Global genetic diversity status and trends: towards a suite of Essential Biodiversity Variables (EBVs) for genetic composition


1Center for Tree Science, The Morton Arboretum, 4100 Illinois Rt 53, Lisle, IL, 60532, USA
2Southwest Fisheries Science Center, NOAA/NMFS, 8901 La Jolla Shores Drive, La Jolla, CA, 92037, USA
3City College of New York, 160 Convent Avenue, New York, NY, 10031, USA
4Research Centre for Ecosystem Resilience, Australian Institute of Botanical Science, The Royal Botanic Garden Sydney, Mrs Macquaries Rd, Sydney, NSW, 2000, Australia
5College of Science and Engineering, Flinders University, University Drive, Bedford Park, SA, 5042, Australia
6School of Biosciences, Cardiff University, Cathays Park, Cardiff, CF10 3AX, Wales, UK
7Department of Primary Industries, New South Wales Fisheries, National Marine Science Centre, 2 Bay Drive, Coffs Harbour, NSW, 2450, Australia
8Wildlife Analysis Unit, Swedish Environmental Protection Agency, Blekelholmserasen 36, Stockholm, SE-106 48, Sweden
9Department of Biology, Graduate Degree in Ecology, Colorado State University, 1878 Campus Delivery, Fort Collins, CO, 80523-1878, USA
10School of Life and Environmental Sciences, Faculty of Science, The University of Sydney, Carslaw Building, Sydney, NSW, 2006, Australia
11Flathead Lake Biological Station, 32125 Bio Station Ln, Polson, MT, 59860, USA
12Exponent, 15375 SE 30th Place, Suite 250, Bellevue, WA, 98007, USA
13School of Natural and Environmental Sciences, Newcastle University, Agriculture Building, Newcastle Upon Tyne, NE1 7RU, UK
14Department of Environmental Studies, Prescott College, 220 Grove Avenue, Prescott, AZ, 86303, USA
15Natural Resources Defense Council, 40 West 20th Street, New York, NY, 10011, USA
16School of Natural Sciences, Massey University, Ōtahuhurohe Campus, Gate 4, Albany Highway, Auckland, Aotearoa, 0745, New Zealand
17Research School of Biology, The Australian National University, Acton, ACT, 2601, Australia
18Research Institute for Nature and Forest, Gaverstraat 4, 9500, Geraardsbergen, Belgium
19Aquatic Ecology, Evolution and Conservation, KULeuven, Charles Deberiotstraat 32, box 2439, 3000, Leuven, Belgium
20Department of Biological Sciences, MacEwan University, 10700 104 Avenue, Edmonton, AB, T5J 4S2, Canada
21College of Marine Science, University of South Florida, 140 7th Avenue South, Saint Petersburg, Florida, 33701, USA
22NatureScot, Great Glen House, Leachkin Road, Inverness, IV3 8NW, UK
23Laboratoire Évolution et Diversité Biologique, Université de Toulouse, CNRS, IRD, UPS, UMR-5174 EDB, 118 route de Narbonne, Toulouse, 31062, France

* Author for correspondence (Tel.: +601 352 264 3484; E-mail: mhunter@usgs.gov).
1Present address: Australian Antarctic Division, Department of Agriculture, Water and the Environment, 203 Channel Highway, Kingston, Tasmania, 7050, Australia
3Group on Earth Observation Biodiversity Observation Network.

Biol. Rev. (2022), 97, pp. 1511–1538 © 2022 The Authors. Biological Reviews published by John Wiley & Sons Ltd on behalf of Cambridge Philosophical Society. This article has been contributed to by U.S. Government employees and their work is in the public domain in the USA. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
ABSTRACT

Biodiversity underlies ecosystem resilience, ecosystem function, sustainable economies, and human well-being. Understanding how biodiversity sustains ecosystems under anthropogenic stressors and global environmental change will require new ways of deriving and applying biodiversity data. A major challenge is that biodiversity data and knowledge are scattered, biased, collected with numerous methods, and stored in inconsistent ways. The Group on Earth Observations Biodiversity Observation Network (GEO BON) has developed the Essential Biodiversity Variables (EBVs) as fundamental metrics to help aggregate, harmonize, and interpret biodiversity observation data from diverse sources. Mapping and analyzing EBVs can help to evaluate how aspects of biodiversity are distributed geographically and how they change over time. EBVs are also intended to serve as inputs and validation to forecast the status and trends of biodiversity, and to support policy and decision making. Here, we assess the feasibility of implementing Genetic Composition EBVs (Genetic EBVs), which are metrics of within-species genetic variation. We review and bring together numerous areas of the field of genetics and evaluate how each contributes to global and regional genetic biodiversity monitoring with respect to theory, sampling logistics, metadata, archiving, data aggregation, modeling, and technological advances. We propose four Genetic EBVs: (i) Genetic Diversity; (ii) Genetic Differentiation; (iii) Inbreeding; and (iv) Effective Population Size ($N_e$). We rank Genetic EBVs according to their relevance, sensitivity to change, generalizability, scalability, feasibility and data availability. We outline the workflow for generating genetic data underlying the Genetic EBVs, and review advances and needs in archiving genetic composition data and metadata. We discuss how Genetic EBVs can be operationalized by visualizing EBVs in space and time across species and by forecasting Genetic EBVs beyond current observations using various modeling approaches. Our review then explores challenges of aggregation, standardization, and costs of operationalizing the Genetic EBVs, as well as future directions and opportunities to maximize their uptake globally in research and policy. The collection, annotation, and availability of genetic data has made major advances in the past decade, each of which contributes to the practical and standardized framework for large-scale genetic observation reporting. Rapid advances in DNA sequencing technology present new opportunities, but also challenges for operationalizing Genetic EBVs for biodiversity monitoring regionally and globally. With these advances, genetic composition monitoring is starting to be integrated into global conservation policy, which can help support the foundation of all biodiversity and species’ long-term persistence in the face of environmental change. We conclude with a summary of concrete steps for researchers and policy makers for advancing operationalization of Genetic EBVs. The technical and analytical foundations of Genetic EBVs are well developed, and conservation practitioners should anticipate their increasing application as efforts emerge to scale up genetic biodiversity monitoring regionally and globally.

Key words: biodiversity monitoring, environmental policy, indicators, metadata, interoperability, molecular ecology

CONTENTS

I Introduction .......................................................... 1513
(1) Genetic data and infrastructure are ready for EBVs ............................................. 1514
(2) What is genetic composition and why is it important for biodiversity policy, conservation, and society? ................................................................. 1515
(3) Aims and structure of this review .................................................................... 1515
II Genetic EBVs .......................................................... 1516
(1) Four proposed Genetic EBVs ........................................................................ 1516
(a) What drives change in genetic composition? ..................................................... 1516
(b) A note on adaptive diversity ........................................................................... 1518
(2) Genetic diversity EBVs: richness and evenness ............................................. 1518
(3) Genetic differentiation EBVs: number of units and their connectivity ............. 1519
(4) Inbreeding EBV ..................................................................................... 1520
(5) Contemporary effective population size EBV ............................................. 1520
(6) Why four Genetic EBVs? ............................................................................ 1521
I INTRODUCTION

Governments, business, and society increasingly recognize that biodiversity contributes to ecosystem resilience and function, sustainable economies, and human well-being. Changes in biodiversity impact ecosystem services, often in ways that are hard to predict. To address this problem, government, non-government, and research organizations aim to quantify biodiversity trends, assess drivers of biodiversity change, measure responses of biological systems to interventions, and define policy options for preserving and restoring biodiversity. These activities require observational data collected across space and time on all levels of biological organization (genes, species and ecosystems). The Group on Earth Observations Biodiversity Observation Network (GEO BON) is tackling this challenge with the Essential Biodiversity Variable (EBV) concept, to help aggregate, harmonize, summarize, and interpret complex biodiversity observation data, especially from diverse data sources (Pereira et al., 2013; Brummitt et al., 2017; Navarro et al., 2017; Schmeller et al., 2017).

EBVs are a set of fundamental metrics that describe genetic, species, population, and ecosystem diversity, building on and harmonizing direct observations (e.g. species occurrences) to serve as a basis for indicators or inputs for modeling and forecasting the status and trends of biodiversity. EBVs track core components of biodiversity change, using highly standardized and scalable biodiversity observations. The concept of Genetic Composition EBVs goes beyond traditional methods and statistics and their practical use in biodiversity monitoring (Schwartz, Luikart & Waples, 2007; Mimura et al., 2017; Leroy et al., 2018); EBVs should provide the most essential information across space, time, and species. According to Pereira et al. (2013), a short list of EBVs that are standardized, recognized, and responsive to natural and anthropogenic drivers will allow for uptake of high-quality, comparable and scalable biodiversity data by an array of endpoint users. Similar to Essential Climate Variables (ECVs; Bojinski et al., 2014), EBVs should be relevant, sensitive to change, generalizable, scalable, feasible, and have substantial accessible databases available (Schmeller et al., 2018). EBVs must consider the fundamental characteristics of biological systems, and methods for collecting and analyzing observation data. EBVs should be applicable to all life forms, at multiple scales, and in all ecosystems. The implementation of EBVs allows for data to be synthesized into products such as indicators (e.g. the Species Habitat Index, https://geobon.org/ebvs/indicators/) for measuring progress on policy commitments, such as the Convention on Biological Diversity (CBD), Sustainable Development Goals, Global Strategy for Plant Conservation, and Intergovernmental Panel on Biodiversity and Ecosystem Services (IPBES; Schmeller et al., 2015; Brummitt et al., 2017; Navarro et al., 2017).

Pereira et al. (2013) proposed six EBV ‘classes’ corresponding to ways that biodiversity can be categorized, observed, and managed: Genetic Composition, Species Populations, Species Traits, Community Composition, Ecosystem Structure, and Ecosystem Function. Each EBV class may comprise multiple EBVs that quantify different key attributes of that level of biodiversity; they should be complementary and comprehensive. For example, two EBVs are proposed in the Species Populations class (species distribution and species abundance; Jetz et al., 2019), while five EBVs are proposed in the Species Traits class (phenology, morphology, reproduction, physiology, and movement; Kissling et al., 2018). The EBV concept was also used for invasive species (Latombe et al., 2017), and an ocean observation framework (Marine BON; Muller-Karger et al., 2018). Schmeller et al. (2018) proposed six criteria for evaluating the suitability of potential EBVs. Progress is ongoing to achieve consensus on suitable EBVs for each class and to define workflows connecting data to EBVs—work that must be accomplished before EBVs can be used operationally (Schmeller et al., 2017).
Genetic composition (within-species genetic variation) is a foundational component of biodiversity. It contributes to function and structure in all ecosystems and to resilience and/or productivity in agriculture, aquaculture, and forestry systems (see Section I.3). Genetic composition can respond to environmental change rapidly and without changes in species abundance (Balkenhol et al., 2016), and in some cases genetic data are cheaper and easier to gather than other methods of biodiversity monitoring (see Carroll et al., 2018; Ferreira et al., 2018; Parsons et al., 2018; Wheat et al., 2016). Nonetheless, observations of genetic composition have not been sufficiently mobilized for, nor integrated into, large-scale biodiversity conservation policy, sustainable development, and decision-making on resource management to date (Laikre et al., 2010; Pierson et al., 2016; Bruford et al., 2017; Thomson et al., 2021), even though it ultimately determines the capacity of populations to adapt to environmental changes and therefore underpins their long-term persistence (Reed & Frankham, 2003; see Section I.3). Although many countries have monitoring programs for species and ecosystem variables, few systematically monitor genetic variation; a recent report found that the status and trends of genetic diversity are rarely included in National Reports to the CBD (Hoban et al., 2021). However, Scotland has committed to regular reporting on genetic diversity, with the first assessment published in 2020 (Hollingsworth et al., 2020), while Sweden has committed to large-scale monitoring of genetic diversity (Posledových, Ekblom & Laikre, 2021).

(1) Genetic data and infrastructure are ready for EBVs

While there are many reviews of genetic tools and statistics, and their practical use in biodiversity monitoring (Schwartz et al., 2007; Mimura et al., 2017; Lévy et al., 2018), there has not yet been an effort to define clearly Genetic Composition EBVs (hereafter Genetic EBVs), summarize the practical framework for collecting and analyzing genetic observations via standardized procedures, and examine the challenges and opportunities that will ensure their usefulness for biodiversity monitoring. It is now timely to do so. Over the last 50 years, genetic data production and methods have advanced rapidly. Genetic data were first collected in the form of isozymes and allozymes from the 1970s to early 1990s, followed by DNA sequences, microsatellites and other technologies from the late 1980s to the present day, and large-scale next generation sequencing assessing genome-wide variation starting in the 2000s and increasing in use to the present. Detailed overviews of the temporal trends, applications and integration of the various types of molecular markers can be found in Schlötterer (2004), Ouborg et al. (2010), Putman & Carbone (2014), Garner et al. (2016), Garrido-Cardenas, Mesa-Valle & Manzano-Agugliaro (2018), and Leigh et al. (2021).

Genetic composition has been assessed for thousands of species globally with increasingly high-resolution population-level genetic data sets available at decreasing costs (Vranckx et al., 2012; Romiguier et al., 2014; Lawrence et al., 2019). Although the changing technology and limited number of systematic monitoring programs does pose problems for spatiotemporal comparisons (see Sections III.1 on obtaining genetic data and V.2 on scale and standardization), recent studies are identifying gaps in genetic marker data availability (Miraldo et al., 2016) and are jointly analyzing genetic summaries obtained across different species, studies and markers using standardization and normalization procedures [e.g. De Kort et al. (2021) for amplified fragment length polymorphism (AFLP)/microsatellites]. Further, hundreds of species now have had their genomes sequenced and assembled, which makes developing genetic markers and interpreting genetic data more efficient in these and related species (e.g. within the same genus). We acknowledge that, at present, genetic data cover only a tiny fraction of all species on Earth and that it will take substantial time and effort to fill the spatiotemporal and taxonomic gaps. However, these gaps and biases are true for other biodiversity data [e.g. detailed data for species’ population sizes and trends is limited to few species (Boakes et al., 2010; Fithian et al., 2013; Hortal et al., 2015; Meyer et al., 2016; Jetz et al., 2019)] and the rapid advances in genetic data accessibility mean that these gaps are swiftly being filled (e.g. the growth in the availability of whole-genome sequences illustrated in fig. 2 in Leigh et al., 2021).

Numerous agencies routinely assess the genetic composition of high-priority species, including via genome sequencing. Advances in sequencing technology, data storage and analysis are opening new frontiers including DNA sequencing ‘in the field’ (Pomerantz et al., 2018; Ovaskainen et al., 2020; see Section V), as well as analyzing DNA from biological collections or paleospecimens to establish pre-human-impact baselines (Bi et al., 2019). Non-invasive (e.g. hair and scat) and environmental (e.g. soil, water) sampling is also increasingly standardized, making genetic data more widely available. There are well-developed, standard data-storage formats and analysis software (often open source, such as R – see Molecular Ecology Resources, Volume 17, Issue 1: Population Genomics with R, and online Supporting Information Fig. S1 and Table S1). Methods for modeling genetic composition and extrapolation in space and time are also well advanced (see Section IV.2).

Alongside data production, advances are occurring in data stewardship. Owing to the research community’s adherence to the 2011 Joint Data Archiving Policy, standardized data formats, and the existence of numerous large-scale data archives (Section III), genetic data are among the most accessible of all biodiversity data, although challenges remain, e.g. missing metadata, incomplete data files, and biases (Toczyszkowski et al., 2021). The ability to collate and aggregate genetic data sets and calculate genetic summary statistics for potential Genetic EBVs has been demonstrated (Section IV). Furthermore, consensus and uptake of metadata standards for biological collections are aligning genetic
data sets with the FAIR principles [Findable, Accessible, Interoperable, and Reproducible (https://www.go-fair.org/fair-principles/; Wilkinson et al., 2016)], applying data schemes like the Minimum Information about any (x) Sequence (MiXS, https://gens.org/mixs/) devised by the Genomics Standards Consortium, and establishing highly searchable metadatabases (Sections III and IV). Accessible and affordable genomic-level data, and digitization and curation of older genetic data (Sections IV and V), provide a pathway towards operational Genetic EBVs. In parallel, all the aforementioned advances have recently favored the emergence of a new field named ‘macrogenetics’, which aims at understanding patterns and drivers of genetic variation across large spatial, temporal, and/or taxonomic scales [hundreds to thousands of species (Blanchet, Prunier & De Kort, 2017; Leigh et al., 2021)]. The fast-growing field of macrogenetics is paving the way for Genetic EBV deployment (see Section IV.1 for details).

(2) What is genetic composition and why is it important for biodiversity policy, conservation, and society?

We define the Genetic EBV class to encompass inherited components, such as variations in DNA sequence or epigenetic modifications (Rey et al., 2020), that determine form and function and vary among individuals or populations of a species. This EBV class has relevance to global policy targets including CBD Aichi Target 13, ‘minimize genetic erosion’ and ‘safeguard genetic diversity’; Sustainable Development Goal (SDG) 2.5, ‘maintain genetic diversity’, SDG 14 on marine resources and SDG 15, on ecosystems and biodiversity; and to Global Strategy for Plant Conservation Target 5, ‘effective management [for] conserving genetic diversity’ and Target 9, ‘70 per cent of genetic diversity ... conserved’. Genetic diversity is mentioned in these policy mechanisms, and its multiple dimensions must be understood, monitored, and managed to achieve sustainable development, benefits, and conservation goals. Genetic EBVs recently have been linked with conservation decision-making in the publication ‘Effective Biodiversity Indicators’ by the UK Government (Henly & Wentworth, 2021) and the Food and Agriculture Organization (FAO) has incorporated the effective population size ($N_e$) as an indicator for forest genetic monitoring (Graudal et al., 2020).

Genetic composition is a measure of within-species diversity, which helps species adapt, maintain fitness, avoid inbreeding depression, and underlies the breadth of species’ niches and the diversity of biotic and abiotic interactions. Genetic diversity is the foundation of the three levels of biodiversity, supporting and complementing species and ecosystems diversity. Genetic diversity provides resilience against abrupt changes and allows species and ecosystems to adapt to changing environments, climates, and other challenges (including diseases). Ultimately, genetic variation allows species to develop into distinct and new lineages. Genetic diversity within and across populations supports ecosystem functions and contributes vital resources to society (e.g. seagrass and mangroves that serve both to protect coastal habitats and as nurseries for fish), and services such as carbon capture (Kettenring et al., 2014; Hollingsworth et al., 2020). Genetic diversity has also been used for millennia to provide thousands of domesticated or harvested species with new adaptations to climate, disease, soil types, food sources, altitude, etc., and to improve sustainability and reduce industrial inputs (e.g. natural defenses against pests). In short, genetic diversity underlies processes upon which all other biodiversity depends.

Genetic diversity data can be observed directly through the analysis of genetic markers (e.g. DNA sequences, microsatellite fragment data, single nucleotide polymorphisms (SNPs), methylation patterns, etc.) in individuals sampled through geographic space and time. Genetic observations are often summarized as allele frequencies at the population level (analogous to species abundances in an ecosystem; Hu, He & Hubbell, 2006), or through summary statistics measured at various hierarchies (individual, population/metapopulation, lineage, or species level; Fig. 1). Such hierarchies can be characterized as alpha diversity (population-level measures at one location), beta diversity (differentiation between locations in space or time), and gamma diversity (across the system of populations).

Note that molecular genetic tools such as DNA sequences are also used to gather observations for taxonomic diversity at and above the species level (e.g. phylogenetics and systemsatics). While genetic data within a species are applied to the Genetic EBV class, data above the species level are appropriate for the Species or Community EBV classes. Molecular genetic data may also be collected for genetic forensics and environmental DNA (eDNA) assessments. Targeted (e.g. species-specific) and metabarcoding eDNA tools can infer species presence using eDNA detected in water, soil, and air samples (Goodwin et al., 2019; Balázs et al., 2020) and are well suited to provide data for the Species Population EBV classes; but note, only within-species data are relevant for the Genetic EBV class.

(3) Aims and structure of this review

In this review we elucidate the major stages and gaps on the route towards enabling large-scale standardized observations of genetic variants. Section II defines Genetic EBVs, explains their calculation and meaning, and assesses how each Genetic EBV meets the criteria of Schmeller et al. (2018): relevance, sensitivity to change, generalizability, scalability, feasibility, and available data. Section III explains the workflow for transforming, aggregating, harmonizing and archiving genetic data underlying Genetic EBVs, including publishing data and metadata standards in interoperable formats for integration with large data sets. In Section IV, we discuss methods and models for interpolation and extrapolation, translating genetic data or proxies for genetic data into patterns of change across space and time, and looking towards EBV development for global application. In Section V, we
discuss opportunities and challenges, and assimilation of Genetic EBVs into regional and global policy instruments. We conclude that the technical and analytical foundation of Genetic EBVs is extremely well developed, but additional steps are needed to scale Genetic EBVs for monitoring programs globally, establish best practices for informatics and modeling, and reconcile data sets with different resolutions.

II GENETIC EBVs

(1) Four proposed Genetic EBVs

Genetic EBVs have been proposed previously but were not thoroughly defined or examined in detail (Pereira et al., 2013; Turak et al., 2017; Kissling et al., 2018). The four previously proposed EBVs were: allelic diversity, population genetic differentiation, co-ancestry, and breed/varietal diversity. Here we refine, revise and clarify these EBVs (Table 1). The first three are proposed to be renamed as: (i) Genetic Diversity, (ii) Genetic Differentiation, and (iii) Inbreeding. The first three EBVs reflect Wright’s F-statistics (Wright, 1969, 1978) measuring the genetic variation among populations, within populations and within individuals, respectively. A new EBV, (iv) Effective Population Size (N_e), also developed by Wright, is added to measure the rate of change (particularly loss) of genetic diversity due to drift in the next generations (Luikart et al., 2010). As opposed to the current state of genetic diversity, N_e informs inferences on future genetic diversity changes and is highly suited for monitoring of conservation management actions. The previous EBV of breed/varietal diversity was integrated into the more comprehensive Genetic Differentiation EBV.

We examine how each EBV reflects the essential characteristics of genetic composition, genetic health, and genetic viability of populations and species. We explain the biological features that each represents, discuss how to interpret change, and review assumptions and cautions. The proposed list should be applicable for wild, semi-wild or managed, and domesticated species, and covers the forces that underpin genetic variation (Section II.1a). A key point is that although each EBV summarizes a different aspect of genetic composition, all Genetic EBVs can typically be calculated from a single data set originating from DNA samples collected from individuals in a target population. The EBVs we propose can be calculated using different genetic markers (e.g. whole-genome sequencing data, SNPs, DNA sequences, microsatellites).

(a) What drives change in genetic composition?

The field of population genetics has a well-established mathematical framework and predictive theory for how processes interact to shape genetic composition, which can enable a high degree of forecasting accuracy. The genetic variation of
Table 1. The four proposed Genetic Composition Essential Biodiversity Variables (EBVs), their definitions, possible values, interpretation, and a ranking from low (+) to high (+++) on various EBV criteria (rows)

<table>
<thead>
<tr>
<th>Genetic EBV</th>
<th>Genetic Diversity</th>
<th>Genetic Differentiation</th>
<th>Inbreeding</th>
<th>Effective Population Size (Ne)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Richness</td>
<td>Diversity</td>
<td>Number of genetic units</td>
<td>Degree of relatedness between pairs of individuals, mating among relatives, or identity by descent</td>
</tr>
<tr>
<td></td>
<td>Evenness</td>
<td>Differentiation between units</td>
<td>Degree of genetic differentiation among populations or units</td>
<td></td>
</tr>
<tr>
<td>Definition</td>
<td>Count of the number of alleles in a population</td>
<td>Expected proportion of heterozygotes in a population at equilibrium</td>
<td>Number of genetic lineages/units within a species</td>
<td></td>
</tr>
<tr>
<td>Examples</td>
<td>Allelic richness, nucleotide diversity (π)</td>
<td><em>H</em>o, <em>H</em>D</td>
<td><em>F</em>ST, <em>G</em>ST, <em>D</em>est</td>
<td></td>
</tr>
<tr>
<td>Range of values</td>
<td>0 to 1</td>
<td>≥1</td>
<td>0 to 1</td>
<td>-1 to 1</td>
</tr>
<tr>
<td>Level of organization</td>
<td>Population</td>
<td>Individual/population</td>
<td>Across species range or specific region</td>
<td>Individual/pairs or families/population</td>
</tr>
<tr>
<td>Interpretation of a change</td>
<td>Decline indicates loss of adaptive capacity</td>
<td>Decline indicates loss of independent evolutionary or demographic units (probable loss of adaptations and long-term persistence)</td>
<td>Increased differentiation indicates increased inbreeding which increases the chance of expression of deleterious alleles, reducing fitness</td>
<td></td>
</tr>
<tr>
<td>Interpretation of a change</td>
<td>Identified across loci</td>
<td><em>F</em>st, <em>F</em>D, <em>F</em>st for loci with two or more alleles.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interpretation of a change</td>
<td>Identified across loci</td>
<td><em>F</em>ST = unbiased genetic differentiation among subpopulations.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relevance to species status</td>
<td>+++</td>
<td>++</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>Sensitivity to change over time</td>
<td>+++</td>
<td>+</td>
<td>+++ (depends on life history and Ne)</td>
<td>+++ (depends on method)</td>
</tr>
<tr>
<td>Spatial scalability</td>
<td>+++</td>
<td>+</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>Feasibility</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+/++ (depends on measure - FROH currently less feasible)</td>
</tr>
<tr>
<td>Data available</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+</td>
</tr>
</tbody>
</table>

*H*o = expected heterozygosity under Hardy–Weinberg assumptions, and *H*D = observed heterozygosity, yielding the probability of randomly drawing two different alleles from the population. *F*ST = genetic differentiation (fixation index) measured as the proportion of the total genetic variance contained in a subpopulation relative to the total genetic variance. *G*ST = genetic differentiation of subpopulations relative to the total population, a generalized index of *F*ST for loci with two or more alleles. *D*est = unbiased genetic differentiation among subpopulations. *F*ROH = inbreeding by autozygosity estimated from runs of homozygosity, i.e. the proportion of the autosomal genome above a specified length, *F*IM = the increase in individual homozygosity relative to mean Hardy–Weinberg expected homozygosity. *t*m = the multilocus outcrossing rate, which is 1 – the self-fertilization rate. *t*o = the degree of outcrossing among unrelated individuals only. *t*e; *t* = the estimated rate of biparental inbreeding. LD, linkage disequilibrium.

++/+ is used for contrast: either positive (++) or negative (+). See Appendix S1 for discussion of alternative spatial scalability ranking. Note that each EBV can be represented by numerous statistics which have individual calculations (Leroy et al., 2018; see Appendix S1).
species and their populations is entirely shaped by four processes: (i) **Mutation** is the random creation of new genetic variants, increasing genetic diversity. (ii) **Genetic drift** is the fluctuation in a population’s allele frequencies due to the random sampling of alleles from generation to generation (e.g. not due to selection). In small populations, these chance statistical effects can lead to rapid change and typically decrease genetic diversity (e.g. “genetic erosion”) and increase among-population genetic differentiation over time (although note that genetic drift can reduce the efficacy of stabilizing selection, which might increase diversity). (iii) **Gene flow** is the exchange of genetic variation among populations via gametes (e.g. pollen, sperm) or whole organisms that mate and produce offspring, and can drive changes in local levels (including local increases) of genetic variation. (iv) **Selection** is a process influencing the differential survival and reproduction of genetic variants across space and time, and can influence the genetic variation within populations and the differentiation between them, depending on the type of selection. Selection is related to specific genomic sites (the genes influencing certain phenotypes, and linked regions), while drift, mutation, and migration can affect diversity at sites across the genome. The influence of these four processes is mediated by \( N_e \), which is roughly related to the number of breeding adults in a genetically ideal population. High \( N_e \) results in more effective selection, more mutations, and lower genetic drift (Charlesworth, 2009). Decades of genetic measurements of these four processes have identified and quantified major detrimental anthropogenic drivers (e.g. harvesting, selective breeding, habitat change, pollution, reduced population size, isolation, and habitat fragmentation), as well as beneficial anthropogenic drivers such as genetic rescue and migration corridors (Pacioni, Wayne & Page, 2019; Almeida-Rocha *et al.*, 2020). Genetic composition is also influenced by life history traits that render some species more sensitive to genetic erosion, or less able to respond to environmental changes through genetic adaptations (e.g. long generation time, small population sizes, inbreeding, self or clonally propagated, restricted dispersal ability) (Romiguier *et al.*, 2014; Schmeller *et al.*, 2018).

While, the genetic EBVs are not presented in order of importance, we begin with Genetic Diversity within a population because this is the unit where natural selection takes place. Nonetheless we emphasize that defining populations (e.g. assessing genetic differentiation or population structure) is usually the first step in evaluating Genetic EBVs. Genetic and other data (environmental, geographic location) can be used to help define populations, a topic covered extensively in the literature (Fenderson, Kovach & Llamas, 2020). Once independent population units are defined, the other EBVs are calculated within each unit. EBVs would typically be calculated on genome-wide or neutral-marker data sets, although see the following discussion on adaptive diversity.

(b) **A note on adaptive diversity**

Genetic loci are often categorized as adaptive (e.g. ‘non-neutral’), or ‘neutral’. Adaptive genetic differentiation, the outcome of natural selection, may confer fitness advantages under local conditions. Genetic composition will differ when measured using adaptive compared to neutral genetic markers (Ralls *et al.*, 2018), which can impact recommended management actions (e.g. Van Oppen *et al.*, 2017). When possible, adaptive loci should be identified and analyzed separately, as comparing population genetic structure assessed using neutral and adaptive markers can provide alternative conservation options (Waples & Lindley, 2018). Identifying adaptive loci under selection is challenging (Hoban *et al.*, 2016), but methods are advancing to account better for loci under selection (Hohenlohe, Funk & Rajora, 2021).

Various correlation methods can be used to identify putative adaptive loci, often through interactions with simple environmental data (Rellstab *et al.*, 2015; Ahrens *et al.*, 2018). Alternatively, loci with a particular, known function (identified with a reference genome and/or field and laboratory experiments) can be analyzed, although such data are available for fewer species. Nonetheless, these loci (assuming sufficiently large SNP datasets, good ecological knowledge and a sequenced, annotated genome) are useful to map local adaptations, model adaptive responses, and determine the number and degree of differentiation among ‘population units’ – the required first step for evaluating Genetic EBVs.

(2) **Genetic diversity EBVs: richness and evenness**

‘Genetic diversity’ describes the level of genetic variability within a population that enables species to adapt and persist. Previously, Pereira *et al.* (2013) referred only to allelic richness, but genetic diversity has two key components: richness and evenness; both should be recorded to capture genetic diversity fully. Both are easy metrics to assess and explain to practitioners (Fig. 2). Richness refers to the count of genetic variants in a population, corrected for sample size (e.g. an expectation for a given sample size). Several statistics fall under this category including number of haplotypes (usually for organellae DNA), number of distinct genetic sequences observed in a sample, and allelic richness [for microsatellites and SNPs, the mean number of genetic variants (alleles) per defined genomic position (locus)]. The number and difference among genetic states are the units that ultimately determine form and function and are the units on which natural selection can act. Genetic richness can only be increased via mutations (or, locally, by gene flow).

Evenness of genetic diversity is often represented by expected heterozygosity (\( H_e \)). It is probably the most widely reported statistic in genetic studies, irrespective of the marker type. It expresses the probability that two randomly drawn alleles from a sample are different [it is identical to the Gini-Simpson diversity index from ecological literature]. Similar to richness, evenness will help a population respond rapidly to environmental change through adaptation. Both richness and evenness have been recorded in thousands of genetic studies (see Section IV).

Richness and evenness are correlated in populations at mutation-drift equilibrium [Nei, Maruyama & Chakraborty,
1975; Fischer et al., 2017), but these metrics respond differently to perturbations. Allelic richness measures respond faster to decreases in population size (similar to loss of rare species) while genetic evenness (heterozygosity) may respond faster to population mixing. Such different responses can be used to make inferences on past demography (population expansions or bottlenecks; Nei et al., 1975). For both measures, a decrease is generally interpreted as a loss of evolutionary potential – the ability of a population or species to adapt to rapid environmental changes such as climate change. Importantly, genetic measures of richness and evenness are not directly comparable when calculated using different genetic markers (although standardization procedures can make them comparable; see De Kort et al., 2021), so interpretation should be directed by the type and number of genetic markers and the number of populations/individuals analyzed (see Section V.2 on data set parameter selection). Alternatively, an integrated framework for measuring diversity, such as Hill numbers, could be explored for unifying different measures of Genetic EBVs, and to investigate variation at different EBV classes across multiple scales (Sherwin et al., 2017; Gaggiotti et al., 2018).

(3) Genetic differentiation EBVs: number of units and their connectivity

Genetic differentiation is defined as a divergence in genetic composition (specifically the frequencies of alleles) between multiple populations of the same species, caused by natural selection, genetic drift, and restricted gene flow (Section II.1a). This is a measure of beta diversity. The Genetic Differentiation EBV includes both the number of distinct populations (or evolutionary lineages) and the levels of differentiation between them; both are needed to capture and interpret ‘differentiation’ fully. Genetic differentiation is sometimes also referred to as population structure, and can be organized hierarchically (populations, regions, management units, genetic clusters, etc.). The previously proposed EBV of cultivated ‘varietal or breed diversity’ in reference to domesticated species could fall under the term ‘genetic differentiation’, which can apply to all species. The term ‘genetic connectivity’ typically connotes gene flow (see Section II.1a), which usually reduces genetic differentiation.

Genetic differentiation can be assessed via clustering methods [e.g. STRUCTURE (Pritchard, Stephens & Donnelly, 2000); discriminant analysis of principal components (DAPC) (Jombart, Devillard & Balloux, 2010); see Table S1] that identify the number and genetic composition of groups, or methods which compute summary statistics of the degree of differentiation among a priori defined groups, usually on a scale of 0 to 1 (see Table 1). It is important to note that measures of genetic differentiation are usually based on a priori, or at times a posteriori, stratification of samples representing prospective units. As such, these units can only represent hypotheses of genetic structuring, which may not correspond to all underlying patterns of differentiation. Thus, this EBV is measured under the assumption that

Fig. 2. A sample of four diploid individuals from a population, with various representations of genetic composition data structures. The workflow process includes genetic sequencing, aligning the sequences from each individual, and polymorphic loci identification. The data from the polymorphic sites (single nucleotide polymorphisms; SNPs) in the sequence can be summarized as a matrix of genotypes for each locus (L1–L4). When these loci are bi-allelic SNPs, the data can be summarized as the Minor Allele Count – the number of occurrences of the least frequent allele at that locus, a convenient summary format for certain statistics and models. A Genetic Diversity EBV for evenness, such as observed heterozygosity ($H_o$), can also be summarized for each individual (rows), or by locus (columns), as illustrated for the SNP matrix. Note that some measures of diversity include invariant sites which are calculated from sequence alignments, not a matrix of SNP genotypes.
population units have been appropriately delineated and provides an essential context within which the other three Genetic EBVs are assessed.

Loss of population units is clearly detrimental from a conservation perspective. An increase in differentiation is also typically detrimental from a manager’s perspective, as it is caused by a decrease in population size and/or a decrease in connectivity (e.g. habitat fragmentation or isolation). However, a decrease in differentiation may sometimes be detrimental, for example when isolated populations adapted to contrasting environments are homogenized, or when a managed species hybridizes with wild populations. In other cases, a decrease in differentiation may be the beneficial outcome of targeted genetic rescue (Whiteley et al., 2015), whereby isolated populations are connected through corridors or translocations. Genetic differentiation is relatively easy to calculate, visualize and explain to conservation managers and policy makers. Unfortunately, many measures of genetic differentiation show slow responses to perturbation in some situations (tens of generations; e.g. Landguth et al., 2010), have undesirable statistical properties (Jost et al., 2018), and are difficult to apply in species with continuous distributions or asymmetric gene flow. Alternative individual-based measures can be used as proxies for gene flow (Shirk, Landguth & Cushman, 2017). Moreover, current differentiation levels may already be a result of habitat fragmentation or other human influences.

Highly genetically differentiated populations, or groups of populations, such as adaptively or morphologically distinct varieties or sets of populations, are often called Evolutionary Significant Units (ESUs). These have been segregated for hundreds of generations and are usually defined by a set of genetically based traits or adaptations (Funk et al., 2012). Such independent genetic units may comprise a breed or cultivar in domesticated species, or an ecotype, subspecies, variety, or stock in non-domesticated species. The number of such units is essential, as loss of ESUs equates to loss of a distinct product of evolutionary history such as unique traits and co-adapted sets of traits. In practice, the units are delineated by significant divergence at multiple loci indicative of independent evolutionary trajectories plus some ecological or trait differences (Fraser & Bernatchez, 2001; de Queiroz, 2007). There is increasing effort to identify and conserve ESUs at national and international scales, promoting evolutionary processes (e.g. Gene Conservation Units in European forestry; Koskela et al., 2013).

(4) Inbreeding EBV

Measurements of inbreeding can be used to determine the level of diversity within individuals. Mating between related individuals constitutes inbreeding. Highly inbred individuals often have lower fitness (termed ‘inbreeding depression’), which can compromise population growth and persistence by reducing demographic rates (e.g. decreased fecundity). Inbreeding metrics indicate near-term risks, and can signal that management action, such as translocating individuals, is urgently needed. When populations are large and/or there is sufficient gene flow within and between populations, inbreeding rates tend to be reduced. This EBV was previously called ‘co-ancestry’ (Pereira et al., 2013) but ‘inbreeding’ is more reflective of the process and is an accessible concept for practitioners. Inbreeding is often a consequence of small population size, increasing rapidly when \( N_e \) is less than approximately 50 and in fragmented populations where gene flow dynamics are impacted (e.g. plant pollination; Breed et al., 2015). However, some species can have naturally high inbreeding levels due to their mating system without any apparent fitness loss and some benefit (e.g. reproductive assurance in plants; Winn et al., 2011).

Inbreeding can be estimated using various methods depending on the timescales and type of data available: path analysis on a pedigree; the degree of homozygosity using whole-genome data (e.g. runs of homozygosity); direct observations of inbreeding via the genotyping of progeny arrays; or the inbreeding coefficient, \( F_{IS} \), which provides evidence of non-random mating according to Hardy–Weinberg equilibrium expectations [i.e. excess in observed homozygous genotypes (Kardos, Luikart & Allendorf, 2013; Kardos et al., 2018)]. In addition to these measures of contemporary inbreeding, one can measure risks of inbreeding. Relatedness, kinship and co-ancestry metrics can identify related individuals in a population and quantify the likelihood of inbreeding before it occurs. Estimates of inbreeding have one major advantage over diversity and differentiation: estimates made with a given method are easy to compare across studies, as they should not be dependent on the type of DNA markers used, as long as the data set has sufficient statistical power.

(5) Contemporary effective population size EBV

Effective population size \( (N_e) \) is an established fundamental parameter for genetic biodiversity monitoring. In a population, \( N_e \) quantifies the amount of genetic change that is occurring and will occur in the future as a result of genetic drift. \( N_e \) determines the equilibrium level of genetic diversity that a population can maintain and thus the long-term genetic diversity in a population (Leffler et al., 2012; Ellegren & Galtier, 2016). \( N_e \) is the most important parameter for assessing the ability of populations to adapt to rapid environmental changes and the likelihood of inbreeding depression and other reactions to perturbations (Hoban, 2014; Tenesa et al., 2007). For example, depending on the scenario, other less-sensitive or well-suited metrics, such as \( H_e \), may have minimal decline between sampling timepoints, while \( N_e \) may decline more rapidly and indicate that rapid loss in genetic variation is predicted in the near future.

\( N_e \) is defined as the number of individuals in a theoretically ideal population (e.g. most adults randomly mate and contribute a Poisson-distributed number of offspring every generation) that will experience the same magnitude of genetic diversity loss (due to random genetic drift) as the real population. \( N_e \) determines the influence of genetic drift based on the
amount of genetic diversity – when $N_e$ is small, genetic diversity is lost from a population faster over time and the random fluctuations in allele frequency caused by genetic drift can overrule the effect of natural selection (Charlesworth, 2009). $N_e$ is widely adopted in practical management due to its relationship to both short-term inbreeding and long-term evolutionary potential (Jamieson & Allendorf, 2012).

$N_e$ can be estimated across different spatial and temporal scales (Ryman, Luikre & Hossjer, 2019). Historical $N_e$ is calculated in terms of coalescence, which translates to an average allelic sampling variance across generations due to past genetic drift (and/or sampling variance) (Sjödin et al., 2005). This provides a retrospective view of $N_e$ equally across the entire genealogy, and limits the suitability of historic $N_e$ for the assessment of change as an EBV. Historic $N_e$ could be used to compare the average $N_e$ for a genealogy across different populations or to characterize the diversity across multiple species in a region; however, these are less informative for monitoring diversity, especially as related to change influenced by environmental drivers (Nadachowska-Bryzsyka, Konczal & Babik, 2022).

Alternatively, in natural resource management situations, the more commonly applied $N_e$ and the primary focus of our EBV, is the contemporary $N_e$ (Nadachowska-Bryzsyka et al., 2022). In conservation situations, where recent demographic changes are the rule rather than the exception, an EBV addressing contemporary $N_e$ can inform inferences on future genetic diversity changes. Using contemporary $N_e$, a prospective view of a population’s genetic diversity in $n$ generations is provided.

Two categories exist for contemporary $N_e$ (hereafter $N_e$) estimators using genetic data – those requiring a single-time-point data set [e.g. linkage disequilibrium (LD), kinship-based methods], and those requiring two time points to be sampled, defined as temporal $N_e$ estimates (Do et al., 2014). The two categories can give concordant estimates, and both are influenced by marker and sample sizes (Ackerman et al., 2017; Hoban et al., 2020; Luikart et al., 2021; Waples, 2021). Temporal methods progressively have been replaced by single-time-point methods since the late 2000s (Waples, 2021). In particular, linkage disequilibrium-based $N_e$ estimates can be generated from a single time point and are more robust to violations in model assumptions and more easily comparable across studies (Waples & England, 2011), but are sensitive to population structure (Neel et al., 2013). The different contemporary $N_e$ estimations are complementary to one another and the use of several estimators can help to reduce uncertainty. There are also multiple kinds of $N_e$ that quantify other parameters and apply to different timescales (e.g. inbreeding, variance, etc.). See Waples (2016) and Ryman et al. (2019) for in-depth reviews of the definitions and interpretation of each measure of $N_e$. When populations are stable and at mutation-drift equilibrium, the historical $N_e$ is the same as the contemporary $N_e$ and the latter reflects the allelic sampling variance of the previous generation(s). In all cases, $N_e$ should be estimated within demographic units (metapopulations or populations).

Effective population size is clearly linked to the census population size ($N_c$) and the variance in reproductive success among individuals; the larger this variance, the smaller resultant $N_e$ for a given $N_c$ (Kimura & Crow, 1963). The effective size $N_e$ is often between 1/2 and 1/10 of the $N_c$ but in some cases it is much smaller. By applying the average ratio for vertebrates of approximately $N_e/N_c = 0.1$, $N_e$ can be inferred from genetically informed effective population size (see Hoban et al., 2020). $N_e$ estimation is sensitive to sample size variance and mixed-age populations and requires careful application and interpretation. However, $N_e$ is one of the best-studied and easily explained metrics for applying biodiversity loss thresholds; $N_e$ below approximately 50 will lead to rapid increases in inbreeding, loss of fitness, and genetic composition change; while $N_e$ below approximately 500 will result in loss of ability to adapt via natural selection (Frankham, 2005; Jamieson & Allendorf, 2012; but see also Frankham, Bradshaw & Brook, 2014 for arguments for doubling these thresholds to 100 and 1000, respectively). $N_e$ is useful to assess conservation management needs and to monitor success of recovery programmes, including through indicator metrics and forecasting (Hoban et al., 2020). From an applied perspective, $N_e$ is a measure of population size, which in many cases is easier (and more cost effective) to determine using genetic tools than through traditional survey efforts.

(6) Why four Genetic EBVs?

The four proposed Genetic EBVs cover the components of Wright’s genetic variation (Fig. 3; Jost, 2006; Daly, Baetens & De Baets, 2018) and together provide a comprehensive description of the impacts of environmental change on genetic composition. As with other classes of EBVs, various Genetic EBVs will sometimes all be affected similarly by a driver (environmental or human factors), while other drivers will lead to idiosyncratic changes. This provides some ability to distinguish among drivers, and to elucidate the underlying mechanistic process, by calculating and comparing multiple EBVs. For example, decreases in census size will likely lead to increases in inbreeding (and genetic drift), as well as associated loss in allelic richness and genetic evenness (heterozygosity). Changes in gene flow will lead to changes in genetic differentiation, average relatedness, allelic richness, and heterozygosity, but each on different timescales (Landguth et al., 2010). Inbreeding and allelic richness typically change relatively quickly after the onset of disturbances, compared to genetic evenness (heterozygosity) and differentiation, which respond more slowly (Keyghobadi et al., 2005; Lowe et al., 2005). Note that interpretation of $N_e$ and genetic differentiation can be sensitive to genetic data set sizes and also to levels of breeding between close relatives (i.e. biparental inbreeding) or selfing (primarily in plants). Note also that hybridization rate was considered as a potential EBV but was determined not to fit our criteria (see Appendix S2).
Important notes on Genetic EBVs as biodiversity metrics

Here we highlight several relevant characteristics of Genetic EBVs. Due to the continuously inherited nature of DNA, Genetic EBVs are useful for assessing populations today and also their history at timescales from recent generations to millions of years ago. Geneticists can use DNA from contemporary samples to infer past changes in migration rates or the rate of change in $N_e$ through models (see Section IV.2).

For example, analysis of DNA in the rarest bat in the UK, Plecotus austriacus, showed a recent 30-fold decline in $N_e$ at a median of 200 years ago, corresponding to changes in farming practices and disappearance of unimproved grasslands (Razgour et al., 2013). Also, because DNA can be obtained from preserved remains up to tens of thousands of years old, the past state of genetic composition can be directly assessed. For example, the genome sequence of a single member of the Wrangel Island population of wooly mammoth (Mammuthus primigenius), the last remaining population before the species went extinct, showed 20% less heterozygosity and 28 times more runs of homozygosity (a metric of inbreeding) than a more ancient, healthy population (Palkopoulou et al., 2015). Also, because DNA can be obtained from preserved remains up to tens of thousands of years old, the past state of genetic composition can be directly assessed.

Another unique aspect of genetic data is that a single genetic or genomic data set can be used for calculating many (or all) Genetic EBVs (assuming sufficient sampling of individuals). This is similar to Species EBVs in which individual observations underlie abundance and extent, but dissimilar to Trait EBVs which require distinct observations using various instruments to cover a range of EBVs. For example, Henry et al. (2009) collected DNA from scat of the Amur tiger (Panthera tigris altaica) population and genotyped 95 individuals at 12 microsatellite markers. This data set revealed strong genetic differentiation between two isolated regions, and very low allelic richness, heterozygosity (evenness) and $N_e$ relative to other tiger populations [see Silva et al. (2020) for a plant example using SNPs]. Small sampling efforts, such as genomic sequences from fewer than 10 individuals, can assess the present state of EBVs and their change over time (Diez-del-Molino et al., 2018).

Lastly, most Genetic EBVs are interpreted in a context of change. As with other EBV classes, they typically require samples across multiple points in space and/or time (with representative sampling and comparable data, e.g. similar sampling and analyses). However, some Genetic EBVs also have relevance based on their ‘state value’ in addition to ‘change’. $N_e$ and inbreeding can be calculated from single-time-point samples and have thresholds in their state value that signal vulnerability (Section II.5). For other EBVs, the
state value has meaning if comparing species with similar life histories but different exposure to drivers such as harvest (see discussion of third category in Section IV.1). Thus, to some extent, Genetic EBVs can be used when only single time points are available, although temporal monitoring is needed to determine the rate of change and degree of change from baselines.

III OBTAINING AND ARCHIVING GENETIC COMPOSITION DATA AND METADATA

(1) Obtaining genetic data

The steps in generating Genetic EBVs include collection of biological samples, laboratory work, computational processing of raw genetic data, analysis/calculation, publishing, archiving, modeling and/or synthesis (Fig. 4). Collecting genetic data begins with a biological sample – some remnant of a living or preserved individual, including ‘invasive’ sampling (direct handling of the organism to obtain a sample) or ‘non-invasive’ sampling (from hair, faeces, shed skin, etc.) (Fig. 5). Samples may be obtained from extant populations, biological collections, or sub-fossils (ancient DNA). Molecular biology techniques are used to analyze DNA sequences extracted from the sample. These data can be presented as sequences of DNA nucleotides, the allele ‘states’ at a locus for each individual, and/or as allele frequencies in populations, from which genetic statistics are derived (Fig. 2). DNA storage, extraction, sequencing and quality control are discussed briefly in Appendix S3. Data sets are often considered ‘genetic’ (tens of genetic markers, usually without a reference genome – as for most legacy data sets) or ‘genomic’ (thousands to millions of markers, often with a reference genome – as for many current data sets). Resolution and precision of EBV values are higher for the latter. Genetic composition can be reliably described for both data types, although reliable estimates of relatedness and inbreeding are best measured using genomic data sets. The growing size and complexity of genomic data sets creates the need for a corresponding growth in computer storage and power (i.e. supercomputers) and statistical programs to allow for

Fig. 4. Steps in generating Genetic Composition Essential Biodiversity Variables (EBVs) include field (or archive) collection of DNA, laboratory work, computational processing of raw data, analysis/calculation, publishing, archiving, modeling and/or synthesis and communication to inform management decisions. GEOME, Genomic Observatories MetaDatabase; Pop, population; QA/QC, quality assurance/quality control.
comprehensive analyses. Advancing the necessary supporting components, such as computing infrastructure and bioinformatics, will facilitate data accessibility and application.

As for many biodiversity observations, genetic composition data can be obtained from different types of scientific studies (e.g., assessing the conservation status of important species; testing ecological hypotheses; stock assessments in forestry or fisheries; taxonomy; forensic investigations). The spatial extent of studies is highly variable; sampling often focuses on a specific region but sometimes includes a species’ entire range. It is critical that samples are collected in an observational design that considers the expected data product, such as prediction of possible future change under different scenarios; nonetheless numerous studies use opportunistic and biased sampling. If the goal is to assess temporal trends, sampling effort should be comparable between time points. Most studies are from a single time point, but some are longitudinal. For example, many national agencies and research institutions sample fish stocks, large mammals (e.g., lynx, wolves, bears), and forest trees regularly to monitor genetic richness, genetic differentiation, and $N_e$. One example is the Columbia River Basin in the USA where state, federal, tribal, and non-profit organizations have genetically monitored multiple salmonid species and populations over several decades with yearly funding of hundreds of millions of USD (Hand et al., 2018). This system is one of the most heavily monitored in the world because of socioeconomic and cultural factors coupled with the critical ecological importance of the basin.

Where the study intent is to calculate Genetic EBVs, ideally each step in the data workflow should be guided by standard operating procedures (SOPs) that are feasible, cost-effective and publicly available to produce an ‘EBV-ready data set’ (Kissling et al., 2018) from which indicators can be derived. SOPs will ensure that EBVs are measured in consistent ways, increase data set interoperability and re-

Fig. 5. Sources of DNA for genetic analyses. Genetic material may be obtained directly from tissue samples from extant populations (biopsy or non-invasive), biological collections (e.g. museums), or sub-fossils (sometimes called ancient DNA) or from the environment (i.e. environmental DNA, eDNA). Older samples may have low-quality and low-quantity DNA, restricting the use of certain Genetic Essential Biodiversity Variables (EBVs); eDNA is challenging to use for Genetic EBVs since the DNA is typically of lower quality and quantity. These examples indicate information typically assessed with these types of data, and do not represent all possibilities. $N_e$, effective population size.
use, and expand the uptake of Genetic EBVs for diverse users. Examples of SOPs are: site choice and sampling methodology (i.e. choice of individuals, relevant spatial scales and temporal frequency), environmental and biological observations to collect, methods of tissue preservation, data quality control, software/analysis pipelines, and data archiving and metadata storage to ensure alignment with FAIR principles (Findable, Accessible, Interoperable, and Reproducible). Several communities have developed genetic SOPs. One of the best known is the International Whaling Commission’s assessment of stock structure which, in cooperation with associated scientists, has developed best practices for sampling genetic data, data analysis, reporting genetic variation measures including $N_e$, over time, defining population units, and assessing genetic data quality (IWC, 2018; Waples et al., 2018). Other research initiatives have established field, laboratory, sequencing and analysis SOPs appropriate for their taxonomic group, and/or genetic methodology [e.g. Oz Mammals Genomic Initiative (Eldridge et al., 2020); Ira Moana Project (Liggins, Noble & the Ira Moana Network, 2021)]. The Marine Biodiversity Observation Network (MBON) is working to develop best practices for ‘omics’ information (Goodwin et al., 2019), including the Global Omics Observatory Network (GLOMICON – https://sites.google.com/view/galomicon/home).

(2) Genetic data archiving practices

It is becoming standard practice to deposit genetic data into searchable and open access repositories – this is mandatory in several leading journals (Joint Data Archiving Policy (https://data.dryad.org/docs/JointDataArchivingPolicy.pdf)). Successful re-use and integration of these data sets, such as to calculate Genetic EBVs, depends on the archive, its interoperability, and the metadata requirements (Section III.3). The International Nucleotide Sequence Database Collaboration repositories [INSDC; Cochrane et al., 2016; primarily a consortium of US, European and Japanese national databases, including the National Center for Biotechnology Information (NCBI)] are the largest publicly available nucleotide databases. These databases host raw DNA sequence data in standard, interoperable formats and they can be queried and programmatically accessed (e.g. ‘rentrez’ R package, ‘genomdb’ R package). While standardized and easily accessible, the genetic data deposited into INSDC does not always include the minimum information required to re-create data sets appropriate for the calculation of Genetic EBVs. Archiving practices may also be inconsistent – the deposited DNA sequences are typically only the new, unique sequences derived from the study rather than all observed sequences, and do not provide frequency information that is needed to calculate EBVs reliably [see Section IV (Pope et al., 2015a; Paz-Vinas et al., 2021)]. For processed genetic data, such as microsatellite and SNP data sets, numerous other open access archives are used, including Dryad, FigShare, and GitHub. Taxonomic and region-specific archives also exist (e.g. CartograTree, DIPNet, see Sections III.3 and IV.1), although it is not always easy to find them. Also, these repositories do not always require standard data file formats or have minimum metadata requirements and therefore these data are not easily searched or programmatically accessed.

Most archived genetic data cannot be operationalized yet. Typically, less than 10% of sequences are georeferenced, and most cannot be geolocated based on text descriptions (Gratton et al., 2017; Mirdalo et al., 2016; Toczydlowski et al., 2021). For example, Theodoridis et al. (2020) examined mammal mitochondrial genetic diversity globally, comprising >150,000 mitochondrial DNA sequences from BOLD and GenBank. They attempted to assign geographic coordinates to sequences without precise geolocation data using GeoNames (http://api.geonames.org), but could do so for just 36% of sequences. In short, even for a very well-studied group (mammals), a majority of data could not be used. In addition, most genetic data sets are not directly linked to non-genetic global biodiversity databases (e.g. the Global Biodiversity Information Facility, GBIF), or environmental databases including widely used satellite and in situ data repositories, although links to these can be made through sufficient spatial data (Kissling et al., 2018). In some cases, the exact georeferences of the sampled organism or population may be purposely withheld, to protect the study organism (e.g. an endangered species, or highly desirable species) or at the request of Indigenous peoples for reasons of cultural sensitivity regarding this information. Nonetheless, standardized practices of quality control and assurance when generating data and metadata, and during data archival, would help to avoid issues with compounding errors during data aggregation.

(3) Metadata

Metadata associated with archived samples, DNA, sequences, or other genetic data are essential for enabling re-use of these resources for calculating and interpreting EBVs. Ideally, genomic records should contain environmental information or a link to an environmental database with high-quality information about local conditions of the sampled organism. At the specimen level, metadata may include collection date, taxonomic designation, morphological measurements, latitude/longitude, age, condition/health, spatial scale, and reproductive status. At the site or population level (by convention termed ‘occurrence’ or ‘event’ level) these may include habitat description or ecoregion classification, co-occurring species, sampling methodology, measured environmental variables (light, soil moisture, etc.), and uncertainty in the geolocation. Metadata regarding the laboratory, sequencing, quality control, and bioinformatic steps may also be recorded, and are important for generating comparable genetic data but are difficult to document and reproduce in a standard way to derive the genetic data set from the raw data. As with laboratory protocols, open text fields for detailed description may be needed.

Despite disparate study objectives, the research community has been active in developing standardized metadata.
fields and vocabularies to capture ecological and environmental data for a sample and/or field sampling event, and laboratory and sequencing methodologies. These metadata standards are supported by the Biodiversity Information Standards community (administering the Darwin Core standard: [https://dwc.tdwg.org/; Wienczek et al., 2012]) and the Genomics Standards Consortium ([https://geome-db.org/; Deck et al., 2017]). If widely adopted, standard metadata fields alongside deposited genetic data would ensure that they are uniformly queryable according to their machine readable metadata (using a web interface or programmatically using the ‘geomdb’ R package) and interoperable with other data aggregators (i.e. INSDC, GBIF). To help coordinate the use of a minimum set of mandatory and/or recommended metadata fields, GEOME’s infrastructure enables ‘Teams’ and ‘Projects’ to design their own metadata templates and rules, based on existing fields and vocabularies (including options to maintain the privacy of some fields, such as exact georeferences for aforementioned reasons). This platform helps users coordinate research projects, ensures the interoperability of data sets, allows discovery of data sets, and will help facilitate the calculation of Genetic EBVs.

Although well developed, uptake and use of the standardized metadata fields and vocabularies by the research community will take time as many researchers remain unaware of suggested metadata fields and large-scale environmental data sets to link to their genetic data. Nonetheless, with the support of leading journals (see Sibbett, Rieseberg & Narum, 2020) and pressure from funding agencies to ensure data is searchable, open access and interoperable, adherence to metadata standards and the use of provided infrastructures will increase, helping ensure these data sets can be used to calculate Genetic EBVs. GEO BON will continue to work with the IUCN, the Society for Conservation Biology, and the EU COST Action G-BiKE to promote broader understanding of these practices in the community.

IV Operationalizing EBVs: From Single Studies to Global Use

In order to better operationalize EBVs, GEO BON has adopted the data structure of a multi-dimensional array as a basic format to represent EBVs (Jetz et al., 2019). Each cell contain an observation or model-based output with the dimensions as species–space–time, at a minimum. There is an expectation that eventually such data structures, termed ‘data cubes’, will have regional to global coverage, across critical timescales (yearly, decadal, centuries). As for most biodiversity data, genetic surveys typically sample a limited range of space and timescales, resulting in sparse and spatially biased data structures. This section covers two vital steps towards operational EBVs: how large numbers of genetic data sets can be compiled and analyzed across space and time, and how modeling efforts are helping to fill in the sparse genetic data landscape.

1 Visualizing EBVs in space and time across species

The well-established population genetic theory describing intrinsic drivers of genetic composition, the relatively small number of genetic statistics, and the community commitment to data archiving have led to numerous compilations of genetic data sets and summary statistics at large spatio-temporal scales. Collectively, this effort has paved the way for large-scale studies to describe regional and global patterns of the four proposed Genetic EBVs and their drivers of change. These large-scale macrogenetic studies (Blanchet et al., 2017; Leigh et al., 2021) can be classified into three categories described below (also see Table 2). Each category represents a major step towards integrating observations into an EBV data object and operationalizing EBVs; they also help develop the knowledge and boundary conditions for simulation models and forecasting of genetic diversity (Section IV.2).

The first category of macrogenetic studies involves integrating data sets from multiple species (tens to thousands) spanning large geographic scales (thousands of square kilometres to global coverage) with extensive environmental databases to identify spatial trends or patterns in EBVs (reviewed in Leigh et al., 2021). A study may test for correlations between genetic diversity and macro-ecological correlates such as latitude (Miraldo et al., 2016), assorted environmental variables (Manel et al., 2020), anthropogenic impact (Millette et al., 2020; Schmidt et al., 2020), species diversity (Theodoridis et al., 2020), or life-history traits (De Kort et al., 2021). Such studies may generate multi-species genetic data from scratch (e.g. Taberlet et al., 2012) and/or extract large amounts of data from the literature or from public repositories like GenBank or BOLD (Leigh et al., 2021). For example, Manel et al. (2020) repurposed 50,588 mitochondrial DNA barcode sequences [COI (cytochrome c oxidase subunit I) gene] from 5426 fish species to explore global determinants of freshwater and marine fish species’ Genetic Diversity EBV (Table 2). The results from these studies can be effectively mapped in the spatial dimension of the GEO BON EBV ‘data cube’. However, numerous statistical considerations apply to macrogenetic studies using data extracted from nucleotide databases, such as inconsistent data depositing practices (see Section III and Paz-Vinas et al., 2021). Further, these studies frequently do not assess true ‘population’ samples, but rather aggregated sequences according to spatial proximity or a grid (e.g. Theodoridis et al., 2020; Millette et al., 2020; Manel et al., 2020; see Paz-Vinas et al., 2021). It is also often hard to reproduce these studies’ results because the data sets are extensively modified once they are obtained from archives (Pope et al., 2015b). Macrogenetic studies based on genetic
Table 2. Examples of three broad categories of large-scale studies that establish a foundation to operationalize Essential Biodiversity Variables (EBVs): taxa examined, data source, EBVs compiled, size of data set, and DNA marker type (mtDNA, mitochondrial DNA; microsat, microsatellite; AFLP, amplified fragment length polymorphism). Category 1 = large scale spatial patterns; Category 2 = temporal change; Category 3 = quantitative relationship between driver and Genetic EBV response

<table>
<thead>
<tr>
<th>Study</th>
<th>Category</th>
<th>Taxa</th>
<th>Data source</th>
<th>Genetic EBV</th>
<th>Number of species</th>
<th>Marker type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manel et al. (2020)</td>
<td>1</td>
<td>Freshwater and marine fishes</td>
<td>GenBank</td>
<td>Genetic Diversity: richness</td>
<td>5426</td>
<td>mtDNA</td>
</tr>
<tr>
<td>Lawrence et al. (2019)</td>
<td>1</td>
<td>Terrestrial vertebrates</td>
<td>From literature</td>
<td>Genetic Diversity: richness and evenness; Genetic Differentiation: differentiation between units</td>
<td>897</td>
<td>Microsat</td>
</tr>
<tr>
<td>De Kort et al. (2021)</td>
<td>1</td>
<td>Animals and plants</td>
<td>From literature</td>
<td>Genetic Diversity: evenness</td>
<td>727</td>
<td>Microsat and AFLP</td>
</tr>
<tr>
<td>Leigh et al. (2019)</td>
<td>2</td>
<td>Animals</td>
<td>From literature</td>
<td>Genetic Diversity: richness and evenness</td>
<td>91</td>
<td>Predominantly microsat</td>
</tr>
<tr>
<td>Jordan et al. (2019)</td>
<td>2</td>
<td>Plants</td>
<td>From literature</td>
<td>Genetic Diversity: richness and evenness</td>
<td>48</td>
<td>Various</td>
</tr>
<tr>
<td>Pinsky &amp; Palumbi (2014)</td>
<td>3</td>
<td>Fish</td>
<td>From literature</td>
<td>Genetic Diversity: richness and evenness Inbreeding</td>
<td>140</td>
<td>Microsat</td>
</tr>
<tr>
<td>Breed et al. (2015)</td>
<td>3</td>
<td>Plants</td>
<td>From literature</td>
<td></td>
<td>40</td>
<td>Predominantly microsat</td>
</tr>
</tbody>
</table>

data extracted from the literature (e.g. Lawrence et al., 2019) may be more reliable for representing population-level genetic data. For instance, De Kort et al. (2021) compiled and mapped genetic evenness (heterozygosity) EBV values for 8386 georeferenced populations (with \( N \geq 10 \) sampled individuals) from 727 animal and plant species to explore the effects of life-history traits and temperature stability on genetic diversity. Further, literature-based macrogenetic studies allow for the assessment of critical information such as the sampling effort needed to obtain a given number of alleles or sequences – information that typically cannot be assessed for data extracted from public repositories (Paz-Vinas et al., 2021; Leigh et al., 2021), although metadata stewardship is continuously improving (see Section III.3; Riginos et al., 2020).

A second category of large-scale macrogenetic studies seeks to document temporal change. Here, researchers analyze time-series observations of the same population or region to quantify changes in genetic composition directly. These time series can come from long-term monitoring, or a combination of contemporary samples and archive, museum or sediment (e.g. sub-fossil) specimens. For example, Leigh et al. (2019) compiled data from 91 species and found that genetic diversity in animals (primarily fish, birds, mammals and insects) has declined about 6% globally. Jordan et al. (2019) found decreased genetic diversity in 52% of restoration plantings (e.g. revegetation) compared to their seed source, an effect most strongly associated with seed obtained from few source sites. These results contrast with analyses by Millette et al. (2020) and Lawrence et al. (2019), emphasizing that some large-scale studies may have less power to detect change due to a limited time span, repurposing data from nucleotide databases, challenges in aggregation, and other reasons (reviewed in Leigh et al., 2021; Table 2).

Unfortunately, true time series of observations of the same population or region with similar sampling are rare, making it difficult to construct a data object with a comprehensive time dimension that spans broad timescales. As with other EBV classes, statistical models will be essential to fill in spatial and temporal gaps (Section IV.2).

A third category seeks to test the effects of specific environmental or human impacts via ‘space for time substitution’ or ‘case–control’ (Table 2). This can help to quantify the impact of a specific factor or process on genetic composition. For example, Pinsky & Palumbi (2014) analyzed microsatellite genetic data from 140 species to document that, on average, overharvested fish populations have 12% lower allelic richness genetic diversity than non-harvested fish. Aguilar et al. (2008) used data from 102 plant species to reveal significantly lower genetic richness and evenness in plant populations after habitat fragmentation, especially after more than 100 years of fragmentation. These and other examples (e.g. DiBattista, 2008), provide information for models (e.g. a minimum distance among populations beyond which significant genetic differences arise; Durrant et al., 2014) that can help populate cells in an EBV data cube, as explained in Section IV.2.

Such efforts at post-hoc data aggregation and mobilization are complemented by coordinated research networks where scientists are compiling their population genetic data sets into curated, standardized archives, supplementary to the general nucleotide databases like GenBank. This involves labour-intensive retroactive application of archiving standards and requirements. For example, the Diversity of the Indo-Pacific Network project compiled more than 200 data sets for Indo-Pacific marine organisms (http://diversityindopacific.net/; Crandall et al., 2019). Similar examples include: Ira Moana
(https://sites.massey.ac.nz/iramoana/), FishGen (www.fishgen.net) and CartograTree (https://treegenesdb.org/ct). Other efforts are planning genetic assessment of many species as a part of a single project, and simultaneously co-sampling them. For example, the IntraBioDiv Consortium sampled 27 alpine plant species with the same strategy, in the same locations, with a common analysis method (https://www.wsl.ch/en/projects/intrabioldiv.html); this data set has been used for many studies (e.g. Taberlet et al., 2012; Hanson et al., 2017). The Oz Mammals Genomics Initiative (https://ozmammalsgenomics.com/) is using standardized methods to generate conservation genomic data sets for 14 threatened Australian mammals (Eldridge et al., 2020; von Takach et al., 2021). All of these projects show that mining databases and the literature, and greater coordination of planning for data interoperability and FAIR practices, has great value for understanding and operationalizing Genetic EBVs at large scales. International groups such as GEO BON are working to network among these groups and share data standards, lessons learned, and best practices for operationalizing EBVs (https://geobon.org/ebvs/working-groups/genetic-composition/).

(2) Predicting Genetic EBVs beyond current observations via modeling

To be spatially comprehensive, EBVs in all classes must rely on a combination of observational and modeled data. For instance, genetic composition data are usually collected from a subset of individuals or locations, which necessitates accurate, spatially predictive models for EBVs at unsampled locations or time points. One simple form of genetic data modeling is spatial interpolation to identify patterns such as clines and hot spots across numerous taxa (Murphy et al., 2008; Paz-Vinas et al., 2018). At large scales, interpolation has been used to provide a global ‘picture’ of genetic status (Vandergast et al., 2011; Miraldo et al., 2016).

Another form of genetic data modeling aims at establishing a statistical relationship between Genetic EBV measurements and specific environmental variables including climate, habitat and biotic interactions (Hand et al., 2015; Jaffé et al., 2019; Carvalho et al., 2019, 2020) or EBVs in other classes like species diversity (Taberlet et al., 2012; Manel et al., 2020). The simplest approach is to determine a relationship between genetic differentiation and Euclidean distance, known as isolation by distance or genetic spatial autocorrelation. This can help infer a minimum distance beyond which populations have significant genetic differentiation (Durrant et al., 2014, Silva et al., 2020). More complex genetic–environmental relationships (e.g. landscape genetics) can identify potential drivers and correlates of change in genetic variation. These analyses project genetic variation into unsampled locations for which data on the driver are available, using the driver as a surrogate of genetic variation (Wasserman et al., 2012; Hanson et al., 2017). This has practical utility for highlighting migration corridors to limit genetic erosion (DiLeo & Wagner, 2016; Monteiro et al., 2019), identifying priority conservation areas aimed at protecting within-species genetic composition [e.g. conservation planning (Paz-Vinas et al., 2018; Hanson et al., 2020)], or modeling population vulnerability and changes in adaptive capacity due to environmental change (Fitzpatrick & Keller, 2015; Martins et al., 2018; Razgour et al., 2018; Vranken et al., 2021). For example, Bay et al. (2018) identified ‘genomic vulnerability’ to climate change in populations of the North American yellow warbler (Setophaga petechia) based on the mismatch between current genomic variation in climate-adapted loci and predicted future conditions.

Modeling approaches have also been used to combine genetic data with species distribution models. Within-species genetic lineages (the Genetic Differentiation EBV) are increasingly used to improve ecological niche model (ENM) predictions of species distributions. ENMs constructed for within-species genetic units offer a more accurate picture of past and present spatial distribution of a species (e.g. Shinneman et al., 2016), as well as future distributions given adaptive variation and potential climatic changes (e.g. Maguire et al., 2018; Razgour et al., 2019), which helps assess vulnerability of a species and possible changes to genetic composition. Meanwhile, ENMs are also used to assess statistical relationships between environmental suitability and heterozygosity, which could be used to map predicted genetic diversity across continuous space (Diniz-Filho et al., 2015).

One remarkable form of genetic modeling is to infer EBV values in the past. Because an individual genome contains information on its ‘ancestral’ genomes, numerous well-developed analytical methods can be used to infer $N_e$ at given time points in the past [e.g. skyline plots, approximate Bayesian computation (ABC)], Fairwise Sequentially Markovian Coalescent (PSMC) models] (Mather, Traves & Ho, 2020). For example, Nater et al. (2015) used modern DNA and ABC to trace the demographic history of orangutan populations, identifying severe $N_e$ declines in response to changes in the extent of rainforest habitats and hunting by early humans. These methods can help establish pre-impact baselines and natural levels of variability in Genetic EBVs, populate the three-dimensional EBV object (e.g. ‘data cube’), and better quantify the impact of drivers through time.

Methods are also well developed for forecasting Genetic EBVs, such as changes in richness due to future habitat change (Alsos et al., 2016). Established population genetic theory (Section II.1a), allows for demographic–genetic simulations (see Hoban, 2014) to assess genetic composition change under drivers such as management strategies (e.g. promoting gene flow among populations, harvest, minimizing mortality in specific regions). Such models can also evaluate the efficiency of different monitoring programs and rank EBVs for sensitivity (Hoban et al., 2014).

It is true that genetic data are currently scattered spatiotemporally and collected in non-uniform ways that can result in bias, as is true of many ecological data sets and all classes of EBVs [see review on impacts of sampling biases in ecological data sets by Hughes et al. (2021); Boakes et al., 2010; Fithian

*Biological Reviews* 97 (2022) 1511–1538 © 2022 The Authors. *Biological Reviews* published by John Wiley & Sons Ltd on behalf of Cambridge Philosophical Society. This article has been contributed to by U.S. Government employees and their work is in the public domain in the USA.
et al., 2015; Hortal et al., 2013; Jetz et al., 2019; Meyer et al., 2016]. Although genetic data are relatively sparse across landscapes, the aforementioned approaches are filling gaps in genetic information and predicting EBVs spatially and temporally, and across multiple species. Further, the development and uptake of these proposed EBVs could improve global standardized data collection, including use of archive repositories for genetic data and uniform metadata. The existing limitations will continue to be overcome through combining additional supporting measures, such as through statistical, modeling, bioinformatic and software guidelines. To help guide and advance genetic data collection for restoring and conserving biodiversity, guidance is being developed to encourage efficient progress and rapid uptake of genomics for spatial conservation planning (e.g. Kershaw et al., 2021) and natural resource and functional ecosystem restoration (e.g. Mohr et al., 2022). A strong focus on education, training and outreach of practitioners and decision makers, especially in relation to operationalizing EBVs, is also needed (Wasserman et al., 2010; Brown et al., 2016).

V FUTURE DIRECTIONS FOR GENETIC EBVs

(1) What is needed for Genetic EBV operationalization?

A number of challenges exist for Genetic EBV operationalization. These include standardizing bioinformatic pipelines, standardizing sampling for adequate characterization of each EBV, cost, scaling issues, and choosing target species. As Navarro et al. (2017, p. 161) explain, the EBV framework as a whole also ‘faces challenges emerging from the lack of global monitoring schemes, the integration of datasets resulting from different collection methods, and technical issues related to data product structure, storage, workflow execution, and legal interoperability’. In Genetic EBVs, different spatial and temporal resolution and measurement units need to be harmonized and standardized to allow for integration of data sets from different sources.

For greater adoption of Genetic EBVs by the community, it is critical to demonstrate their value (as we attempt to achieve here), and provide education, training, outreach, and guidance, and also develop necessary supporting components, such as bioinformatics and computing infrastructure. Further, clear protocols and decision trees providing guidance on particular methods or software to use for a given problem, question, or data set (while being mindful of their caveats) could improve uptake. Last, adequate funding to support these goals can allow for increased and sustainable utilization. We and others are continuing to make progress in these areas and are committed to focus on these goals. With these in place, the likelihood of Genetic EBV application should increase.

(2) Challenges: scale, standardization and costs

There is a general lack of systematic global biodiversity data collection and monitoring. Similar to populations and traits, monitoring of genetic change will require combining data sets collected from different species in different locations across time and with different methods – a substantial statistical challenge (see Section IV). Comparisons of Genetic EBVs across taxa will also need to account for correlations with phylogenetic or trait variation (Romiguier et al., 2014).

Just as species or traits may be observed through a variety of measurements, each with its own error rates and biases, genetic information is measured with different markers such as microsatellites, SNPs, or DNA sequences, and there is a need for harmonization across these measurements, particularly to incorporate temporal change. For example, it is not currently possible to compare Genetic EBV values from microsatellites to values from SNPs, without data on both marker types for the same samples. Looking forward, most data sets will soon be sequence-based (e.g. candidate gene resequencing, whole-genome resequencing, reduced-representation sequencing) which are likely to be more robust for long-term monitoring as the data are forward-compatible (see Galla et al., 2019; Wright et al., 2019). Additionally, standardized quality control and assurance practices could ensure that the large volumes of data and metadata harmonized for EBVs do not lead to ‘small sources of systematic or random error [that] can cause spurious results or obscure real effects’ (Laurie et al., 2010, p. 591).

Another challenge is a lack of standardized minimum sampling size and spatiotemporal sampling guidelines (although a single strategy is impossible as it must adapt depending on organisms and locations), although some recognized sampling rules-of-thumb have been established and validated [e.g. Hale, Burg & Steeves (2012) for microsatellites; Nazarenko et al. (2017) for SNPs]. A larger challenge is likely a current lack of standard bioinformatic pipelines and decision-making guidelines, e.g. standard thresholds for filtering genetic data. Genetic differentiation and genetic diversity in reduced-representation sequencing data sets can be affected by bioinformatic processing such as SNP calling (e.g. Wright et al., 2019; Graham et al., 2020). SNP calling will depend on whether a reference-based versus de novo approach is used, with reference-based approaches being preferred when possible (Shafer et al., 2017; Luikart et al., 2018). The rapidly decreasing cost of whole-genome resequencing will eventually diminish these problems, however. While there will not be a single analytical pipeline that will be appropriate for calculating all Genetic EBVs for all species, having a set of standard methods that are agreed upon, helpful guides for key choices, and standards and templates for documenting bioinformatic pipelines (with well-understood values and rules), will aid with transparency, repeatability, and integration (Tahsin et al., 2016; Gratton et al., 2017, Duruz et al., 2019).

A final challenge is the high cost of gathering and processing genetic data. DNA extraction and sequencing are expensive (e. US$10–50 per sample for laboratory supplies...
and US$10–100 per sample for personnel time) and require specialized equipment and personnel training. Funding and opportunities for laboratory work and data analysis are limited or absent in some regions, including in many countries within critical biodiversity hotspots. An additional challenge is choosing appropriate representative species and ecosystems to monitor with genetic data. The best choice will be species for which a change in genetic variation serves as a signal of change in other species (indicator species). When efforts and funding are constrained, we suggest that a few representatives from each group and ecosystem be chosen, including foundation/keystone species. Species selection should also consider the amount of pre-existing ‘genomic resources’, such as annotated reference genomes for the genus. Species can also be selected for genetic monitoring based on the representation of key areas: conservation priority, cultural importance, ecosystem services, wild harvesting, and important game species (as in Hollingsworth et al., 2020).

The provisioning of sample access, and uptake and application of EBVs, will be determined by conservation practitioners, including Indigenous peoples as stewards of the natural environment and genetic resources. The Nagoya Protocol (https://www.cbd.int/abs/) to the CBD acknowledges the link of biodiversity conservation with Indigenous peoples. In many nations, Indigenous peoples will be responsible for authorizing access to genetic samples and derived data. Geneticists need to work closely with Indigenous communities (McCarty et al., 2021) to ensure sharing of benefits from research activity and aligning research with Indigenous worldviews, such as by following the CARE Principles [Collective Benefit, Authority to Control, Responsibility and Ethics (https://www.gida-global.org/care; Liggins, Hudson & Anderson, 2021a); Carroll et al., 2020].

(3) Examples of integrating Genetic EBVs into legislation or policy at large scales

National conservation laws and policies, such as the U.S. Endangered Species Act (ESA), seek to preserve biodiversity through the conservation of individual species ( Drozdowski, 1995; Waples et al., 2013). Under the ESA, those species, subspecies, or ‘distinct population segments’ (DPSs) are listed when they are determined to be ‘endangered’ (in danger of extinction), or ‘threatened’ (likely to become endangered in the foreseeable future). Although these policies are often criticized for the lack of quantitative metrics in their language for guidance (Waples et al., 2013), inferences from genetic analyses are frequently utilized to identify units to conserve, determine the level at which they should be listed, and evaluate listed species for delisting (Fallon, 2007; Funk et al., 2019). Although there have been efforts to streamline this process and the listing criteria, genetic data are used inconsistently (Kelly, 2010; Coates, Byrne & Moritz, 2018). The establishment of internationally recognized and easily understood Genetic EBVs will provide quantitative metrics making ESA listing more objective and transparent. Globally, Key Biodiversity Areas now integrate genetic diversity and differentiation in their declaration (IUCN, 2016).

Another national initiative is Scotland’s pragmatic assessment of genetic composition in representative species via ‘scorecards’. The scorecards use structured expert reviews to assess several categories of risks: loss of diversity (through population declines, loss of functional variation and loss of divergent lineages); maladaptive hybridization; and low reproduction or constraints on adaptation. The scorecards also assess ex situ conservation representation (zoos, seed banks, botanic gardens) and current conservation actions. For each species, there is a rating of risk (along with an evaluation of uncertainty) and of the effectiveness of mitigation. A similar scorecard approach was used in Beckman et al. (2019) which focused on all U.S. species of Quercus (oaks), an emblematic, well-known genus of ecological, socioeconomic and cultural importance. Scorecard approaches, including genetic issues and genetic data, could easily be adapted to include Genetic EBVs.

The European Union (EU) Habitats Directive is the centerpiece of biodiversity conservation legislation in the Union. It focuses both on the preservation of habitats and species of conservation concern and requires each member state to define and pursue favorable reference conditions for habitats and species. Since 2012, the directive explicitly mentions that the criteria for a favorable reference population should include genetic information (population structure, genetic connectivity, etc.), and that this favorable reference population is by definition larger than the minimum viable population size (Evans & Arvela, 2011; Laikre et al., 2009). However, general guidelines on how populations should be defined are not provided and are interpreted differently by each member state. Several member states are working to generally implement genetic criteria in their evaluations (Mergay, 2012), including the $N_e$ EBV. Other EU initiatives that incorporate genetic approaches, and could benefit from EBVs include the UK Strategy for Forest Genetics Resources (Trivedi et al., 2019) and EUFORGEN (de Vries et al., 2015).

(4) Future directions and opportunities

We have focused on the use of genetic data sets, and models that build on such data sets, for EBVs. Nonetheless, large parts of the world lack genetic data observations, necessitating a discussion of proxies or substitutes for genetic data. Suitable proxies could also overcome challenges surrounding the cost of obtaining genetic data. While there are currently no proxies that can fully replace genetic sampling to obtain Genetic EBVs, some show promise, including geographic range loss, decrease in number of adult individuals, or fragmentation rates (see Section IV.2). Decades of theoretical work and hundreds of studies have shown strong genetic composition responses to these proxies as they are connected to basic genetic processes (Sections II.1a and IV). For instance, Khoury et al. (2019) advocate that safeguarding genetic diversity can be based on the percentage of geographic range found in protected areas (in situ) or represented
in seed banks (ex situ). However, genetic composition responses are mediated through species- and situation-specific factors, and responses are highly non-linear, such that this relationship may be only weakly predictive (Alós et al., 2012). Future research effort is needed towards establishing statistical relationships with proxies, identifying how species traits mediate these relationships, and developing statistical frameworks and models (see Section IV.2). Simulations can test how well proxies relate to genetic diversity, i.e. the simulation of loss of range size. Another potential proxy is pedigree data or studbooks, which allow direct inference of $N_e$, relatedness/inbreeding, and loss of genetic evenness (heterozygosity) via available software (e.g. PMx), although this is only applicable to domesticated species, zoo populations, and a few highly monitored natural populations.

The genetic data obtained for Genetic EBVs can also provide information useful to other EBV classes (Gurgel et al., 2020). Genetic data can be used to infer population size (either by mark–recapture or via $N_e$) for the Species Abundance EBV in the Species Populations class, or to delimit species occurrence (Species Population or Community Composition EBVs) through eDNA sampling (see Section I.3). Coordination of sampling is another useful synergy; genetic sample collection could occur concurrently with Species Populations and Traits EBV data collection to facilitate cost-effectiveness and allow cohesive assessments across multiple EBVs, an important consideration for Biodiversity Observation Networks (BONs). There are opportunities across EBV classes to coordinate development of standardized sampling protocols, data-storage structures, and modeling and harmonization techniques. Lastly, we have little knowledge of how Genetic EBVs and other EBV classes co-vary in response to environmental change (although see Razgour et al., 2019), including whether EBV classes exhibit related changes to response or response time (e.g. lags in response). Indicators or other high-level reporting metrics could also integrate multiple classes to illustrate a comprehensive picture of biodiversity.

The final phase of EBV operationalization is policy implementation. EBVs have great potential to form the basis of indicators for policy (Navarro et al., 2017). Unfortunately, currently recommended CBD indicators do not track within-species genetic variation and are not built upon Genetic EBVs (Appendix S4; Laïkre et al., 2020; Hoban et al., 2020). Still, there are emerging examples of Genetic EBVs or similar concepts in use in policy frameworks that may inform and engender further efforts across the globe (see Section V.3).

Genetic and genomic methods continue to advance rapidly, creating opportunities for refining Genetic EBVs in the future, particularly focused on quantifying uncertainties and creating open data frameworks. Decreasing costs and ongoing research efforts are rapidly increasing the amount and quality of reference genome data available, and knowledge of genomic processes, genes and their functions. Long-read sequencing technologies like PacBio and Oxford Nanopore technology will soon allow monitoring via whole-genome sequencing for species of small to moderate genome size. To facilitate such studies, tissue samples should be stored in biorepositories, e.g. biobanks, for use in future advanced analyses. Developments in transcriptomics and metabonomics will lead to advances in mechanistic models and prediction of specific genomic responses to environmental change. Technological advances such as synthetic biology [e.g. CRISPR (clustered regularly interspaced short palindromic repeats) modification] may lead to transformative ways to manage and conserve genetic variation. Additionally, rapid sequencing technology for use ‘in the field’ promises to allow near-immediate genetic data collection for urgent questions (e.g. for prioritizing harvest from specific source populations in fish and in wildlife management). Rapid, ‘pocket-size’ genomic sequencers such as the Nanopore cost less than US$1000 and plug into a laptop. Although genetic composition is not currently amenable to remote-sensing technologies such as satellites, autonomous collection via in situ recording infrastructure (such as the U.S. National Ecological Observatory Network, NEON), hair traps, or automated sampling devices may soon allow DNA capture, sequencing and analysis in the field (Turon et al., 2020). Drones with increased autonomy allow sampling in elusive species or inaccessible sites (e.g. cliffs, oceans), and regularly repeated sampling. Advances in eDNA analysis are also anticipated, although due to eDNA quantities and degradation, it might not be possible to use eDNA for all EBVs (Section I.3).

Technology for collecting environmental data that are necessary for interpreting and modeling genetic data are also advancing. ‘Biologging’, a technique for live registration of physiological parameters, coupled with epigenetics, will help to determine the importance of genetic variants for individual phenotypes and ecosystem functions (e.g. Tree Talker© is a micro-device for measuring water transport in trees, diametrical growth, and spectral characteristics; Valentini et al., 2019). Coupling these data with Genetic EBVs will help to determine underlying causal factors behind genetic, species, and ecosystem diversity, enable us to predict future responses, and help to identify adaptive genetic variants that can respond to climate and environmental change.

VI CONCLUSIONS

Here we summarize our review and describe concrete steps for advancing operationalization of Genetic EBVs.

1. We propose four Genetic EBVs: (i) Genetic Diversity, (ii) Genetic Differentiation, (iii) Inbreeding, and (iv) Effective Population Size. These EBVs offer a viable means of monitoring the essential characteristics of genetic composition, genetic health, and genetic viability of wild and domesticated populations and species, at regional to global scales.

2. The technical and analytical foundation of these Genetic EBVs is well developed, and workflows for...
archiving, obtaining and analyzing genetic data are maturing. There are highly standardized data structures and file formats for storage, and well-developed open-access pipelines for calculating standard statistics underlying the EBVs. Numerous statistical and computational developments have led to highly advanced models for inferring and predicting genetic change, which are complemented by increasingly large-scale genetic data set aggregation.

(3) Despite the robust underpinnings of these Genetic EBVs, additional developments and research efforts are required. Specifically, progress towards increased standardization and shared resources will lead to routine, robust uptake and operationalization of Genetic EBVs in biodiversity assessments and policy. The next steps are:

(a) Improved and upscaled monitoring of temporal changes in genetic composition. Standardized temporal measures of genetic composition remain scarce due to limited longitudinal genetic monitoring, biases in location and taxonomy of the available data, and still unsolved statistical challenges in aggregating and harmonizing existing genetic data sets in large-scale analyses.

(b) Standardization of genetic methods. Standard laboratory and informatic methodologies, and guidelines for consistent, routine archiving and reuse of genetic data, still require improvements to make all genetic data interoperable.

(c) Better recording and use of genetic metadata. There are broad opportunities for increased use of genetic archival databases and biological and environmental metadatabases, and for better links between genetic and environmental repositories and observatories to improve data searching and inferences regarding the drivers of changes in Genetic EBVs. There is a need to develop better proxies in the absence of genetic data.

(d) Development of genetic resources and technology. Further work towards development of key genetic resources (reference genomes), especially for IUCN threatened species and indicator species and those that underpin biodiversity (foundation species), as well as increased deployment of genetic technologies to additional countries and ecosystems, will greatly aid Genetic EBV calculations.

(4) Genetic EBVs capture the genetic status and trajectory of life on Earth. Progress is being made towards integration of genetic variation observations into major international regulations and legislation. International initiatives are working towards a framework for monitoring Genetic EBVs globally, including GEO BON, the IUCN Conservation Genetics Specialist Group and the EU COST Action G-BiKE.

(5) The EBVs proposed here are centered on genetic population processes and will remain relevant, even as innovative calculation methods are developed and with advances in genomic technologies. Implementation of our recommendations for data curation, sharing and integration, alongside increasingly detailed assessments of Genetic EBVs, can lead to robust assessment of genetic status and improved conservation and management of the world’s biomes into the future.

VII. ACKNOWLEDGEMENTS

We would like to thank the organizations and groups that have contributed to this project, including NIMBioS, GEO BON, G-BiKE, and IUCN. We would like to acknowledge John Robinson and Sarah Pearson for their early input to this review. This article is based on work from COST Action G-BiKE, CA 18134, supported by COST (European Cooperation in Science and Technology), www.cost.eu. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the United States Government.

VIII. REFERENCES


EBVs to monitor global genetic biodiversity


climate-associated genetic variation in Mexican populations of Quercus Rubra. Evolutionary Applications 11, 1842–1858.


Morton, J. J., Hoppe, J. & Breed, M. F. (2022). Is the genomics “cart” before the restoration ecology “horse”? Insights from qualitative interviews and trends from the literature. bioRxiv 2021.08.05.453206. https://doi.org/10.1101/2021.08.05.453206


IX. Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1. Workflow for genetic data processing with examples of common, standard file formats and software used to manipulate and edit files.

Table S1. Examples of common, standard analysis methods and software to measure Essential Biodiversity Variables (EBVs).

Appendix S1. Additional discussion on ranking the scalability of EBVs.

Appendix S2. Consideration of Hybridization as an EBV.

Appendix S3. Obtaining genetic data.

Appendix S4. Brief discussion of indicators.

(Received 23 October 2020; revised 25 February 2022; accepted 2 March 2022; published online 12 April 2022)