How genomics can help biodiversity conservation

The European Reference Genome Atlas Consortium

Published in:
Trends in Genetics

DOI:
10.1016/j.tig.2023.01.005

Publication date:
2023

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
How genomics can help biodiversity conservation


The availability of public genomic resources can greatly assist biodiversity assessment, conservation, and restoration efforts by providing evidence for scientifically informed management decisions. Here we survey the main approaches and applications in biodiversity and conservation genomics, considering practical factors, such as cost, time, prerequisite skills, and current shortcomings of applications. Most approaches perform best in combination with reference genomes from the target species or closely related species. We review case studies to illustrate how reference genomes can facilitate biodiversity research and conservation across the tree of life. We conclude that the time is ripe to view reference genomes as fundamental resources and to integrate their use as a best practice in conservation genomics.

The value of integrating genomics into conservation

We are in the midst of the sixth mass extinction, a biodiversity crisis with devastating consequences on ecosystem functioning and health, evolutionary heritage, and the adaptive potential of species, ultimately posing a major threat to humanity [1,2]. Although genetic diversity has long been recognized as fundamental to all levels of biological organization (individuals, populations, species, communities, and ecosystems), genomics is often neglected in biodiversity assessments and conservation efforts [3]. This gap between our ability to generate data and study genomic diversity and the use of genomics in conservation may be due to several factors [4]. These include prioritizing limited funds to address anthropogenic threats [5], limited knowledge transfer and collaboration between genomics and conservation practitioners, stakeholders, and politicians [6], and a general lack of explicit references to measuring, monitoring, and preserving genetic diversity in the most relevant international regulations, such as those issued by the Convention for Biological Diversity [7] (although the situation may have recently started to change, https://www.cbd.int/article/cop15-cbd-press-release-final-19dec2022).

Biodiversity preservation critically depends on addressing key conservation issues. These include taxonomic identification and biodiversity monitoring associated with ecosystem protection and restoration (e.g., for invasive species management). At the same time, human activities exert significant demographic pressures on habitats and endangered species. This requires managing small populations, restoring and increasing genetic diversity of target species/populations, and supporting species adaptation to a changing environment. Genomic data can help tackle these issues as they allow us to characterize and monitor genetic diversity through a wide array of

Highlights

Genomics provides effective tools to characterize biodiversity, but the full implementation of genomic techniques in practical conservation is still limited. We review some of the main approaches in biodiversity genomics available to conservationists and genomicists.

High-quality, long-read sequencing and bioinformatic technologies facilitate genome sequencing and assembly for any species. We summarize how reference genomes, in conjunction with population genomic data, can contribute to biodiversity monitoring, conservation, and restoration efforts.

Over the past decade, many initiatives to generate reference genomes spanning the tree of life have emerged worldwide. We call for increased integration of reference genomes and population genomics data into interdisciplinary conservation efforts to fully unlock the potential of genomics in safeguarding global biodiversity.
emerging tools. The novel insights that can be obtained from genomic data have led to the formation of several national and international initiatives aiming to expand the genomic resources available for non-model species (Table S1 in the supplemental information online). At the same time, scientists and practitioners are collaborating to: (i) standardize protocols for detecting and monitoring species genetic diversity and their adaptive potential, and (ii) integrate genetic and evolutionary knowledge into conservation planning [3]. These actions are critical to promoting transboundary management to ensure the persistence of populations and species and, ultimately, the continued provision of nature-based ecosystem services. Recently, in a forum article [9], we argued for a new era in conservation genomics underpinned by reference genomes (see Glossary), also taking advantage of the increasing ease and widespread interest in generating them. Here, we expand this perspective in more detail, illustrated with recent examples. We also provide an overview on how genomics can help biodiversity conservation, which should be helpful to both conservationists who are not experts in genomics and to genomists without a conservation biology background interested in engaging with conservation genomics research.

Genomic approaches in biodiversity research

High-throughput genomic sequencing technologies have lately evolved from mainly generating short (50–300 bp) to much longer (>10 000 bp) DNA sequencing reads. Genetic and genomic approaches commonly used in biodiversity research include DNA barcoding/metabarcoding, reduced representation DNA techniques, transcriptome sequencing (RNA-Seq), and whole-genome (re-)sequencing (Figure 1). Successful execution of these approaches depends on the quantity and quality of the available biological material, laboratory and bioinformatic skills, feasibility and costs, and on the quality and completeness of the available reference genome database.

We review these common approaches, including their specific merits and limitations (Table 1). Since the quality of the starting biological material is a major consideration in applying genomic approaches, we also review approaches that use noninvasive sampling.

DNA barcoding and metabarcoding

DNA barcoding has become a standard, efficient genetic approach for species identification and biodiversity monitoring [10]. DNA barcoding sequences informative DNA loci with universal or taxon-specific primers that anneal to conserved flanking regions. Initially, DNA barcoding was based on the mitochondrial cytochrome c oxidase subunit I (COI) gene and mainly focused on animals, particularly invertebrates. Over the years, additional DNA loci have been utilized to barcode vertebrates (e.g., 12S, 16S, and Cytb in mtDNA), plants (rbcL and matK in cpDNA), fungi (ITS in rDNA), protists and nematodes (18S in rDNA), and bacteria (16S in rDNA) [10]. DNA metabarcoding combines the principles of DNA barcoding with next-generation sequencing (NGS), enabling the analysis of complex samples containing a mixture of specimens and/or species [11]. Metabarcoding has been widely used in biodiversity assessment and monitoring (e.g., species turnover during ecosystem restoration [12], mapping of ecological networks [13], or detection of invasive species [14]). A key advantage of metabarcoding is bulk sampling and sequencing, circumventing costly sorting and processing of samples into individual specimens (Table 1), thereby enabling high-throughput ecosystem-wide assessments and monitoring in most environments [15]. Moreover, metabarcoding is an appropriate approach to sequence environmental DNA, often degraded into short fragments in an environmental medium [16]. However, the short length of the DNA regions targeted in barcoding and metabarcoding can often limit accurate characterization of the genetic and taxonomic diversity in a community, failing to discern closely related taxa or taxa with introgressed nuclear genes or organellar genomes. Genome skimming circumvents some of the experimental biases in metabarcoding, potentially allowing more accurate metagenomic estimates of biodiversity [17,18] and wildlife forensic investigations [19]. Genome skimming has been conducted in situ, enabling rapid, field-based
assessments of non-model organisms (e.g., [20]) (Table 1). Use of genome skimming can also provide increased information content compared with barcoding and can be applied for sequencing of challenging specimens, such as museum specimens whose DNA may be too degraded and challenging for Sanger sequencing [21].

Reduced genomic representation
Reduced representation DNA sequencing (RRS) approaches are predominant in studies aimed at non-model species [22]. RRS approaches provide genome-wide data in large sample sizes at a comparably reduced cost (Table 1). A small fraction of the genome is reproducibly targeted in each specimen, either using restriction endonucleases and size selection (e.g., RADseq and related methods [23]) or captured by hybridization using probes (e.g., ultra-conserved elements or custom baits) or a combination of both (e.g., HyRAD [24], Rapture [25]). Although RRS approaches only capture a small portion of the genome, they provide sufficient genome-wide data to estimate genetic diversity, inbreeding, effective population size, population structure and assignment, gene flow, phylogeographic patterns, and phylogenetic relationships [22,26]. However, combining RRS-derived data from different studies necessitates identical experimental protocols, which limits replicability. Moreover, RRS approaches utilizing restriction endonucleases (and PCR) may be subject to allelic dropout [23]. Although RRS approaches can be conducted without a reference genome, alignment to a reference genome improves inferences obtained from RRS data [27]. Furthermore, a reference genome provides genome coordinates for most SNPs, thereby facilitating the identification of linked loci, which is key to many population genetic inferences [28].

Gene expression
Gene expression data (usually RNA-Seq) have given rise to a new conservation framework by characterizing genetic variation in natural populations through functional variation [29] and rapid responses of individuals or populations to environmental change [30]. Differences in gene expression have been linked to life history traits and population dynamics, aiding in the identification of candidate genes potentially affecting eco-evolutionary processes [31]. Gene expression data have provided insights into responses to pesticide exposure [32] and susceptibility or resistance to diseases [33]. Gene expression data have also been used to predict range shifts and identify vulnerable populations or adaptive phenotypes [34]. Further, gene expression data have been used to identify entire gene networks in the absence of prior knowledge of the genes involved [35] (Table 1). Analyzing gene expression data is more challenging than DNA-based analyses, because RNA is more susceptible to degradation than DNA and because transcription varies across cell and tissue types, sex, age, physiological and life stage, and even according to circadian rhythms. Consequently, gene expression studies require careful consideration of confounding biological factors [36]. As for RRS approaches, gene expression data are most informative when aligned against a reference genome [37], which provides an independent functional annotation to support the assignment of reads to genes [38,39].

In parallel with genomic assessments of DNA and RNA, epigenomics is being increasingly applied in biodiversity conservation research [40]. Epigenetic modifications may play a key role in phenotypic and biological innovations during major ecological transitions, thus potentially enabling rapid evolutionary responses [41,42]. Identifying the epigenetic modifications that modulate phenotypic variation can therefore potentially identify species that are vulnerable to environmental changes and be informative for monitoring the evolution of invasive species [43]. These approaches rely entirely on the availability of a reference genome to identify epigenetic signatures.

Whole-genome sequencing
Whole-genome sequencing (WGS) data offer unparalleled power and resolution in analyses of demographic history, admixture and introgression, recombination and linkage disequilibrium,
genetic load, natural selection, and species diversification [44–47]. Elucidating evolutionary processes may require analyses of linkage disequilibrium between many physically distant loci [48]. WGS data enable detection of genomic regions under selection, mutations in regulatory elements, rare variants and structural variation, and investigating the genetic basis and architecture of phenotypic traits (e.g., disease susceptibility/resistance in endangered species) [49]. WGS data can also provide unique insights into phylogenetic relationships, evolutionary history, extinct and cryptic taxa, and ancestral intraspecific genetic diversity and structure, which are valuable in assessing temporal genomic erosion and guiding conservation and restoration efforts [50,51]. Since samples of extinct or endangered species usually are rare, WGS data maximize the genomic information obtained from each sample. The highly fragmented DNA in museum and subfossil specimens (often <100 bp) requires the availability of reference genomes for WGS read mapping and downstream evolutionary analyses [52].

A cost-efficient approach for population genomic studies is Pool-Seq (i.e., deep sequencing of multiple pooled individuals) [53] (Table 1). Pool-Seq has been applied to investigate genome-wide patterns of genetic diversity, signatures of adaptation, population differentiation, and genotype–environment associations [54–56]. In some cases, Pool-Seq is the only possible or practical approach, for example, with microscopic organisms. It facilitates surveys of multiple populations and conditions, with large numbers of individuals (>50) sampled in each, and can significantly reduce costs in species with a small to intermediate-sized genome (<1 Gb). Despite the cost-saving advantages, Pool-Seq entails some analytical and statistical challenges [53]. The robustness and accuracy of results directly depend on the pool size [53] and can be reduced in complex systems with patterns of weak population structure [57].

Noninvasive genomic sampling
Noninvasive or minimally invasive sampling of biological material (e.g., from feces, feather, or hair [58,59]) is commonly used in wildlife monitoring as the main source of genetic material to provide insights into the ecology of endangered species [60]. Noninvasive samples typically contain low amounts of poor-quality DNA, often contaminated with exogenous DNA [59]. The technical challenges specific to noninvasive samples have been addressed through approaches that maximize the amount of retrieved endogenous DNA, such as: (i) SNP genotyping arrays [61]; (ii) multiplex PCR amplicon sequencing [62]; or (iii) target sequence capture using RNA/DNA oligonucleotides [63], which may be developed from existing information obtained using RRS approaches or reference genomes generated from high-quality DNA specimens.

Table 1. Examples of methods used in population genomic studies.

What can reference genomes bring to biodiversity conservation?
Since reference genomes are a prerequisite or enhance many of the analyses outlined earlier, their availability is desirable for any biodiversity genomics project [64]. Indeed, reference genomes are becoming commonplace across the entire tree of life. This is due to recent technological developments (Box 1) that enable the generation of accurate, chromosome-level genome assemblies for non-model organisms [9], and to multiple concerted international initiatives devoted to systematically generating reference genomes at scale (Table S1 in the supplemental information online). Next, we summarize how the availability of reference genomes substantially improves the use of genomics in biodiversity conservation.

The key to the whole trove of genomic information
Reference genomes provide the backbone and framework to map and annotate intra- and interspecific genomic variation [65–67]. They also provide a context to genetic variants, regions of low heterozygosity, patterns of linkage disequilibrium, and site frequency spectra (SFS) along chromosomes. In turn, these can be used to identify candidate genes associated with

---

Glossary

- **Genotype**: A set of genes present in an organism.
- **Phenotype**: The observable characteristics of an organism.
- **Gene**: A segment of DNA that codes for a protein.
- **Exon**: A segment of DNA that is transcribed into mRNA.
- **Intron**: A segment of DNA that is not transcribed into mRNA.
- **Exon-intron boundary**: The location where an intron begins or ends.
- **Spliceosome**: A macromolecular complex that recognizes and splices introns from pre-mRNA.
- **Human genome**: The complete set of genes present in the human species.
- **Genome**: The complete set of genes present in a cell or organism.
- **Chromosome**: A linear array of genes that carries genetic information.
- **Telomere**: A repetitive sequence of DNA that caps the ends of chromosomes.
- **Altitude**: The height of an object above sea level.
- **Water**: A clear, colorless substance that is the source of most of the liquid on Earth.
- **Carbon**: A chemical element that is a component of many compounds in the world.
- **Sugar**: A type of carbohydrate that provides energy for living organisms.
- **Amino acid**: A building block of proteins.
- **Protein**: A macromolecule that is composed of amino acids.
- **DNA**: Deoxyribonucleic acid, the genetic material that is found in cells.
- **RNA**: Ribonucleic acid, a type of nucleic acid that is involved in the expression of genetic information.
- **Gene expression**: The process by which genetic information is translated into a product, such as a protein.
- **Mutation**: A change in the DNA sequence that can alter the function of a gene.
- **Genetic variation**: Differences in the DNA sequence among individuals of the same species.
- **Allele**: One of two or more forms of a gene that occur in a population.
- **Linkage**: The tendency for certain genetic traits to be inherited together due to their close proximity on the same chromosome.
- **Segregation**: The process by which alleles are separated during meiosis.
- **Recombination**: The exchange of genetic material between homologous chromosomes.
- **Heterozygous**: Having two different alleles of a gene.
- **Homozygous**: Having two identical alleles of a gene.
- **Dominant**: One allele that completely masks the effects of the other allele in a heterozygous individual.
- **Recessive**: One allele that is only expressed when the individual is homozygous.
- **Genetic load**: The total number of deleterious genes in a population.
- **Effective population size**: A measure of genetic diversity that takes into account the effect of inbreeding and natural selection.
- **Hybridization**: The process of between members of different species.
- **Inbreeding**: The mating of closely related individuals, typically within a closed population.
- **Inbreeding depression**: The reduction in fitness of inbred individuals compared to non-inbred individuals.
- **Genomic erosion**: The loss of genetic diversity due to genetic drift, mutation, and natural selection.
- **Genetic drift**: The change in allele frequencies in a population due to random events.
- **Mutation**: A change in the DNA sequence that can alter the function of a gene.
- **Genome**: The complete set of genes present in a cell or organism.
- **Gene expression**: The process by which genetic information is translated into a product, such as a protein.
- **Mutation**: A change in the DNA sequence that can alter the function of a gene.
- **Genetic variation**: Differences in the DNA sequence among individuals of the same species.
- **Allele**: One of two or more forms of a gene that occur in a population.
- **Linkage**: The tendency for certain genetic traits to be inherited together due to their close proximity on the same chromosome.
- **Segregation**: The process by which alleles are separated during meiosis.
- **Recombination**: The exchange of genetic material between homologous chromosomes.
- **Heterozygous**: Having two different alleles of a gene.
- **Homozygous**: Having two identical alleles of a gene.
- **Dominant**: One allele that completely masks the effects of the other allele in a heterozygous individual.
- **Recessive**: One allele that is only expressed when the individual is homozygous.
- **Genetic load**: The total number of deleterious genes in a population.
- **Effective population size**: A measure of genetic diversity that takes into account the effect of inbreeding and natural selection.
- **Hybridization**: The process of between members of different species.
- **Inbreeding**: The mating of closely related individuals, typically within a closed population.
- **Inbreeding depression**: The reduction in fitness of inbred individuals compared to non-inbred individuals.
- **Genomic erosion**: The loss of genetic diversity due to genetic drift, mutation, and natural selection.
Functional and adaptive genetic variation

Reference genomes facilitate the identification of functional genetic variation underlying phenotypic traits, fitness variation, and adaptive potential [71–73]. Such analyses are mostly successful at identifying variants with large effects [71,74,75], likely representing a small proportion of the adaptive functional variation. However, large-effect variants can underlie traits that are responsible for resilience to anthropogenic threats such as diseases or climate change. Such traits, identified, for example, in genome-wide association studies (GWAS), can be selected for in captive breeding or translocation programs [Box 2: Case study 3 on the ash dieback]. Those programs must then carefully consider issues such as maintaining genome-wide diversity and evolutionary potential [76–78] and genetic adaptation to captivity [79]. Most phenotypic traits are polygenic (i.e., under the control of multiple interacting loci with small effects). These loci are, in turn, likely subject to gene regulatory networks difficult to fathom with DNA-based tests aimed at detecting adaptive variation at individual loci. To this end, statistical tools have been developed to detect contributions from multiple loci across the entire genome that may underlie the function and heritability of trait variation [80], as well as more direct approaches such as massively parallel reporter assays [81]. Reference genomes facilitate the estimation of the relative fitness of alleles and genotypes in threatened species [82], which may not be amenable to experimental approaches, for example, by WGS comparisons of individuals with different phenotypes [83]. Notwithstanding the analytical challenges, WGS data, combined with phenotypic and fitness data, enable a thorough understanding of adaptive variants [74] that drive responses (or the lack thereof) to current and future anthropogenic threats [84], such as ongoing environmental change [85]. Incorporating insights of adaptive potential into species distribution models helps modeling climate change effects on endangered populations [86,87], such as genotype–phenotype associations underlying species resilience, which can guide assisted breeding programs [88] (Box 2: Case study 4 on the European beech). Identifying adaptive variation may also aid in defining conservation units that ensure the preservation of evolutionary heritage and adaptive potential [89].

Inbreeding and genetic load

Assessing detrimental fitness effects associated with inbreeding (inbreeding depression) is a central topic in conservation genetics [90,91]. The availability of annotated reference genomes is invaluable for characterizing inbreeding and the genetic architecture of inbreeding depression, in particular the number of loci involved, their effect size, and the contributions of deleterious recessive alleles [92,93]. This knowledge can guide the selection of founders, breeders, and individuals for reintroduction in ex situ conservation management and help design purging strategies, or inform monitoring of deleterious variants and genomic erosion in the wild [94]. Runs of homozygosity (ROH) analyses are useful for distinguishing historical and contemporary inbreeding, pinpointing candidate loci contributing to inbreeding depression (particularly loci with large effects), and selecting captive breeders with different ROH to minimize inbreeding.
depression [90,91,95]. Reference genomes facilitate predicting the putative functional significance and fitness consequences of specific allelic variants that have drifted to high frequency or fixation in small wild populations and estimating the accumulated genetic load and dynamics of deleterious variation [51,96–98]. Such in-depth reference genome-based analyses are being carried out in an increasing number of endangered species (e.g., Box 2: Case study 5 on the Iberian lynx, the crested ibis [99], and the Sumatran rhinoceros [92]). In perspective, reference genomes combined with WGS data provided invaluable insights into the long-term management of translocated populations (Box 2: Case study 6 on the Florida panther).

**Outbreeding and hybridization**

Outbreeding can also result in a reduction of fitness, known as outbreeding depression, due to chromosomal or genic incompatibilities, epistatic interactions, disruption of interactions between coadapted genes, or introduction of variants that are maladaptive to local environmental conditions. In risk assessments of outbreeding depression, WGS and reference genomes can be invaluable in enabling detailed analyses of adaptive divergence and structural variation across the genome [46,82]. This knowledge is important for reducing the chances of outbreeding depression when selecting source populations and individuals for translocations to reinforce a population threatened with extinction [100]. Human-mediated hybridization and introgression have increased dramatically worldwide. In association with the perturbation and homogenization of environments, hybridization can become a major conservation concern, threatening biodiversity and evolutionary heritage. Genome-wide data make it possible to characterize introgression patterns, dynamics, and admixture proportions more effectively than traditionally possible with...
Table 1. List of genomic approaches with application and comparison of raw sequencing costs (i.e., costs of sample collection, researcher time, and analysis are not included)

<table>
<thead>
<tr>
<th>DNA barcoding/metabarcoding</th>
<th>Genome skimming</th>
<th>Reduced representation</th>
<th>Transcriptome sequencing</th>
<th>Whole-genome resequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td>What genome do you get?</td>
<td>None</td>
<td>Organelle, k-mer representation of nuclear</td>
<td>None</td>
<td>Nonrepetitive genome, depends on coverage</td>
</tr>
<tr>
<td>Cost in dollars (as of date)$^a$</td>
<td>$5 per sample (Sanger/NGS)</td>
<td>$50</td>
<td>$50</td>
<td>$100-$400</td>
</tr>
<tr>
<td>What type of samples are needed</td>
<td>Fresh tissue samples, museum specimens, noninvasive samples</td>
<td>Fresh tissue samples, museum specimens</td>
<td>Fresh tissue samples, museum specimens, noninvasive samples</td>
<td>Tissue-specific, live/fresh, flash frozen/RNA buffer</td>
</tr>
<tr>
<td>Genetic diversity$^b$</td>
<td>Yes, but limited</td>
<td>Yes, typically organelle based</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Population structure$^b$</td>
<td>Yes, but weak to detect shallow/cryptic genetic structure; economical for detailed spatial sampling</td>
<td>Yes, typically organelle based</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Phylogenetic information</td>
<td>Yes, but barcode based</td>
<td>Yes, typically organelle based</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Introgression event</td>
<td>No</td>
<td>No</td>
<td>Yes, but no individual genes</td>
<td>Yes, but limited detection power</td>
</tr>
<tr>
<td>QTL mapping</td>
<td>No</td>
<td>No</td>
<td>Yes, but low resolution</td>
<td>Yes, expression QTL (eQTL)</td>
</tr>
<tr>
<td>Natural selection signal detection</td>
<td>No</td>
<td>Yes, on organelle genes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Gene structure study</td>
<td>No</td>
<td>Yes, on organelle genes</td>
<td>Potentially, if reference genome available</td>
<td>No, Yes, if reference genome available</td>
</tr>
<tr>
<td>Gene family analyses</td>
<td>No</td>
<td>Yes, on organelle genes</td>
<td>Potentially, if reference genome available</td>
<td>Yes, if reference genome available</td>
</tr>
<tr>
<td>Genome rearrangement study</td>
<td>No</td>
<td>Yes, on organelle genes</td>
<td>Potentially, if reference genome available</td>
<td>No, Yes, if reference genome available</td>
</tr>
<tr>
<td>Functional genomic study$^b$</td>
<td>No</td>
<td>Yes, on organelle genes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Genome size estimation</td>
<td>No</td>
<td>Yes, typically organelle genome</td>
<td>No</td>
<td>No, Depending on coverage</td>
</tr>
<tr>
<td>Linkage disequilibrium</td>
<td>No</td>
<td>No</td>
<td>Yes, usually, not always</td>
<td>Limited, Yes</td>
</tr>
<tr>
<td>Demographic reconstructions (MSMC)$^e$</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Demographic reconstructions from SFS</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>GWAS</td>
<td>No</td>
<td>No</td>
<td>Yes, but low resolution</td>
<td>Yes, Transcriptome WAS</td>
</tr>
</tbody>
</table>

Whole-genome sequencing

<table>
<thead>
<tr>
<th>Short reads</th>
<th>Linked reads (10X/stLFR/Tell-Seq)</th>
<th>Long reads + scaffolding technologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>What genome do you get?</td>
<td>Very fragmented, contig N50 &lt;10 kb</td>
<td>Contig N50 &gt;200 kb, scaffold N50 &gt;10 Mb</td>
</tr>
<tr>
<td>Cost in dollars (as of date)$^a$</td>
<td>$100-$600</td>
<td>$2000</td>
</tr>
<tr>
<td>What type of samples are needed</td>
<td>Fresh tissue samples, museum specimens</td>
<td>Frozen upon collection</td>
</tr>
</tbody>
</table>

(continued on next page)
only dozens of markers [101]. Introgression rates in different genomic regions can vary. Whole-genome data mapped to reference genomes are useful to identify admixture (or migrant) tracts along individual genomes and thus to understand local effects of introgression and the age of admixture events [102,103].

**Genome editing: engineering adaptation, gene drives, and de-extinction**

Genome-editing tools, such as CRISPR-Cas9 [104], allow precise modification of genes and genomes of living organisms. Genome editing can be applied as a tool to mediate locus-specific genetic rescue in endangered species threatened with high frequencies of deleterious

### Table 1. (continued)

<table>
<thead>
<tr>
<th>Whole-genome sequencing</th>
<th>Short reads</th>
<th>Linked reads (10X/stLFR/Tell-Seq)</th>
<th>Long reads + scaffolding technologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetic diversity*</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Population structure*</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Phylogenetic information</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Introgression event</td>
<td>Yes, but limited detection power</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>QTL mapping</td>
<td>Yes, but low resolution</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Natural selection signal detection</td>
<td>Partially, if reference genome available</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Gene structure study</td>
<td>Partially, if reference genome available</td>
<td>Yes, miss few regions</td>
<td>Yes</td>
</tr>
<tr>
<td>Gene family analyses</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Genome rearrangement study</td>
<td>No</td>
<td>Yes, but fragmented karyotype</td>
<td>Yes, can trace karyotype evolution</td>
</tr>
<tr>
<td>Functional genomic study†</td>
<td>Yes, but missing many regions</td>
<td>Yes, but missing some regions</td>
<td>Yes</td>
</tr>
<tr>
<td>Linkage disequilibrium</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Demographic reconstructions (MSMC)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Demographic reconstructions from SFS</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>GWAS</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

*aBased on a 2 Gb vertebrate genome.
*Noting that single genomes deliver estimates from two haplotypes only.
*Noting that single specimens do not deliver to this aim well.
†Integrated with ChIP-seq, Hi-C, RNAseq for regulatory element annotation.
*Multiple sequentially Markovian coalescent.

Box 1. Genomes, population genomics, and pangenomes

Reference genomes, point representations of the structure and organization of a species’ genome, are expected to revolutionize conservation genomics [9]. Currently, a combination of single-molecule long-read sequencing (SMRT or nanopore sequencing) or linked reads (e.g., TELL-seq or stLFR) for contigging, optical maps, and/or proximity ligation reads (e.g., SC-seq or Hi-C) for scaffolding appears to be a general strategy capable of generating phased chromosome-scale reference genomes across the tree of life [124,125] (see Table 1 in main text). When sufficient high molecular weight DNA for long-read sequencing remains problematic to obtain (e.g., with small organisms, albeit solutions are rapidly becoming available [126]), an alternative cost-effective strategy is represented by using linked reads only [127]. To gain more accurate insights into a species’ genomic diversity, the genomes of large numbers of individuals can be resequenced. Such multiple genomes do not have to be assembled de novo (de novo assembly) and may be derived from cheaper and quicker short-read sequencing and subsequent mapping to an existing assembly [128]. Additionally, the availability of multiple reference genomes per species provides a more complete genome characterization [129,130] and allows estimating a species pangenome [131]. Individuals of a species are part of a pangenomic architecture, sharing a larger proportion of their genes as a core set, whilst differing in a smaller proportion of dispensable genes [132]. The analysis of such data helps to differentiate adaptive and neutral evolutionary processes, to define populations, ecotypes, and species, and identify core and accessory genomic regions, gene flow, hybridization, and incomplete lineage sorting [133].
Box 2. Case studies

1. Atlantic cod

The Atlantic cod (Gadus morhua; Figure I) comprises several cryptic ecotypes differing in feeding, spawning, and migratory behaviors and, consequently, in population dynamics relevant for fisheries management [134]. The Atlantic cod was one of the first non-model organisms with a chromosome-anchored genome assembly [135]. Combined with population genomic analyses, the genome assembly enabled identification of four large chromosomal inversions that discriminate between the migratory and nonmigratory ecotypes in Norwegian and Icelandic waters and are associated with adaptation to environmental conditions [134,136]. Each of the four inversion regions contains hundreds of genes with a high level of divergence between the inverted and noninverted haplotypes. The four inversions originated independently up to 1.7 million years ago and have been maintained due to strong selection [136,137].

2. Horseshoe crabs

Horseshoe crabs (Figure II) belong to the chelicerate order Xiphosura. This enigmatic order comprises only four extant species (Limulus polyphemus, Tachypleus gigas, Tachypleus tridentatus, and Carcinocorpus rotundicauda), all considered ‘living fossils’ since their morphology is virtually identical to that of their Triassic ancestors. Recent studies have generated chromosome-level assemblies of three of these species, and analyses suggest three rounds of whole-genome duplication [138,139]. Several gene families, particularly those involved in innate immunity, have undergone extensive tandem duplication. These expanded gene families may be important components of the innate immune system of horseshoe crabs, whose amebocytes are presently exploited for detecting endotoxin contamination [139]. These genomic resources are of value for breeding programs and conservation [138].

3. Ash dieback

In the mid-1990s, European ash trees (Fraxinus excelsior; Figure IC) started dying across Europe from a new disease, ash dieback, which spread rapidly and was determined to be caused by Hymenoscyphus fraxineus, a fungal pathogen new to Europe. Genome sequencing of this fungus along with isolates from East Asia confirmed that the European outbreak was likely caused by the introduction of imported Asian Fraxinus mandshurica. Importantly, the fungal genomes showed that at most two independent introductions of H. fraxineus had occurred [140]. The availability of the fungal genomes will aid surveillance for new introductions. While no ash trees appeared to be fully resistant, threatening extinction of this iconic species, analysis of Danish forestry germplasm identified individuals that had reduced susceptibility. An international consortium sequenced the European ash genome [141] and used it to map polygenic traits associated with reduced susceptibility [142], promising to assist in directed breeding to restock ash forests across the continent.

4. European beech

In the summers of 2015 and 2019, around two-thirds of European beech trees (Fagus sylvatica; Figure ID) were damaged or killed by extreme drought. It is critical to keep beech woods healthy, as they are home to over 6000 other species of animals and plants. However, not all the trees in each forest responded in the same way, with severely damaged trees often sitting less than 5 m from fully healthy ones. This suggests that the genetic makeup of a tree, rather than its local environment, determines how well it copes with drought. After the beech genome was assembled [143], over 400 European beech trees from pairs of neighboring trees that had responded differently to the droughts were analyzed [88]. More than 80 regions of the genome differed between healthy and damaged trees. Subsequently, a genetic test was developed which can quickly and inexpensively predict how well an individual beech tree might cope with drought. Now this genetic test can be used to select and reproduce trees that are better adapted to drought.

5. Iberian lynx

One of the first whole genomes sequenced for a highly endangered species was that of the Iberian lynx (Lynx pardinus; Figure IE). Whole-genome resequencing analyses revealed that the species underwent a history of severe population declines, leading to one of the lowest genomic diversities ever reported [85]. Historical genetic data allowed the quantification of recent diversity loss and validated the mixing of the two differentiated remnant populations [144]. Subsequent genome-wide studies focused on functional variation have shown a higher genetic load in the more bottlenecks lynx populations [145], but also the purging of highly deleterious variants in Iberian lynx with respect to the more abundant Eurasian lynx [146]. These studies have also generated a catalog of potentially deleterious variation that can now be tested for association with reduced fitness and genetic diseases. The access to genome-wide variation also allowed the selection of a set of highly efficient and informative SNPs [147], which is now being extensively applied in combination with noninvasive samples to monitor and manage genetic diversity across ex situ, remnant, and reintroduced populations. These concerted conservation efforts aided by genetic information succeeded in avoiding the imminent extinction and in reverting the negative population trends of the species by the turn of the century.

6. Florida panther

In the early 1990s, isolated Florida panthers, a population of puma (Puma concolor; Figure IF) in the Big Cypress National Preserve, showed low reproductive success and multiple signs of inbreeding. Therefore, pumas from Texas were translocated to support the isolated Florida population. Genetic diversity and reproductive success increased and the consequences of severe inbreeding, such as undescended testicles, heart failures, cowlick, and kinked tails, vanished [148]. A recent study explored the genomic consequences of the long-term isolation of this and other puma populations using a reference genome and a geographically broad panel of resequenced individuals [86]. The genome of a Florida panther descended from translocated Central American individuals had long tracts of homozygosity despite recent outbreeding. This suggests that sustaining diversity in small and isolated populations will require either repeated translocations or restoration of landscape connectivity.

7. Freshwater sponge

Relatively few chromosome-level reference genomes are available for non-model invertebrate species, which hampers the implementation of conservation or biodiversity genomics approaches in a group of species that spans more than 95% of the animal tree of life. The freshwater sponge Ephydatia muelleri (Figure IQ) is a common species in Palearctic lakes and freshwater streams that can even colonize drinking water systems, is transported by migratory birds and plays important ecological roles where it...
mutations or to increase resistance to infectious diseases and resilience to anthropogenic environmental change [105]. Modifications could even involve gene drives to assist the spread of deleterious mutations through invasive populations [106,107], although this is still controversial due to concerns of possible unintended outcomes [108,109] and lack of an international regulatory framework for safe and responsible use [110]. Reference genomes are necessary for both the editing and gene drive approaches. An even more technically challenging prospect is resurrecting extinct species, or more accurately, functional proxies of extinct species [111]. Species driven to extinction by humans may be ethically justifiable candidates for de-extinction, especially keystone species for extant ecosystems with beneficial cascading effects [112]. De-extinction is controversial since the causes of extinction may still be present [113] or because they could divert funding and public attention from ongoing, critical conservation.
efforts of threatened species [114]. Cloning via somatic cell nuclear transfer (SCNT) could allow de-extinction of recently extinct mammal species where well-preserved tissues have been cryopreserved [111]. Likewise, SCNT can be used to resurrect extinct genetic variation to facilitate genetic rescue in highly inbred, endangered species [115]. Obviously, for the vast majority of species driven extinct by humans, what remains at most is degraded and fragmented DNA. In such cases, reference genomes from extant relatives are crucial for attempts at reconstructing a genome sequence of the extinct species from short-read WGS data [111].

Beyond single species: structure and function of communities

There is increasing interest in genomics tools usable at the community level [116]. Reference genomes are particularly helpful in metagenomic and metatranscriptomic analyses, where the DNA or RNA from an entire community sample is sequenced. Metagenomics targets community composition and its functional potential [117], whereas metatranscriptomics provides a temporal snapshot of community activity through gene expression [118]. Both approaches depend on the availability of annotated reference genomes. Metagenomics and metatranscriptomics are frequently applied to microbial community samples, benefiting from a more extensive reference genome availability, in particular for known bacteria, archaea, and fungi [119]. Indeed, in the absence of reference genomes, most non-bacterial sequences remain unidentified. The accuracy and information content of the genomes is highly relevant for metagenomic assignment [120]. Reference genomes will allow a more complete characterization of communities when applying metagenomic or metatranscriptomic approaches [121], facilitating monitoring and managing taxonomic and functional diversity in entire ecosystems.

Concluding remarks

Accelerating global biodiversity loss due to rapid environmental change and other anthropogenic impacts drives the need for urgent conservation efforts. Genome assemblies provide a fundamental framework to interpret and protect biodiversity. While many questions remain (see Outstanding questions), we have reviewed how biodiversity conservation can benefit from genomic data built upon reference genomes. We presented case studies that, among many others,

Box 3. Connecting genomic initiatives: from genomes to conservation actions

Genomics can help address main ecological questions that are central to understanding and maintaining biodiversity and ecosystem services. In Europe, in recent years there have been three main political initiatives to support biodiversity research and conservation: the Biodiversity Strategy for 2030, the European Partnership on Biodiversity, and the European Green Deal. These initiatives all advocate for the importance of genetics in ecosystem preservation and restoration. To address this, several initiatives have started to bring genomics applications for biodiversity to a European perspective. Notable examples of such initiatives are the European Reference Genome Atlas (ERGA) and the Genomic Biodiversity Knowledge for resilient Ecosystems (G-BiKE). ERGA is a consortium involving nearly 50 EU countries and international partners, recently funded through a Horizon Europe call on Biodiversity and Ecosystem Services, with the aim of generating high-quality, nearly error-free reference genomes representing the genetic makeup of European eukaryotic biodiversity, ranging from endangered species, species of importance for ecosystem function and stability, to key species for agriculture, forestry, and fisheries, but also non-model organisms from under-represented taxa that make up a huge proportion of biodiversity. Besides a diverse yet coordinated production pipeline, strong components of ERGA are the application of reference genomes to conservation genomics endeavors, ethical, legal, and social implications on the use of genomic resources, as well as citizen science initiatives. G-BiKE is a network funded by the European Cooperation in Science and Technology (COST) program involving more than 110 researchers and practitioners from 39 European countries with the aim of establishing the use of genomic data as a standard tool for monitoring and managing wild and ex situ populations of plants and animals. This is achieved by organizing workshops, training schools, short-term scientific missions, and virtual mobility grants. The ultimate goal of G-BiKE is to integrate genetic diversity monitoring into EU policy and planning on biodiversity conservation, including the missions of the European Green Deal, the Biodiversity Strategy for 2030, the Natura 2000 network, and Habitats Directive. Recently, G-BiKE has enlarged its scope by playing an active role in the process leading to the post-2020 Global Biodiversity Framework of the Convention on Biological Diversity. A joint venture between ERGA and G-BiKE is anticipated to advocate for the incorporation of genomic data into European biodiversity and ecosystem services protection programs.

Outstanding questions

A considerable fraction of biodiversity is endangered, but some taxa are more accessible and studied than others. How do we obtain high-quality DNA and generate genomic information from very rare and small (e.g., flies, meiobiota, gastrotriches) or challenging taxa (e.g., conifers due to large genome size, or uncultured fungi)?

How can we promote both taxonomic and geographic coverage to contribute to the global endeavor of a balanced conservation of biodiversity?

We are now able to estimate many genomic aspects relevant to conservation, but we still need to integrate them with geographical and environmental data. How do we combine these factors and translate them into adaptive potential and extinction risk in a predictive quantitative manner that can inform conservation efforts?

While various efforts already exist to address ethical, legal, and social issues, there are still legitimate concerns about the rights of local human communities over biodiversity resources. How can we provide open access to reference genomes, while also ensuring benefit sharing for those communities?

Biodiversity protection and climate change mitigation are strongly co-dependent. How can biodiversity research be directed towards sustainable generation and analysis of genomic data?

Compiling reference genomes and omics data (e.g., transcriptomic, epigenomic) for endangered species creates ‘digital Noah’s arks’ that complement the more fragile and perishable frozen zoos. How can we create momentum for the concerted construction of these multi-omics digital arks that may be vital for de-extinction efforts in the future?

Reference genomes and genome editing will increasingly allow identification and modification of genomic regions to increase fitness, persistence, and adaptive potential of endangered species. How can we
demonstrate how genomic resources enrich and enhance integrated conservation efforts. Biodiversity conservation must explicitly consider genomic diversity to optimize strategies that minimize the loss of fitness, thus aiding in maintaining population viability, and to preserve evolutionary potential for adaptive responses to environmental change and diversification in general. To this end, we advocate for promoting reference genome-based approaches in conservation research and encourage knowledge transfer between the research community, conservation practitioners, and society (Box 3) [122]. As the Human Genome Project strongly impacted the biomedical sciences over the past two decades [123], we anticipate that the availability of reference genomes across the tree of life will provide a solid, quantitative, and comparable foundation for biodiversity assessments, conservation, and restoration.

Acknowledgments
We thank Cecilia Helmerson, Antonio Rivas, Nelson Lauzon, and Sally Leys, for the pictures of the Atlantic cod, the Iberian lynx, the European beech, and the freshwater sponge, respectively. We also thank Fabien Condamine, Love Dalén, Richard Durbin, Bruno Fosso, Roderic Guigó, Marc Hanikenne, Alberto Pallavicini, Olga Vinnere Pettersson, Xavier Turon, and Detlef Weigel for their contributions to the manuscript, as well as the whole ERGA community for making this possible.

Declaration of interests
No interests are declared.

Supplemental information
Supplemental information associated with this article can be found online https://doi.org/10.1016/j.tig.2023.01.005.

References
7. Hoban, S. et al. (2021) Genetic diversity is considered important but interpreted narrowly in country reports to the Convention on Biological Diversity: current actions and indicators are insufficient. Biol. Conserv. 261, 108633
36. Wang, B. et al. (2019) Rewiring the transcriptome studies: an insight into the emergence of single-molecule transcriptomics sequencing. Front. Mol. Biosci. 6, 284
61. Harmoinen, J. et al. (2021) Reliable wolf-dog hybrid detection in Europe using a reduced SNP panel developed for non-invasively collected samples. BMC Genomics 22, 473
139. Shingate, P. et al. (2020) Chromosome-level assembly of the horseshoe crab genome provides insights into its genome evolution. Nat. Commun. 11, 2322