The era of reference genomes in conservation genomics

Trends in Ecology & Evolution, 2022. 10.1016/j.tree.2021.1011.1008.

Giulio Formenti*, Kathrin Theissinger*, Carlos Fernandes*, Iliana Bista, Aureliano Bombarely, Christoph Bleidorn, Claudio Ciofi, Angelica Crottini, José A. Godoy, Jacob Höglund, Joanna Malukiewicz, Alice Mouton, Rebekah A. Oomen, Sadye Paez, Per J. Palsbøll, Christophe Pampoulie, María J. Ruiz-López, Hannes Svardal, Constantina Theofanopoulou, Jan de Vries, Ann-Marie Waldvogel, Guojie Zhang, Camila J. Mazzoni, Erich D. Jarvis, Miklós Bálint+, and The European Reference Genome Atlas (ERGA) Consortium‡

‡ This work originated from a collective effort within the ERGA Consortium:

Čiampor, F., Hoglund, J., Palsbøll, P., Ruiz-López, M.J., Zhang, G., Jarvis, E., Aghayan, S.A., Alioto, T.S., Almudi, I., Alvarez, N., Alves, P.C., Amorim, I.R., Antunes, A., Arribas, P., Baldrian, P., Berg, P.R., Bertorelle, G., Böhne, A., Bonisoli-Alquati, A., Boštjančić, L.L., Boussau, B., Breton, C.M., Buzan, E., Campos, P.F., Carreras, C., Castro, L.F., Chueca, L.J., Conti, E., Cook-Deegan, R., Croll, D., Cunha, M.V., Delsuc, F., Dennis, A.B., Dimitrov, D., Faria, R., Favre, A., Fedrigo, O.D., Fernández, R., Ficetola, G.F., Flot, J.F., Gabaldón, T., Galea Agius, D.R., Gallo, G.R., Giani, A.M., Gilbert, M.T.P., Grebenc, T., Guschanski, K., Guyot, R., Hausdorf, B., Hawlitschek, O.,

Heintzman, P.D., Heinze, B., Hiller, M., Husemann, M., Iannucci, A., Irisarri, I., Jakobsen, K.S., Jentoft, S., Klinga, P., Kloch, A., Kratochwil, C.F., Kusche, H., Layton, K.K.S., Leonard, J.A., Lerat, E., Liti, G., Manousaki, T., Marques-Bonet, T., Matos-Maraví, P., Matschiner, M., Maumus, F., Mc Cartney, A.M., Meiri, S., Melo-Ferreira, J., Mengual, X., Monaghan, M.T., Montagna, M., Mysłajek, R.W., Neiber, M.T., Nicolas, V., Novo, M., Ozretić, P., Palero, F., Pârvulescu, L., Pascual, M., Paulo, O.S., Pavlek, M., Pegueroles, C., Pellissier, L., Pesole, G., Primmer, C.R., Riesgo, A., Rüber, L., Rubolini, D., Salvi, D., Seehausen, O., Seidel, M., Secomandi, S., Studer, B., Theodoridis, S., Thines, M., Urban, L., Vasemägi, A., Vella, A., Vella, N., Vernes, S.C., Vernesi, C., Vieites, D.R., Waterhouse, R.M., Wheat, C.W., Wörheide, G., Wurm, Y., Zammit, G.

Abstract

Progress in genome sequencing now enables the large-scale generation of reference genomes. Various international initiatives aim to generate reference genomes representing global biodiversity. These genomes provide unique insights into genomic diversity and architecture, thereby enabling comprehensive analyses of population and functional genomics, and are expected to revolutionize conservation genomics.

Keywords: conservation genetics, biodiversity conservation, European Reference Genome Atlas, ERGA

Conservation, genomics, and reference genomes

In 2020, both the United Nations Biodiversity Summit and the European Environment Agency emphasized the accelerating global loss of biodiversity (*https://www.un.org/pga/75/united-nations-summit-on-biodiversity/*; *https://www.eea.europa.eu/highlights/latest-evaluation-shows-europes-nature*). We are in the sixth mass extinction, and while the primary route to preserve biodiversity comprises protection and restoration of species, habitats and ecosystems, genomics provides a rapidly expanding array of novel tools to characterize biodiversity and assist such conservation efforts. The need for immediate actions that aid to reverse the current biodiversity decline has prompted national and international initiatives aimed at expanding the genomic reference resources available for biodiversity research and conservation across the tree of life (**Box 1**). Many of these efforts collectively contribute to the Earth BioGenome Project (EBP), with the aim of cataloguing and characterizing the genomes of all of Earth's eukaryotic biodiversity. A large and inclusive community of scientists has recently gathered as the European hub of the EBP to promote the generation of a European Reference Genome Atlas (ERGA, www.erga-biodiversity.eu). This initiative is building a pan-European open access infrastructure to streamline ethical and legally compliant sample and metadata collection [1], sequencing, assembly [2], **annotation** [3], and release in public archives of high-quality genomic information, creating reference genomes for a wide variety of eukaryotic species (**Box 1**).

Reference genomes (see Glossary), by which we mean highly contiguous, accurate, and annotated genome assemblies, greatly enhance genomic studies, both experimentally and analytically [4]. A reference genome is a point representation of the structure and organization of a species' genome. Similar to type specimens in taxonomy, reference genomes serve as the standard for subsequent genomic studies [5]. To unravel the genomic diversity of species, multiple conspecific individuals can be **resequenced** and aligned to available reference genomes, rather than assembled *de novo*. Thus, reference genomes provide a comprehensive and fundamental basis onto which additional genomic variation can be mapped, to characterize and ultimately aid preserving genetic diversity [4]. To this end, special attention should be paid to the origin of the

individuals used as reference, which if excessively divergent from the populations under study may compromise the analyses. To overcome this issue, multiple conspecific genomes [6] can now be summarized in a species' **pangenome** [7].

Reference genomes have been, until recently, available only for a handful of model organisms. Thanks to the coordinated and standardized efforts of international genome initiatives, the situation is rapidly changing. Recent technological advances appear to provide a general strategy for generating chromosome-scale reference genomes for all organisms across the tree of life [8]. These advances rely on a combination of single-molecule **long-read** sequencing (SMRT or nanopore sequencing) and/or linked reads (e.g. TELL-seq or stLFR) for **contig assembly**, **optical mapping** and/or **proximity ligation** (e.g. 3C-seq or Hi-C) for **scaffolding** [2]. Decreasing costs, improved scalability, and increasing quality of sequencing technologies, combined with better algorithms and increased computational power [8], facilitate the establishment of reference genomes across the full spectrum of life biodiversity. Importantly, reference genomes are fundamental for a comprehensive and accurate characterization of genomic information, for instance of structural features, that cannot be revealed by fragmented genomes and **reduced-representation sequencing** approaches (Figure 1). While analytical methods and interpretations are constantly evolving, reference genomes, particularly when coupled with resequencing data, should become a standard resource in conservation genomics.

Key contributions of reference genomes in conservation genomics

The full spectrum of genomic diversity

Reference genomes provide a view of the genome's architecture, comprising genic and intergenic regions. These include repetitive regions, some of which are challenging to assemble, such as **segmental duplications**, centromeres and telomeres, **satellites**, and **mobile elements**. Population genomics guided by reference genomes aids the identification of classical genetic variants such as SNPs and **CNVs**, as well as **structural variants** that have proven particularly difficult to detect with fragmented and incomplete

reference genomes alone, and yet may be important for adaptation to environmental change [9].

Inbreeding and deleterious mutations

Assessments of inbreeding have long informed conservation and breeding programs, guiding genetic crosses and translocation of individuals. While often estimated from few loci, understanding the genetic architecture and accurately quantifying **inbreeding depression** requires a genome-wide perspective, e.g., the number of genes involved, the presence of alleles with large effects, the role of deleterious recessive alleles, and the **heterozygote advantage** [10]. While several questions remain, multiple studies have showcased the power of population genomics guided by reference genomes to identify **runs of homozygosity** to estimate inbreeding, and the dynamics and fate of deleterious variation in threatened species (e.g. [11]).

Outbreeding and introgression

Mating between individuals from genetically distinct lineages may lead to **outbreeding depression**, due to chromosomal or genic incompatibilities, epistatic interactions, disruption of interactions between co-adapted genes, or the introduction of maladaptive variants into local populations. Population genomics guided by reference genomes greatly facilitates the disentanglement of these phenomena. **Hybridization** is a common evolutionary process that can promote through **introgression** the spread of adaptive variation and speciation. Anthropogenic hybridization and introgression, however, are often major threats to biodiversity and evolutionary heritage. Reference genomes facilitate characterization of introgression patterns and dynamics as well as **admixture** proportions, and particularly of introgressed tracts along individual genomes [12].

Local adaptation and genetic rescue

The use of reference genomes in population genomics facilitates the identification of traits under natural selection, and thus the basis and architecture of local adaptations and ultimately, speciation. Reference genomes provide functional and genomic contexts for regions influenced by selection, enabling association

of such loci with phenotypes important for adaptation and resilience. Identifying locally adapted genes can inform definitions of conservation units and identify optimal source populations for translocations to facilitate **genetic rescue** [13].

Phylogenetic diversity and phylogenomics

Phylogenetic diversity is essential for ecosystem stability and resilience, and it is used to delineate evolutionarily distinct components of biodiversity that are conservation priorities (e.g., ***EDGE* species**) [14]. Genome-scale analyses, based on hundreds or thousands of loci, have become the gold standard for phylogenetic inference and help to capture the evolutionary histories of target taxa. Reference genomes serve as the basis for **phylogenomic** analyses, as they greatly improve **orthology** inference at the DNA and protein levels, while also allowing inferences based on genome organisation.

Structure and function of communities

Reference genomes are particularly important in **metagenomics** and **metatranscriptomics**, where total DNA, or cDNA derived from RNA, from entire communities is sequenced to understand their composition, abundance, function, and dynamics. Facilitated by the availability of reference genomes, metagenomics and metatranscriptomics have been mostly applied to microbial community samples. Eukaryotic reference genomes allow assigning DNA/cDNA reads to higher taxa within environmental samples, leading to a more complete characterization of communities, and thus support biomonitoring and management of taxonomic and functional diversity in entire ecosystems.

A collective effort to conserve biodiversity

The age of reference-guided genomics for non-model organisms has begun.. Consequently, conservation efforts need to account for genomic diversity to optimise management strategies. Accounting for genomic diversity will aid in maintaining population viability and preserving adaptive potential to respond to environmental change. The availability of reference genomes will provide a solid, quantitative and comparable foundation for biodiversity assessments, conservation, management, and restoration.

Acknowledgments

We 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 its constant support.

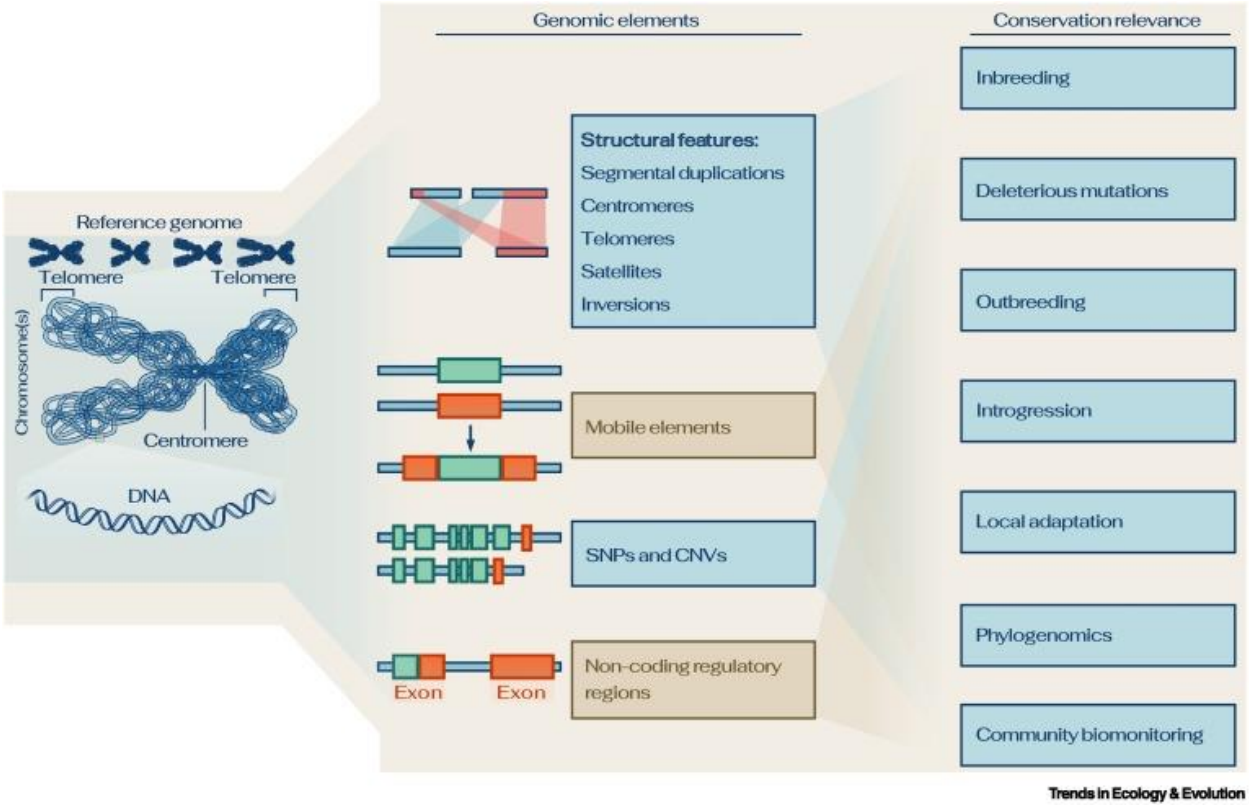

Figure 1. Reference genomes offer an (almost) complete record of a species' genome. They characterize genomic information more thoroughly than fragmented genomes can do. Importantly, they reveal structural features which often remain elusive for fragmented genome sequences. These features are relevant for conservation genomics applications.

Box 1: Sequencing the tree of life

International initiatives aimed at generating genomic resources, and particularly reference genomes, have flourished in recent years. Some focus on specific taxa, such as the Vertebrate Genomes Project, and 1000 Fungal Genomes project. Others focus on geographic regions, such as the California Conservation Genomics Project, the Darwin Tree of Life for Britain and Ireland, on applications such as Translational Biodiversity Genomics in Germany, or on ecological systems, such as the Aquatic Symbiosis Genomics project. Collectively part of the Earth Biogenome Project, in Europe these initiatives are organized under the umbrella of the European Reference Genome Atlas.

A genome atlas of European biodiversity

ERGA is a pan-European scientific response to the current threats to biodiversity. Approximately one fifth of the ~200,000 eukaryotic species present in Europe can be inferred to be at risk of extinction.

ERGA aims to generate reference genomes of European eukaryotic species across the tree of life, including threatened, endemic, and keystone species, as well as pests and species important to agriculture, fisheries, and ecosystem function and stability. ERGA builds upon current genomic consortia in EU member states, EU Associated Countries, representatives of other countries within the European bioregion, and international collaborators. These reference genomes will address fundamental and applied questions in conservation, biology and health. ERGA seeks to alert the European Union (EU) about the potential of conservation genomics, and particularly, the role of reference genomes in biodiversity assessment, conservation strategies and restoration efforts.

Glossary

Term	Definition
Admixture	Production of new genetic combinations in hybrid populations through recombination.
Annotation	Identification of coding sequences and other features in a genome.
Assembly, *de novo*	The process of generating a genome sequence from individual sequencing reads without the use of an existing reference.
Assembly, chromosome-level	Contiguous sequence of all chromosomes, often aided by genetic maps or other techniques.
Copy Number Variation (CNV)	Copy Number Variation (CNV) of a specific genomic segment. Refers to genome sections that are repeated, with the number of repeats being variable across individuals.
EDGE species	Evolutionary Distinct and Globally Endangered (EDGE) species of high conservation priority.

Genetic rescue	Mitigation strategy for restoring intraspecific genetic diversity and reducing extinction risks in small, isolated or inbred populations through induced gene flow.
Heterozygote advantage	Refers to a heterozygous genotype with a higher relative fitness as compared to a homozygous dominant or homozygous recessive genotype.
Hybridization	Interbreeding of individuals from genetically distinct lineages.
Inbreeding depression	Reduced fitness in offspring as a result of inbreeding, i.e., production of offspring from the mating of closely related individuals.
Introgression	Gene flow between hybridizing populations or species by backcrossing hybrids with one or both parental populations.
Metagenomics / Metatranscriptomics	Sequencing of DNA or RNA-derived cDNA extracted from environmental & bulk samples.
Mobile genetic elements	Genetic material that can move within a genome and be

	transferred between species.
Optical mapping	Technology to build long-range ordered sets of genomic segments that can improve fragmented genome assemblies.
Orthology	Sequence homology derived from a speciation (not duplication) event.
Outbreeding depression	Reduced fitness of offspring from matings between genetically divergent individuals.
Pangenome	The entire set of DNA sequences (or genes) of a species represented by the *core* genome and the *accessory* genome.
Phylogenomics	The inference of the phylogenetic relationship among different lineages of organisms from genome-wide data.
Proximity ligation	Library preparation method that captures the three-dimensional structure of chromatin through DNA cross-linking.
Reduced-representation sequencing	Reduced genome representation approaches to generate genome-wide high-throughput sequencing data

Read, short and long	Nucleic acids (D/RNA) sequences less than 500 base pairs (short reads) or longer than 10,000 base pairs (long reads).
Reference genome	Contiguous and accurate genome assembly representative of a species with coordinates of genes and other important features annotated. See [8] and https://www.earthbiogenome.org/assembly-standards for current definitions of reference genome quality.
Runs of Homozygosity (ROH)	Long tracts of the genome with very little or no heterozygous sites that can inform about recent and past population events and can be used to estimate individual inbreeding levels.
Satellites	Tandemly repeated non-coding DNA.
Segmental duplications	Low-copy repetitions, in tandem or interspersed and resulting from duplication events, of long DNA sequences
Structural variation	Regions of a chromosome presenting structural changes including the insertion, deletion, inversion or translocation of DNA.

References

[1] Shaw F, Etuk A, Minotto A, Gonzalez-Beltran A, Johnson D, et al. COPO: a metadata platform for brokering FAIR data in the life sciences. F1000Res 2020;9:495.

[2] Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature 2021;592:737–46.

[3] Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2021. Nucleic Acids Res 2021;49:D884–91.

[4] Brandies P, Peel E, Hogg CJ, Belov K. The value of reference genomes in the conservation of threatened species. Genes 2019;DOI: 10.3390/genes10110846:(https://www.mdpi.com/journal/genes).

[5] Ballouz S, Dobin A, Gillis JA. Is it time to change the reference genome? Genome Biol 2019;DOI: 10.1186/s13059-019-1774-4:(https://genomebiology.biomedcentral.com/).

[6] Valiente-Mullor C, Beamud B, Ansari I, Francés-Cuesta C, García-González N, et al. One is not enough: on the effects of reference genome for the mapping and subsequent analyses of short-reads. PLoS Comput Biol 2021;DOI: 10.1371/journal.pcbi.1008678:(https://journals.plos.org/ploscompbiol/).

[7] Llamas B, Narzisi G, Schneider V, Audano PA, Biederstedt E, et al. A strategy for building and using a human reference pangenome. F1000Research 2019;DOI: 10.12688/f1000research.19630.1:(https://f1000research.com/).

[8] Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature 2021. doi:10.1038/s41586-021-03451-0.

[9] Mérot C, Oomen RA, Tigano A, Wellenreuther M. A roadmap for understanding the evolutionary significance of structural genomic variation. Trends Ecol Evol 2020;35:561–72.

[10] Kardos M, Taylor HR, Ellegren H, Luikart G, Allendorf FW. Genomics advances the study of inbreeding depression in the wild. Evol Appl 2016;9:1205–18.

[11] Dussex N, van der Valk T, Morales HE, Wheat CW, Díez-del-Molino D, et al. Population genomics of the critically endangered kākāpō. Cell Genomics 2021:100002.

[12] Rogers J, Raveendran M, Harris RA, Mailund T, Leppälä K, et al. The comparative genomics and complex population history of *Papio* baboons. Sci Adv 2019;DOI: 10.1126/sciadaau6947:(https://advances.sciencemag.org/).

[13] Flanagan SP, Forester BR, Latch EK, Aitken SN, Hoban S. Guidelines for planning genomic assessment and monitoring of locally adaptive variation to inform species conservation. Evol Appl 2018;11:1035–52.

[14] Owen NR, Gumbs R, Gray CL, Faith DP. Global conservation of phylogenetic diversity captures more than just functional diversity. Nat Commun 2019;DOI: 10.1038/s41467-019-08600-8:(https://www.nature.com/ncomms/).

[15] Dahlberg A, Genney DR, Heilmann-Clausen J. Developing a comprehensive strategy for fungal conservation in Europe: current status and future needs. Fungal Ecol 2010;3:50–64.