Opportunities and challenges of macrogenetic studies

Deborah M. Leigh[1,21,23], *Charles B. van Rees*[2,22,23], *Katie L. Millette*[3,4,21,23], *Martin F. Breed*[5,20,21], *Chloé Schmidt*[6,21], *Laura D. Bertola*[7,8,20,21], *Brian K. Hand*[2,21], *Margaret E. Hunter*[9,20,21], *Evelyn L. Jensen*[10,21], *Francine Kershaw*[11,21], *Libby Liggins*[12,21], *Gordon Luikart*[2,20,21], *Stéphanie Manel*[13,21], *Joachim Mergeay*[14,15,20,21], *Joshua M. Miller*[16,21], *Gernot Segelbacher*[17,20,21], *Sean Hoban*[18,20,21] ✉ *and Ivan Paz-Vinas*[19,21] ✉

Abstract | The rapidly emerging field of macrogenetics focuses on analysing publicly accessible genetic datasets from thousands of species to explore large-scale patterns and predictors of intraspecific genetic variation. Facilitated by advances in evolutionary biology, technology, data infrastructure, statistics and open science, macrogenetics addresses core evolutionary hypotheses (such as disentangling environmental and life-history effects on genetic variation) with a global focus. Yet, there are important, often overlooked, limitations to this approach and best practices need to be considered and adopted if macrogenetics is to continue its exciting trajectory and reach its full potential in fields such as biodiversity monitoring and conservation. Here, we review the history of this rapidly growing field, highlight knowledge gaps and future directions, and provide guidelines for further research.

Phylogeography
The study of how historical events have helped to shape the current geographical distribution of genetic lineages within and among closely related species.

Biogeography
The study of the spatiotemporal distribution of species, communities and ecosystems.

Macroecology
The study of broad-scale ecological patterns and processes, including topics such as metabolic scaling, extinction risk and diversity gradients.

Macrophysiology
The study of variation in physiological traits for multiple species over large geographic and temporal scales and the ecological implications of this variation.

✉*e-mail:*
shoban@mortonarb.org;
ivanpaz23@gmail.com

https://doi.org/10.1038/s41576-021-00394-0

The end of the twentieth century saw a transition in major biology disciplines. Studies shifted from a local focus to studying patterns and processes across entire continents and biomes[1–4]. Phylogeography expanded on the earlier field of biogeography[5,6] by focusing on the current distributions of genetic lineages and the historical processes that generated them[4,7,8]. Macroecology investigated the underlying processes and mechanisms generating patterns of species abundance, distribution and diversity across different scales[1,9], while macrophysiology investigated global patterns of intraspecific variation in physiological traits (such as mammalian basal metabolic rate[10] and plant leaf characteristics[11]) and their global environmental and geographical predictors[2,3,12]. These fields vastly improved the understanding of natural systems, with implications for addressing the challenges arising from human-mediated global change[12,13].

Until recently, analyses in population genetics, evolutionary biology and molecular ecology rarely reached broad taxonomic, spatial or temporal scales. Studies characterizing intraspecific genetic variation (IGV), which encompasses within-population genetic diversity and among-population genetic differentiation, usually included one or a handful of taxa, sampled at few locations and distributed irregularly across ecosystems[14,15]. Comparative population genetics and phylogeography studies and meta-analyses examining predictors of genetic patterns from even tens of species have long remained scarce[16] owing to the cost of molecular methods, the logistical constraints inherent to vast sampling designs[7] and the relatively late adoption of meta-analytical approaches by evolutionary disciplines[17]. Since the early 2010s, however, advances in molecular biology, data availability, open-access software, remote sensing, biostatistics, spatial ecology, comparative phylogeography and landscape genetics have paved the way for a new field that has been named macrogenetics[14].

Macrogenetic studies explore the patterns and predictors of IGV across many taxa (dozens to thousands) at broad taxonomic, spatial, and/or temporal scales[14]. IGV is a key biodiversity facet that reflects the evolutionary history, biological connectivity, adaptability and viability of populations, species and communities[15,18]. It may be represented by metrics such as allelic richness, heterozygosity, nucleotide diversity, and fixation index or by more complex metrics informed by population genetics theory such as gene flow estimations or effective population sizes (N_e). Predictors that have been studied include latitude[19–21], environmental factors such as climatic variables or habitat productivity[22,23], species characteristics such as life-history traits[19,24,25], and anthropogenic factors, including land use or urbanization[20,26,27]. Examining IGV from a macrogenetic perspective helps us to understand the evolutionary processes that generate, maintain and erode biodiversity, offering a precise and rapid account of biodiversity changes in

Intraspecific genetic variation
(IGV). Genetic variation observed at the DNA level within a species, including within-population genetic diversity and among-population genetic differentiation. It can be measured using many metrics, including gene diversity, allelic richness and nucleotide diversity.

Landscape genetics
The study of the effects of the environment including recent global change (such as climate or land-use change) on genetic patterns, and of how species will adapt to these changes on ecological timescales.

Fixation index
A metric indicating the nearness of fixation (from 0 to 1) of a subpopulation (S) relative to the total sampled population (T), which is frequently used to assess genetic differences among populations.

response to large-scale environmental impacts, and ultimately improving biodiversity conservation at the within-species and community scales.

In this Review, we first identify three classes of macrogenetic studies determined by how underlying genetic data are obtained (FIG. 1). We describe the transition from population genetic studies of single (or few) species (or populations) to studies analysing genetic data from dozens to thousands of species (for example, 17,082 vertebrate species[27]) at continental to global scales. We then summarize major hypotheses and findings of macrogenetic studies and highlight important challenges of the most recently emerged class of macrogenetic studies (that is, Class III; FIG. 1), which is repurposing the vast amounts of genetic and predictor data currently available to contribute to our understanding of evolutionary biology and inform biodiversity conservation. We propose future directions, including how to expand and exploit macrogenetic findings, and discuss the role of genomics in shaping the field moving forwards.

Classes of macrogenetic studies

To better describe common difficulties and best practices for macrogenetic studies, we differentiate among three classes of macrogenetic studies (FIG. 1). Class I macrogenetic studies generate raw genetic data for several species sampled simultaneously in the same study area (for example, REFS[28–30]; FIG. 1). The multispecies genetic data collected typically share a similar sampling design and molecular markers and involve a restricted collaborative group of researchers (such as the IntraBioDiv Consortium[30]). Class I studies are typically limited in taxonomic and spatial breadth because they require substantial field and laboratory work and associated funding.

Class II macrogenetic studies compile IGV summary statistics (such as heterozygosity, allelic richness, fixation index) from published datasets to evaluate patterns across multiple populations and species in relation to spatial and/or other environmental factors, including through the use of meta-analysis techniques[25,31,32] (FIG. 1). Class II studies tend to have a greater number of species, taxonomic representation (generally 50–100 species; Supplementary Table 1, but see REF.[25]) and spatial coverage (continental to global) compared to Class I. Although the mixed origins and data types of Class II studies may impart challenges in data analysis and interpretation (for example, lack of standardization of IGV metrics across studies, variation in sampling designs or marker types considered), authors can refer to the original publications for context, hence offering opportunities to incorporate important information relevant to the data (such as the marker type used to estimate IGV, the delineation of local populations and original sample sizes[25]) to overcome such issues.

Class III macrogenetic studies extract, aggregate and re-analyse previously collected genetic data from public repositories (such as GenBank, Barcode of Life Data system (BOLD)[33] or DRYAD) and harvest data on potential predictors from global databases (such as WorldClim[34], Global Biodiversity Information Facility (GBIF) or TRY[35]) to conduct multispecies IGV analyses (FIG. 1). Class III studies use raw data (like Class I); however, the large size of the aggregated datasets (for example, >10,000 sequences from >1,000 species[20]) means authors rarely refer to the original publication for context (unlike Class II) and important information is often lost. Class III studies face novel and important data challenges due to the variety of data origins, data archive issues, and wide taxonomic and spatial coverage. However, these studies are growing in number, scale and type of repurposed genetic data[19–21,23,26,36] and are tackling groundbreaking questions. For these reasons, this article will focus mainly on Class III macrogenetic studies.

It is noteworthy that the classes of macrogenetic studies will likely change with the field. Studies can bridge multiple classes when the underlying genetic data are extracted from multiple sources (Supplementary Table 1). Furthermore, our classifications include a small number of early studies arising from landscape genetics, comparative phylogeography and cross-taxonomic population genetic assessments (Supplementary Table 1) as these fields have independently derived ways to conduct IGV investigations at large scales[37].

The emergence of macrogenetics

Macrogenetics has similarities to other 'macro' fields, such as macroecology, although it has a distinct history. Macroecology emerged from population and community ecology, disciplines that were historically based on empirical observations and field experiments. Unifying theory (such as the unified neutral theory of biodiversity and

Author addresses

[1]Swiss Federal Research Institute WSL, Birmensdorf, Switzerland.
[2]Flathead Lake Biological Station, Polson, MT, USA.
[3]Department of Biology, McGill University, Montreal, QC, Canada.
[4]Secretariat of the Convention on Biological Diversity (SCBD), Montreal, QC, Canada.
[5]College of Science and Engineering, Flinders University, Bedford Park, SA, Australia.
[6]Department of Biological Sciences, University of Manitoba, Winnipeg, Manitoba, Canada.
[7]City College of New York, New York, NY, USA.
[8]Department of Biology, Section for Computational and RNA Biology, University of Copenhagen, Copenhagen, Denmark.
[9]U.S. Geological Survey, Wetland and Aquatic Research Center, Gainesville, FL, USA.
[10]School of Natural and Environmental Sciences, Newcastle University, Newcastle Upon Tyne, UK.
[11]Natural Resources Defense Council, New York, NY, USA.
[12]School of Natural and Computational Sciences, Massey University Auckland, Auckland, Aotearoa, New Zealand.
[13]CEFE, Univ Montpellier, CNRS, EPHE-PSL University, IRD, Montpellier, France.
[14]Research Institute for Nature and Forest, Geraardsbergen, Belgium.
[15]Aquatic Ecology, Evolution and Conservation, KULeuven, Leuven, Belgium.
[16]Department of Biological Sciences, University of Alberta, Edmonton, Canada.
[17]Chair of Wildlife Ecology and Management, University Freiburg, Freiburg, Germany.
[18]Center for Tree Science, The Morton Arboretum, Lisle, USA.
[19]Laboratoire Evolution & Diversité Biologique, Université de Toulouse, Centre National pour la Recherche Scientifique, Institut de Recherche pour le Développement, Université Toulouse III Paul Sabatier, UMR5174 EDB, Toulouse, Toulouse, France.
[20]Conservation Genetics Specialist Group, International Union for Conservation of Nature (IUCN), Gland, Switzerland.
[21]Group on Earth Observation Biodiversity Observation Network (GEO BON) Genetic Composition Working Group, Montreal, QC, Canada.
[22]Group on Earth Observation Biodiversity Observation Network (GEO BON) Species populations and Freshwater BONs, Montreal, QC, Canada.
[23]These authors contributed equally: Deborah M. Leigh, Charles B. van Rees, Katie L. Millette.

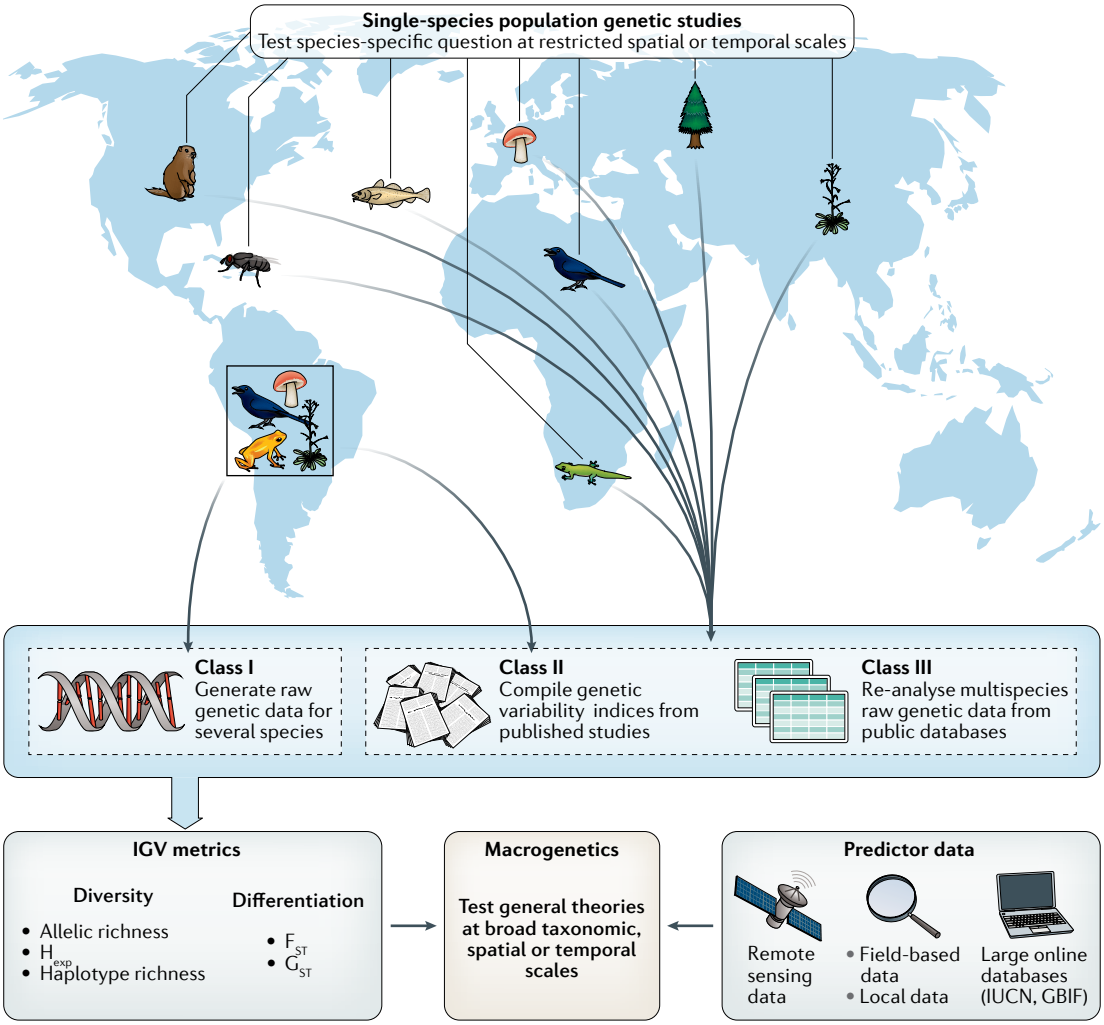


Fig. 1 | **What makes a macrogenetic study?** Macrogenetic studies investigate genetic patterns across many species and across large spatial and/or temporal scales. Genetic data is either generated in situ for multiple species (Class I), extracted and summarized from the literature (Class II), or aggregated from data repositories for many populations of multiple species (Class III; although not all studies fit neatly in just one category, see Supplementary Table 1). Genetic data is analysed in parallel with environmental, ecological or socio-economic predictors that can be based on local data (such as field-based data for Class I studies) and/or large-scale data from online databases or remote sensing (notably for Class II and III studies). By upscaling traditional single species population genetic investigations, macrogenetic studies are investigating patterns and predictors of intraspecific genetic variation worldwide. F_{ST}, fixation index; GBIF, Global Biodiversity Information Facility; G_{ST}, coefficient of genetic differentiation; H_{exp}, expected heterozygosity; IGV, intraspecific genetic variation; IUCN, International Union for Conservation of Nature.

biogeography[38]) was subsequently derived to help understand the processes and mechanisms underlying the observed macroecological patterns. Macrogenetics has taken longer to emerge and it is rooted in population genetics, which has an early and well-established theoretical foundation. Conceived in the first half of the twentieth century[39–41], population genetics theory outlines how IGV varies under several population models such as the Wright–Fisher model[40,41] or the stepping-stone model[42]. Fifty years ago, the coalescent theory[43] and the neutral theory of molecular variation[44] further contributed to our understanding of population genetics and IGV as did the subsequent rapid growth of disciplines such as phylogeography[4,7,8]. As empirical data accumulated in the 1970–1980s, population geneticists began testing theoretical predictions in natural populations[45,46] and progressive developments in molecular biology, including PCR[47], Sanger sequencing[48] and high-throughput genotyping techniques such as restriction site-associated DNA sequencing[49], increased the capacity to produce massive amounts of DNA sequence data cheaply and quickly[14,15] (FIG. 2).

Important developments in obtaining genetic data were not limited to molecular techniques. Advances in non-invasive or minimally invasive sampling[50] expanded our ability to sample natural populations (such as DNA from faeces, urine, skin, mucus, hair, soil, water or air samples), while the development of DNA barcoding (and metabarcoding) techniques for species identification[51,52] fostered the generation of DNA sequences for many species. Lastly, recent improvements in DNA extraction and analysis from historical or ancient samples have opened

Effective population sizes (N_e). A concept that helps represent how fast a given population is expected to lose genetic diversity; it is often only 10–20% of the population census size.

Unified neutral theory of ecology and biogeography A model inspired by the neutral theory of molecular evolution that explains species biodiversity patterns assuming ecologically equivalent species.

Wright–Fisher model A selectively neutral mathematical model that describes allele frequency change across discrete generations in an idealized population.

Stepping-stone model A statistical model of metapopulation connectivity in which each subpopulation can only exchange migrants with its nearest neighbours. This constraint leads to a pattern of genetic isolation by distance.

Coalescent theory A theory developed to model how allele copies sampled from a population originate from (coalesce in) a common ancestor and used to develop neutral expectations and infer the demographic history of populations.

Neutral theory of molecular evolution A model of evolution that assumes that most genetic diversity at the molecular level in populations and species is the result of neutral (non-selective) processes such as genetic drift and mutation.

Restriction site-associated DNA sequencing A genotyping method whereby thousands of short regions (100–300 bp) of DNA surrounding a restriction enzyme site are sequenced and variants are identified.

DNA barcoding A method of identifying what species a DNA sample belongs to by comparing a particular DNA sequence with a database containing reference sequences of many species.

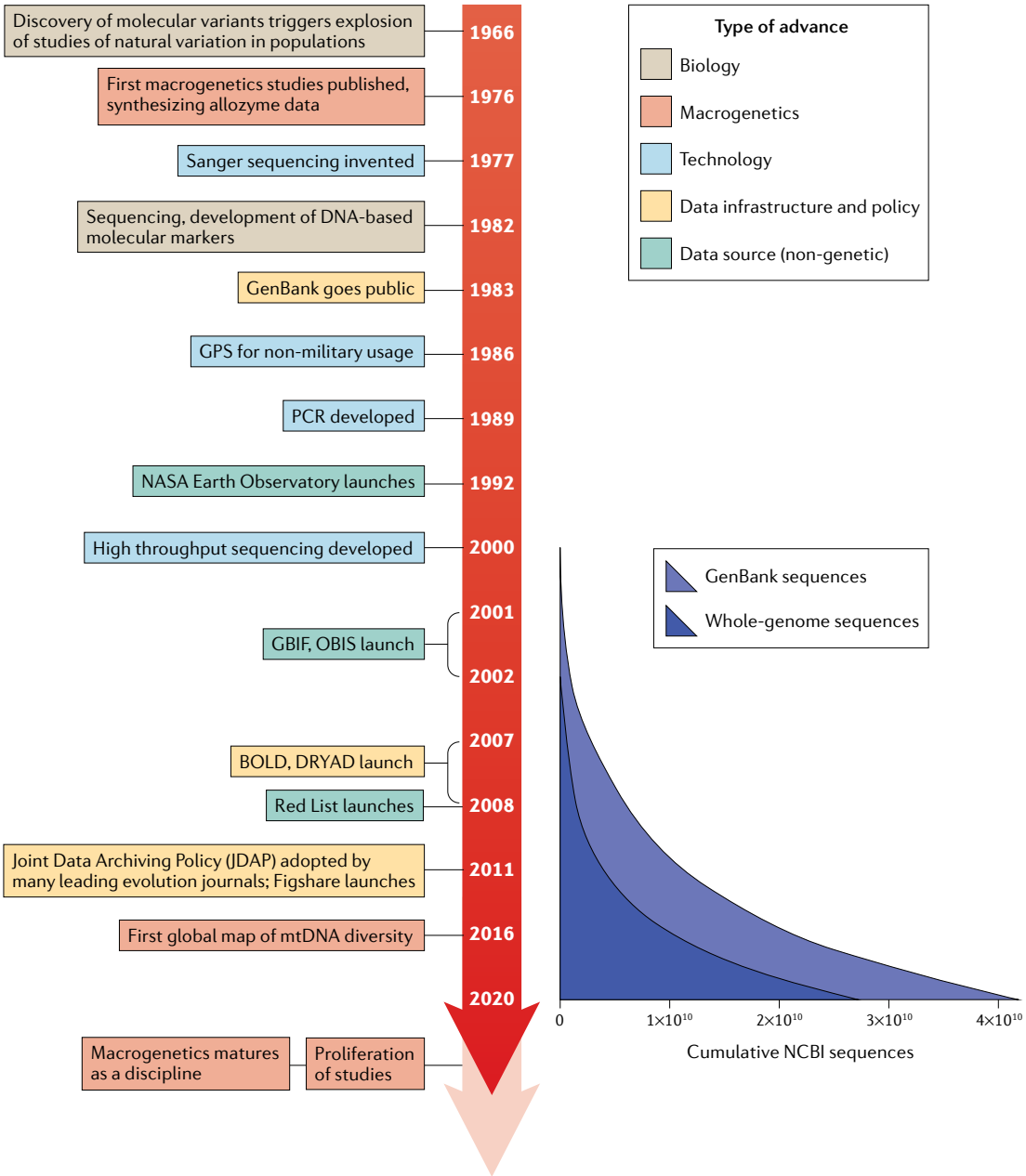


Fig. 2 | **Timeline of key advances underlying the emergence of macrogenetics.** Pinpointed are key advances in biology, technology, data infrastructure and data policy that have facilitated the emergence of macrogenetics. Also shown is the accumulation of publicly accessible genetic data (cumulative number of archived sequences and whole-genome sequences in GenBank since its first release; data source: NCBI website, last accessed December 7, 2020). This accumulation of accessible data is fundamental to Class III macrogenetic studies. BOLD, Barcode of Life Data system; GBIF, Global Biodiversity Information Facility; GPS, Global Positioning System; mtDNA, mitochondrial DNA; OBIS, Ocean Biodiversity Information System.

museums and herbaria to geneticists, expanding the spatial and temporal span of available genetic data[53,54]. Combined, these technical improvements led to a rapid increase in the production of DNA sequences (FIG. 2; Supplementary Figure 1A) and facilitated the emergence of macrogenetics[14].

Improvements in genetic data quality, storage, accessibility, sharing and re-use policies have been critical for the development of Class III macrogenetic studies (FIG. 2). The release of GenBank in 1982 constituted one of the earliest bioinformatics open data community projects worldwide[55]. GenBank (in partnership with the International Nucleotide Sequence Database Collaboration) archives an exponentially increasing amount of genetic data from all organisms. Other curated genetic databases (for example, BOLD) have fostered the accumulation of DNA sequences for specific loci such as *COI* for animals or *rbcL* (also known as *cbbL*) for plants. More recently, general-purpose data repositories, such as Dryad or Figshare, have made refined

COI
A mitochondrial DNA gene sequence that encodes cytochrome C oxidase subunit I and is frequently used for species identification via DNA barcoding in Metazoa.

rbcL (or *cbbL*)
A plant chloroplast gene sequence that encodes ribulose bisphosphate carboxylase large chain and is frequently used for species identification via DNA barcoding in plants.

Interoperability
In the context of genetic and genomic data, refers to the ability of different datasets to be connected and integrated at present and in the future owing to standardized formats, storage, metadata and accessibility.

Species–genetic diversity correlation concept
A concept that suggests patterns of species and intraspecific genetic diversity are correlated because they both are influenced by the same underlying processes (such as stochasticity, selection, dispersal, speciation or mutation) and environmental variation.

genotypic microsatellite and SNP datasets more accessible, though search, aggregation and re-use of this data can be challenging (see 'Best practices for macrogenetic studies').

Data deposition was further boosted by policies mandating accessibility of scientific data such as the Joint Data Archiving Policy[56] and the Findable, Accessible, Interoperable, and Reusable (FAIR) guiding principles[57]. These principles emphasize the importance of the interoperability and reproducibility of genetic datasets and have prompted the curation of associated spatial and ecological metadata (for example, in the Genomics Observatories Metadatabase (GEOME)[58]).

Macrogenetic studies, especially those in Class III, synthesize available genetic data along with existing knowledge and methods from multiple disciplines, including landscape genetics, biogeography, comparative phylogeography, remote sensing, community ecology and biostatistics, among others. Therefore, the growth of macrogenetics relies on advancements in informatics (FIG. 2), especially flexible open-source programming languages for statistical computing and data mining such as R[59]. The huge increase in the number of multidisciplinary R packages (Supplementary Fig. 1), fostered by open-source code-sharing platforms such as GitHub, now allows researchers to collaborate and conduct an increasing number of tasks using a unique environment (Supplementary Fig. 2).

Advances in concepts and tools of landscape genetics were particularly beneficial to macrogenetics by facilitating the testing of environmental predictors of IGV patterns at various spatial scales[60]. Landscape genetic studies have revealed how to improve the interpretation of complex multivariate correlations among genetic and environmental predictors[61], progress our understanding of the effect of key environmental predictors such as climate change variables[62], and take a temporal view when investigating predictor effects on IGV[63].

Progress on comparative phylogeography has also advanced macrogenetics. Comparative phylogeography often handles large-scale multispecific genetic data to illustrate how biogeographical and historical processes (for example, climate-induced demographic changes or vicariance owing to geological features or events, typically thousands to millions of years ago) determine the current spatial distribution of genetic lineages, generally among clades or species[64,65]. Although similar, macrogenetics differs from comparative phylogeography by broadly examining all scales of IGV (spatial, temporal and taxonomic scales[14]) and by having a broader focus on recent predictors of observed patterns rather than evolutionary history. Nevertheless, comparative phylogeography has advanced macrogenetics by facilitating large-scale public genetic data repurposing and aggregation methods[66,67]. It has also helped develop sensitivity analyses aimed at testing sample size issues[68] and has shown the value of integrating trait-based data into explorations based on large-scale genetic data[66].

Finally, the emergence of Class III macrogenetic studies also required robust data to test potential explanatory variables, covering the same large spatiotemporal and taxonomic scales considered for IGV assessment. The development of powerful Geographic Information Systems and of scientific research and journals (including *Scientific Data* and *GigaScience*, among others) dedicated to data have facilitated the accessibility of spatialized data from a range of disciplines, including remote sensing, climatology, community ecology, and social and economic sciences. This includes data on current and past land use (for example, from LUCAS[69] or Copernicus Global Land Cover[70]), variables relating to climate (for example, from WorldClim[34] or CHELSA) and geoscience (for example, from NASA's Earth observatory platform), functional traits (for example from TRY[35] or PanTHERIA[71]), community taxonomic composition[72], and species distributions (for example, from GBIF or International Union for Conservation of Nature (IUCN) global databases).

Macrogenetic patterns and predictors

Main hypotheses tested by macrogenetic studies. Correlations between species diversity and genetic diversity (as described by the species–genetic diversity correlation concept[73]) is a popular hypothesis that has been explored from a macrogenetic perspective (TABLE 1). These correlations are hypothesized to be generated by two main, though not exclusive, mechanisms. First, elevated genetic diversity underlying functional traits in foundation species (such as phenology or biochemical composition in plant leaves) may promote high species diversity in the surrounding community via biotic interactions (for example, with herbivorous insects and pollinators), subsequently increasing community stability[15,73]. Second, the same variables that influence species richness and community taxonomic composition patterns (such as geographical isolation[74]) may similarly affect IGV[15,73]. Early investigations of species–genetic diversity correlations that were based on a handful of species[74,75] have been upscaled in recent macrogenetic studies to tens (for example, REF.[28]) and thousands of species[22,23,76] (see Supplementary Table 1 for detailed examples). Global-scale Class III studies based on mitochondrial loci detect positive relationships between species and genetic diversities in mammals[23] and fish[22]. By contrast, a recent preprint highlights a putative negative relationship between nuclear genetic diversity and species richness in North American mammals, with resource availability and habitat heterogeneity governing diversity at both genetic and species levels[77].

The latitudinal gradient in species diversity is one of the oldest patterns recognized in ecology[78]. There has been considerable interest in using macrogenetic studies to test for latitudinal patterns in IGV (for example, REFS[20,21,25]; TABLE 1). Two general hypothesized predictors for this pattern prevail. First, long-term climatic stability in the tropics may facilitate large and stable population sizes with high genetic diversity at low latitudes compared to temperate zones[79]. Second, high temperatures at low latitudes may drive IGV via increased metabolic rates, mutation rates and faster generation times[79]. Whether these processes drive congruent patterns across nuclear and organellar loci and for neutral and adaptive IGV remains unclear[80,81]. Indeed, some macrogenetic studies reveal high IGV in mitochondrial and nuclear adaptive DNA markers at low latitudes[19–21,27]. By contrast,

Table 1 | **Summary of the main scientific questions, hypotheses and findings of published macrogenetic studies**

Question	Hypothesis	Summary of main findings	Marker types	Genetic metrics	Study number reference in Supplementary Table 1
What is the relationship between genetic diversity and species richness (positive, negative, null) and what processes drive it?	Genetic diversity and species richness are shaped by the same processes (including evolutionary speed, evolutionary time or ecological limits) acting simultaneously at both levels	mtDNA diversity is positively correlated with species richness; relationships with neutral nuclear diversity are unclear	Allozymes, mtDNA, SSRs	Div	2,19,42,48,50,53,54,61
Are there gradients in genetic diversity and what are the causes?	Environmental factors thought to underlie species richness gradients create parallel gradients in genetic diversity; can be explained by evolutionary speed and energy-richness hypotheses	mtDNA (non-neutral) genetic diversity increases towards the tropics; no clear latitudinal gradient detected in neutral nuclear genetic diversity	mtDNA, SSRs	Div	7,19,29,30,42,48, 50,52–54,58
Why does genetic diversity differ between species?	Life-history traits that influence N_e should cause differences between species	Life-history and ecological traits that influence N_e affect neutral nuclear genetic diversity more so than historical factors	Genomic data, SSRs, AFLPs, allozymes, mtDNA	Div	3–6,8–11,13,14,24, 28,32,34,35,45,47,49,56
What factors influence population structure and differentiation?	Traits that affect dispersal and gene flow will cause consistent differences in population structure across species and higher taxonomic levels	Isolation by environment is more common than isolation by distance; dispersal ability, habitat specialization, habitat connectivity, abundance, mating system, geographic range size and N_e influence patterns of genetic differentiation	SSRs, allozymes, AFLPs, mtDNA, cpDNA	Diff	5,11,12,20,25,27,28, 33,34,36,39,43,51
What is the relationship between genetic diversity and population size in natural populations?	More abundant species will have larger N_e, thus higher genetic diversity	Wildlife N_e are often much lower than census population size; diversity measured with neutral nuclear DNA markers is positively related to population size	SSRs, allozymes, mtDNA	Div	15–17,46
What are the relationships between genetic diversity, adaptive potential and conservation status?	Neutral genetic diversity is correlated with genome-wide diversity and should be positively correlated with adaptive potential; threatened species will have lower genetic diversity and adaptive potential	Molecular markers are only weakly correlated with adaptive potential; most species are negatively affected by genetic factors before extinction	SSRs, SNPs	Div	5,11,12,20,25,27,28, 33,34,36,39,43,51
Are there broad-scale patterns in adaptive genetic diversity and what drives them?	MHC plays a key role in vertebrate parasite resistance; therefore, positive selection is expected to shape MHC-adaptive genetic diversity	Adaptive IGV is higher at low latitudes, in small body mass mammals and under positive selection	*MHC* gene	Div	60
How does human activity affect genetic variation?	Fragmentation in urban habitats reduces population size and connectivity, reducing neutral genetic diversity and increasing genetic differentiation	Neutral nuclear genetic diversity in wild populations has decreased since the nineteenth century; human land-use change consistently negatively affects mammalian nuclear genetic diversity, with species-specific effects in amphibians and birds; no consistent effects detected in mtDNA diversity	SSRs, SNPs, mtDNA	Div Diff	5,21,22,26,40,41, 43,44,54,55,57

AFLP, amplified fragment length polymorphism; cpDNA, chloroplast DNA; Diff, genetic differentiation metrics (such as estimates of genetic differentiation, including the fixation index (F_{ST}), coefficient of genetic differentiation (G_{ST}) and Φ_{ST}); Div, genetic diversity metrics (such as heterozygosity or allelic richness); IGV, intraspecific genetic variation; mtDNA, mitochondrial DNA; N_e, effective population size; SSRs, simple sequence repeats (also known as microsatellites).

others suggest a lack of, or even a negative, relationship between nuclear genetic diversity and latitude, emphasizing the importance of possible fine-scale drivers and putative discordance among molecular markers[25,76,77,82] (TABLE 1; Supplementary Table 1).

Environment and life history are also key factors affecting IGV patterns although disentangling their respective interacting effects on IGV with latitudinal and biogeographic factors is challenging. Classes I and II macrogenetic studies were among the first to

broadly examine relationships between IGV and species life-history and ecological traits[45,83], demonstrating that traits influencing N_e (such as body size and trophic level) can predict levels of nuclear IGV across taxa (TABLE 1). A Class III study of 8,955 species[19] showed that range size, elevation and latitude were stronger predictors of mitochondrial IGV than life-history traits and that life history contributed minimally to mitochondrial IGV in Nearctic amphibians[24]. Adaptive IGV in genes in the major histocompatibility complex (MHC) locus was shown to vary with latitude and gradients were related to life-history traits such as body mass and litter size in 93 terrestrial mammals[21]. Class III studies are also able to leverage georeferenced data to study spatial IGV patterns and how they interact with environmental predictors. In fish, for example, sea surface temperature contributed most to mitochondrial IGV patterns in marine species, whereas hydrographical region and average slope were among the dominant factors for freshwater species[22]. Mitochondrial genetic diversity has also been shown to co-vary with past rapid climate change and to increase with mean annual[84] or interannual precipitation variability[23].

Understanding the effects of anthropogenic activity on IGV is an important area of macrogenetic investigation (TABLE 1). Anthropogenic activities are expected to decrease genetic diversity by disrupting gene flow, lowering N_e and thereby increasing the strength of genetic drift[85]. Class II approaches have explored the impact of human land-use change on IGV, reporting both general and species-specific effects[31,32,86–88]. Class III studies also have heterogeneous and taxon-specific results. Mitochondrial IGV was negatively associated with human disturbance in amphibians and mammals[20], while human density and land use had negative effects on insects and fish, respectively[27]. Additional studies suggest mammalian mitochondrial IGV may be unrelated to human land use[23,82], yet neutral nuclear IGV in mammals but not in birds or amphibians seems to be consistently lower in urban environments[26,89]. Both positive and negative temporal trends in mitochondrial IGV patterns across taxa during the past four decades were reported[27], while a Class II study spanning >100 years found global trends of genetic diversity loss since the 1800s[88]. However, human disturbance does not reduce IGV in all populations, especially in stocked species (such as harvested fish or ungulates) or species living in or around human settlements (for example, bed bugs, pigeons, rats or ornamental plants[85]).

Variability across macrogenetic studies. We see variable and contradictory patterns across macrogenetic studies, which is not surprising considering the variety of approaches used in each study. Notable patterns of IGV are emerging as the number of studies increases, especially with regard to marker types[84] and taxonomic groups[82]. This observation underscores two important future considerations. First, the biological characteristics of genetic marker types (nuclear versus organellar and neutral versus non-neutral) must be considered fully in study design, analysis and interpretation[81]. Nuclear and organellar DNA have varying relationships with population size and demography and neutral IGV correlates more strongly with population size and demographic structure than do non-neutral loci[90–92]. Second, macrogenetics must increasingly consider how species traits contribute to diverging IGV patterns. The heterogeneous impacts of human activity on IGV may be related to species and local population characteristics such as dispersal abilities and population connectivity[93], which current global datasets may not detect.

Current macrogenetic studies demonstrate a lack of strong global relationships between IGV patterns and predictors. Often, less than 20% of IGV is explained by chosen predictors (for example, REFS[22,26]). It is unclear if this phenomenon is biological, driven by weak predictors or is an artefact of the analytical constraints (detailed in 'Best practices for macrogenetic studies'). First, weak IGV patterns may arise because of the cumulative effect of multiple predictors acting asynchronously over time or space. The impact of one predictor on a population's gene pool may be obscured by the effect of others acting at another time or place on the same gene pool. Second, a mismatch between mutation rates, generation intervals, evolutionary history and selection timescales may limit the detectable effects of predictors on IGV. Specifically, temporally and spatially consistent predictors (such as latitude and species-specific life-history traits) or past biogeographical processes (such as glacial cycles and colonizations) might have deeply affected IGV because they can act over multiple generations. By contrast, predictors acting at more recent eco-evolutionary timescales (such as land-use change) might have acted upon too few generations to induce detectable effects on IGV. This is particularly true for environmental change predictors, which can experience delayed evolutionary responses[94]. Third, mismatches between the distribution of genetic data and the geographic area affected by fine-scale predictors (such as river slope and human population density) might also obscure predictor effects.

Best practices for macrogenetic studies

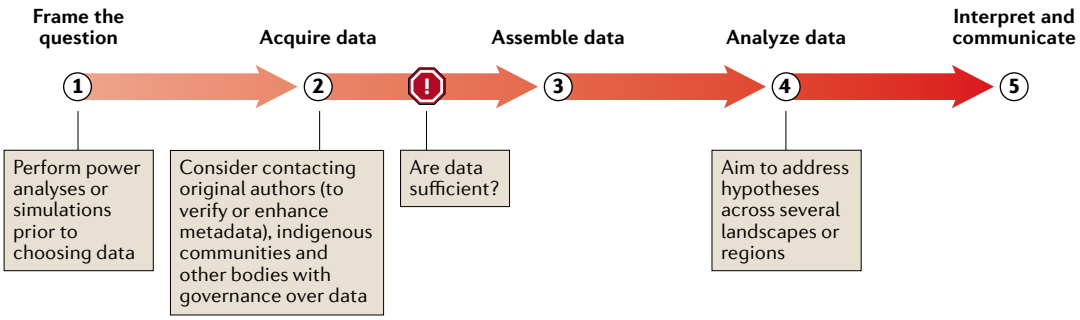
In this section, we review conceptual and methodological challenges in macrogenetics and provide guidelines for future research. Given the nature of this emerging field, some of our recommendations are based on rationales or theoretical knowledge. Nevertheless, the future needs of macrogenetics include testing these guidelines across a range of data types and taxonomic groups. We organize key points according to an idealized workflow for Class III macrogenetic studies (BOX 1).

Framing the question and acquiring data. Data from Class III studies are generally aggregated from different sampling events and locations to obtain analytical units with sufficient sample sizes or time series lengths, with the ultimate goal of creating a dataset that is representative and appropriate for the research question. However, aggregating genetic samples across space or time introduces the risk of evaluating spatiotemporal resolutions that do not match the scale at which population genetic processes operate[95]. Spatially aggregated data may not correspond to biologically meaningful units (such as populations or synchronous generations) or reflect

Box 1 | **Workflow for Class III macrogenetic studies**

We recommend a framework of best practices and important questions that authors should consider when conducting a Class III macrogenetic study (though many of them can also be useful when conducting Classes I and II studies). These criteria are essential to assess data quality, suitability and the robustness of findings, which are core to the correct interpretation and communication of results.

1. Frame the question
   - Marker: What is the intraspecific genetic variation (IGV) metric of interest (for example, neutral diversity, adaptive diversity)?
   - Scope: What taxa will you work with? What is the temporal extent?
   - Scale: Are sampling units species, populations or grid cells? Are covariate data available at desired resolution? Are appropriate temporal baselines available?
2. Acquire data
   - Databases: Are there systematic biases in the data (for example, only unique variants archived; see Table 2)?
   - Systematic search and filtering: Develop pertinent keywords and inclusion criteria (for example, exclude invasive species). Set quality thresholds (such as minimums for sequences or species per grid cell, individuals per population). Are required data metadata available (for example, georeferenced)? Are there cultural sensitivities or other ethical considerations in reusing the data?
3. Assemble data
   - Aggregating data: It is appropriate to combine different markers, taxa or data from varied spatiotemporal coverage?
   - Georeferencing: Develop consistent georeferencing methods (such as coordinates in metadata, GeoNames search).
   - IGV metrics: Ensure the IGV metric is appropriate (for example, standardized for variable sample size, is it comparable across species?)
4. Analyse data
   - Confounds and biases: Account for factors that could bias analyses (such as species traits, phylogeny). Check and control for spatial or temporal autocorrelation.
5. Interpret and communicate
   - Restrict inference: Be careful not to overgeneralize results based on the data used. Does interpretation line up with the original question, with sufficient evidence?
   - Open data: Macrogenetics relies on open data. It is important that supporting information, datasets and core data are shared publicly.



sub-scale biological processes, which could potentially lead to erroneous conclusions[95]. A first step in Class III studies should be to estimate the capacity to detect the effects of predictors on IGV across the hypothesized spatiotemporal scales (BOX 1).

Power analyses and exploratory simulations can validate whether available datasets are suitable in sample size (for example, REF.[19]) or marker type and number (for example, REF.[96]). Conducting analyses at different spatial scales (for example, by varying distance thresholds when aggregating georeferenced genetic data[21,27]) will help evaluate the robustness of IGV patterns to uneven sampling effort and uncertainties in predictor data. Rarefaction analyses, in which the amount of genotype or sequencing data are varied to calculate IGV[26,27], or randomized subsampling procedures testing for the effect of sample size on IGV[24], will help determine the minimum data necessary for each analytical unit (that is, populations or other data aggregates) to accurately capture the effects of the predictors of interest. Genetic data simulators (such as fastsimcoal[97] and Nemo[98]) or specific tools (such as the R package HACSim[99]) can also be used to design sampling schemes prior to data collection. Genetic simulations could indeed evaluate genetic marker or sample size suitability as well as the expected effects of confounding predictors.

When conducting global macrogenetic investigations, we also recommend applying minimum coverage (for example, 30–50%[100]) of each taxon's geographic range to accurately capture species-wide spatial IGV patterns (for example, by sampling populations persisting in different habitat types or both the edge and central portions of the taxon's distribution). Covering a minimum proportion of the distribution increases the probability that genetic structure is correctly detected and not obscured or biased by sampling issues. We also recommend using minimum sample sizes to ensure that analytical units are informative and can be compared

Shifting baselines
The phenomenon whereby each generation of humans loses perception of biodiversity change by assuming that the biological state they observed at early stages of their lives or careers was the norm. Working under these misassumptions could fuel the use of incorrect baselines in temporal studies.

across space, taxa and/or time. Ideally, aggregates should represent local populations, metapopulations or otherwise genetically connected biogeographically meaningful divisions (for example, drainages for freshwater organisms[22,95]). Merging disconnected gene pools will conceal within-population effects of genetic drift: independently drifting populations may lose considerable IGV but when examined as a group will show no loss, on average, as a result of random fixation of different alleles across each population[101].

Assembling the data. When aggregating samples for analysis, the dispersal abilities (that is, the functional connectivity, *sensu*[102]), generation times and lifespan of species should be incorporated into analyses (TABLE 2). For example, pooling samples beyond species-specific mean dispersal distances increases the risk of merging distinct gene pools into a single sample. Sensitivity analyses can be used to validate aggregation decisions, for example, by assessing trends with and without high-dispersing or migratory species[26]. Linking IGV to environmental predictors can also be complicated by partial sampling of the range of a species (such as sampling only the breeding range of a migratory species): conditions experienced by the species outside of the sampled area (for example, habitat loss in the non-breeding range) may affect IGV without a clear link to environmental conditions within the study area. For temporal studies, the number of generations elapsed should be considered instead of years (TABLE 2) as the time for populations to reach genetic equilibrium after a disturbance depends on generations rather than elapsed time[103]. Taxa with long generation intervals and lifespans may show weaker associations to selective or demographic changes[104] than species that have multiple generations per year and the effect of overlapping generations should also be considered when assessing genetic turnover (TABLE 2).

Researchers should also be cognisant of locus-specific effects[80]. While mitochondrial DNA (mtDNA) data are abundant for animals, the maternal inheritance and differing selection pressures of mtDNA make it less illustrative of genome-wide IGV than nuclear DNA data[81,90,105]. For macrogenetic questions concerned with neutral, genome-wide IGV, it is best practice to use nuclear DNA[26] or combine inferences from multiple marker types (including neutral and adaptive gene markers)[90,105]. *COI* sequences, for example, should be used to address questions about interspecific animal phylogeographic patterns and not IGV (for example, REFS[67,96]). However, we note that the lack of easily accessible nuclear genetic data from centralized databases currently limits their use. Variation in mtDNA sequences is interesting from other perspectives, especially given their potential importance for climatic adaptation[106]. Accordingly, mtDNA loci, not limited to *COI* and *CYTb*, can be treated as indicators of adaptive IGV (as in REF.[21] for *MHC* genes) or to address questions specific to mitochondrial biology[81].

Analysing the data. Systematic biases in data acquisition may be introduced by inconsistent archiving practices, taxonomic biases and sampling gaps in publicly available data (especially for *COI*[95,107]; TABLE 2). For example, it is common for only unique, consensus or newly discovered haplotypes to be deposited in genetic repositories, producing an inaccurate depiction of IGV across space and time[95]. Data often come predominantly from North America and Europe[88] and easily sampled species, leading to under-sampled biodiversity-rich regions[20] (Table S1). Baseline IGV values collected in the past 30–40 years may also mask important IGV changes over time due to shifting baselines[108]. Indeed, significant levels of IGV may have already been lost before the first collection date of samples. Including historical, ancient or museum DNA where available may provide a better reference (BOX 2). Lastly, when examining temporal IGV trends, authors must recognize that local extinctions equate to 100% local IGV losses that are not easily accounted for because of the lack of genetic samples at present times.

It is also essential in macrogenetic studies to identify and address potential confounding factors such as species-specific traits, historical glaciation patterns, phylogeny and environmental heterogeneity (for example, through increased use of advanced statistical frameworks; see 'Future macrogenetic studies'). Explanatory factors such as International Union for Conservation of Nature (IUCN) Red List status, non-equilibrium population trends and invasive status should be included as predictors or random variables in model design. Spatial structure in genetic data (such as isolation by distance) should also be accounted for in statistical models that use covariates or spatially explicit random effects (for example, by using distance-based Moran's Eigenvector Maps[26] or generalized additive models[82]). As discussed above, biases arising from spatial heterogeneity and autocorrelation in sampling effort should also be considered[109] and summary statistics or analyses that are robust to variable sample size (for example, by using rarefaction, interpolation or extrapolation using coverage estimates) should be chosen (TABLE 2) to minimize noise effects associated with repurposed genetic data.

Appropriately interpreting and communicating results. Macrogenetic studies are often analytically challenging, especially when data are repurposed and synthesized, making transparency and reproducibility paramount. To ensure data are reusable and reproducible, researchers should carefully annotate their analytical pipeline, curate important metadata, and make scripts and all final datasets available upon publication. When possible, researchers should incorporate traditional ecological knowledge in identifying predictors and acknowledge traditional land holders[58,110,111]. It is also important to avoid overextending the conclusions drawn from macrogenetic studies, which could have far-reaching conservation implications[95,112]. The recent emergence of macrogenetics, the heterogeneous and scale-dependent conclusions found to date, and the diversity of potential confounding factors necessitate a careful approach in future studies.

Future macrogenetic studies

For macrogenetics to fulfil its potential, several data and knowledge gaps require filling and there is a need for improved cross-method collaboration and integration (BOX 3).

Table 2 | **Constraints, challenges and recommendations for future macrogenetic studies**

Constraints/limitations	Resulting challenges	Recommendations	Refs
Aggregating genetic samples across biologically unrealistic spatial and temporal scales	IGV patterns are obscured by combining samples from genetically divergent populations that have different evolutionary histories, N_e or selective environments	Ensure that grouped samples are from interbreeding populations (for example, test for cryptic substructure or immigrants)	19,20,22,95
		Group only samples collected within comparable, biologically realistic periods (for example, synchronous generations) when building spatial and/or temporal analytical units for IGV representation	
		Include spatiotemporal metadata when depositing novel genetic data	
Years as temporal unit in time series	Temporal samples separated by calendar years may not be independent if their generations overlap (for example, as for species with long generation times and lifespan); this may result in temporal autocorrelation and risk creating spurious genetic signals	Sample non-overlapping generations	95,146
		Use generations instead of years as the unit of time span separating temporal samples	
		If possible, include historical and/or ancient samples (for example, from museums or herbaria) to increase the temporal scale of the baseline	
Short or recent time frame for temporal analyses of IGV	Timeframes that cover too few generations may capture little genetic change by having low statistical power; time series based on relatively recent samples may miss the period of impact (owing to shifting baselines; Box 2)	Increase sample size to improve the signal-to-noise ratio, use high-diversity markers	94,147,148
		Where possible, use pre-impact samples such as ancient DNA or samples from museums or herbaria	
		Consider temporal frequency of sampling for robust trend analyses	
Use of only one marker or of an unsuitable marker for IGV characterization	A single marker gives only that marker's perspective on evolution; mitochondrial markers (such as COI) and other plasmid markers are unsuitable for overall IGV assessments (for example, owing to maternal inheritance for mitochondrial loci); neutral markers are unsuitable for assessing adaptive IGV; adaptive markers and genes are unsuitable for overall IGV estimations	When possible, use multiple markers, including nuclear DNA	19,20,22,26,36, 67,90,95,96,105
		Consider analysing single mitochondrial markers (such as COI) as phylogenetic or phylogeographic data, rather than IGV	
		Use loci mapped to a reference genome for genomics questions (for example, local adaptation) when useful and feasible	
		Use the most appropriate marker (for example, nuclear markers for overall IGV assessments; functional genes for adaptive IGV)	
Small or unequal sample sizes	Small or unequal samples violate assumptions of associated statistical models and obscure temporal trends or spatial genetic variance	Conduct power analyses to explore the effects of varying sample sizes and/or time points	19,20,22,26,67,96
		Retain only populations with equal sample numbers for population genetic analyses	
		Use rarefaction-based IGV metrics (such as allelic richness) or other procedures that correct for unequal sample sizes	
		Test for sampling effects (such as bias and low precision)	
Inclusion of invasive, harvested, stocked, supplemented, bottlenecked, or hybridized populations or species	Introduces variability in population dynamics and IGV and creates patterns that are unrepresentative of natural populations, generating noise and/or bias	Remove heavily managed species or run models both with and without these samples	19,20,22,149
		Include IUCN Red List status or other species management descriptor as a variable in models	
		Deposit metadata on status of population with genetic data	
		Remove (or statistically control for) bottlenecked populations	
Taxonomic and geographic bias	Species and locations in North America and Europe are over-represented and other regions are severely under-represented, obscuring global patterns and predictors of IGV	Exercise caution in interpreting results of global analyses and qualify results to readers and stakeholders	20,88,150
		Call for increased sampling efforts to improve global coverage	
		Facilitate capacity-building in under-represented regions	
		Sub-sample data to test for the effect of regional and/or phylogenetic over-representation (for example, rarefaction)	
Use of nucleotide diversity (π) as a metric of IGV	π has low sensitivity to change over short temporal scales	Use haplotype diversity or richness as markers of IGV	151

COI, cytochrome c oxidase subunit I; IGV, intraspecific genetic variation; IUCN, International Union for Conservation of Nature; N_e, effective population size.

Wallacean shortfall
The scientific knowledge gap on the geographical species distributions, driven by the unequal global species presence or absence in formal survey efforts.

Improving publically available genetic data. Poor metadata (for example, lack of location and sampling date) necessitates the exclusion of many genetic datasets from Class III macrogenetic studies (for example, studies have reported that 40–73% of suitable data were subsequently excluded from analysis owing to insufficient or undisclosed metadata[20,22,23,26,113]). Such gaps are significantly worse in some taxa (for example, 95% of amphibian sequences have no sampling year[27,113]). It is essential to enrich these datasets and rapidly expand the scope and robustness of macrogenetic studies. The minimum metadata required for genetic sequence data has been defined as the MIxS standards by the Genomic Standards Consortium[114] and includes the sampling date, geographic location (preferably decimal latitude and longitude) and environment (for example, biome, feature or material, depending on the nature of the investigation).

Instead of improved public metadata availability, the scope and robustness of several macrogenetic studies have been achieved through study-specific retroactive metadata enrichment. For example, many macrogenetic studies[20,22,23,26] converted metadata place names to spatial coordinates using dedicated tools (such as GeoNames). Alternatively, public archives can be directly improved through large manual enrichment initiatives (for example, REF.[96]). The uptake of better metadata stewardship practices[115] and infrastructures (such as GEOME[58]) should minimize the amount of missing metadata within future novel genetic datasets.

Class III macrogenetic studies should be expanded to understudied phylogenetic groups, including plants, fungi and algae. These groups present different or variable genomic features (such as ploidy), reproductive strategies, and evolutionary histories (such as genome duplications) that may drive differences in local diversity patterns and the relative importance of predictors. Temporal macrogenetic plant studies are also important because of the widespread decline of native plant species[116] but no Class III plant macrogenetic study exists to date. There are more than 108,000 plant species (about one-third of all known plant species) with sequences in GenBank[117], representing a huge opportunity for macrogenetic assessments. However, plants are challenging to integrate into the macrogenetics framework because of their widely variable lifespans and the challenges that chloroplast DNA barcodes pose to researchers (for example, the lack of a single universal barcode locus for all plants and variable mutation rates within and across species[118]). Taxonomic expansion can be advanced through exploiting taxonomic-specific databases (for example, CartograTree for trees[119]) or through initiatives to improve taxon-specific metadata.

To circumvent issues arising from geographic data gaps, future macrogenetic studies should also test hypotheses over smaller regions with a large quantity of high-quality data, rather than over global extents (supported by regional databases such as MacroPopGen[120]). Smaller spatial scales, such as continents or subcontinents, may be more appropriate for testing certain predictors and can help reduce confounding biological influences (such as biogeographic and disturbance histories in North America[26]) or technical issues (such as mismatches between data and predictor resolution) that can obscure macrogenetic trends over larger spatial scales. Smaller spatial scales can facilitate improvement in predictor data, allowing a transition from repurposing sparse low-resolution global environmental data (such as weather station-based climatic variables[34]) to fine-resolution data that is more representative of the conditions a species experiences (for example,

Box 2 | **Baselines to test hypotheses**

Temporally focused macrogenetic studies aim to test how genetic diversity has changed over time[27,88] or across generations[152]. To successfully measure change, a 'baseline' is required that represents conditions before the impact being studied. However, there is relatively limited sequencing data available prior to the 1990s, which is long after the start of many important stressors. Therefore, good baselines are unlikely to be found in existing publicly available datasets. Many temporal studies also use a single historical baseline[88] but this can drive misleading results if genetic trends are variable over time.

Poor baselines can have large effects

The appropriate baseline will depend on the question being asked. The baseline for 'what impact has highway construction had on these populations' (for example, REF.[153]) will be different from 'to what extent have humans cumulatively impacted the genetic diversity of all wildlife' (for example, REF.[27]). As intraspecific genetic variation (IGV) is not lost linearly[149] and stressors often have a gradual and cumulative cross-generational impact, the time span to observe change may be longer than anticipated and should be measured as generations elapsed and not simply years.

An ideal study to examine temporal genetic diversity change in a single species would measure >30 individuals[154] (fewer if genome-wide SNP data is available) from a population at a time point pre-dating the impact of interest and measure the same number of samples from the current population. Samples at multiple time points before and after the impact would better capture natural variation in genetic indices[155]. However, very few sets of such temporal samples exist. Archived samples are limited in number, irregular in time, and geographically and taxonomically biased even in the best-represented areas and taxa[126]. To evaluate the usefulness of available samples, simulations can test whether the time frame, frequency of sampling, sample size and genetic markers have sufficient power to detect different degrees of change in genetic indices (see REF.[156]; suitable software includes Serial SimCoal[157] and Nemo[98]).

Analysing time series of IGV in a macrogenetic study has further challenges. In addition to the issues common to all macrogenetic studies (uneven sampling and differing quality and quantity of data across populations and species), temporal studies will likely have a variable quality of historical baselines (owing to sample quality, quantity and evenness) and span unequal numbers of generations across species. The effects of such patchy datasets are unclear but they could obscure IGV patterns and potential drivers, resulting in misleading conclusions[95]. Again, simulations could be used to explore these effects and help outline idealized temporal macrogenetic datasets.

Collecting data for the future

Building baselines for the future through standardized sampling initiatives across large geographical (or environmental) gradients, followed by archiving of such samples in museums is essential for the future of macrogenetics. Although there is increasing use of historical collections for sequencing, very few contemporary samples are concurrently collected and archived, which is causing a temporal gap that may render important baselines inaccessible to future research. To accommodate contemporary samples, increased investment in museum infrastructure is needed. Such investments would also facilitate macrogenetic studies, for example, by linking digital sample records with existing sequencing data and pre-emptively categorizing other samples as available for sequencing. Metadata from museum, herbaria and other archived specimens can also support macrogenetic studies by outlining historical species distributions and providing life-history data (such as phenotype) for elusive species.

Standardized field sample expeditions or local initiatives will also ensure macrogenetic studies can expand into new geographic areas and biomes. As there may be poor existing knowledge, systematic sampling will help ensure the species distributions are accurately captured (addressing the Wallacean shortfall) and will limit biases that could easily arise from any large taxonomic knowledge gaps (the Linnean shortfall)[158].

Box 3 | **Extending the reach of macrogenetics with complementary approaches**

The classes of macrogenetic studies need greater integration with each other and with simulations, predictive modelling and systematic conservation planning (see Box figure).

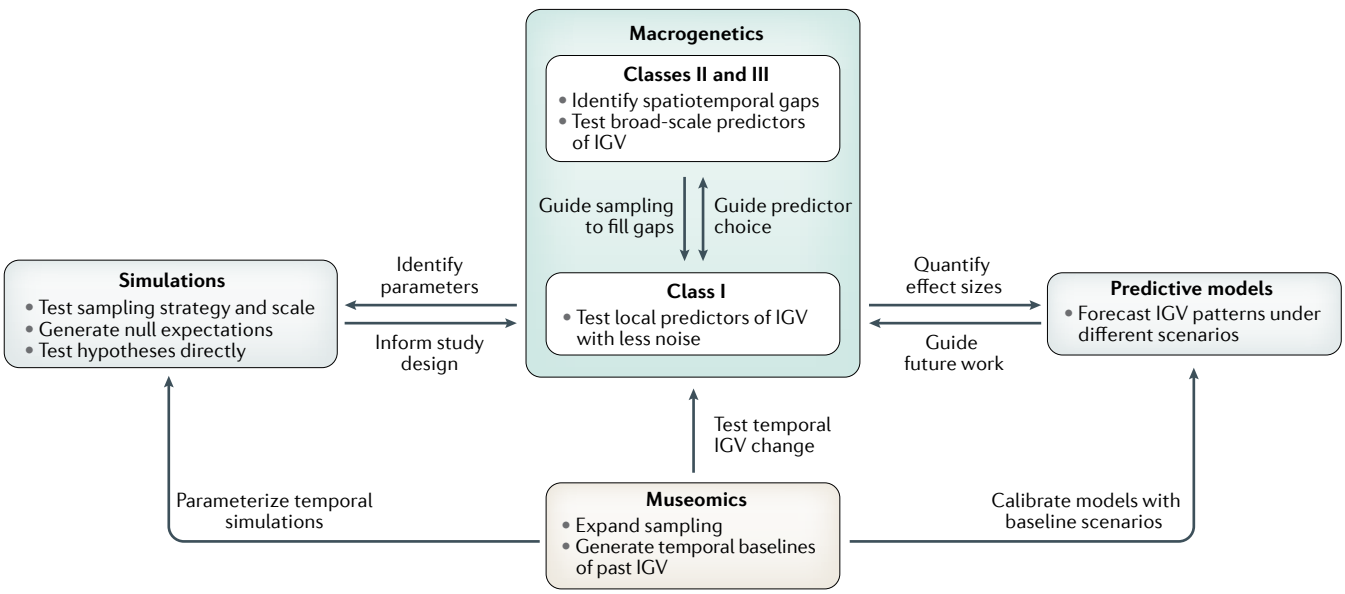
Integrating classes of macrogenetic studies. Class III studies have highlighted taxonomic and geographic gaps in available genetic data, thus identifying high-priority areas for future Class I studies, which provide increased power through standardized sample sizes, molecular markers, explanatory variables and sampling locations. Class I studies better account for shared geological and climatic settings and ensure the collection of high-quality predictor data. For instance, the IntraBioDiv consortium sampled 27 alpine plant species[30], helping test the influence of soil type, temperature, species diversity and traits at high spatial resolution on intraspecific genetic variation (IGV)[28,29]. This approach should be replicated in numerous regions of the world. Environmental DNA will soon allow multispecies IGV parameter estimation (for example, REFS[159,160]) facilitating novel Class I studies. Autonomous environmental DNA sampling techniques (for example, drones or autonomous underwater vehicles[161]) can help close sampling gaps and facilitate the assessment of cryptic and understudied taxa.

Class II studies provide vital information and should be considered when planning Class III studies; they have often already identified putative predictors and influential traits for IGV and the impact of factors such as harvest[87] or habitat fragmentation[32] on IGV (see Supplementary Table 1). For example, habitat fragmentation was found to reduce IGV and outcrossing rates, especially in fragments isolated for more than 50 years[86] (see also REFS[31,162]). Class II studies thus provide expected effect sizes for simulations and study design planning.

Integrating museomics. Museomics integrates historical samples into temporal macrogenetic analyses, hence providing baselines (BOX 2) to compare with contemporary values or to be used for starting values in simulations. Space-for-time substitutions do not replace true baselines because they are confounded by natural variation in IGV across a species' range. In one example, museum collections spanning ~100 years were used to evaluate country-wide trends of genetic erosion in two species of butterflies[53]. Integrating museomics into all three Classes is needed but will be a statistical challenge if samples are scattered and baselines are unequal.

Integrating simulations. Besides their potential for power analysis and study design[99,163], simulations can be used to test macrogenetic hypotheses in silico[164]. For example, simulations can examine if large-scale phenomena such as Late Quaternary glaciations (that is, range shifts) drive IGV patterns across species[96,165]. They can also be used to test network complexity effects on spatial IGV patterns[131]. Individual-based, realistic simulations can also identify putative traits (such as dispersal and mating system) driving IGV patterns. However, simulations are not suitable for testing multispecies–IGV relationships or the evolutionary speed[23] or Red Queen hypothesis because they lack sufficient realism (such as complex multispecies interactions) and scale. Parameters and starting values for future simulations can come from Class II studies. Simulations can also be used to forecast IGV changes based on the results from all Classes of studies, including climate change effects, species diversity erosion and human impacts.

Integrating predictive models. Conservation and management (including spatial conservation planning, policy and management) can be informed by all Classes of macrogenetic studies via predictive models. As yet, however, findings from macrogenetic studies have rarely been applied to conservation problems directly, although they have informed conservation policy documents (for example, the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) evaluations of Nature's contribution to People (NCP)[88]) and ecosystem restoration planning[152,162]. Macrogenetic studies have exciting potential to help prioritize and protect highly imperilled regions or taxa, especially if applied to larger datasets than has been done so far (for example, 6–20 species[18,138]).



Linnean shortfall
The scientific knowledge gap for described species, that is, the gap between the number of formally described species and the greater number that actually exist.

microclimatic data[121]). Such predictor data improvements will increase study power and help determine if weak macrogenetic trends are biologically meaningful or a product of poor predictor data.

More spatially targeted studies may further enable explicit hypothesis-testing approaches to be used; for instance, making use of geographic or environmental gradients that are repeated across the land (for example, altitudinal gradients) and in the oceans (for example, bathymetric gradients) to provide study replicates. Such macrogenetic approaches may allow us to understand if, and why, different trends are found across markers, taxa and species with different life-history traits. Nevertheless, it is also important to expand the geographical focus beyond high gross domestic product countries via rapidly increasing sequencing efforts and capacity-building[122]. Field expeditions deploying metagenomic (or metabarcoding) approaches with

standardized sampling protocols at large scales (such as the Tara expedition for oceanic microorganisms[123]) could also expand macrogenetics into new, historically underexplored, ecosystems.

Box 4 | **Towards macrogenomics**

It is becoming increasingly affordable to produce genome-wide marker data and make it available in public archives. Therefore, we predict the rapid advancement of all classes of macrogenomics studies in the coming years. The advantages of macrogenomics are clear but the transition to genomic data also poses new challenges.

Advantages of macrogenomics.
Access to numerous, densely distributed markers with associated genomic data will allow researchers to accurately characterize genome-wide intraspecific genetic variation (IGV)[166], ensuring that researchers can adjust their marker set to their question; that is, to target putatively adaptive variation that is functional or under selection, candidate genes and/or orthologues, or neutral markers (such as non-coding regions). This increased accuracy and precision of measures of IGV will allow researchers to identify finer patterns, such as subtle genetic structure, and directly examine loci under selection. Statistical power will also increase with the higher marker number and may allow researchers to sample fewer individuals per location[167], reducing costs or enabling more sampling locations.

Challenges of transitioning to macrogenomics.

- Computational cost: Genomics remains computationally costly and requires expertise for data production and analysis, limiting the accessibility of macrogenomics to research groups with access to large computer clusters, reliable internet and substantial local file storage, potentially exacerbating geographic data disparity. Access to cloud computing and local capacity-building would help address this challenge.
- Expertise: Bioinformatics and genomic data expertise is imperative for macrogenomic study success because there are no standard genomic data processing pipelines; thus, researchers must have an understanding of the effects of using different analysis steps[168] and data types[169] on analytical outcomes. For example, the different sequencing or genotyping-by-sequencing methods used across projects (such as pooled sequencing, restriction site-associated DNA sequencing or low coverage resequencing) and over time (such as changes in read length or technology[170]) may produce false biological trends known as 'batch effects' (for example, REF.[171]). Batch effects can be tested for by including technical variables (such as sequencing chemistry) in downstream analyses.
- Software: Different variant detection and genotype calling methods (for example, cohort/joint or genotype likelihood calling[168]) and filters (for example, minor allele frequency spectrum[172]) may obscure biological trends or render data incomparable. As sequencing costs fall, the transition towards whole-genome sequencing will reduce differences in sequencing approaches but the software challenges will remain. Reprocessing the raw sequencing read files in a standard pipeline may be necessary to address this but is computationally costly. Therefore, subsets of data (for example, 25% of individuals) could be reprocessed at first and compared to the original variant files to identify if major biological differences are present. If data are reanalyzed, all parameter choices in the data analysis workflow (such as software version and variant filters) should be recorded to facilitate analysis replication when additional data become available[173]. Simulations can also be used to evaluate the impact of different variant filters (for example, REF.[172]). Simulations may be particularly valuable to Class III studies if inconsistent data types have been archived (that is, SNP calls, filtered SNP calls, or sequencing reads) rendering universal reprocessing impossible.
- Reference genomes: Variation in reference genome quality, specifically contig length or continuity[174], heavily impacts analyses using sliding windows or length-based statistics such as runs of homozygosity. Calculating metrics as proportions of contig size may account for these effects[175].
- Genome variation: Variation in genome ploidy across many taxonomic groups or across the life cycle of a species, gene duplication, transposable elements, mutation rate, marker linkage and chromosome size may introduce a number of biases with currently unknown consequences. Several of these genomic processes or phenomena show a strong taxonomic signal. Disentangling their biological effects from systematic errors may be possible by checking if predictors correlate with genome assembly quality, examining only well-characterized (often exonic) genomic regions, and filtering SNPs that violate Hardy–Weinberg equilibrium, have spurious heterozygosity, or show extreme sequencing depth.

Expanding and complementing macrogenetics. Multispecies genetic studies encompassing a large spatial scale (that is, Class I studies) will play an increasing role in the near future as will macrogenomics (BOXES 3,4). These approaches can address macrogenetic and/or macrogenomic hypotheses more directly and have high power when sampling across predictor gradients or cross-category pairings[124]. While costs have limited these studies to tens of species previously, consortiums and large government projects are overcoming this limitation. For example, the California Conservation Genomics Project is examining genomic patterns in 200 species with 150 samples each, a study scale that was previously unimaginable. Incorporation into long-term or governmental species monitoring will advance macrogenetics while improving the monitoring itself[125]. In turn, this will help increase and strengthen available temporal genetic data by establishing genetic baselines for future evaluations of IGV change.

Sequencing of historical or ancient samples will also be necessary to address temporally relevant macrogenetic hypotheses[54,95]. Public databases often start in or after the 1980s, which is too late to establish pre-impact baselines; thus, we need to look at sample archives that date back as far as the 1800s (BOXES 2,3). Unfortunately, historical DNA samples are often degraded and require substantial expertise, specialized facilities, large natural history collections and substantial financial support (which is typically only possible in high gross domestic product countries)[126]. Such studies are often opportunistic and geographically biased yet data from all countries and regions, particularly those rich in biodiversity (such as equatorial countries[20]), are needed for testing macrogenetic predictors. A global museomics initiative that enables the acquisition of skills and equipment for sequencing historical samples from herbaria, museums and other natural history collections is important for capacity-building[122,127]. Scientists should also consider sharing or advertising biological samples on platforms such as Otlet, CryoArks or GEOME and have plans for long-term (50+ years) archiving in museums, herbaria or other collections[128]. Collaborative research infrastructures for natural science collections such as the Distributed System of Scientific Collections (DiSSCO) can help streamline efforts even further.

To achieve a better understanding of macrogenetic patterns and processes, future studies should use cross-method approaches (BOX 3). Coupling macrogenetics with simulations is one such approach. Simulations could be parameterized using data (such as temperature or human density effects) from macrogenetic studies to forecast the future effects of environmental change or extreme events. Furthermore, simulations can be used to test some macrogenetic hypotheses without any real genetic data (BOX 3).

It is not yet clear if an absence of consistent and strong macrogenetic patterns is driven by actual weak predictors or is because of poor data and scaling issues. The above recommendations and complementary approaches, coupled with increased use of advanced statistical frameworks such as causal analyses[129,130], machine learning procedures[19,24,131] or approximate

Red Queen hypothesis
An evolutionary hypothesis that states that antagonistically interacting species constantly co-evolve in order to adapt to each other's attack and defence strategies.

Pooled sequencing
A method of high-throughput DNA sequencing in which DNA extracts from groups of individuals are pooled together for sequencing, rather than each individual being sequenced independently.

Museomics
DNA sequencing of historical specimens archived in museums, herbaria and other natural history collections. It typically refers to samples that may be decades to centuries old.

Bayesian computations[63,132] will allow us to explore the drivers of macrogenetic patterns and to predict how they might shift in response to environmental change, hence expanding the range of questions that can be tackled by macrogenetics. Causal analyses can be used to understand biological processes underlying macrogenetic patterns. They can discern direct and indirect relationships between IGV predictors and dependent variable networks[129,130]. For example, this framework could assess how different processes interact to affect both IGV patterns and other biodiversity facets (such as species diversity patterns[77]). Machine learning procedures can capitalize on large datasets to predictively model macrogenetic patterns. For instance, random forests[133] can estimate which IGV metric is the most sensitive to a specific process[131] and determine the relative importance of IGV predictors from large sets of life-history, geographic and historical predictors[19,24]. Finally, approximate Bayesian computation may help disentangle which processes generate IGV patterns observed for multiple species[131]. This can also improve evolutionary parameter estimation (such as N_e and migration rates) or allow the identification of past demographic trends (such as demographic expansions or bottlenecks[66]), generating valuable information to calibrate IGV forecast models (BOX 3).

Applications to biodiversity conservation. Macrogenetic studies can make great contributions to biodiversity conservation by facilitating initiatives to 'upscale' monitoring through global databases, essential biodiversity variables, and big data analytics[134]. Assessments of multispecific IGV patterns over regional to global scales will help pinpoint IGV hotspots[20,135] and support the incorporation of IGV into conservation planning and protected area design[18,136,137]. Macrogenetic studies assessing IGV across time and space will provide valuable, large-scale monitoring opportunities to detect early genetic signals of climate change effects that might otherwise escape detection or delay in manifestation[36]. Macrogenetic datasets might also be used to predict potential future climatic or anthropogenic effects on IGV by overlaying projected human development on IGV maps to forecast genetic diversity loss[36] or by modelling exposure to climate change at the intraspecific genetic level[138]. Macrogenetic studies may also soon permit quantitative identification of the most harmful drivers of biodiversity loss (for example, REFS[27,88]), which could guide conservation policy and provide much-needed conservation evidence[139]. Ultimately, observations of large-scale IGV patterns must be linked to policy, including national legislation and global initiatives such as the Convention on Biological Diversity[125,140] or the United Nations Decade on Ocean Science[141]. For this to occur, macrogeneticists must not only publish scientific papers but also extend the communication of their studies to the general public, industry and policy-makers[142].

Conclusions

The field of macrogenetics has rapidly expanded thanks to open data initiatives, open-source data analysis, technological innovation and the strong foundations of evolutionary biology disciplines. These factors have enabled the field to test core hypotheses and map broad-scale IGV patterns. However, some discordance among the findings of recent studies highlights the need to overcome challenges presented by several conceptual and methodological factors. Limitations and challenges can — and should — be overcome in future studies to move the field forward. Capacity-building, comprehensive benefit-sharing and appropriate data stewardship (for example, considering the FAIR principles and the CARE principles for indigenous data governance[57,143]) are each important to resolve current geographical and taxonomic data biases. There are several trade-offs associated with each step necessary for improvement (such as cost, labour and necessary expertise), making it imperative to keep macrogenetics an open, collaborative field.

Genetic diversity is a key biodiversity facet, yet it has been long-neglected in conservation policy, planning and management[18,125,140]. Macrogenetics has great potential to support a global biodiversity monitoring system and inform biodiversity conservation. Steps are now needed to disseminate macrogenetic insights to end-users, including conservation practitioners and policy-makers. As the field progresses, we hope to see broader taxonomic coverage and integration of species-specific traits. Given that most statistical and analytical tools used in macrogenetics are repurposed from other disciplines (including population genetics and comparative phylogeography, among others), we also expect — and encourage researchers to undertake — in-depth technical explorations of which tools are the best suited to analyse upscaled macrogenetic datasets as well as the development of new macrogenetics-oriented tools. We hope to soon see global Class I macrogenomic studies, explorations of adaptive IGV (for example, REF.[21]), multispecies patterns of intraspecific epigenetic variation[144] and large-scale museomics studies. In the future, we hope for interdisciplinary integration of all 'macro' disciplines to develop a unified framework for the exploration of multifaceted biodiversity patterns across large scales (for example, REF.[145]). These studies will offer unprecedented detail of global patterns of IGV, address previously intractable hypotheses and spark the development of new hypotheses while helping to curb global biodiversity loss.

Published online: 18 August 2021

1. Brown, J. H. & Maurer, B. A. Macroecology: the division of food and space among species on continents. *Science* **243**, 1145–1150 (1989).
2. Gaston, K. J., Robinson, D. & Chown, S. L. Macrophysiology: large-scale patterns in physiological traits and their ecological implications. *Funct. Ecol.* **18**, 159–167 (2004).
3. Chown, S. L. & Gaston, K. J. Macrophysiology–progress and prospects. *Funct. Ecol.* **30**, 330–344 (2016).
4. Avise, J. C. *Phylogeography: the History and Formation of Species* (Harvard University Press, 2000).
5. Ebach, M. C. *Origins of Biogeography*. Vol. 13 (Springer, 2015).
6. Brundin, L. On the real nature of transantarctic relationships. *Evolution* **19**, 496–505 (1965).
7. Beheregaray, L. B. Twenty years of phylogeography: the state of the field and the challenges for the Southern Hemisphere. *Mol. Ecol.* **17**, 3754–3774 (2008).
8. Hickerson, M. J. et al. Phylogeography's past, present, and future: 10 years after Avise, 2000. *Mol. Phylogenet. Evol.* **54**, 291–301 (2010).
9. Gaston, K. J. & Blackburn, T. M. A critique for macroecology. *Oikos* **84**, 353–368 (1999).
10. Lovegrove, B. G. The zoogeography of mammalian basal metabolic rate. *Am. Nat.* **156**, 201–219 (2000).

11. Reich, P. B., Walters, M. B. & Ellsworth, D. S. From tropics to tundra: Global convergence in plant functioning. *Proc. Natl Acad. Sci. USA* **94**, 13730–13734 (1997).
12. Chown, S. L. & Gaston, K. J. Macrophysiology for a changing world. *Proc. Biol. Sci.* **275**, 1469–1478 (2008).
13. Kerr, J. T., Kharouba, H. M. & Currie, D. J. The macroecological contribution to global change solutions. *Science* **316**, 1581–1584 (2007).
14. Blanchet, S., Prunier, J. G. & De Kort, H. Time to go bigger: Emerging patterns in macrogenetics. *Trends Genet.* **33**, 579–580 (2017).
This study coined the term 'macrogenetics' and illustrated, through three study examples, how shifting toward macrogenetics should generate new perspectives and theories concerning genetic diversity patterns.
15. Blanchet, S. et al. A river runs through it: the causes, consequences, and management of intraspecific diversity in river networks. *Evol. Appl.* **13**, 1195–1213 (2020).
16. Frankham, R. Resolving conceptual issues in conservation genetics: the roles of laboratory species and meta-analyses. *Hereditas* **130**, 195–201 (2004).
17. Arnqvist, G. & Wooster, D. Meta-analysis: synthesizing research findings in ecology and evolution. *Trends Ecol. Evol.* **10**, 236–240 (1995).
18. Paz-Vinas, I. et al. Systematic conservation planning for intraspecific genetic diversity. *Proc. Biol. Sci.* **285**, 20172746 (2018).
19. Pelletier, T. A. & Carstens, B. C. Geographical range size and latitude predict population genetic structure in a global survey. *Biol. Lett.* **14**, 20170566 (2018).
20. Miraldo, A. et al. An anthropocene map of genetic diversity. *Science* **353**, 1532–1535 (2016).
This paper is thought to be the first published study to massively repurpose public mtDNA sequences to explore global genetic patterns (100,791 sequences from >4,500 terrestrial mammal and amphibian species).
21. Yiming, L. et al. Latitudinal gradients in genetic diversity and natural selection at a highly adaptive gene in terrestrial mammals. *Ecography* **44**, 206–218 (2021).
This study found that adaptive IGV is higher at low latitudes and in smaller mammal species using repurposed MHC gene data from 93 mammal species.
22. Manel, S. et al. Global determinants of freshwater and marine fish genetic diversity. *Nat. Commun.* **11**, 692 (2020).
This study repurposed 58,565 public mtDNA sequences from 5,912 freshwater and marine fish to explore the effects of environmental drivers (temperature, species diversity) on intraspecific genetic diversity.
23. Theodoridis, S. et al. Evolutionary history and past climate change shape the distribution of genetic diversity in terrestrial mammals. *Nat. Commun.* **11**, 2557 (2020).
This study revealed a negative effect of past rapid climate change and a positive effect of interannual precipitation variability in shaping the genetic diversity of terrestrial mammals using 46,965 mtDNA sequences.
24. Barrow, L. N., da Fonseca, E. M., Thompson, C. E. P. & Carstens, B. C. Predicting amphibian intraspecific diversity with machine learning: Challenges and prospects for integrating traits, geography, and genetic data. *Mol. Ecol. Resour.* https://doi.org/10.1111/1755-0998.13303 (2020).
25. De Kort, H. et al. Life history, climate and biogeography interactively affect worldwide genetic diversity of plant and animal populations. *Nat. Commun.* **12**, 516 (2021).
This study found weak support for latitudinal IGV gradients, taxonomic-specific effects of temperature stability and life-history traits, and higher IGV in animals compared to plants using microsatellite and amplified fragment length polymorphism data from 8,386 local populations from 727 animal and plant species.
26. Schmidt, C., Domaratzki, M., Kinnunen, R. P., Bowman, J. & Garroway, C. J. Continent-wide effects of urbanization on bird and mammal genetic diversity. *Proc. Biol. Sci.* **287**, 20192497 (2020).
This study used archived microsatellite data from 85 studies (66 species) to explore the effects of urbanization in mammals and birds.
27. Millette, K. L. et al. No consistent effects of humans on animal genetic diversity worldwide. *Ecol. Lett.* **23**, 55–67 (2020).
The authors of this article conducted spatial and temporal analysis of the effects of humans on animal genetic diversity worldwide, by repurposing 175,247 mtDNA sequences from >17,000 animal species.
28. Taberlet, P. et al. Genetic diversity in widespread species is not congruent with species richness in alpine plant communities. *Ecol. Lett.* **15**, 1439–1448 (2012).
This paper reports a Class I macrogenetic study based on amplified fragment length polymorphism genetic data from 27 alpine plant species that tested whether genetic and species diversities co-vary.
29. Manel, S. et al. Broad-scale adaptive genetic variation in alpine plants is driven by temperature and precipitation. *Mol. Ecol.* **21**, 3729–3738 (2012).
30. Gugerli, F. et al. Relationships among levels of biodiversity and the relevance of intraspecific diversity in conservation – a project synopsis. *Perspect. Plant. Ecol. Evol. Syst.* **10**, 259–281 (2008).
31. Schlaepfer, D. R., Braschler, B., Rusterholz, H.-P. & Baur, B. Genetic effects of anthropogenic habitat fragmentation on remnant animal and plant populations: a meta-analysis. *Ecosphere* **9**, e02488 (2018).
32. González, A. V., Gómez-Silva, V., Ramírez, M. J. & Fontúrbel, F. E. Meta-analysis of the differential effects of habitat fragmentation and degradation on plant genetic diversity. *Conserv. Biol.* **34**, 711–720 (2020).
33. Ratnasingham, S. & Hebert, P. D. N. Bold: the barcode of life data system. *Mol. Ecol. Notes* **7**, 355–364 (2007).
34. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**, 1965–1978 (2005).
35. Kattge, J. et al. TRY plant trait database–enhanced coverage and open access. *Glob. Change Biol.* **26**, 119–188 (2020).
36. Theodoridis, S., Rahbek, C. & Nogues-Bravo, D. Exposure of mammal genetic diversity to mid-21st century global change. *Ecography* **44**, 817–831 (2021).
37. Rissler, L. J. Union of phylogeography and landscape genetics. *Proc. Natl Acad. Sci. USA* **113**, 8079–8086 (2016).
38. Hubbell, S. P. The unified neutral theory of biodiversity and biogeography (Princeton University Press, 2001).
39. Haldane, J. B. S. A mathematical theory of natural and artificial selection, Part V: selection and mutation. *Math. Proc. Camb. Philos. Soc.* **23**, 838–844 (1927).
40. Wright, S. Evolution in Mendelian populations. *Genetics* **16**, 97–159 (1931).
41. Fisher, R. A. On the dominance ratio. *Proc. R. Soc. Edinburgh* **42**, 321–341 (1922).
42. Kimura, M. & Weiss, G. H. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* **49**, 561–576 (1964).
43. Kingman, J. F. C. The coalescent. *Stoch. Process. Their Appl.* **13**, 235–248 (1982).
44. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).
45. Soulé, M. E. in *Molecular Evolution* (ed. Ayala, F. J.) 60–77 (Sinauer Associates, 1976).
46. Brown, A. H. Isozymes, plant population genetic structure and genetic conservation. *Tag. Theor. Appl. Genet. Theor. Angew. Genet.* **52**, 145–157 (1978).
47. Mullis, K. et al. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb. Symp. Quant. Biol.* **51**, 263–273 (1986).
48. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA* **74**, 5463–5467 (1977).
49. Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A. & Johnson, E. A. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res.* **17**, 240–248 (2007).
50. Carroll, E. L. et al. Genetic and genomic monitoring with minimally invasive sampling methods. *Evol. Appl.* **11**, 1094–1119 (2018).
51. Hebert, P. D. N., Cywinska, A., Ball, S. L. & deWaard, J. R. Biological identifications through DNA barcodes. *Proc. Biol. Sci.* **270**, 313–321 (2003).
52. Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol. Ecol.* **21**, 2045–2050 (2012).
53. Gauthier, J. et al. Museomics identifies genetic erosion in two butterfly species across the 20th century in Finland. *Mol. Ecol. Resour.* **20**, 1191–1205 (2020).
54. Wandeler, P., Hoeck, P. E. A. & Keller, L. F. Back to the future: museum specimens in population genetics. *Trends Ecol. Evol.* **22**, 634–642 (2007).
55. Strasser, B. J. The experimenter's museum: GenBank, natural history, and the moral economies of biomedicine. *Isis* **102**, 60–96 (2011).
56. Whitlock, M. C. Data archiving in ecology and evolution: best practices. *Trends Ecol. Evol.* **26**, 61–65 (2011).
57. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
58. Deck, J. et al. The Genomic Observatories Metadatabase (GeOMe): A new repository for field and sampling event metadata associated with genetic samples. *PLoS Biol.* **15**, e2002925 (2017).
59. R Core Team. R: a language and environment for statistical computing, *R Foundation for Statistical Computing* http://www.r-project.org/index.html (2021).
60. Manel, S. & Holderegger, R. Ten years of landscape genetics. *Trends Ecol. Evol.* **28**, 614–621 (2013).
61. Prunier, J. G., Colyn, M., Legendre, X., Nimon, K. F. & Flamand, M. C. Multicollinearity in spatial genetics: separating the wheat from the chaff using commonality analyses. *Mol. Ecol.* **24**, 263–283 (2015).
62. Stanley, R. R. E. et al. A climate-associated multispecies cryptic cline in the northwest Atlantic. *Sci. Adv.* **4**, eaaq0929 (2018).
63. Fenderson, L. E., Kovach, A. I. & Llamas, B. Spatiotemporal landscape genetics: investigating ecology and evolution through space and time. *Mol. Ecol.* **29**, 218–246 (2020).
64. Daza, J. M., Castoe, T. A. & Parkinson, C. L. Using regional comparative phylogeographic data from snake lineages to infer historical processes in middle America. *Ecography* **33**, 343–354 (2010).
65. Riddle, B. R. Comparative phylogeography clarifies the complexity and problems of continental distribution that drove A. R. Wallace to favor islands. *Proc. Natl Acad. Sci. USA* **113**, 7970–7977 (2016).
66. Carstens, B. C., Morales, A. E., Field, K. & Pelletier, T. A. A global analysis of bats using automated comparative phylogeography uncovers a surprising impact of Pleistocene glaciation. *J. Biogeogr.* **45**, 1795–1805 (2018).
67. Smith, B. T., Seeholzer, G. F., Harvey, M. G., Cuervo, A. M. & Brumfield, R. T. A latitudinal phylogeographic diversity gradient in birds. *PLoS Biol.* **15**, e2001073 (2017).
68. Smith, B. T. et al. The drivers of tropical speciation. *Nature* **515**, 406–409 (2014).
69. Ballin, M., Barcaroli, G., Masselli, M. & Scarnò, M. *Redesign Sample for Land Use/Cover Area Frame Survey (LUCAS) 2018* (EU Publications, 2018).
70. Buchhorn, M. et al. Copernicus global land cover layers — Collection 2. *Remote. Sens.* **12**, 1044 (2020).
71. Jones, K. E. et al. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals: Ecological Archives E090-184. *Ecology* **90**, 2648–2648 (2009).
72. Tedesco, P. A. et al. A global database on freshwater fish species occurrence in drainage basins. *Sci. Data* **4**, 170141 (2017).
73. Vellend, M. & Geber, M. A. Connections between species diversity and genetic diversity: species diversity and genetic diversity. *Ecol. Lett.* **8**, 767–781 (2005).
74. Fourtune, L., Paz-Vinas, I., Loot, G., Prunier, J. G. & Blanchet, S. Lessons from the fish: a multi-species analysis reveals common processes underlying similar species-genetic diversity correlations. *Freshw. Biol.* **61**, 1830–1845 (2016).
75. Bertin, A. et al. Genetic variation of loci potentially under selection confounds species-genetic diversity correlations in a fragmented habitat. *Mol. Ecol.* **26**, 431–443 (2017).
76. Lawrence, E. R. & Fraser, D. J. Latitudinal biodiversity gradients at three levels: linking species richness, population richness and genetic diversity. *Glob. Ecol. Biogeogr.* **29**, 770–788 (2020).
77. Schmidt, C., Dray, S. & Garroway, C. J. Genetic and species-level biodiversity patterns are linked by demography and ecological opportunity. *bioRxiv* https://doi.org/10.1101/2020.06.03.132092 (2021).
78. Hillebrand, H. On the generality of the latitudinal diversity gradient. *Am. Nat.* **163**, 192–211 (2004).
79. Pontarp, M. et al. The latitudinal diversity gradient: novel understanding through mechanistic eco-evolutionary models. *Trends Ecol. Evol.* **34**, 211–223 (2019).

80. Toews, D. P. L. & Brelsford, A. The biogeography of mitochondrial and nuclear discordance in animals. *Mol. Ecol.* **21**, 3907–3930 (2012).
81. Schmidt, C. & Garroway, C. J. The conservation utility of mitochondrial genetic diversity in macrogenetic research. *Conserv. Genet.* **22**, 323–327 (2021).
82. Gratton, P. et al. Which latitudinal gradients for genetic diversity? *Trends Ecol. Evol.* **32**, 724–726 (2017). **This response to Miraldo et al.[20] identified a limitation of that article in that it did not account for the decay of genetic similarity with distance and represents the first critique of the downsides of the macrogenetic approach and the need for rigorous statistics.**
83. Loveless, M. D. & Hamrick, J. L. Ecological determinants of genetic structure in plant populations. *Annu. Rev. Ecol. Syst.* **15**, 65–95 (1984).
84. Hu, Y. et al. Spatial patterns and conservation of genetic and phylogenetic diversity of wildlife in China. *Sci. Adv.* **7**, eabd5725 (2021).
85. Johnson, M. T. J. & Munshi-South, J. Evolution of life in urban environments. *Science* **358**, eaam8327 (2017).
86. Aguilar, R., Quesada, M., Ashworth, L., Herrerias-Diego, Y. & Lobo, J. Genetic consequences of habitat fragmentation in plant populations: susceptible signals in plant traits and methodological approaches. *Mol. Ecol.* **17**, 5177–5188 (2008).
87. Pinsky, M. L. & Palumbi, S. R. Meta-analysis reveals lower genetic diversity in overfished populations. *Mol. Ecol.* **23**, 29–39 (2014).
88. Leigh, D. M., Hendry, A. P., Vázquez-Domínguez, E. & Friesen, V. L. Estimated six per cent loss of genetic variation in wild populations since the industrial revolution. *Evol. Appl.* **12**, 1505–1512 (2019). **This study estimated the magnitude of the loss of genetic variation over a century-scale using microsatellite data from 91 species.**
89. Schmidt, C. & Garroway, C. J. The population genetics of urban and rural amphibians in north America. *Mol. Ecol.* https://doi.org/10.1111/mec.16005 (2021).
90. Bazin, E., Glémin, S. & Galtier, N. Population size does not influence mitochondrial genetic diversity in animals. *Science* **312**, 570–572 (2006).
91. Galtier, N., Nabholz, B., Glémin, S. & Hurst, G. D. D. Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Mol. Ecol.* **18**, 4541–4550 (2009).
92. Allio, R., Donega, S., Galtier, N. & Nabholz, B. Large variation in the ratio of mitochondrial to nuclear mutation rate across animals: implications for genetic diversity and the use of mitochondrial DNA as a molecular marker. *Mol. Biol. Evol.* **34**, 2762–2772 (2017).
93. Almeida-Rocha, J. M., Soares, L. A. S. S., Andrade, E. R., Gaiotto, F. A. & Cazetta, E. The impact of anthropogenic disturbances on the genetic diversity of terrestrial species: a global meta-analysis. *Mol. Ecol.* **29**, 4812–4822 (2020).
94. Landguth, E. L. et al. Quantifying the lag time to detect barriers in landscape genetics. *Mol. Ecol.* **19**, 4179–4191 (2010).
95. Paz-Vinas, I. et al. Macrogenetic studies must not ignore limitations of genetic markers and scale. *Ecol. Lett.* **24**, 1282–1284 (2021).
96. Crandall, E. D. et al. The molecular biogeography of the Indo-Pacific: testing hypotheses with multispecies genetic patterns. *Glob. Ecol. Biogeogr.* **28**, 943–960 (2019).
97. Excoffier, L. & Foll, M. fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27**, 1332–1334 (2011).
98. Guillaume, F. & Rougemont, J. Nemo: an evolutionary and population genetics programming framework. *Bioinformatics* **22**, 2556–2557 (2006).
99. Phillips, J. D., French, S. H., Hanner, R. H. & Gillis, D. J. HACSim: an R package to estimate intraspecific sample sizes for genetic diversity assessment using haplotype accumulation curves. *PeerJ Comput. Sci.* **6**, e243 (2020).
100. Gratton, P. et al. A world of sequences: can we use georeferenced nucleotide databases for a robust automated phylogeography? *J. Biogeogr.* **44**, 475–486 (2017).
101. Kimura, M. On the probability of fixation of mutant genes in a population. *Genetics* **47**, 713–719 (1962).
102. Baguette, M. & Van Dyck, H. Landscape connectivity and animal behavior: functional grain as a key determinant for dispersal. *Landsc. Ecol.* **22**, 1117–1129 (2007).
103. Crow, J. F. & Aoki, K. Group selection for a polygenic behavioral trait: estimating the degree of population subdivision. *Proc. Natl Acad. Sci. USA* **81**, 6073–6077 (1984).
104. Lanner, R. Why do trees live so long? *Ageing Res. Rev.* **1**, 653–671 (2002).
105. Nabholz, B., Mauffrey, J.-F., Bazin, E., Galtier, N. & Glemin, S. Determination of mitochondrial genetic diversity in mammals. *Genetics* **178**, 351–361 (2008).
106. Lasne, C., Heerwaarden, B., Sgrò, C. M. & Connallon, T. Quantifying the relative contributions of the X chromosome, autosomes, and mitochondrial genome to local adaptation. *Evolution* **73**, 262–277 (2019).
107. Phillips, J. D., Gillis, D. J. & Hanner, R. H. Incomplete estimates of genetic diversity within species: implications for DNA barcoding. *Ecol. Evol.* **9**, 2996–3010 (2019).
108. Humphries, P. & Winemiller, K. O. Historical impacts on river fauna, shifting baselines, and challenges for restoration. *BioScience* **59**, 673–684 (2009).
109. Stoffel, M. A. et al. Demographic histories and genetic diversity across pinnipeds are shaped by human exploitation, ecology and life-history. *Nat. Commun.* **9**, 4836 (2018).
110. Collier-Robinson, L., Rayne, A., Rupene, M., Thoms, C. & Steeves, T. Embedding indigenous principles in genomic research of culturally significant species: a conservation genomics case study. *N. Z. J. Ecol.* **43**, 3389 (2019).
111. Des Roches, S., Pendleton, L. H., Shapiro, B. & Palkovacs, E. P. Conserving intraspecific variation for nature's contributions to people. *Nat. Ecol. Evol.* **5**, 574–582 (2021).
112. Gonzalez, A. et al. Estimating local biodiversity change: a critique of papers claiming no net loss of local diversity. *Ecology* **97**, 1949–1960 (2016).
113. Pope, L. C., Liggins, L., Keyse, J., Carvalho, S. B. & Riginos, C. Not the time or the place: the missing spatio-temporal link in publicly available genetic data. *Mol. Ecol.* **24**, 3802–3809 (2015).
114. Yilmaz, P. et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* **29**, 415–420 (2011).
115. Sibbett, B., Rieseberg, L. H. & Narum, S. The genomic observatories metadatabase. *Mol. Ecol. Resour.* **20**, 1453–1454 (2020).
116. Eichenberg, D. et al. Widespread decline in Central European plant diversity across six decades. *Glob. Change Biol.* **27**, 1097–1110 (2020).
117. Cornwell, W. K., Pearse, W. D., Dalrymple, R. L. & Zanne, A. E. What we (don't) know about global plant diversity. *Ecography* **42**, 1819–1831 (2019).
118. Li, X. et al. Plant DNA barcoding: from gene to genome. *Biol. Rev.* **90**, 157–166 (2015).
119. Vasquez-Gross, H. A. et al. CartograTree: connecting tree genomes, phenotypes and environment. *Mol. Ecol. Resour.* **13**, 528–537 (2013).
120. Lawrence, E. R. et al. Geo-referenced population-specific microsatellite data across American continents, the MacroPopGen Database. *Sci. Data* **6**, 14 (2019). **This paper reports a compilation of georeferenced vertebrate microsatellite data, summary statistics and meta-data across the Americas for 897 species and 9,090 genetically distinct populations.**
121. Zellweger, F., De Frenne, P., Lenoir, J., Rocchini, D. & Coomes, D. Advances in microclimate ecology arising from remote sensing. *Trends Ecol. Evol.* **34**, 327–341 (2019).
122. Barber, P. H. et al. Advancing biodiversity research in developing countries: the need for changing paradigms. *Bull. Mar. Sci.* **90**, 187–210 (2014).
123. Bork, P. et al. Tara Oceans. Tara Oceans studies plankton at planetary scale. Introduction. *Science* **348**, 873–873 (2015).
124. Lotterhos, K. E. & Whitlock, M. C. The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Mol. Ecol.* **24**, 1031–1046 (2015).
125. Hoban, S. et al. Genetic diversity targets and indicators in the CBD post-2020 Global Biodiversity Framework must be improved. *Biol. Conserv.* **248**, 108654 (2020).
126. Holmes, M. W. et al. Natural history collections as windows on evolutionary processes. *Mol. Ecol.* **25**, 864–881 (2016).
127. Boukhdoud, L. et al. First DNA sequence reference library for mammals and plants of the Eastern Mediterranean Region. *Genome* **64**, 39–49 (2021).
128. Colella, J. P. et al. The Open-Specimen movement. *BioScience* **71**, 405–414 (2020).
129. Wright, S. Correlation and causation. *J. Agric. Res.* **20**, 557–585 (1921).
130. Fourtune, L. et al. Inferring causalities in landscape genetics: an extension of Wright's causal modeling to distance matrices. *Am. Nat.* **191**, 491–508 (2018).
131. Paz-Vinas, I., Loot, G., Stevens, V. M. & Blanchet, S. Evolutionary processes driving spatial patterns of intraspecific genetic diversity in river ecosystems. *Mol. Ecol.* **24**, 4586–4604 (2015).
132. Beaumont, M. A., Zhang, W. & Balding, D. J. Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035 (2002).
133. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
134. Proença, V. et al. Global biodiversity monitoring: From data sources to Essential Biodiversity Variables. *Biol. Conserv.* **213**, 256–263 (2017).
135. Větrovský, T. et al. A meta-analysis of global fungal distribution reveals climate-driven patterns. *Nat. Commun.* **10**, 5142 (2019).
136. Hanson, J. O. et al. Conservation planning for adaptive and neutral evolutionary processes. *J. Appl. Ecol.* **57**, 2159–2169 (2020).
137. Xuereb, A., D'Aloia, C. C., Andrello, M., Bernatchez, L. & Fortin, M. Incorporating putatively neutral and adaptive genomic data into marine conservation planning. *Conserv. Biol.* **35**, 909–920 (2021).
138. Carvalho, S. B., Torres, J., Tarroso, P. & Velo-Antón, G. Genes on the edge: a framework to detect genetic diversity imperiled by climate change. *Glob. Change Biol.* **25**, 4034–4047 (2019).
139. Adams, W. M. & Sandbrook, C. Conservation, evidence and policy. *Oryx* **47**, 329–335 (2013).
140. Laikre, L. et al. Post-2020 goals overlook genetic diversity. *Science* **367**, 1083.2–1085 (2020).
141. Thomson, A. I. et al. Charting a course for genetic diversity in the UN Decade of Ocean Science. *Evol. Appl.* **14**, 1497–1518 (2021).
142. Hoban, S. M. et al. Bringing genetic diversity to the forefront of conservation policy and management. *Conserv. Genet. Resour.* **5**, 593–598 (2013).
143. Carroll, S. R. et al. The CARE principles for indigenous data governance. *Data Sci. J.* **19**, 43 (2020).
144. Fargeot, L. et al. Patterns of epigenetic diversity in two sympatric fish species: genetic vs. environmental determinants. *Genes* **12**, 107 (2021).
145. Gaggiotti, O. E. et al. Diversity from genes to ecosystems: a unifying framework to study variation across biological metrics and scales. *Evol. Appl.* **11**, 1176–1193 (2018).
146. Waples, R. S., Antao, T. & Luikart, G. Effects of overlapping generations on linkage disequilibrium estimates of effective population size. *Genetics* **197**, 769–780 (2014).
147. Waples, R. S. & Yokota, M. Temporal estimates of effective population size in species with overlapping generations. *Genetics* **175**, 219–233 (2007).
148. Antao, T., Pérez-Figueroa, A. & Luikart, G. Early detection of population declines: high power of genetic monitoring using effective population size estimators. *Evol. Appl.* **4**, 144–154 (2011).
149. Cornuet, J. M. & Luikart, G. Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* **144**, 2001–2014 (1996).
150. Phillips, J. D., Gwiazdowski, R. A., Ashlock, D. & Hanner, R. An exploration of sufficient sampling effort to describe intraspecific DNA barcode haplotype diversity: examples from the ray-finned fishes (Chordata: Actinopterygii). *DNA Barcodes* **3**, 66–73 (2015).
151. Tajima, F. The effect of change in population size on DNA polymorphism. *Genetics* **123**, 597–601 (1989).
152. Jordan, R., Breed, M. F., Prober, S. M., Miller, A. D. & Hoffmann, A. A. How well do revegetation plantings capture genetic diversity? *Biol. Lett.* **15**, 20190460 (2019).
153. Holderegger, R. & Di Giulio, M. The genetic effects of roads: a review of empirical evidence. *Basic. Appl. Ecol.* **11**, 522–531 (2010).
154. Hale, M. L., Burg, T. M. & Steeves, T. E. Sampling for microsatellite-based population genetic studies: 25 to 30 individuals per population is enough to accurately estimate allele frequencies. *PLoS One* **7**, e45170 (2012).
155. Jackson, T. M., Roegner, G. C. & O'Malley, K. G. Evidence for interannual variation in genetic structure of Dungeness crab (*Cancer magister*) along the California Current System. *Mol. Ecol.* **27**, 352–368 (2018).
156. Hoban, S. et al. Comparative evaluation of potential indicators and temporal sampling protocols for monitoring genetic erosion. *Evol. Appl.* **7**, 984–998 (2014).

157. Anderson, C. N. K., Ramakrishnan, U., Chan, Y. L. & Hadly, E. A. Serial SimCoal: a population genetics model for data from multiple populations and points in time. *Bioinformatics* **21**, 1733–1734 (2005).
158. Hortal, J. et al. Seven shortfalls that beset large-scale knowledge of biodiversity. *Annu. Rev. Ecol. Evol. Syst.* **46**, 523–549 (2015).
159. Elbrecht, V., Vamos, E. E., Steinke, D. & Leese, F. Estimating intraspecific genetic diversity from community DNA metabarcoding data. *PeerJ* **6**, e4644 (2018).
160. Shum, P. & Palumbi, S. R. Testing small-scale ecological gradients and intraspecific differentiation for hundreds of kelp forest species using haplotypes from metabarcoding. *Mol. Ecol.* https://doi.org/10.1111/mec.15851 (2021).
161. Yamahara, K. M. et al. In situ autonomous acquisition and preservation of marine environmental DNA using an autonomous underwater vehicle. *Front. Mar. Sci.* **6**, 373 (2019).
162. Breed, M. F. et al. Mating patterns and pollinator mobility are critical traits in forest fragmentation genetics. *Heredity* **115**, 108–114 (2015).
163. Hoban, S., Gaggiotti, O. & Bertorelle, G. Sample Planning Optimization Tool for conservation and population Genetics (SPOTG): a software for choosing the appropriate number of markers and samples. *Methods Ecol. Evol.* **4**, 299–303 (2013).
164. Peck, S. L. Simulation as experiment: a philosophical reassessment for biological modeling. *Trends Ecol. Evol.* **19**, 530–534 (2004).
165. Reid, B. N., Naro-Maciel, E., Hahn, A. T., FitzSimmons, N. N. & Gehara, M. Geography best explains global patterns of genetic diversity and postglacial co-expansion in marine turtles. *Mol. Ecol.* **28**, 3358–3370 (2019).
166. Kardos, M., Luikart, G. & Allendorf, F. W. Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees. *Heredity* **115**, 63–72 (2015).
167. Willing, E.-M., Dreyer, C. & van Oosterhout, C. Estimates of genetic differentiation measured by F_{ST} do not necessarily require large sample sizes when using many SNP markers. *PLoS One* **7**, e42649 (2012).
168. Shafer, A. B. A. et al. Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods Ecol. Evol.* **8**, 907–917 (2017).
169. Cariou, M., Duret, L. & Charlat, S. How and how much does RAD-seq bias genetic diversity estimates? *BMC Evol. Biol.* **16**, 240 (2016).
170. De-Kayne, R. et al. Sequencing platform shifts provide opportunities but pose challenges for combining genomic data sets. *Mol. Ecol. Resour.* **21**, 653–660 (2021).
171. Leigh, D. M., Lischer, H. E. L., Grossen, C. & Keller, L. F. Batch effects in a multiyear sequencing study: false biological trends due to changes in read lengths. *Mol. Ecol. Resour.* **18**, 778–788 (2018).
172. Linck, E. & Battey, C. J. Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Mol. Ecol. Resour.* **19**, 639–647 (2019).
173. Benestan, L. M. et al. Conservation genomics of natural and managed populations: building a conceptual and practical framework. *Mol. Ecol.* **25**, 2967–2977 (2016).
174. Feng, S. et al. Dense sampling of bird diversity increases power of comparative genomics. *Nature* **587**, 252–257 (2020).
175. Brandies, P., Peel, E., Hogg, C. J. & Belov, K. The value of reference genomes in the conservation of threatened species. *Genes* **10**, 846 (2019).

Acknowledgements
We acknowledge the support of the GEO BON Genetic Composition Working Group in the development of this manuscript. We thank B. Dauphin, L. Beheregaray, L. Di Santo, W. C. Funk, J. Fant, A. MacDonald, A. Strand, C. Grueber and C. Richards for their insightful comments on early versions of the manuscript. D.M.L. is funded by the SNSF grant IZHRZ0_180651, "Dynamics of virus infection in mycovirus-mediated biological control of a fungal pathogen". This research was funded in part by a USGS Northwest Climate Adaptation Science Center award G17AC000218 to C.Bv.R. I.P-V. works in a laboratory supported by the 'Laboratoire d'Excellence' (LABEX) entitled TULIP (ANR-10-LABX-41). S.H. was supported by National Science Foundation grant 1759759. L.L. is supported by a New Zealand Rutherford Discovery Fellowship (RDF-20-MAU-001). M.F.B. is funded by Australian Research Council (ARC) grants LP190100051, LP190100484, DP210101932 and DP180100668. CS was funded by an NSERC Discovery Grant to Colin J. Garroway.

Author contributions
D.M.L., C.B.v.R., K.L.M., M.F.B., C.S., S.H. and I.P-V. contributed to all aspects of the article. L.D.B., B.K.H., M.E.H., E.L.J., F.K., L.L., G.L., S.M., J.M., J.M.M. and G.S. researched data for the article, made substantial contributions to discussions of the content, and reviewed and/or edited the manuscript before submission.

Competing interests
The authors declare no competing interests.

Disclaimer
Any use of trade, firm or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Peer review information
Nature Reviews Genetics thanks B.C. Carstens, D. Nogues-Bravo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary information
The online version contains supplementary material available at https://doi.org/10.1038/s41576-021-00394-0.

RELATED LINKS
California Conservation Genomics Project: https://www.ccgproject.org/
Chelsa: https://chelsa-climate.org/
CryoArks: https://www.cryoarks.org/
Distributed System of Scientific Collections (DiSSCO): https://www.dissco.eu
DRYAD: https://datadryad.org/
Figshare: https://figshare.com/
GenBank: https://www.ncbi.nlm.nih.gov/genbank/
GeoNames: https://www.geonames.org/
GitHub: https://github.com
Global Biodiversity Information Facility (GBIF): https://www.gbif.org/
International Nucleotide Sequence Database Collaboration INSDC: http://www.insdc.org/
International Union for Conservation of Nature (IUCN): https://www.iucn.org
MIxS standards: https://gensc.org/mixs/
NASA's Earth observatory platform: https://Earthobservatory.nasa.gov/
Otlet: https://otlet.io