mutations are at very low frequency, as expected during population expansion, the scaling of genetic diversity with area is approximately 1 (z_{MAR} = 1.025 [CI95%: 0.878 - 1.173]).

IV. The mutations-area parameters for diverse species

Every dataset was retrieved online either from the published article in the form of VCF or fastq files, or provided by the study authors upon request. All datasets were first transformed into PLINK files using PLINK v1.9 (Purcell *et al.*, 2007). For computational efficiency, and since we showed random subsampling does not appear to affect calculations of z_{MAR} (Section III.3), we conducted all analyses with up to 10,000 randomly selected SNPs for each species sampled genome-wide, or in the largest chromosome for those species with large genomes. We aim to use mostly unfiltered SNP datasets to avoid ascertainment biased toward intermediate frequency SNPs, and therefore we did not apply a MAF filter for any analyses. By default, PLINK transforms SNP matrices into biallelic (if multiallelic, it takes the two most common alleles). Although the preservation of structural genetic variation may also be relevant and may have important consequences in adaptation (Mérot *et al.*, 2020), we do not expect dramatic differences in their scaling relationship compared to biallelic SNPs, as their SFS are relatively similar (Structural variants may show a skew to lower frequency, resulting in steeper z_{MAR}. By excluding those, our analyses may be conservative). In order to properly characterize the geographic distribution of a mutation using all available geo-tagged individuals, we filtered for genotyping rate (plink --geno), and the final value is reported per dataset.

Species-specific details for dataset processing or homogenization are described below.

- The 1001 Arabidopsis Genomes Consortium (1001 Genomes Consortium, 2016) generated a WGS Illumina sequencing dataset of *Arabidopsis thaliana* comprising 1,135 individuals and 11,769,920 SNPs. These included recently colonized regions such as North America or Japan. Analyses of z_{MAR} were calculated only for the native range, which comprises most of the species diversity (>99%) and 1001 individuals. For computational efficiency, we conducted analyses using randomly sampled SNPs from chromosome 1, as we did not observe any difference when sampling from other chromosomes. Targeted analyses of geographic subregions such as 50 individuals from the distribution core (Germany) or 50 individuals from the warm edge (Spain), or a recently colonized area with 100 individuals from North America, were conducted to understand the stability of the z_{MAR} estimate as well as to understand how z_{MAR} behaves in very recent populations.

- Lucek & Willi (Lucek and Willi, 2021) recently published a dataset of WGS Illumina sequencing 108 *Arabidopsis lyrata* individuals from North America, which the authors directly shared as a VCF. We retrieved the latitude/longitude data from the supplemental material. We applied a genotyping rate filter ending with a dataset of 0.955431 genotyping rate. 10,000 SNPs were subsetted at random from the genome-wide data.

- Kreiner et al. (Kreiner *et al.*, 2019) WGS Illumina sequenced 165 individuals of *Amaranthus tuberculatus*. 155 individuals contained latitude and longitude information and were kept for the analyses. The genotyping rate was 0.98162 and we subsetted randomly 10,000 SNPs.

- Supple et al. (Supple *et al.*, 2018) generated a dataset of *Eucalyptus melliodora* of 275 individuals from 36 broadly distributed populations. The dataset was produced by Illumina sequence Genotyping by Sequencing (GBS) libraries digested with ApeKI as in Elshire et al. (2011). The authors provided the dataset in PLINK format. Genotyping rate was 0.769807 but we did not apply a further filter to avoid reducing the total number of variants. We conducted analyses with all 9378 SNPs. The genotyping rate in this dataset is likely not problematic as the total number of GPS locations is 36, with multiple individuals sampled closely. This sampling scheme probably allows to characterize an allele's distribution correctly despite the lower genotyping rate.

- Vallejo-Marin et al. (Vallejo-Marín *et al.*, 2021)(2021) generated a GBS dataset of 521 Mimulus plants, with 286 samples being *Mimulus guttatus* from its native distribution. Libraries for Genotyping-By-Sequencing were prepared with PstI enzyme as described in

Twyford & Friedman (2015) and sequenced using Illumina. The authors provided the dataset in VCF format. After applying a filtering for missingness, we ended up with a genotyping rate of 0.904192 and 1,498 SNPs, which were used for the analyses.

- Lovell & MacQueen (Lovell *et al.*, 2021)(2021) generated a WGS Illumina sequencing dataset of Switchgrass, *Panicum virgatum*, of a collection of 732 individuals and 33,905,044 variants. 576 individuals were from natural collections and had latitude/longitude data. The dataset contains also other collections such as cultivars, which were not used to build the MAR. The genotyping rate was 0.976393 and analyses were conducted with 10,000 SNPs drawn from the largest chromosome.

- MacLachlan et al. (MacLachlan *et al.*, 2021)(2021) generated a SNP chip dataset of *Pinus contorta* comprising 929 trees with latitude and longitude information and 32,449 SNPs. Genotyping was conducted with the AdapTree lodgepole pine Affymetrix Axiom 50,298 SNP array and data was provided in the supplemental material of the paper along with custom scripts to parse the data. These were transformed into PLINK. The genotyping rate was 0.959146, and analyses were conducted with 10,000 randomly drawn SNPs. The fact that this dataset was created with ascertained SNPs likely generates a frequency bias. In Fig. SIV.1, one can see that this may be a problem to calculate z_{MAR}, as the mutations~area graph appears nonlinear and rapidly saturates. This can happen if SNPs are common variants, as they are discovered immediately with very few samples.

- Tuskan et al. (Tuskan *et al.*, no date)(2015) WGS Illumina sequenced 882 *Populus trichocarpa* trees. The dataset includes 28,342,826 SNPs. The authors provided the dataset as a VCF along with latitude/longitude coordinates. This dataset was downsampled to the first chromosome. The genotyping rate was 0.921191, and 10,000 SNPs were randomly sampled for analyses.

- The Anopheles gambiae 1000 Genomes Consortium (Anopheles gambiae 1000 Genomes Consortium *et al.*, 2017) (Phase 2) produced Whole-Genome Illumina sequencing data for 1142 wild-caught mosquitoes of *Anopheles gambiae*. The data is available through https://www.malariagen.net/data as VCF and latitude/longitude coordinate files. The VCF was filtered for genotyping rate ending up at a 0.998895 rate. For efficiency, 10,000 randomly-selected SNPs from the VCF of the largest chromosome 2L were used for analyses downstream.

- Fuller et al. (Fuller *et al.*, 2020) WGS Illumina sequenced 253 coral individuals of *Acropora millepora* in 12 reefs. The dataset was downloaded as fastq files from the published online material, and SNPs were called as described in the supplemental material ending with 17,931,448, which were filtered to achieve a genotyping rate of 0.935709 for a total of 2,512 SNPs, which were used in the analyses.

- Ruegg et al. (2021, https://github.com/eriqande/ruegg-et-al-wifl-genoscape) generated a dataset of 219 songbirds *Setophaga petechia*, for which 199 could be matched with geographci coordinates. SNPs were ascertained from several publications using RAD seq and Fluidigm 96.96 IFC described in the repository, from which data was retrieved. A total of 349,014 SNPs were parsed using their custom scripts and we transformed them into PLINK files. A genotyping rate filter was applied ending with a 0.96061 rate and 195,700 SNPs.

- Kingsley et al. (Kingsley *et al.*, 2017) produced a dataset of 80 *Peromyscus maniculatus* deermice, for which 78 could be matched with geographic locations. The SNP dataset was produced using MY-select capture followed by Illumina sequencing (see publication for details). Links to the dataset were provided in the supplemental material. The dataset was already in PLINK format, including a total of 14,076 variants which were filtered to achieve a genotyping rate of 0.940411 for 2,946 SNPs, which were used in subsequent analyses.

- Smeds et al. (Smeds *et al.*, 2021)(2021) produced a WGS Illumina sequencing dataset and combined it with pre-existing datasets (see publication) for a total of 349 local dog breeds and wolves, of which 230 were *Canis lupus* from natural populations. A VCF was provided by the authors, which was transformed into a PLINK file, with a total of 1,517,226 SNPs, and a genotyping rate of 100%. A sample of 10,000 SNPs was randomly selected for subsequent analyses. Individuals were not precisely geo-tagged, but countries of origin were reported. The average latitude/longitude of the country of collection was used. This dataset is not ideal for z_{MAR}, as calculations are based on only 5 distinct geographic locations. Instead, in the main text, we ended up presenting the results from a second dataset of 107 geo-tagged grey wolves from Schweizer et al.(Schweizer *et al.*, 2016) capturing and resequencing 1040 genes. This provided us with 13,092 SNPs at 0.993061 calling rate, and a better geographic resolution.

- The 1000 Genome Consortium (1000 Genomes Project Consortium *et al.*, 2015) created WGS Illumina sequencing for over 2,504 humans and 24 unique geographic locations. We downloaded chromosome 1 from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/datacollections/1000G2504highcoverage/working/20190425NYGCGATK/ and gathered the population locations from https://www.internationalgenome.org/data-portal/population. To conduct analyses, we subsampled 10,000 SNPs at genotping rate 0.991069.

- Palacio-Mejia (Palacio-Mejía *et al.*, 2021) used WGS for 591 *Panicum hallii* individuals to sequence at low coverage, producing 45,589 SNPs. Because stringent filters of calling rates would lead to very small SNP sets, we settled on a genotyping rate of 0.825824 for 242 variants, all of which were used for downstream analyses. The authors shared an unfiltered VCF.

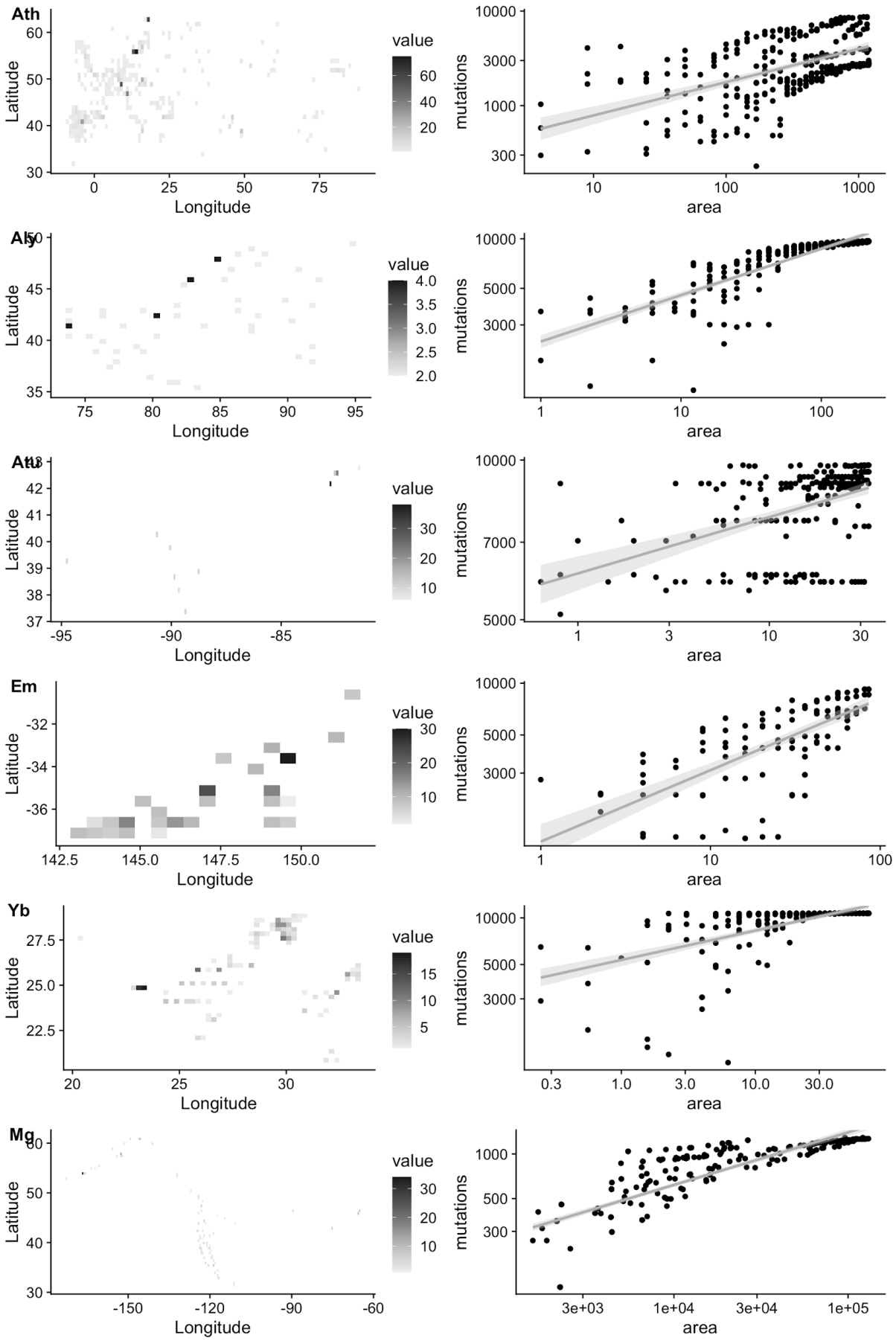
- Royer et al. (Royer, Streisfeld and Smith, 2016) produced a SNP dataset using RAD-Seq based Genotyping-By-Sequencing of 290 *Yucca brevifolia* (Joshua Tree) individuals. A total of 10,695 SNPs with a genotyping rate of 0.897501 were used for the analyses. The data was available at Dryad https://datadryad.org/stash/dataset/doi%253A10.5061%252Fdryad.7pj4t

- Kapun et al. (Kapun *et al.*, 2021) produced a WGS dataset of pooled *Drosophila melanogaster*, sequencing ~80 pooled individuals from each of 271 populations as part of the European "Drosophila Evolution over Space and Time" (DEST) project. A total of 5,019 shared SNPs with a genotyping rate of 0.937697 were used for analyses. The dataset was available through https://dest.bio/.

- Di Santo et al. (Di Santo *et al.*, 2021) studied the highly-threatened species *Pinus torreyana*. They used Genotyping-by-Sequencing of 242 individuals of the last remaining populations and shared the data directly with us. From a total set of 166,564 SNPs with a genotyping rate of 0.964632, 10,000 were randomly selected for our analyses.

- von Seth et al. (von Seth *et al.*, 2021) studied the highly-threatened species *Dicerorhinus sumatrensis*. They used Illumina WGS of 16 individuals of the last remaining populations and shared the data directly with us. In total, this comprises a set of 8,870,513 SNPs, with a genotyping rate of 0.854862, which we did not further filter due to the small number of individuals. For computational efficiency we selected 10,000 SNPs from the largest chromosome.

Information and results per species are gathered in Table 1 and its extended version, Table SIV.1, and the average z_{MAR} across species are provided in Table SIV.2.


Ath
Latitude
Longitude
value
mutations
area
Aly
Atu
Em
Yb
Mg

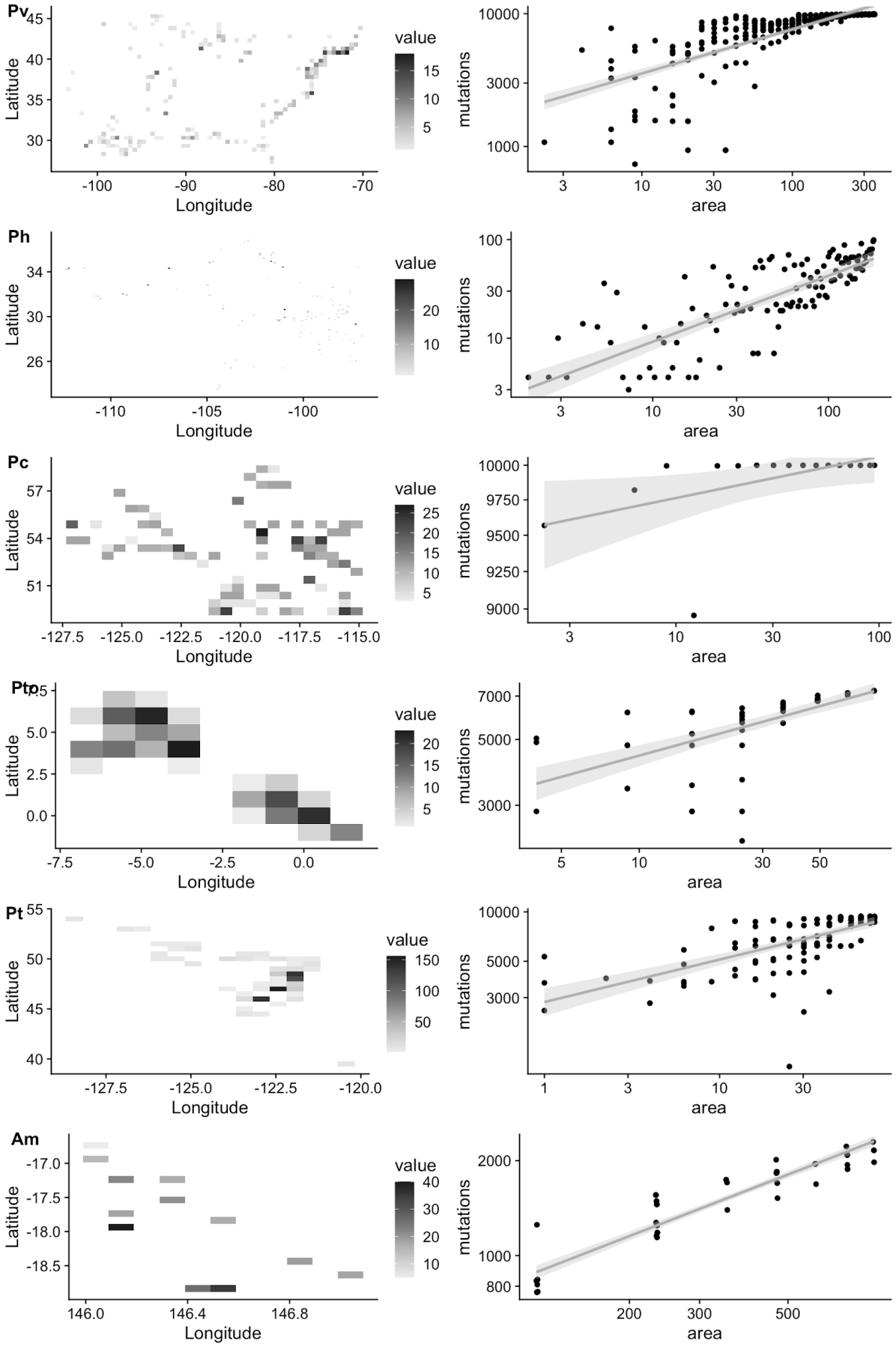


Fig. SIV.1 | MAR summaries across species.

For each species we plot (left) the map of sample density in space and the Mutations-Area Relationship. Acronyms correspond to the species names as first letter of genus and first letter (or second if redundancy exists) of the epitope (e.g. Ath = Arabidopsis thaliana).

Table SIV.1 | The mutations-area relationship across species. Extended Table 1

The Mutations-Area Relationship (MAR) fit with Area = Individuals and the scaled version. In the main text areas to protect 90% of genetic diversity per species are provided given the scaled z. Here, we also provide the average estimated area based on % of grid cells per species to be transformed from 2015 to 2050 using the LUH² dataset, the area where at least 10% of grid will be transformed, and the genetic extinction equivalents to those area transformations.*

Species (study)	**SFS** mod [ΔAIC]	**MAR (A=N)** z_N [CI95%]	**MAR scaled** z* [CI95%]	**LUH² change '50**	**LUH² >10% change '50**	**LUH² extinct '50**	**LUH² >10% extinct '50**
Arabidopsis thaliana	logN (85.8)	0.431 (0.423 - 0.439)	0.312 (0.305–0.32)	4.58	13.54	1.12	3.43
Arabidopsis lyrata	logN (9592.4)	0.254 (0.238 - 0.27)	0.15 (0.136–0.165)	0.79	2.64	0.19	0.64
Amaranthus tuberculatus	logN (7317.5)	0.244 (0.237 - 0.251)	0.142 (0.135–0.148)	4.86	11.13	1.19	2.79
Eucalyptus melliodora[VU]	logN (157.5)	0.531 (0.526 - 0.536)	0.402 (0.397–0.406)	3.82	7.77	0.93	1.92
Yucca brevifolia[CA]	logN(33300)	0.141 (0.128 - 0.155)	0.049 (0.037–0.062)	*0.74*	*0*	*0.18*	*0*
Mimulus guttatus	logN (580.8)	0.342 (0.331 - 0.353)	0.231 (0.221–0.241)	3.78	NA	0.92	NA
Panicum virgatum	logN (8345.2)	0.226 (0.215 - 0.237)	0.126 (0.116–0.136)	8.07	27.65	2	7.47
Panicum hallii	logN (86)	0.805 (0.702 - 0.908)	0.651 (0.558–0.744)	3.78	11.36	0.92	2.85
Pinus contorta	Wei (19413.7)	0.019 (0.018 - 0.02)	-	*1.95*	*5.54*	*0.47*	*1.36*
Pinus torreyana[CR]	logN(766156)	0.239 (0.232 - 0.245)	0.105 (0.099–0.11)	*25.4*	*NA*	*6.79*	*NA*
Populus trichocarpa	logS (0)	0.268 (0.257 - 0.28)	0.164 (0.154–0.175)	4.68	17.28	1.14	4.45
Anopheles gambiae	logS (0)	0.221 (0.209 - 0.233)	0.121 (0.11–0.132)	9.95	21.96	2.48	5.78
Acropora millepora	logN (452.3)	0.403 (0.395 - 0.41)	0.287 (0.28–0.293)	72.73	84.69	26.79	36.26
Drosophila melanogaster	logN(33300)	0.445 (0.433 - 0.458)	0.324 (0.313–0.336)	0.95	NA	0.23	NA
Setophaga petechia	Wei (640401.9)	0.169 (0.139 - 0.199)	0.074 (0.047–0.101)	5.55	15.14	1.36	3.86
Peromyscus maniculatus	logN (1449.7)	0.844 (0.769 - 0.919)	0.68 (0.613–0.748)	5.61	13.68	1.38	3.47
Dicerorhinus sumatrensis[CR]	w (107864.2)	0.474 (0.449 - 0.498)	0.123 (0.106–0.14)	*0.25*	*NA*	*0.06*	*NA*
Canis lupus	logN (85.8)	0.29 (0.28 - 0.301)	0.183 (0.174–0.193)	0.23	NA	0.06	NA
Homo sapiens	logN (9592.4)	0.395 (0.339 - 0.451)	0.28 (0.229–0.331)	28.81	40.13	7.83	11.58

Extended acronyms:
logN: log Normal distribution. logS: log Series distribution. Wei: Weibull distribution.

Table SIV.2 | Mean z_{MAR} across species.

We selected those species that did not show artifacts in Fig. SIV.1 or whose z_{MAR} overlapped with 0 to calculate a species-wide mean. These species were A. thaliana, A. lyrata, A. tuberculatus, E. melliodora, M. guttatus, P. virgatum, P. trichocarpa, A. millepora, S. petechia, *and* P. maculatus

	z_{MAR}	z_{MAR} *(A=N)*	z_{MAR} *scaled*
mean	0.30	0.36	0.25
median	0.26	0.34	0.23
IQR	0.12	0.19	0.19

Although we could not see any obvious patterns relating z_{MAR} with certain groups of species (Table 1), we wondered whether any life history trait of the species analyzed could explain the variation we observed (see Table SIV.3 of traits). An ANOVA did not show any significant relationship. Because we know theoretically this parameter must be related to the degree of dispersal ability of genotypes of a species relative to the whole species geographic range, we expect traits involved in determining these to be good predictors. Future work should be necessary to validate this, as the sample size (n=19) may not permit enough power to detect these expected patterns.

Table SIV.3 | Traits, life history, and other characteristics of the analyzed species.

Species	RedList	Known Decline	Kingdom	Reproduction	Pollination	Mobility	AreaRange
Arabidopsis thaliana	NO	NO	Plantae	Selfing	Selfing	Sessile	27337467.4
Arabidopsis lyrata	NO	NO	Plantae	Outcrossing	Vector	Sessile	2791301.4
Amaranthus tuberculatus	LC	NO	Plantae	Outcrossing	Vector	Sessile	804124.8
Eucalyptus melliodora	VU	NO	Plantae	Outcrossing	Wind	Sessile	948699.3
Yucca brevifolia	LC	YES	Plantae	Outcrossing	Vector	Sessile	1213454.4
Mimulus guttatus	LC	NO	Plantae	Outcrossing	Vector	Sessile	25138310.6
Panicum virgatum	LC	NO	Plantae	Outcrossing	Wind	Sessile	6291400.2
Panicum hallii	NO	NO	Plantae	Outcrossing	Wind	Sessile	2188807.4
Pinus contorta	LC	NO	Plantae	Outcrossing	Wind	Sessile	886182.2
Pinus torreyana	CR	YES	Plantae	Outcrossing	Wind	Sessile	30781.95
Populus trichocarpa	LC	NO	Plantae	Outcrossing	Wind	Sessile	1119664.1
Drosophila melanogaster	NO	NO	Animalia	Outcrossing	Activemating	Fly	115208408
Anopheles gambiae	NO	NO	Animalia	Outcrossing	Activemating	Fly	19959809.9
Acropora millepora	NT	YES	Animalia	Outcrossing	Activemating	Fly	26725.9
Setophaga petechia	LC	NO	Animalia	Outcrossing	Activemating	Fly	7027395.2
Peromyscus maniculatus	LC	NO	Animalia	Outcrossing	Activemating	Mobile	22609152.6
Dicerorhinus sumatrensis	CR	YES	Animalia	Outcrossing	Activemating	Mobile	3335605.58
Canis lupus	LC	NO	Animalia	Outcrossing	Activemating	Mobile	19102403.5
Homo sapiens	NA	NA	NA	NA	NA	NA	80763121.8

Table SIV.4 | Association of traits, life history, and other characteristics with z_{MAR}.

Acronyms: NO=not assessed but likely non-threatened, LC=low concern, VU=vulnerable, CR=critically endangered

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
RedList	4	0.0952396	0.0238099	0.5580988	0.7040464
KnownDecline	1	0.0275537	0.0275537	0.6458527	0.4580865
Kingdom	1	0.0011684	0.0011684	0.0273876	0.8750400
Reproduction	1	0.0003238	0.0003238	0.0075890	0.9339612
Pollination	1	0.0375975	0.0375975	0.8812784	0.3909509
Mobility	1	0.1600627	0.1600627	3.7518370	0.1104995
AreaRange	1	0.0174745	0.0174745	0.4095989	0.5503439
Residuals	5	0.2133125	0.0426625	NA	NA

V. An estimate of global genetic diversity extinction

Using the approach described in section II.4, we generated a number of estimates either per ecosystem or per species. All estimates below tried to be conservative, and thus we always used the scaled z_{MAR} values (section II.3.2.)

V.1 Estimates of ecosystem area losses

Table SV.1 | Millennium Ecosystem Assessment land cover transformation.
Source: https://www.millenniumassessment.org

Ecoystem	Area (million km2)	Earth surface (%)	Protected areas (%)	Area transformed (%)
MARINE	349.3	68.6	0.3	NA
COASTAL	17.2	4.1	7.0	NA
COASTAL TERRESTRIAL	6.0	4.1	4.0	11.00
COASTAL MARINE	11.2	2.2	9.0	NA
INLAND WATER	10.3	7.0	12.0	11.00
FOREST/WOODLAND	41.9	28.4	10.0	42.00
FOREST/WOODLAND TROPICAL	23.3	15.8	11.0	34.00
FOREST/WOODLAND TEMPERATE	6.2	4.2	16.0	67.00
FOREST/WOODLAND BOREAL	12.4	8.4	4.0	25.00
DRYLAND	59.9	40.6	7.0	18.00
DRYLAND HYPERARID	9.6	6.5	11.0	1.00
DRYLAND ARID	15.3	10.4	6.0	5.00
DRYLAND SEMIARID	22.3	15.3	6.0	25.00
DRYLAND SUBHUMID	12.7	8.6	7.0	35.00
ISLAND	7.1	4.8	17.0	17.00
ISLAND STATES	4.7	3.2	18.0	21.00
MOUNTAINS	35.8	24.3	14.0	12.00
MOUNTAINS 300-1000	13.0	8.8	11.0	13.00
MOUNTAINS 1000-2500	11.3	7.7	14.0	13.00
MOUNTAINS 2500-4500	9.6	6.5	18.0	6.00
MOUNTAINS 4500PLUS	1.8	1.2	22.0	0.30
POLAR	23.0	15.6	42.0	0.38
CULTIVATED	35.3	23.9	6.0	47.00
CULTIVATED PASTURE	0.1	0.1	4.0	11.00
CULTIVATED CROPLAND	8.3	5.7	4.0	62.00
CULTIVATED MIXED	26.9	18.2	6.0	43.00
URBAN	3.6	2.4	0.0	100.00
GLOBAL	510.0	NA	4.0	38.00

Ecosystem transformation has been tracked over the decades. We extracted ecosystem transformations from the Millennium Ecosystem Assessment (Millennium Ecosystem Assessment, 2005), which estimated ecosystem transformations from presumably native systems to cultivated or urban areas by GLC2000 land cover dataset (Table SV.1). The forest/woodland is calculated as percentage change between potential vegetation from WWF ecoregions to the current actual forest/woodland areas from GLC2000. These provide bulk ecosystem reductions, not for a given species, but may be a good proxy for an average across species.

Table SV.2 | IPBES land cover transformation.
Source: https://ipbes.net

Region	Area (million km2)	MSA (%) 2010	MSA (%) 2010 SSP2	MSA (%) 2010 SSP1	MSA (%) 2050 SSP3
North America	20	65	56	NA	NA
Central and South America	18	65	53	NA	NA
Middle East and Northern Africa	11	81	77	NA	NA
Sub-Saharan Africa	24	70	56	NA	NA
Western and Central Europe	6	37	29	NA	NA
Russian region and Central Asia	21	73	65	NA	NA
South Asia	5	44	35	NA	NA
China region	11	56	49	NA	NA
Southeast Asia	7	55	43	NA	NA
Japan, Korea and Oceania	8	71	57	NA	NA
Polar	2	96	91	NA	NA
World	132	66	56	62	54

The Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) recently used a PBL satellite products from the Netherlands Environmental Assessment Agency (https://www.pbl.nl/en/nature-and-biodiversity) to study the % of area ecosystem transformation in the world. This provides an updated estimate to the Millennium Assessment as well as projections under several Shared Socioeconomic Pathways (1-3) for 2050. These were reported per region as of 2010, and for projections to 2050 (scenario SSP2). Instead of direct area, the metric is a composite of land use information to predict Mean Species Abundance (MSA), a measure of the size of populations of wild organisms as a percentage of their inferred abundance in their natural state (% MSA).

A global transformation metric can also be captured by the most updated land use transformation data, the Land Use Harmonization 2 (release v2e for 2015-2011 and release v2h for baseline 1850-2015) (Hurtt *et al.*, 2020). Baseline transformation of primary ecosystems was calculated subtracting the total area covered by primary forest (primf) and primary non-forest (primn) variables between year 1850 layer (roughly pre-industrial baseline) and the present, 2015, as $1\text{-}A_{2015} / A_{1850}$ (Table SV.3). Analyses that use projections to mid-21st century were conducted similarly as in (Theodoridis, Rahbek and Nogues-Bravo, 2021), summing over all transitions from primary forest (primf), primary non-forest (primn), secondary forest (secdf) and secondary non forest (secdn) lands to any other category for all years within the 2015-2050 period (see Table SIV.1).

Table SV.3 | Land Use Harmonization 2 from 1850 to 2015
Source: https://luh.umd.edu/data.shtml

	Area %
Primary forest transformed	43
Primary non-forest transformed	50

Finally, we searched for timely estimates of forest reduction (based on vegetation cover) reported in the Global Forest Watch website: https://www.globalforestwatch.org/dashboards/global/ (accessed June 2021). From 2002 to 2020, there has been a global tree cover loss of 10%, with an annual tree cover loss of 0.6-1.1%.

Although these are not direct area transformations, we also used the IUCN Red List resource (https://www.iucnredlist.org, Table SIV.3 shows status of the species analyzed here), which includes guides to categorize species as vulnerable, endangered, critically endangered, and extinct, and has conducted extensive assessments across thousands of species (Table SV.4).

Table SV.4 | IUCN Red List categories of extinction risk and number of species.
Source: www.iucnredlist.org, January 2021

IUCN Red List Category	Description	Criterion of area reduction (>%)	# species assessed
EX	Extinct	100	164
CR	Critically Endangered	80	4674
EN	Endangered	50	8593
VU	Vulnerable	30	8459
NC	No concern (all other)	0	32237

V.2 A global estimate of genetic extinction

Taking the estimates and standard error of z_{MAR} across species, and the world's reduction of ecosystems we can calculate the fraction of genetic diversity reduction following the extinction MAR equation (section II.4), giving a range of estimates (Table SV.5).

Table SV.5 | Estimates of average expected genetic extinction for different ecosystems.
Assuming ecosystem transformation approximately translates into average species distribution reduction, and using the ranges of z_{MAR} from Table 1 of the main text, we project the average genetic extinction using the Mutations Area Relationship.

System	*Area transformed (%)*	*Genetic extinction % (mean z based)*	*Genetic extinction % (min z based)*	*Genetic extinction % (max z based)*
COASTAL TERRESTRIAL	*11*	*3*	*0.9*	*7.7*
INLAND WATER	*11*	*3*	*0.9*	*7.7*
FOREST/WOODLAND	*42*	*13.5*	*4*	*31.1*
FOREST/WOODLAND TROPICAL	*34*	*10.5*	*3*	*24.7*
FOREST/WOODLAND TEMPERATE	*67*	*25.5*	*7.9*	*53.1*
FOREST/WOODLAND BOREAL	*25*	*7.4*	*2.1*	*17.8*
DRYLAND	*18*	*5.1*	*1.5*	*12.7*
DRYLAND HYPERARID	*1*	*0.3*	*0.1*	*0.7*
DRYLAND ARID	*5*	*1.4*	*0.4*	*3.4*
DRYLAND SEMIARID	*25*	*7.4*	*2.1*	*17.8*
DRYLAND SUBHUMID	*35*	*10.8*	*3.2*	*25.5*
ISLAND	*17*	*4.8*	*1.4*	*11.9*
MOUNTAINS	*12*	*3.3*	*0.9*	*8.4*
MOUNTAINS 300-1000	*13*	*3.6*	*1*	*9.1*
MOUNTAINS 1000-2500	*13*	*3.6*	*1*	*9.1*
MOUNTAINS 2500-4500	*6*	*1.6*	*0.5*	*4.1*
MOUNTAINS 4500+	*0.3*	*0.1*	*0*	*0.2*
POLAR	*0.4*	*0.1*	*0*	*0.3*
GLOBAL	*38*	*11.9*	*3.5*	*27.9*

Using the Land Use Harmonized 2 dataset, we also create per-species predictions based on the % transformation of each of the sampled regions per species (Table SV.1). An important approach in the future

Utilizing tree cover from the Global Forest Watch, which estimates 0.6-1.1% of transformation per year across Canada, United States and Australia, an extrapolation of extinction to 50 years for tree species will correspond to 5.36-9.86% genetic extinction.

Assuming that the calculated z_{MAR} estimates are representative of plant species, we conducted an experiment to create a distribution of % of genetic extinction in threatened species. We then used the number of species in each IUCN category (Table SV.4) for a total of 54,127 plant species. For plant species, one of the evaluation criteria of percentage of population loss likely translates faithfully to area reduction in the species. The proportion of species per category then gives a discrete probability distribution of the ranges of percentage of area loss: P(0-29%)=0.596 P(30-49%)=0.156, P(50-79%)=0.159, P(80-99%)=0.086, P(99%-100%)=0.003). Using a simulation-based sampling approach, we drew 350,000 random area reductions A_t / A_{t-1} from the previous distribution and a z_{MAR} from the mean and variance of our estimates from Table 1 for plants. These were plugged into the extinction equation (Section II.4) to calculate the percentage of genetic extinction of these 350,000 random draws. The resulting distribution had a median and interquartile range of: : 16.73 % [7.38-30.57].

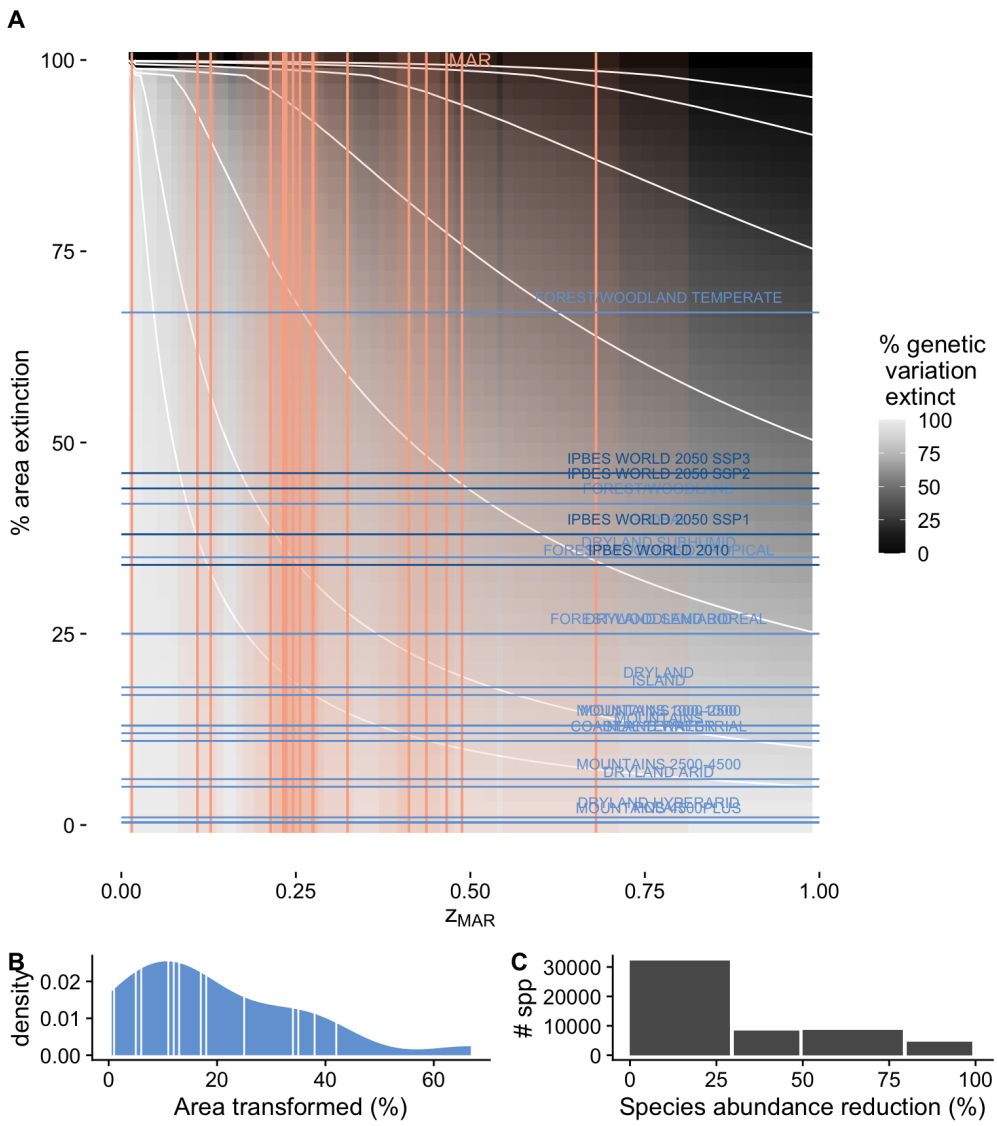


Fig. SV.1 | The parameter space of genetic diversity extinction, extended

(A) The theoretical space of genetic diversity extinction. z_{MAR} values (using area, unscaled for samples, differently from Fig 2) computed for species analyzed here are marked as orange vertical lines, with confidence intervals as orange shading. Blue horizontal lines correspond to ecosystem transformations from the Millennium Assessment (light blue) and IPBES Assessment (dark blue) (B) The distribution of percentage of area transformed across ecosystems, with averages per ecosystem marked in the distribution as well as horizontal lines in (A). (C) The number of species of each of the IUCN categories and the most optimistic range of area or abundance reduction for each of the category brackets.

V.3 Community ecology simulations and MAR

To test whether intermediate levels of MAR would be expected across species in entire ecosystems, we conducted community assembly simulations of ~100-500 species following the Neutral Theory of Biodiversity (Hubbell, 2001; Mérot *et al.*, 2020) and coalescent simulations (Kelleher, Etheridge and McVean, 2016) using the software MESS (Overcast *et al.*, 2020). These simulations are computationally demanding and could not run in a complete 2D spatial grid. Instead, they were simulated in a mainland-island system, with islands of increasing areas. The community forms by species colonizing an empty island according to Hubbell's Unified Neutral Theory of Biodiversity and Biogeography (UNTB), where all species are equally likely to colonize and persist in the local community. Continued colonization and migration to the local community continues to bring in new species that may or may not survive, while also continuously bringing in individuals of species already in the local community. The community assembly process ends when the community has

reached an equilibrium denoted as the balance between local extinction and new species dispersing into the area (Hubbell 2001). Once the forward-time process has ended, we simulate the coalescent history of each species backward in time. For this, MESS considers the population size, divergence time, and migration rates of the meta and local communities. These coalescent simulations provide us with genetic data and ultimately diversity estimates for each species in the community.

We simulated 100 MESS communities, and for each community the size of the local community was varied from 1K to 100K. We varied the size of communities to emulate variation in area occupied by a given community because we assume as the number of individuals in a community increases from 1,000 to 100,000, so does the area occupied. All other parameters were kept consistent across each of these community simulations, and most remained at their default value. The parameters changed were the length of the sequences simulated for the coalescent-based simulations, which was fixed at 10,000 bp, and the migration rate, which was fixed at 0.01.

The simulation output was used to then compute a single z_{SAR} for the system as $S=cA^{zSAR}$, and one z_{MAR} for each species in the same way, $M=cA^{zmar}$. This resulted in a distribution of z_{MAR} from Fig. SV.2. This confirmed that we can recover typical z_{SAR} and z_{MAR} values from completely stochastic neutral yet spatially structured systems such as species in communities and mutations in populations of a species.

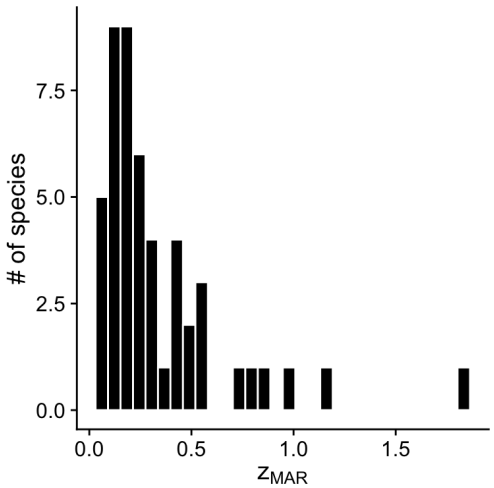


Fig. SV.2 | z_{MAR} calculated from MESS eco-evolutionary simulations
Using the MESS framework of a mainland-island model with different island sizes, z_{MAR} per species is recovered. The stochastic nature of the simulations results in each species having different abundances and migration histories that change the scaling value. Values over 1 are likely noisy estimates.

V.4 The nested species and genetic extinction process

Finally, we worried that our estimates of V.2 would be mistaken as overestimates. In fact, we believe these may be underestimated. Recent policy proposals for the United Nations' Sustainability Goals emphasize that the target of protecting 90% of species genetic diversity for all species cannot leave the already-extinct species behind (Díaz *et al.*, 2020) (That is, one cannot protect 90% of species and leave 10% to become extinct to meet this goal). This clearly exemplifies a problem in conservation biology that what researchers can study is (most of the time) what has escaped extinction, and therefore if we do not account for extinct species in our overall estimates of genetic extinction we may naively think ecosystems have not suffered genetic extinction (i.e. in the extreme scenario, an ecosystem that has lost but one abundant species may not really appear genetically eroded if such species is in good shape).

We then created spatial simulations in R where 1,000 species are distributed in 100x100 grid cells following a UNTB abundance distribution and then proceeded with an edge extinction of the ecosystem (see Fig. SV.3 for a cartoon).

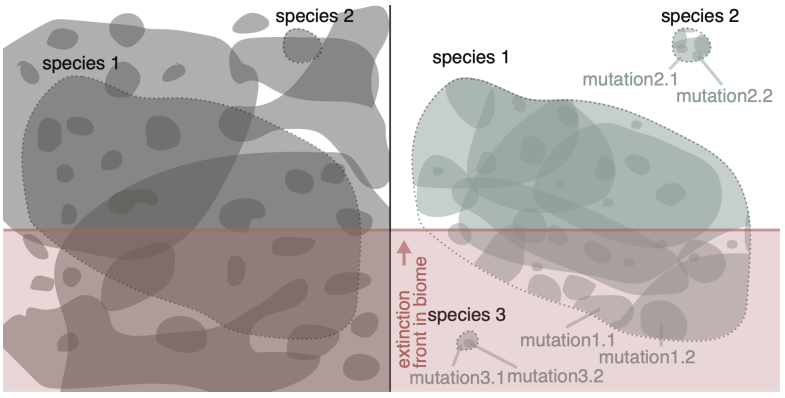


Fig. SV.3 | Cartoon of nested extinction of species and genetic diversity.
An ecosystem with multiple species within it (left), distributed in space, with few species broadly distributed and many narrowly distributed. Moving one level of biological organization lower, mutations within species (right) are also spatially distributed with many narrowly distributed. As extinction happens (red line moving bottom to top), all species below the red line go extinct, but only the mutations within species 1 below the line go extinct, while mutations above the line remain. Species 3 has already become extinct, and therefore also all the mutations within it.

Two extreme types of distributions of species can be imagined: species are randomly placed in space, or species are found mostly in perfectly contiguous ranges (We ended up using as an example a simulation with 85% of the individuals of a species found in a core square continuous distribution and 15% found outside that core in fragmented observations, as this scenario produced the canonical SAR of z~0.3). Spatial structure interestingly creates two extreme distributions of area reductions across species (Fig. SV.4): random placement of cell habitats essentially show that the average area reduction per ecosystem is followed by most species, while autocorrelated placement of cell habitats create a U distribution in area reductions, where at the beginning of the extinction process most species have not experienced any impact (Fig. SV4.B left) but at the end of ecosystem reduction virtually all species are already extinct (Note we may be at the beginning of SV4.B process given the data from IUCN, Fig 2.C).

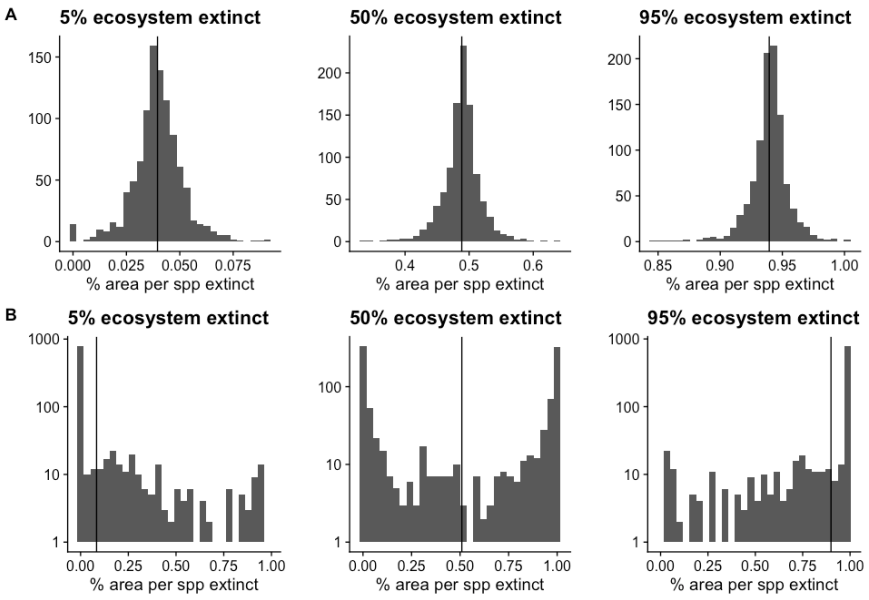


Fig. SV.4 | The distribution of genetic extinction in simulated ecosystems with 1000 species
Two ecosystems of 100x100 cells with 1000 species. Species are either randomly distributed in cells (A) or spatially autocorrelated with occupying mostly contiguous cells (B). As the extinction process wiped out part of the ecosystem (snapshots are provided at 5%, 50%, and 95%), the area loss per species is tracked. In (A) the average area lost per species is roughly the total reduction of the ecosystem, whereas in (B) the distribution is U shaped (note the log-scaled y-axis). While in (B) the mean area lost in the distribution correctly captures the area loss of the ecosystem, per species losses are highly uneven.

We ended up following up the more realistic autocorrelated scenario. With a latitudinal sweeping extinction of the ecosystem, we then aimed to track the percentage of genetic extinction per species *i*. For extant species this would follow the MAR relationship (section II.4), with an assumed constant $z_{MAR} = 0.3$ for simplicity. For extinct species (100% of their area reduced), we considered genetic diversity extinction was 100%. The compound total genetic extinction would then just be the sum of those $X_{Tot} = \sum_{i=1}^{1000} X_i$. (Of course, in reality species may vary in their genome-wide diversity average, and we could for instance use Watterson's Θ_W (see section II.2) to scale $\sum_{i=1}^{1000} \Theta_{Wi} X_i$). Interestingly, if we calculate the *z* of the slope of compound genetic extinction across species it is much larger than MAR or SAR alone: 0.6 (Fig. SV.5).

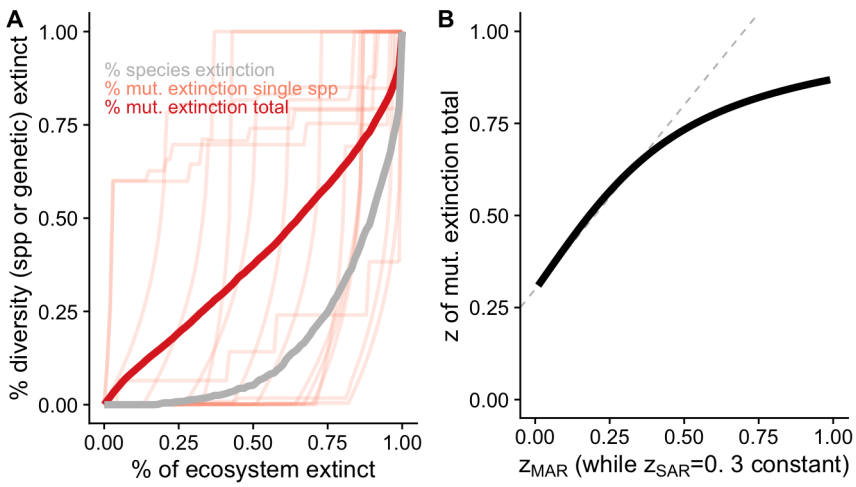


Fig. SV.5 | Numeric simulation of nested genetic extinction.
(A) Simulating the extinction of an ecosystem with 1,000 species that follow a log-normal species abundance curve. Extinction of the ecosystem creates a curve of species extinction z~0.3 (grey). Likewise, each species trajectory (light red, 15 species drawn randomly) follows a simulated z_{MAR}~0.3 extinction as they lose area. Because species' distributions are smaller than the whole ecosystem, those distributed closer to the start of the extinction front lose area first, while those

distributed farthest from the extinction front only lose area when the ecosystem is almost wiped out. Because genetic diversity loss is both due to complete extinction of species as well as area reduction of extant species, the compound extinction curve (red) follows the faster extinction dynamics. (B) Holding z_{SAR}=0.3 constant, and varying z_{MAR} shows that the compound genetic extinction across species is close to the sum of both z slopes, but it saturates at ca. 0.85 (grey dotted lines show $z_{MAR} + z_{SAR}$).

V.5 Ongoing efforts in characterizing genetic diversity loss

There are many global calls to start monitoring genetic diversity (Hoban et al. 2021) in standard ways to maximize genetic information in germplasms (Lockwood et al. 2007), and ancient DNA techniques in combination to historical specimens can enable large temporal screens of genomics (Lang et al. 2018), but to our knowledge only two studies have tried to meta-analyze public datasets to empirically estimate genetic diversity loss across species:

Recently (Leigh *et al.*, 2019) gathered genetic diversity estimates mostly of animals in studies where the same population had been sampled in the past and in the present (7 years (±439 SD) or 27 generations (±44 SD). Eighty-eight estimates were based on microsatellite markers, nine on single nucleotide polymorphisms, and two on restriction enzyme-based markers. The authors quantified a 5.4%–6.5% (±18.8% SD) decline in genetic diversity from present vs historic samples. The MAR estimates of genetic diversity loss fall well within this range, and appear consistent with the order of magnitude quantified using the MAR relationship. Estimates, however, are not directly comparable to the MAR forecasting, for multiple reasons: (1) Leigh et al. used mostly microsatellites (which are intendedly ascertained to be highly variable and thus should not follow the neutral SFS nor MAR). (2) Because this empirical quantification intends to track the same population over time to understand genetic diversity decline, whereas MAR studies the genetic diversity decline by the loss of populations. This could make samples biased toward populations in environments experiencing little change over the sampled time periods; for example, commercially important fish made up 31% of the populations sampled. The exclusion of taxa experiencing habitat loss may have led to a conservative estimate of genetic diversity reduction. The pattern observed by Leigh and colleagues may be related to the feedback effect that losing other populations or gene flow among populations could create in the focal population equilibrium (see section VI, reasons for underestimation).

A second study found no significant trends between the current amount of human impact in a given area and the diversity in the mitochondrial cytochrome c oxidase subunit I (COI) gene of a wide variety of bird, fish, insect, and mammal species (Millette *et al.*, 2020). The lack of signal of this study could be due to the gene studied (Paz-Vinas *et al.*, 2021), which is under purifying selection (Pentinsaari et al 2016), and then may not capture genome-wide neutral diversity (Kardos *et al.*, 2021). Rather than a temporal study, Millette et al's resembles a sensitivity analysis, where one may expect areas that have been more disturbed historically by humans will have lower levels of genetic diversity.

We are positive that empirical studies like these will be important in the future to validate MAR projections, especially temporal studies like Leigh et al., once new ecosystem-wide genomic resources become more and more available new whole-genome datasets (Lewin *et al.*, 2018; Meyer *et al.*, 2019; Shaffer and Toffelmier, 2020)

VI. Limitations and outlook

In this last section we discuss some potential limitations of an inherently simple scaling law, and what approaches could be used to address those and improve genetic extinction projections.

VI.1 Reasons for overestimations

Many researchers have proposed that SAR likely overestimates extinction (He and Hubbell, 2011; Rahbek and Colwell, 2011). For instance:

- Ignoring that a diversity-area relationship can be defined outwards, inwards, or focusing on endemisms can have an impact (He and Hubbell, 2011; Rahbek and Colwell, 2011; Storch, Keil and Jetz, 2012). To address this, we confirmed relative consistency between inward, outward, and random placement MAR, and proposed that the EMAR may not be that appropriate to study extinction (or at least does not agree with our simulation).
- Species may persist in altered habitats, like some animals are known to do (Pereira and Daily, 2006). We have focused some of the estimates in this study on plants, but further extensions could be applied as described by Pereira and Daily (Pereira and Daily, 2006) in the future.
- SAR is not a mechanistic model (Harte, Smith and Storch, 2009). We have derived its ranges of possible values and averages analytically and are beginning to understand how evolutionary forces shape MAR. Realistic simulations can help understand in a process-based framework how populations (and their MAR) react to extinction (continuous space simulations with progressive area reductions appear to fit well with the MAR predictions before extinction, section III.2.6).
- There is a scale dependence in the SAR slope (Storch, Keil and Jetz, 2012). Since power laws are typically fit with large-scale datasets and used to predict local scale extinctions, predictions could be overestimated.

VI.2 Reasons for underestimations

While the simplicity of power laws to make predictions of extinction may lead to overestimations, there are also reasons to believe MAR would underestimate extinction.

- The use of ascertained genetic markers underestimates z_{MAR} and therefore the degree of genetic extinction with area shrinkage. This is clear in the pre-selected-only marker dataset of *Pinus contorta* (Fig. SII.2), but (Lockwood, Richards and Volk, 2007)
- The use of z_{MAR} that scales down for minimum average $z_{MAR} > 0$ for small sample sizes but assumes maximum average z_{MAR} can be 1 (section II.3.2). This would effectively lead to smaller z_{MAR}, and thus underestimation of extinction.
- When species shrink in area, the effective population size of the remaining population decreases, increasing drift and moving towards a lower diversity equilibrium. This reactive process is not captured by the phenomenological MAR relationship.
- Although sequencing methods have an error rate that misreads true nucleotide sequences, this rate is typically extremely low (many sequencing projects described here used Illumina HiSeq series, which has a 0.112% error rate, or about 1 misread nucleotide in 1000). This could intuitively lead to overestimates in mutations in space but in fact, the mis-reading of DNA ends up causing an underestimation. This is because bioinformatic software that transforms raw data into SNP variant tables errs towards the conservative direction, often not calling mutations that have been observed very few times, and thus likely under-representing rare mutations (Czech and Exposito-Alonso, 2021).
- The nested extinction (section V.3).

VI.3 Final notes

Ultimately, to make accurate predictions of genetic extinction and increased extinction risk of whole species, very detailed data per species will be required: census sizes, genome size, migration in meta-populations, mating system, detailed maps of genetic makeups, and finescale area transformations. This could enable mechanistic models projected forward-in-time such as discussed in section II.3.6. The production of new genomic datasets across entire ecosystems should further help create maps of genetic diversity at high resolution to track losses (Parks *et al.*, 2013; Miraldo *et al.*, 2016; Li *et al.*, 2021).

Our aim in this work has been to try to be as conservative as possible in extinction estimates (using area calculations that produce lower z_{MAR} values, scaling them for low sample bias, using lower estimates of ecosystem transformation, etc.). However we run into the danger of under-estimating extinction. As described in V.4. with

Even if
to then sharply collapse in late stages of the whole-species extinction event (Ehrlich and Walker, 1998)

In the meantime, we believe MAR is a quantitative and scalable first-approximation of genetic extinction that would just require accurate understanding of abundance or area reductions and minimal information about population structure or mating/dispersal/range relationships. Given that scaling relationships are already applied by conservation policy (IPBES, 2019), and given that assumptions and limitations are understood, we expect MAR to become a relevant tool to project a dimension of biodiversity so far mostly invisible or unaddressable in large conservation projections.

VII. Supplemental References

1000 Genomes Project Consortium *et al.* (2015) 'A global reference for human genetic variation', *Nature*, 526(7571), pp. 68–74. doi: 10.1038/nature15393.

1001 Genomes Consortium (2016) '1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana', *Cell*, 166(2), pp. 481–491. doi: 10.1016/j.cell.2016.05.063.

Alonso, D. and McKane, A. J. (2004) 'Sampling Hubbell's neutral theory of biodiversity', *Ecology letters*, 7(10), pp. 901–910. doi: 10.1111/j.1461-0248.2004.00640.x.

Anopheles gambiae 1000 Genomes Consortium *et al.* (2017) 'Genetic diversity of the African malaria vector Anopheles gambiae', *Nature*, 552(7683), pp. 96–100. doi: 10.1038/nature24995.

Booker, T. R., Yeaman, S. and Whitlock, M. C. (2021) 'The WZA: A window-based method for characterizing genotype-environment association', *bioRxiv*. doi: 10.1101/2021.06.25.449972.

Buffalo, V. (2021) 'Quantifying the relationship between genetic diversity and population size suggests natural selection cannot explain Lewontin's paradox', *eLife*, 10. doi: 10.7554/eLife.67509.

Campos, P. R. A., de Oliveira, V. M. and Rosas, A. (2010) 'Epistasis and environmental heterogeneity in the speciation process', *Ecological modelling*, 221(21), pp. 2546–2554. doi: 10.1016/j.ecolmodel.2010.07.023.

Czech, L. and Exposito-Alonso, M. (2021) 'grenepipe: A flexible, scalable, and reproducible pipeline to automate variant and frequency calling from sequence reads', *arXiv*. Available at: http://arxiv.org/abs/2103.15167.

Díaz, S. *et al.* (2020) 'Set ambitious goals for biodiversity and sustainability', *Science*, 370(6515), pp. 411–413. doi: 10.1126/science.abe1530.

Di Santo, L. N. *et al.* (2021) 'Reduced representation sequencing to understand the evolutionary history of Torrey pine (Pinus torreyana Parry) with implications for rare species conservation', *bioRxiv*. doi: 10.1101/2021.07.02.450939.

Ehrlich, P. and Walker, B. (1998) 'Rivets and redundancy', *Bioscience*, 48, p. 387. Available at: https://go.gale.com/ps/i.do?id=GALE%7CA20924871&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=00063568&p=HRCA&sw=w.

van Etten, R. J. H. &. J. (2012) 'raster: Geographic analysis and modeling with raster data'. Available at: http://CRAN.R-project.org/package=raster.

Exposito-Alonso, M. *et al.* (2018) 'The rate and potential relevance of new mutations in a colonizing plant lineage', *PLoS genetics*, 14(2), p. e1007155. doi: 10.1371/journal.pgen.1007155.

Exposito-Alonso, M. *et al.* (2019) 'Natural selection in the *Arabidopsis thaliana* genome in present and future climates', *Nature*, 573(7772), pp. 126–129. doi: 10.1038/s41586-019-1520-9.

Fan 樊海英, H. *et al.* (2019) 'Genetic Diversity-Area Relationships across Bird Species', *The American naturalist*, 194(5), pp. 736–740. doi: 10.1086/705346.

Fisher, R. A. (1931) 'XVII.—The Distribution of Gene Ratios for Rare Mutations', *Proceedings of the Royal Society of Edinburgh*, 50, pp. 204–219. doi: 10.1017/S0370164600044886.

Fisher, R. A., Corbet, A. S. and Williams, C. B. (1943) 'The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population', *The Journal of animal ecology*, 12(1), pp. 42–58. doi: 10.2307/1411.

Franklin, I. R. and Frankham, R. (1998) 'How large must populations be to retain evolutionary potential?', *Animal conservation*, 1(1), pp. 69–70. doi: 10.1111/j.1469-1795.1998.tb00228.x.

Fuller, Z. L. *et al.* (2020) 'Population genetics of the coral Acropora millepora: Toward genomic prediction of bleaching', *Science*, 369(6501). doi: 10.1126/science.aba4674.

Hahn, M. W. (2018) *Molecular Population Genetics*. Oxford University Press. Available at: https://play.google.com/store/books/details?id=3BDkswEACAAJ.

Haller, B. C. and Messer, P. W. (2019) 'SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model', *Molecular biology and evolution*, 36(3), pp. 632–637. doi: 10.1093/molbev/msy228.

Harte, J., Smith, A. B. and Storch, D. (2009) 'Biodiversity scales from plots to biomes with a universal species-area curve', *Ecology letters*, 12(8), pp. 789–797. doi: 10.1111/j.1461-0248.2009.01328.x.

He, F. and Hubbell, S. P. (2011) 'Species-area relationships always overestimate extinction rates from habitat loss', *Nature*, 473(7347), pp. 368–371. doi: 10.1038/nature09985.

Hubbell, S. P. (2001) *The Unified Neutral Theory of Biodiversity and Biogeography*. Monographs in Population Biology.

Hudson, R. R., Slatkin, M. and Maddison, W. P. (1992) 'Estimation of levels of gene flow from DNA sequence data', *Genetics*, 132(2), pp. 583–589. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1427045.

Hurtt, G. C. *et al.* (2020) 'Harmonization of global land use change and management for the period 850–2100 (LUH2) for CMIP6', *Geoscientific model development*, 13(11), pp. 5425–5464. doi: 10.5194/gmd-13-5425-2020.

IPBES (2019) *Global assessment report of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*. Bonn, Germany: IPBES Secretariat, p. 1753. doi: 10.5281/zenodo.3831674.

Kapun, M. *et al.* (2021) 'Drosophila Evolution over Space and Time (DEST) - A New Population Genomics Resource', *bioRxiv*. doi: 10.1101/2021.02.01.428994.

Kardos, M. *et al.* (2021) 'The crucial role of genome-wide genetic variation in conservation', *bioRxiv*. doi: 10.1101/2021.07.05.451163.

Kelleher, J. *et al.* (2018) 'Efficient pedigree recording for fast population genetics simulation', *PLoS Comput. Biol.* doi: 10.1371/journal.pcbi.1006581.

Kelleher, J., Etheridge, A. M. and McVean, G. (2016) 'Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes', *PLoS computational biology*, 12(5), p. e1004842. doi: 10.1371/journal.pcbi.1004842.

Kingsley, E. P. *et al.* (2017) 'The ultimate and proximate mechanisms driving the evolution of long tails in forest deer mice', *Evolution; international journal of organic evolution*, 71(2), pp. 261–273. doi: 10.1111/evo.13150.

Kreiner, J. M. *et al.* (2019) 'Multiple modes of convergent adaptation in the spread of glyphosate-resistant Amaranthus tuberculatus', *Proceedings of the National Academy of Sciences of the United States of America*, 116(42), pp. 21076–21084. doi: 10.1073/pnas.1900870116.

Kyriazis, C. C., Wayne, R. K. and Lohmueller, K. E. (2020) 'Strongly deleterious mutations are a

primary determinant of extinction risk due to inbreeding depression', *Evolution Letters*, n/a(n/a). doi: 10.1002/evl3.209.

Lande, R. (1993) 'Risks of Population Extinction from Demographic and Environmental Stochasticity and Random Catastrophes', *The American naturalist*, 142(6), pp. 911–927. doi: 10.1086/285580.

Lande, R. (1995) 'Mutation and Conservation', *Conservation biology: the journal of the Society for Conservation Biology*, 9(4), pp. 782–791. doi: 10.1046/j.1523-1739.1995.09040782.x.

Leigh, D. M. *et al.* (2019) 'Estimated six per cent loss of genetic variation in wild populations since the industrial revolution', *Evolutionary applications*, 12(8), pp. 1505–1512. doi: 10.1111/eva.12810.

Lewin, H. A. *et al.* (2018) 'Earth BioGenome Project: Sequencing life for the future of life', *Proceedings of the National Academy of Sciences of the United States of America*, 115(17), pp. 4325–4333. doi: 10.1073/pnas.1720115115.

Li, C. *et al.* (2021) 'Genome Variation Map: a worldwide collection of genome variations across multiple species', *Nucleic acids research*, 49(D1), pp. D1186–D1191. doi: 10.1093/nar/gkaa1005.

Li, H. and Durbin, R. (2011) 'Inference of human population history from individual whole-genome sequences', *Nature*, 475(7357), pp. 493–496. doi: 10.1038/nature10231.

Lockwood, D. R., Richards, C. M. and Volk, G. M. (2007) 'Probabilistic models for collecting genetic diversity: Comparisons, caveats, and limitations', *Crop science*, 47(2), pp. 861–866. doi: 10.2135/cropsci2006.04.0262.

Lovell, J. T. *et al.* (2021) 'Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass', *Nature*. doi: 10.1038/s41586-020-03127-1.

Lucek, K. and Willi, Y. (2021) 'Drivers of linkage disequilibrium across a species' geographic range', *PLoS genetics*, 17(3), p. e1009477. doi: 10.1371/journal.pgen.1009477.

Lynch, M. and Lande, R. (1998) 'The critical effective size for a genetically secure population', *Animal conservation*, 1(1), pp. 70–72. doi: 10.1111/j.1469-1795.1998.tb00229.x.

MacArthur, R. H. (1957) 'On the relative abundance of bird species', *Proceedings of the National Academy of Sciences of the United States of America*, 43(3), pp. 293–295. doi: 10.1073/pnas.43.3.293.

MacLachlan, I. R. *et al.* (2021) 'Genome-wide shifts in climate-related variation underpin responses to selective breeding in a widespread conifer', *Proceedings of the National Academy of Sciences of the United States of America*, 118(10). doi: 10.1073/pnas.2016900118.

Marshal, D. R. and Brown, A. D. H. (1975) 'Optimum sampling strategies in genetic conservation', in Frankel, O. H. and Hawkes, J. G. (eds) *Crop genetic resources for today and tomorrow*. Cambridge University Press. Available at: https://www.researchgate.net/publication/280057199.

Matthews, T. J. *et al.* (2019) 'sars: an R package for fitting, evaluating and comparing species–area relationship models', *Ecography*, 42(8), pp. 1446–1455. doi: 10.1111/ecog.04271.

McGill, B. J. *et al.* (2006) 'Rebuilding community ecology from functional traits', *Trends in ecology & evolution*, 21(4), pp. 178–185. doi: 10.1016/j.tree.2006.02.002.

Mérot, C. *et al.* (2020) 'A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation', *Trends in ecology & evolution*, 35(7), pp. 561–572. doi: 10.1016/j.tree.2020.03.002.

Meyer, R. S. *et al.* (2019) 'The California environmental DNA "CALeDNA" program', *bioRxiv*. doi:

10.1101/503383.

Millennium Ecosystem Assessment (2005) *Millennium ecosystem assessment*. Millennium Ecosystem Assessment. Available at: http://chapter.ser.org/europe/files/2012/08/Harris.pdf.

Millette, K. L. *et al.* (2020) 'No consistent effects of humans on animal genetic diversity worldwide', *Ecology letters*, 23(1), pp. 55–67. doi: 10.1111/ele.13394.

Miraldo, A. *et al.* (2016) 'An Anthropocene map of genetic diversity', *Science*, 353(6307), pp. 1532–1535. doi: 10.1126/science.aaf4381.

MOTOMURA and I (1932) 'A statistical treatment of ecological communities', *Zoological Magazine*, 44, pp. 379–383. Available at: https://ci.nii.ac.jp/naid/10009662952/ (Accessed: 29 September 2021).

Novembre, J. and Stephens, M. (2008) 'Interpreting principal component analyses of spatial population genetic variation', *Nature genetics*, 40(5), pp. 646–649. doi: 10.1038/ng.139.

Orr, H. A. and Unckless, R. L. (2014) 'The population genetics of evolutionary rescue', *PLoS genetics*, 10(8), p. e1004551. doi: 10.1371/journal.pgen.1004551.

Overcast, I. *et al.* (2020) 'A unified model of species abundance, genetic diversity, and functional diversity reveals the mechanisms structuring ecological communities', *bioRxiv*. doi: 10.1101/2020.01.30.927236.

Palacio-Mejía, J. D. *et al.* (2021) 'Geographic patterns of genomic diversity and structure in the C4 grass Panicum hallii across its natural distribution', *AoB plants*, 13(2), p. lab002. doi: 10.1093/aobpla/plab002.

Parks, D. H. *et al.* (2013) 'GenGIS 2: Geospatial Analysis of Traditional and Genetic Biodiversity, with New Gradient Algorithms and an Extensible Plugin Framework', *PloS one*, 8(7), p. e69885. doi: 10.1371/journal.pone.0069885.

Paz-Vinas, I. *et al.* (2021) 'Macrogenetic studies must not ignore limitations of genetic markers and scale', *Ecology letters*, 24(6), pp. 1282–1284. doi: 10.1111/ele.13732.

Pereira, H. M. and Daily, G. C. (2006) 'Modeling biodiversity dynamics in countryside landscapes', *Ecology*, 87(8), pp. 1877–1885. doi: 10.1890/0012-9658(2006)87[1877:mbdicl]2.0.co;2.

Petkova, D., Novembre, J. and Stephens, M. (2016) 'Visualizing spatial population structure with estimated effective migration surfaces', *Nature genetics*, 48(1), pp. 94–100. doi: 10.1038/ng.3464.

Pimm, S. L. *et al.* (1995) 'The future of biodiversity', *Science*, 269(5222), pp. 347–350. doi: 10.1126/science.269.5222.347.

Prado, P. I., Miranda, M. and Chalom, A. (2018) *sads: R package for fitting species abundance distributions*. Github. Available at: https://github.com/piLaboratory/sads (Accessed: 12 May 2021).

Preston, F. W. (1948) 'The commonness, and rarity, of species', *Ecology*, 29(3), pp. 254–283. doi: 10.2307/1930989.

Preston, F. W. (1962) 'The canonical distribution of commonness and rarity: Part I', *Ecology*, 43(2), p. 185. doi: 10.2307/1931976.

Purcell, S. *et al.* (2007) 'PLINK: a tool set for whole-genome association and population-based linkage analyses', *American journal of human genetics*, 81(3), pp. 559–575. doi: 10.1086/519795.

Rahbek, C. and Colwell, R. K. (2011) 'Biodiversity: Species loss revisited', *Nature*, pp. 288–289. doi:

10.1038/473288a.

Rockman, M. V. (2012) 'The QTN program and the alleles that matter for evolution: all that's gold does not glitter', *Evolution; international journal of organic evolution*, 66(1), pp. 1–17. doi: 10.1111/j.1558-5646.2011.01486.x.

Royer, A. M., Streisfeld, M. A. and Smith, C. I. (2016) 'Population genomics of divergence within an obligate pollination mutualism: Selection maintains differences between Joshua tree species', *American journal of botany*, 103(10), pp. 1730–1741. doi: 10.3732/ajb.1600069.

Schweizer, R. M. *et al.* (2016) 'Targeted capture and resequencing of 1040 genes reveal environmentally driven functional variation in grey wolves', *Molecular ecology*, 25(1), pp. 357–379. doi: 10.1111/mec.13467.

Seebens, H. *et al.* (2015) 'Global trade will accelerate plant invasions in emerging economies under climate change', *Global change biology*, 21(11), pp. 4128–4140. doi: 10.1111/gcb.13021.

Seebens, H. *et al.* (2017) 'No saturation in the accumulation of alien species worldwide', *Nature communications*, 8, p. 14435. doi: 10.1038/ncomms14435.

von Seth, J. *et al.* (2021) 'Genomic insights into the conservation status of the world's last remaining Sumatran rhinoceros populations', *Nature communications*, 12(1), p. 2393. doi: 10.1038/s41467-021-22386-8.

Shaffer, H. B. and Toffelmier, E. (2020) 'California Conservation Genomics Project First Year Annual Report'. Available at: https://escholarship.org/content/qt2sc7s29z/qt2sc7s29z.pdf.

Simons, Y. B. *et al.* (2018) 'A population genetic interpretation of GWAS findings for human quantitative traits', *PLoS biology*, 16(3), p. e2002985. doi: 10.1371/journal.pbio.2002985.

Smeds, L. *et al.* (2021) 'Whole-genome analyses provide no evidence for dog introgression in Fennoscandian wolf populations', *Evolutionary applications*, 14(3), pp. 721–734. doi: 10.1111/eva.13151.

Storch, D., Keil, P. and Jetz, W. (2012) 'Universal species-area and endemics-area relationships at continental scales', *Nature*, 488(7409), pp. 78–81. doi: 10.1038/nature11226.

Supple, M. A. *et al.* (2018) 'Landscape genomic prediction for restoration of a Eucalyptus foundation species under climate change', *eLife*, 7. doi: 10.7554/eLife.31835.

Theodoridis, S., Rahbek, C. and Nogues‐Bravo, D. (2021) 'Exposure of mammal genetic diversity to mid‐21st century global change', *Ecography*, (ecog.05588). doi: 10.1111/ecog.05588.

Thomas, C. D. *et al.* (2004) 'Extinction risk from climate change', *Nature*, 427(6970), pp. 145–148. doi: 10.1038/nature02121.

Tokeshi, M. (1990) 'Niche Apportionment or Random Assortment: Species Abundance Patterns Revisited', *The Journal of animal ecology*, 59(3), pp. 1129–1146. doi: 10.2307/5036.

Tokeshi, M. (1993) 'Species Abundance Patterns and Community Structure', in Begon, M. and Fitter, A. H. (eds) *Advances in Ecological Research*. Academic Press, pp. 111–186. doi: 10.1016/S0065-2504(08)60042-2.

Tuskan, G. *et al.* (no date) *Populus Trichocarpa Genome-Wide Association Study (GWAS) Population SNP Dataset Released*. Available at: https://doi.ccs.ornl.gov/ui/doi/55 (Accessed: 2 June 2021).

Vallejo-Marín, M. *et al.* (2021) 'Population genomic and historical analysis suggests a global invasion

by bridgehead processes in Mimulus guttatus', *Communications biology*, 4(1), p. 327. doi: 10.1038/s42003-021-01795-x.

Willi, Y. (2013) 'Mutational meltdown in selfing Arabidopsis lyrata', *Evolution; international journal of organic evolution*, 67(3), pp. 806–815. doi: 10.1111/j.1558-5646.2012.01818.x.

Wright, S. (1943) 'Isolation by Distance', *Genetics*, 28(2), pp. 114–138. Available at: https://www.ncbi.nlm.nih.gov/pubmed/17247074.