

1 Contextualising samples: Supporting reference genomes  
2 ~~for~~ of European biodiversity through sample and  
3 associated metadata collection  
4  
5

6 Authors

7 **Astrid Böhne**, Leibniz Institute for the Analysis of Biodiversity Change, Museum Koenig Bonn,  
8 Centre for Molecular Biodiversity Research, Adenauerallee 127, 53113 Bonn, Germany;  
9 [a.boehne@leibniz-lib.de](mailto:a.boehne@leibniz-lib.de), ORCID 0000-0002-1284-3115.

10 **Rosa Fernández**, Metazoa Phylogenomics Lab, Biodiversity Program, Institute of Evolutionary  
11 Biology (IBE, CSIC-UPF), Passeig marítim de la Barceloneta 37-49, 08003, Barcelona, Spain.  
12 [rosa.fernandez@ibe.upf-csic.es](mailto:rosa.fernandez@ibe.upf-csic.es), ORCID 0000-0002-4719-6640.

13 **Jennifer A. Leonard**, Conservation and Evolutionary Genetics Group, Estación Biológica de  
14 Doñana (EBD-CSIC), Avda. Americo Vespucio 26, 41092, Sevilla, Spain. [jleonard@ebd.csic.es](mailto:jleonard@ebd.csic.es),  
15 ORCID 0000-0003-0291-7819.

16 **Ann M. McCartney**, Genomics Institute, University of California Santa Cruz, Santa Cruz, CA,  
17 USA, [anmmccar@ucsc.edu](mailto:anmmccar@ucsc.edu) 0000-0003-3191-3200

18 **Seanna McTaggart**, Earlham Institute, Norwich Research Park, Norwich, Norfolk, NR4 7UZ,  
19 United Kingdom; [seanna.mctaggart@earlham.ac.uk](mailto:seanna.mctaggart@earlham.ac.uk), ORCID 0000-0003-0977-4785

20 **José Melo-Ferreira**, (1) CIBIO, Centro de Investigação em Biodiversidade e Recursos  
21 Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto, 4485-661  
22 Vairão, Portugal; (2) Departamento de Biologia, Faculdade de Ciências, Universidade do Porto,  
23 4099-002 Porto, Portugal; (3) BIOPOLIS Program in Genomics, Biodiversity and Land Planning,  
24 CIBIO, Campus de Vairão, 4485-661 Vairão, Portugal.  
25 [jmeloferreira@cibio.up.pt](mailto:jmeloferreira@cibio.up.pt), ORCID 0000-0003-4473-1908.

26 **Rita Monteiro**, Leibniz Institute for the Analysis of Biodiversity Change, Museum Koenig Bonn,  
27 Centre for Molecular Biodiversity Research, Adenauerallee 127, 53113 Bonn, Germany.  
28 [r.monteiro@leibniz-lib.de](mailto:r.monteiro@leibniz-lib.de), ORCID 0000-0003-1374-4474.

29 **Rebekah A. Oomen**, (1) Centre for Ecological & Evolutionary Synthesis, University of Oslo,  
30 Blindernveien 31, 0371 Oslo, Norway, (2) Natural History Museum, University of Oslo, P.O. Box  
31 1172, Blindern, 0318 Oslo, Norway, (3) Centre for Coastal Research, University of Agder,  
32 Universitetsveien 25, 4630 Kristiansand, Norway. [rebekahoomen@gmail.com](mailto:rebekahoomen@gmail.com), ORCID 0000-  
33 0002-2094-5592.

34 **Olga Vinnere Pettersson**, Science for Life Laboratory - Sweden (SciLifeLab), National  
35 Genomics Infrastructure, Uppsala University, P.O. Box 815, SE-752 37 Uppsala, Sweden.  
36 [olga.pettersson@scilifelab.uu.se](mailto:olga.pettersson@scilifelab.uu.se), ORCID 0000-0002-5597-1870.

37 **Torsten H. Struck**, Natural History Museum, University of Oslo, P.O. Box 1172, Blindern, 0318  
38 Oslo, Norway. [t.h.struck@nhm.uio.no](mailto:t.h.struck@nhm.uio.no) ORCID 0000-0003-3280-6239.

## 39 Acknowledgements

40 We thank all members of the ERGA SSP committee and the committee's meeting participants  
41 for their support to the SSP and ERGA mission. In particular, we thank Alice Minotto and Felix  
42 Shaw from COPO, Josephine Burgins and Joana Pauperio from the EBI/EMBL, as well as Luisa  
43 Marins (Leibniz Institute for Zoo and Wildlife Research) ~~and Rita Monteiro (Leibniz Institute for~~  
44 ~~the Analysis of Biodiversity Change)~~ for their help in implementing the ERGA manifest. We thank  
45 the Samples Working Group of the Darwin Tree of Life Project for a fruitful exchange on  
46 metadata collection and standards. We thank ERGA's Data Analysis Committee for access to  
47 questionnaire data used in Figure 1. We acknowledge the essential work of Giulio Formenti and  
48 Alice Mouton, members of the ERGA Pilot Project coordination team, in making this work  
49 possible by contributing to build the necessary sample metadata collection infrastructure,  
50 including the early establishment of the sample manifest collection process, the ERGA manifest  
51 Github repository, as well as with their constant coordination effort as part of the ERGA Pilot  
52 Project daily activities. We especially thank the ERGA chairs for fruitful exchanges and their  
53 continuous support during the establishment phase of ERGA.

54 R. Oomen was supported by the James S. McDonnell Foundation 21st Century Postdoctoral  
55 Research Fellowship, the Natural Sciences and Engineering Research Council of Canada  
56 Postdoctoral Research Fellowship, and the Research Council of Norway (Earth BioGenome  
57 Project Norway; Project no. 326819). R. Fernández acknowledges support from the following  
58 sources of funding: Ramón y Cajal fellowship (grant agreement no. RYC-2017-22492 funded by  
59 MCIN/AEI /10.13039/501100011033 and ESF 'Investing in your future'), project PID2019-  
60 108824GA-I00 funded by MCIN/AEI/10.13039/501100011033, and by the European Research  
61 Council (ERC) under the European's Union's Horizon 2020 research and innovation programme  
62 (grant agreement no. 948281). O. Vinnere Pettersson is supported by RFI/VR and Science for  
63 Life Laboratory, Sweden. S. McTaggart was supported by the Biotechnology and Biological  
64 Sciences Research Council (BBSRC), part of UK Research and Innovation, through the Core  
65 Capability Grant BB/CCG1720/1 and the Earlham Institute Strategic Programme Grant  
66 Decoding Biodiversity BBX011089/1 and BBS/E/ER/230002B at the Earlham Institute. J. Melo-  
67 Ferreira acknowledges support from FCT, Fundação para a Ciência e a Tecnologia  
68 (2021.00150.CEECIND contract and project HybridChange, PTDC/BIA-EVL/1307/2020, via  
69 Portuguese national funds). T. H. Struck acknowledges funding from the Research Council  
70 Norway (project number 300587). A. Böhne acknowledges support from the German Research  
71 Foundation DFG (grant numbers DFG 497674620 and DFG 492407022) and the Leibniz  
72 Association. A. Böhne, R. Monteiro, R. Oomen, T. Struck, R. Fernandez, S. McTaggart, J. Melo-  
73 Ferreira, J. A. Leonard and O. Vinnere Pettersson were funded by Horizon Europe under the  
74 Biodiversity, Circular Economy and Environment (REA.B.3); co-funded by the Swiss State  
75 Secretariat for Education, Research and Innovation (SERI) under contract number 22.00173;  
76 and by the UK Research and Innovation (UKRI) under the Department for Business, Energy and  
77 Industrial Strategy's Horizon Europe Guarantee Scheme. We would also like to acknowledge  
78 the contributions of the biodiversity genomics initiatives that contributed data concerning their  
79 metadata acquisition processes including Pine Biotech Omiclogics, Vertebrate Genomes  
80 Project, Bio and Emerging Technology Institute, The Ira Moana Project, Aotearoa Genomic Data  
81 Repository, Diversity of the Indo-Pacific Network, Earlham Institute, the Yoder Lab, Catalan

82 BioGenome Project, African BioGenome Project, Squalomix University of Port Harcourt, Global  
83 Genome Initiative, Genomic Observatories Database, Beenome 100 Project, National History  
84 Museum of Los Angeles County, Threatened Species Initiative, Bioplatforms Australia, Darwin  
85 Tree of Life, Aquatic Symbiosis Project, and InverOmics.  
86 Traditional Knowledge and Biocultural Label and Notice development: The implementation of  
87 the Labels and Notices and the development of the supporting guidance documentation was  
88 funded through the European Open Science Cloud (RDA\_OSF\_EOSC-228) in partnership with  
89 the Global Indigenous Data Alliance, RDA and Local Contexts.

## 90 Abstract

91 The European Reference Genome Atlas (ERGA) consortium aims to generate a reference  
92 genome catalogue for all of Europe's eukaryotic biodiversity. The biological material underlying  
93 this mission, the specimens and their derived samples, are provided through ERGA's pan-  
94 European network. To demonstrate the community's capability and capacity to realise ERGA's  
95 ambitious mission, the ERGA Pilot project was initiated. In support of the ERGA Pilot effort to  
96 generate reference genomes for European biodiversity, the ERGA Sampling and Sample  
97 Processing committee (SSP) was formed by volunteer experts from ERGA's member base. SSP  
98 aims to aid participating researchers through i) establishing standards for and collecting of  
99 sample/ specimen metadata; ii) prioritisation of species for genome sequencing; and iii)  
100 development of taxon-specific collection guidelines including logistics support. SSP serves as  
101 the ~~sample provider's~~ entry point [for samplesamples providers](#) to the ERGA genomic resource  
102 production infrastructure and guarantees that ERGA's high-quality standards are upheld  
103 throughout sample collection and processing. With the volume of researchers, projects,  
104 consortia, and organisations with interests in genomics resources expanding, this manuscript  
105 shares important experiences and lessons learned during the development of standardised  
106 operational procedures and sample provider support. The manuscript details our experiences in  
107 incorporating the FAIR and CARE principles, species prioritisation, and workflow development,  
108 which could be useful to individuals as well as other initiatives.

## 109 I. The Sampling and Sample Processing committee of 110 ERGA

111 The European Reference Genome Atlas ([ERGA, Mazzoni et al. 2023](#)) consortium, the European  
112 node of the [Earth BioGenome Project](#) (EBP; Lewin et al. 2022), aims to generate a publicly  
113 available reference genome catalogue for all European eukaryotic biodiversity (Formenti et al.  
114 2022; Theissingner et al. 2023). ERGA has the potential to catapult the fields of biodiversity  
115 conservation, evolution, ecology, and others to a new sphere analogous to how the first complete  
116 sequence of the human genome surged the fields of medical genetics, genomics, anthropology,  
117 and others (Formenti et al. 2022; Theissingner et al. 2023). It is akin to the appearance of the first  
118 natural history collections dating back as far as the 1800s that still lay the foundations for many  
119 new and important insights today.

120 ERGA is led by its chair and two co-chairs in cooperation with the ERGA council (a team  
121 consisting of two elected representatives of each member country). To support the multitude of  
122 ERGA tasks, [several scientific and Science+ committees](#) have been established, ~~one of which~~  
123 ~~is the Sampling and Sample Processing committee (SSP)~~. ERGA's first project - [the ERGA Pilot](#)  
124 [\(McCartney et al. 2023\)](#), tested a distributed genomics infrastructure while fuelling the ERGA  
125 committees. The Pilot Project is a community effort without a dedicated funding source, which  
126 will result in the production of over 98 genomes from 34 provider countries, connecting close to  
127 400 involved ERGA members.

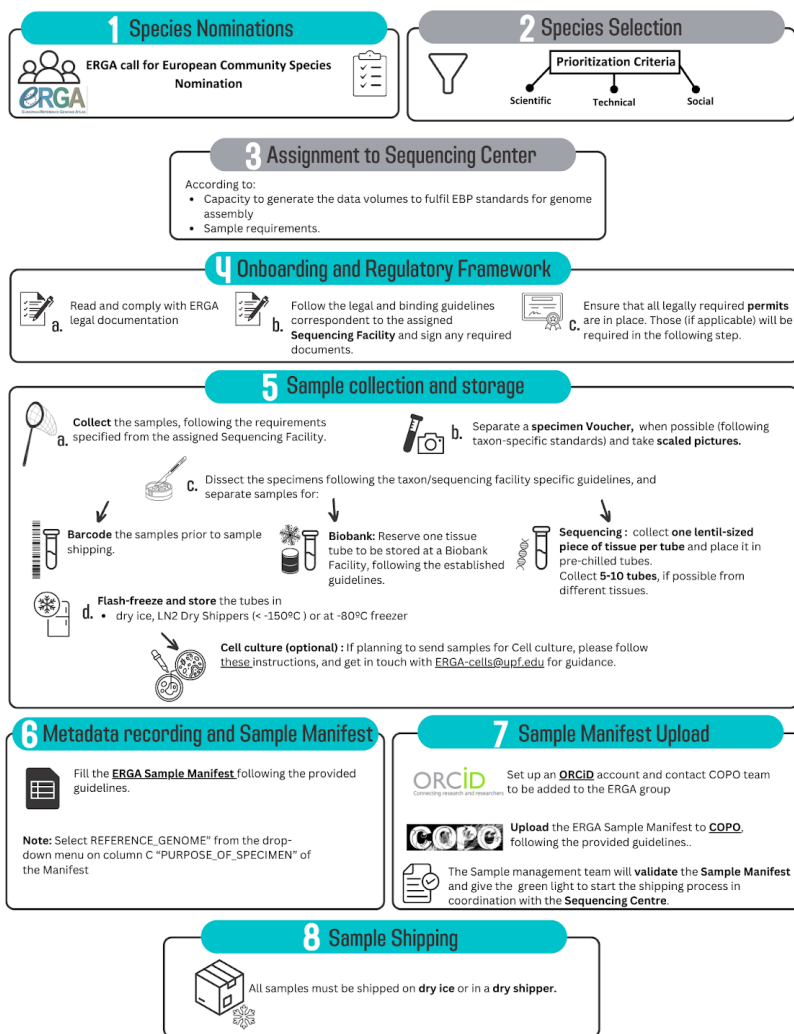
128 [The Sampling and Sample Processing committee \(SSP\)](#) ~~SSP~~ is a ~~working-committeegroup~~ of  
129 volunteer expert ERGA members tasked with developing guidelines to support sampling and  
130 sample processing. Specifically, the SSP's initial responsibilities included i) establishing  
131 standards and mechanisms to collect sample/specimen metadata; ii) prioritising species  
132 collection; and iii) developing taxon-specific collection guidelines for the biological material  
133 underlying ERGA's mission. The specimens and their derived samples are provided through  
134 ERGA's large network of biodiversity partners spread across Europe (Box 1).

135 The SSP serves as the sample provider's entry point into ERGA's distributed genomic  
136 infrastructure and helps ensure standardised sample processing. As ERGA was maturing,  
137 additional SSP tasks emerged: iv) providing guidance to sample providers for the compliance  
138 with legal obligations in collaboration with ERGA's [ELSI committee \(Ethical, Legal, and Social](#)  
139 [Issues\)](#) and v) sample provision - facilitating sample shipping between sample providers and  
140 sequencing centres.

141 As the number of EBP-associated projects across the globe gradually increases, we share here  
142 the experiences we gained whilst developing the operational procedures and sample provider  
143 support systems for the first continent-wide, distributed, genomics infrastructure. We hope our  
144 lessons can be useful to other large consortia who are pursuing the shared mission of  
145 sequencing all of life. Our experience in tackling [FAIR](#) (Findable, Accessible, Interoperable,  
146 Reusable) and [CARE](#) (Collective benefit, Authority to control, Responsibility, Ethics) data  
147 principles, species prioritisation, and workflow development may also be of use to smaller  
148 initiatives.

## 149 II. The sample flow within ERGA

**Box1.** The scheme shows the ERGA workflow in the Pilot project. Species were initially nominated by the ERGA community (1), accompanied by a comprehensive form containing questions used for Species Selection (2), based on several exclusion, prioritisation and feasibility criteria. Species were distributed to the participating Sequencing Partners (3), which were responsible to contact the Genome Team lead (often the sample provider) to organise all necessary onboarding and regulatory requirements and documentation and agreed to generate reference genomes that fulfil [EBP quality metrics](#) (4). Samples were collected, vouchered, and several tubes of subsamples were prepared for sequencing as arranged with the sequencing partner and collaborating research groups (5). Sample providers were also encouraged to barcode the samples prior to sequencing and to store corresponding material in local biobanking facilities. Metadata was recorded using the ERGA sample manifest following established guidelines (6), uploaded to the metadata brokering platform COPO and validated by the Pilot sample management team (7). After confirmation that all the required documentation and metadata was in place, samples were shipped assuring a cold chain to the designated sequencing facility (8).



150 Reference genome production within a multinational consortium like ERGA involves many  
 151 partners spanning dozens of countries. To manage diverse expectations, ensure efficient task  
 152 execution, streamline communication, and safeguard fair attribution, ERGA has implemented  
 153 the formation of multidisciplinary ‘Genome Teams’ (Supplementary File 1). These include all  
 154 contributors to the production of a reference genome (i.e., researchers, stakeholders, and rights  
 155 holders) from the field to the final data analysis. The Genome Team lead’s (in the ERGA Pilot  
 156 known as the sample ambassador) initial responsibilities include providing all necessary  
 157 documentation, data, and metadata for a sample to enter the sequencing workflow (Box 1). Most  
 158 often, this function is filled by the sample provider. All members of the Genome Team agree to  
 159 adhere to [ERGA’s Sample Code of Practice](#) as well as [ERGA’s Code of Conduct](#). The SSP  
 160 committee serves as an important touch point for the Genome Team lead, providing advice and  
 161 guidance on sampling requirements, metadata standards, legal compliance, and vouchering  
 162 strategies.

163 **Selecting species for biodiversity genomics - species prioritisation**  
 164 **in ERGA’s initial phase**

165 Reference genome sequencing initiatives require implementing prioritisation criteria, given  
 166 resource and technical limitations that prevent sequencing all targeted species immediately.  
 167 Scientific, technical, and social criteria can govern such species prioritisation.

168  
 169 **Table 1** Non-exhaustive list of criteria for species prioritisation for genome sequencing projects

Criteria	Scientific criteria	Technical criteria	Social criteria
Examples	taxonomic representation/targets	sample availability including voucher specimen	importance to local communities
	conservation status	specimen/sample size (amount of biological material and therefore DNA and/or RNA)	cultural significance
	value of genome for specific field of interest (e.g., biomedicine, biotechnology, agriculture)	sampling and handling logistics	inclusiveness targets concerning countries and individuals
	Taxonomic certainty	genome characteristics (estimated genome size and ploidy)	community engagement

170  
 171  
 172 For initiating ERGA as a continent-wide genomic infrastructure network, a pool of candidate  
 173 species for reference genome generation was solicited that were representative of the diversity  
 174 of species and scientists across the consortium. To this aim, the ERGA community was asked  
 175 to propose species through an initial simple ERGA species suggestion form resulting in 276  
 176 nominations. Subsequently, nominating persons were contacted to complete a comprehensive

177 form (Supplementary File 2) containing 117 questions and commenting fields. The form included  
178 questions related to taxonomic identity, genome properties, voucher availability, habitat of  
179 species in question, sampling strategy, species conservation status, permits to obtain material  
180 for genome sequencing, sample properties (e.g., sex, amount, preservation quality, and tissue  
181 type), and species identification certainty. The refined species nomination form was open for 26  
182 days and received 155 submissions.

183 After excluding species that already had available reference genomes, SSP implemented a  
184 prioritisation process based on country of origin and a simple scoring system, attributing a score  
185 of 1 to 3 in eight categories (Table 2). Higher priority was given to species that: i) had a genome  
186 size smaller than 1Gb, ii) were readily available, iii) could be freshly collected and for which  
187 biological material could be flash frozen, iv) could deliver >1g of tissue (if the organism permitted)  
188 and had well-established extraction protocols that allowed isolating chemically pure HMW DNA,  
189 v) could deposit a specimen voucher, vi) had no ambiguity risk in species identification, vii) had  
190 all permits present or were not needed (a formal documentation for either of the solutions was  
191 requested), and viii) had no export restrictions (if applicable).

192 After ranking the species according to this scoring system, each proposing country was given  
193 the opportunity to refine their selection of species and to propose three final species considering  
194 three predefined target categories (endangered/iconic, marine/freshwater and pollinator) to  
195 match the available resources. At that point, ERGA had no centralised funding so feasibility was  
196 strongly determined by the availability of sufficient funds to support genome sequencing for a  
197 particular species. The project relied on resources contributed by participating ERGA members,  
198 institutions, and sequencing centres, with some additional support from industrial sponsors, that  
199 was used to supplement equity deserving genome teams in order to improve wide access to  
200 participation. As an extension to the selected list, standalone species were also included under  
201 the ERGA umbrella if they were completely funded by independent resources.

202 The circulation of the list of nominated species within ERGA resulted in cross-country  
203 collaborations especially for species proposed by more than one country, fostering exchange  
204 and reducing costs and redundancies.

205 The species selection and prioritisation process resulted in 98 selected species  
206 (<https://goat.genomehubs.org/projects/ERGA>), from 15 phyla (Figure 1B) and 34 countries or  
207 regions (Figure 1). With six of the seven selection scores relating to feasibility (including legal),  
208 this was the most prominent criterion, while the other criteria (i.e., conservation status, scientific  
209 relevance, socioeconomic relevance, taxonomic gaps, and community engagement) played only  
210 an indirect role via the subjective selection by the ERGA council members. [ERGA has planned  
211 to implement unbiased species selection procedures in the future to alleviate the dominance of  
212 feasibility as selection criterion \(see section V below\).](#)

213 Both the initial and the final list of selected species showed a predominance of chordates,  
214 arthropods, and tracheophytes. Given that the initial pool of species was suggested by the ERGA  
215 community, this predominance may reflect the organism-bias of the biodiversity genomics  
216 community at large (see below). This taxon bias remained despite the dynamic nature of the  
217 taxonomic composition, as some species were removed due to sampling or sequencing

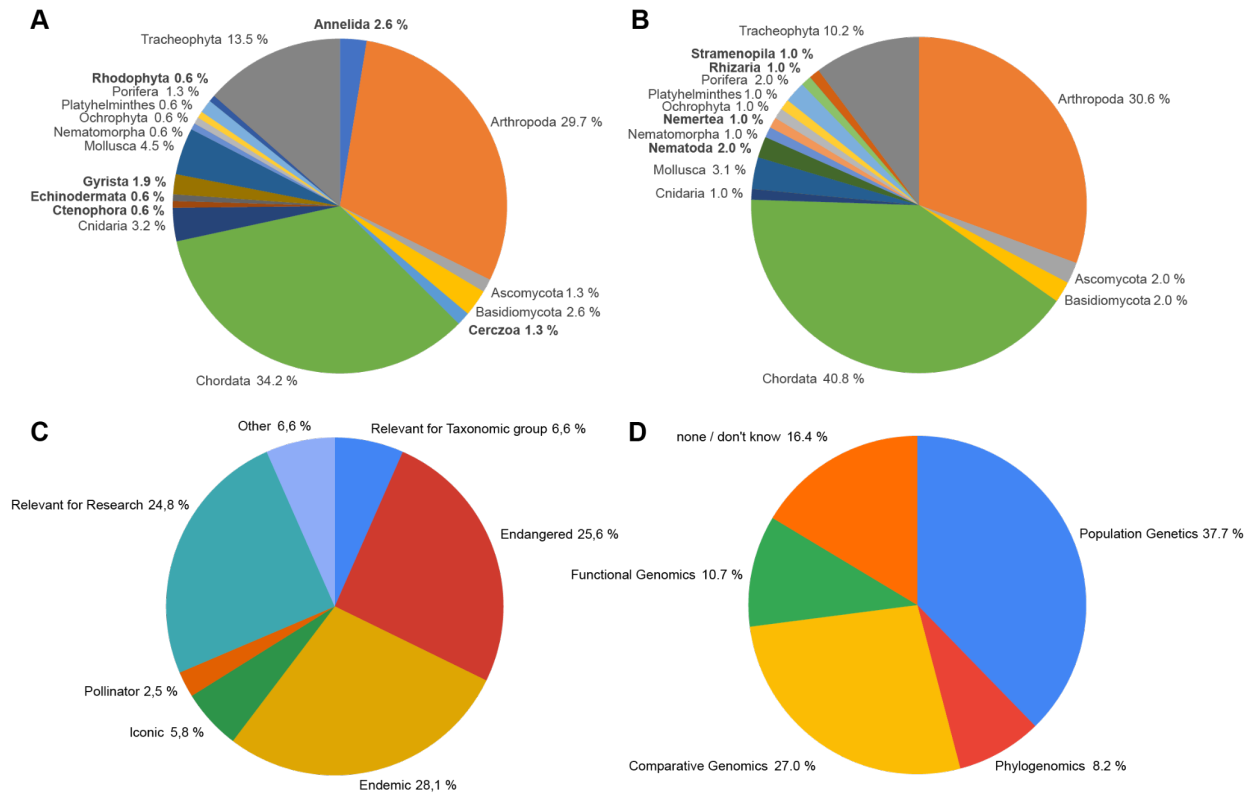


218 technical barriers whilst others were added to increase representation and participation across  
 219 ERGA's diverse members. A total of 37% of the species were considered for the category  
 220 endangered/iconic, and 12% were pollinators (as one example of scientific relevance and a  
 221 target group of the Biodiversity Strategy of the European Commission). Most of the reference  
 222 genomes were generated because the species are endemic (28%), endangered (26%) (and  
 223 therefore the genome could be leveraged to inform conservation plans in the future) or to be  
 224 used to answer specific scientific questions for research purposes (25%) (Figure 1C); and †The  
 225 most popular planned downstream analyses involve population genomics (38%) or comparative  
 226 genomics (27%) (Figure 1D) (data from a questionnaire to species ambassadors, done by  
 227 ERGA's Data Analysis Committee, DAC, in the framework of Mc Cartney et al. (2023)).

228 Regarding inclusiveness, of the 18 Widening countries represented in the ERGA council 17 had  
 229 at least one species included in the final list of generated reference genomes. The representation  
 230 of ITC (Inclusiveness Target Countries) and Widening countries with 44% and 50 % of the 34  
 231 countries suggesting species is good overall~~good overall~~ good. However, only 36% or 42 % of  
 232 the final species came from ITC or Widening countries, respectively.

233  
 234 **Table 2** Feasibility criteria scoring for species suggested as sequencing targets of the ERGA Pilot  
 235 Project

Category	1	2	3
<b>Genome size</b>	<1Gb	1-3Gb	>3Gb
<b>Sample Availability</b>	Until end April 2020	May-June 2020	July 2020 or after
<b>Sample Preservation</b>	Freshly collected, flash frozen, -80°C, no preservative, never thawed	in-between 1 and 3 (to be evaluated by sequencing centre)	Not freshly collected and/or thawed several times, and/or not kept in -80°C
<b>Sample Size</b>	>1g	100mg-1g	<100mg
<b>Suitability for HMW DNA</b>	Already extracted or taxon known to work well (e.g., vertebrates)	Not tested and not known for the taxon (can be checked with sequencing centres)	Inhibitors known to make DNA extraction and/or sequencing very challenging
<b>Voucher &amp; SpeciesID</b>	Voucher kept in collection and no ambiguity in species identification		No voucher and/or ambiguous species identification
<b>Sampling Permits</b>	Yes or Not needed (documentation required either way)	Pending	No when needed or No documentation
<b>Export Regulations</b>	No restrictions between countries where sample will be handled or entire sequencing performed within country	Indexed to conservation status or Nagoya regulations to be clarified	No possibility for obtaining needed permits



236  
237  
238  
239  
240  
241  
242

**Figure 1** Pie charts of the number of species per phylum that were suggested for the ERGA Pilot Project at the beginning (A) and that are on the list of genomes realised or in production as of April 25th 2023 (B). The phyla are indicated together with the percentage of species per phylum. Phyla, which are different between A and B, are highlighted in bold. Additionally, the criterion for choosing the species (C) and the planned downstream analyses (D) are provided in percentages.

243  
244

## II. FAIR and CARE principles, Metadata Collection and Brokering

245

### FAIR and CARE principles

246  
247  
248  
249  
250  
251  
252  
253

As the number of initiatives working towards complete reference genomes for all of eukaryotic life are increasing, so too is the demand for freshly collected, wild specimens. This provides an opportune and pertinent moment to revisit biodiversity genomic metadata standards to ensure they are both scientifically comprehensive and also align with current ethical, legal and social standards for data governance. Ensuring that data are findable, accessible, interoperable and reusable (FAIR) is fast becoming a central dogma of the biodiversity genomics community (Wilkinson et al. 2016)<sup>1</sup>. Throughout the metadata standard development process (see next section), SSP intentionally and carefully aligned all ontologies to the FAIR principles to safeguard

<sup>1</sup>FAIR was introduced by Wilkinson et al. (2016), which has since been accessed 580,000 times and cited 5,636 times

254 that all ERGA data would have a maximised scientific potential, increased re-usability, and  
255 greater longevity.

256 Indigenous Peoples and Indigenous knowledge systems have, and continue to be, treated as  
257 subordinate and outside of western science, specifically when considering contextual metadata  
258 (Turner 2022). This has had the systematic consequence of severing the connection between  
259 Indigenous Peoples and Local Communities with their samples and data. To mitigate the  
260 manifestation of this exclusion within ERGA, SSP developed new metadata ontologies to  
261 support the disclosure of Indigenous rights and interests by Indigenous Peoples [by sample](#)  
262 [providers](#). This purposeful inclusion and recognition of Indigenous Peoples and their rights  
263 actualises the CARE principles of Indigenous data governance (Carroll et al. 2021) whilst  
264 simultaneously working in complementary fashion to the FAIR principles. By creating this space  
265 at the entry point into ERGA processes, i.e., sample provisioning, SSP provided an opportunity  
266 for Indigenous Peoples and knowledge systems to permeate throughout the process of  
267 reference genome production and beyond (Figure 2). By operationalizing the FAIR and CARE  
268 principles across the metadata ontologies developed, ERGA members are supported to  
269 responsibly and openly share data.

## 270 ERGA Manifest for Metadata Collection and Brokering

271 Developing consortium-wide procedures for metadata collection is an opportunity to set a  
272 minimum standard of excellence, and ensures consistency across datasets. This approach is  
273 also a challenge since an unintentional exclusion of an important metric will lead to its systematic  
274 erasure from all data produced by the consortium. To support ERGA's sampling process, SSP  
275 implemented the consortium's first metadata standard, the [ERGA manifest](#), and its supporting  
276 documentation (standard operating procedure (SOP)). This SOP and manifest were built on pre-  
277 existing standards that were developed for an established reference genome production  
278 initiative, [Darwin Tree of Life](#) (Lawniczak et al., 2022; Shaw et al., 2022), which followed the  
279 [Darwin Core standard](#). The manifest supports ERGA's goal to collect standardised, high-quality  
280 metadata that remains linked to the genome across the relevant repositories. The highly detailed  
281 SOP facilitates completing the ERGA manifest by the Genome Team lead who is responsible to  
282 provide information on: 1) sample identifiers (e.g., field and tube numbers, Genome Team lead),  
283 2) taxonomic details, 3) sample type (e.g., life stage, organism part), 4) the sequencing partner,  
284 5) sample collection event, 6) taxonomic identification and uncertainty, 7) sample preservation,  
285 8) DNA barcoding, 9) biobanking and vouchering, 10) regulatory compliances including  
286 Indigenous rights and traditional knowledge, and 11) other relevant comments from the Genome  
287 Team representative.

288 [The SOP explains every data point asked for, links to explanatory resources such as tutorial](#)  
289 [videos, and help contacts.](#)  
290 [Expert members of SSP, i.e., sample managers, help genome teams upon request with filling in](#)  
291 [metadata fields and choosing appropriate terms in case of doubt. Sample managers can also](#)  
292 [check investigate-manifests prior to submission to avoid frustrating periods of trial and error for](#)

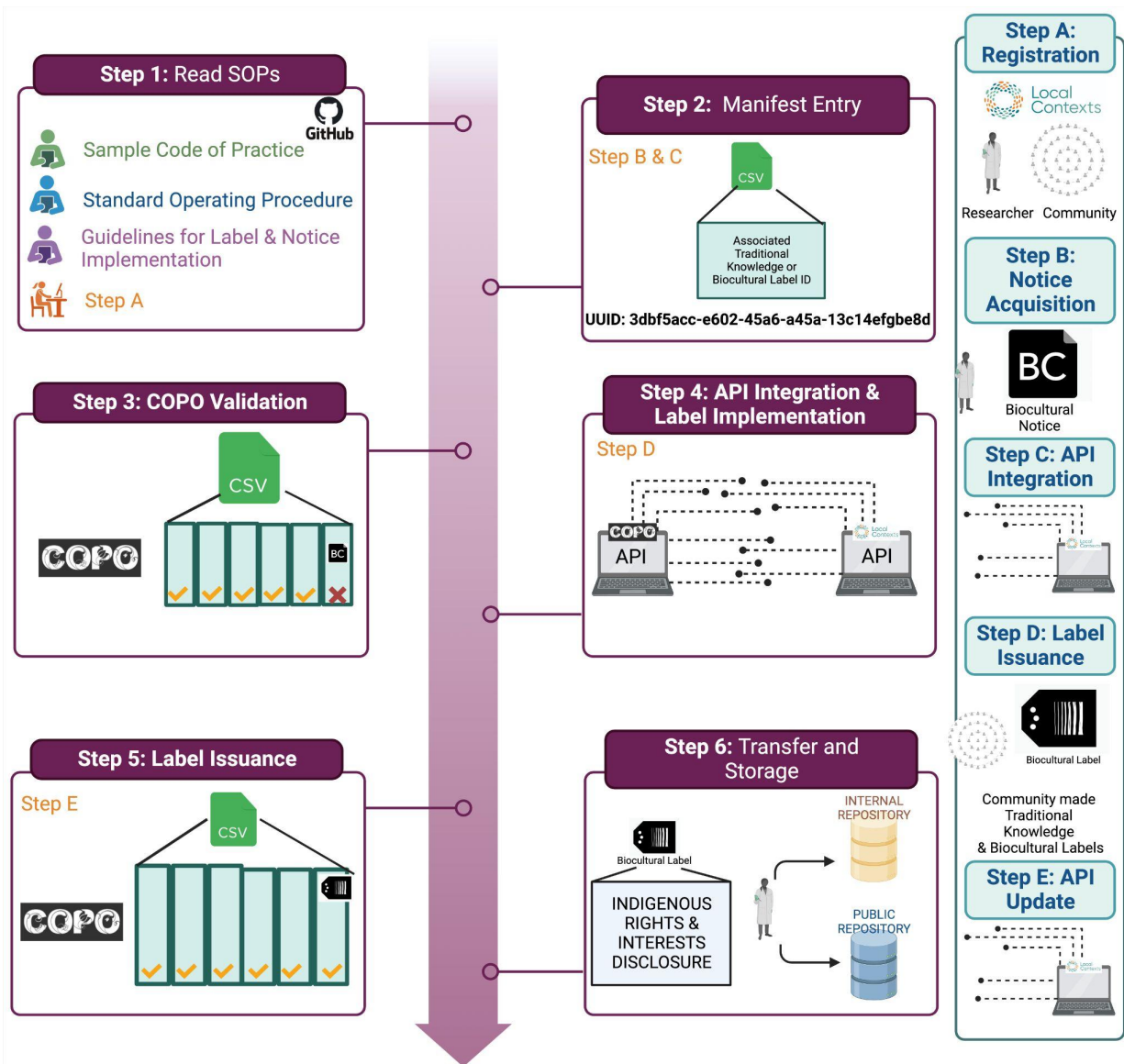
293 [sample providers. Based on continuous user feedback, the SOP is updated twice a year under](#)  
294 [constant revision to facilitate metadata collection for genome teams.](#)

295 Upon upload of the manifest through the metadata brokering platform [COPO](#) (Shaw et al., 2020),  
296 metadata fields are validated against predefined standards and checklists to ensure terms and  
297 formats meet both ERGA and data repository expectations. [Guidance to this process is provided](#)  
298 [through a visual guide on the COPO help webpage.](#)

299 Upon manifest validation by the sample managers (~~part of SSP~~), an indicated set of mandatory  
300 metadata fields are brokered to the [European Nucleotide Archive](#) (ENA) under a dedicated  
301 [BioSample](#) entry ultimately connecting the digital sequence data to standardised sample  
302 metadata.

303 To mitigate the risk of missing information important to specific taxonomic groups or habitats due  
304 to own bias (see below), SSP included diverse team members when developing the manifest  
305 and planned for bi-annual updates of the metadata protocol so that accidental exclusions could  
306 be fixed in a timely manner and allow ~~for~~ sufficient implementation and testing [time](#) for front- and  
307 backend [development](#). ~~AnySuch and other~~ issues with the manifest encountered by the  
308 community can be raised in the ERGA manifest GitHub or by contacting the SSP directly. The  
309 ERGA Pilot allowed the SSP committee to test the ERGA manifest on a broad variety of  
310 organisms by a pan-European network of researchers. Guidance for understanding and  
311 implementing the collection of metadata and vouchers was extensively requested [and provided](#)  
312 [by SSP members. Finalisation of the ERGA manifest and its SOP was achieved through](#)  
313 [discussions with other ERGA committees, especially ELSI, and the ERGA coordination.](#) The  
314 ERGA metadata collection is a semi-automated process that is highly scalable, preparing ERGA  
315 for an anticipated increased sample workflow. Validation of the sample manifest is the  
316 checkpoint of transitioning to the sequencing workflow. [In addition, the SOP and manifest are](#)  
317 [under constant critical revision based on user feedback aiming to simplify the collection process.](#)

318 The SSP data collection process links biological material, metadata, and sequence information  
319 in a maximally automatised fashion over open access databases and throughout the genome  
320 workflow from collection through nucleic acid extraction, sequencing, assembly and annotation  
321 steps. While open access genomic information is already a highly appreciated resource,  
322 comprehensive metadata enhances its value by making it more reusable. It is crucial that the  
323 metadata, sample(s), and derived sequence data are linked from the outset, because the  
324 opportunity to link them declines substantially with time (Crandall et al. 2022).



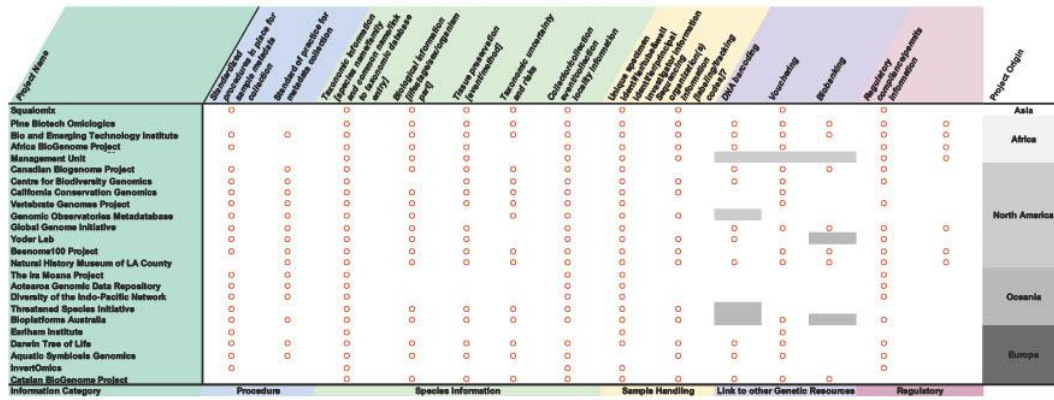
325  
326

**Figure 2** ERGA's Biocultural and Traditional Knowledge Labels and Notices implementation protocol.

## 327 Status Quo of metadata collection amongst biodiversity initiatives

328 To gain an understanding of the diversity and interoperability between the various metadata  
 329 collection procedures being implemented within the community, SSP conducted a survey across  
 330 global biodiversity genomics projects (Figure 3). A total of 24 initiatives that are actively  
 331 generating high-quality reference genomes for non-human species responded, spanning Africa,  
 332 North America, Oceania, Europe and Asia<sup>2\*</sup>.

<sup>2</sup> Notably, the lowest amounts of survey responses were obtained from Asia (the authors note that this is certainly due to our inability to identify appropriate contact points and does not reflect a lower number of biodiversity projects in this continent)



346 **Figure 3** Results summary from the metadata survey conducted across 24 biodiversity initiatives  
 347 worldwide. Red circles within a cell indicate presence, and empty cells indicate absence.

348  
 349 The results indicate that overall, 83% of responding initiatives have a standardised metadata  
 350 collection procedure in place and 67% have an associated SOP to support and guide  
 351 researchers in the metadata submission process. In terms of species-specific metadata  
 352 collection, initiatives prioritise the collection of taxonomic (100%), collection information (96%),  
 353 biological information (75%) and tissue preservation (75%) over providing more fine-grained  
 354 information on the taxonomic uncertainty or risks associated with the species being sampled  
 355 (59%). Almost all initiatives (96%) collected unique specimen and tube/well identifiers as well as  
 356 the associated principal investigators whereas just 67% required information about the  
 357 sequencing facility.

358 The amount of metadata collected about other associated genetic resources from the species  
 359 sample was relatively low. For instance, only 55% of the 20 projects collect DNA barcoding  
 360 information within their metadata. Further, just 65% of initiatives collect vouchers and 33%  
 361 collect cryopreserved samples and require this information as part of their standard metadata  
 362 collection processes. Finally, 42% of initiatives required some kind of disclosure of regulatory  
 363 compliance and just 33% of projects required metadata concerning associated Indigenous rights  
 364 and interest.

### 365 Scaling Legal Compliance

366 SSP also focussed on creating an infrastructure that supports and promotes legal as well as  
 367 ethical and scientifically sound sample collection. As an initial safeguard, SSP supported ERGA  
 368 to develop a document of best practices for ethical and legal sample collection ([ERGA Code of](#)  
 369 [Conduct](#)). All researchers participating in the Pilot were required to agree to these practices in  
 370 advance of making their metadata manifest submission. These practices detailed expectations  
 371 surrounding local, regional, national, and international permitting in addition to how to ethically  
 372 collect samples to minimise harm.

373 Further, the ERGA manifest contained seven metadata fields regarding the regulation and permit  
 374 requirements for each sample. [These questions comprise comprehensively all permit forms that](#)

375 [could be required to obtain a sample for genome sequencing: i\) initial question if regulatory](#)  
376 [compliance is required and adhered to, ii\) Applicability of traditional knowledge or biocultural](#)  
377 [rights with subsequent collection of rights definition, project ID provided by the Local Context](#)  
378 [Hub and contact information iii\) Request for ethics permit applicability, definition and permit iv\)](#)  
379 [Request for sampling permit applicability, definition and permit and v\) Request for Nagoya](#)  
380 [Protocol permit applicability, definition and permit. This comprehensive request for applicability](#)  
381 [and documentation of compliance raises awareness also for sample providers to respect all](#)  
382 [regulations.](#)

383 In partnership with COPO, ERGA required the mandatory upload of permits during the manifest  
384 submission process. Expert personnel within ERGA were alerted when a permit had been  
385 uploaded into the directory and, where possible, confirmed the appropriate permits had been  
386 obtained.

## 387 The importance of vouchers for biodiversity genomics


388 Voucher specimens in natural history collections are benchmarks against which we compare the  
389 world around us. They illuminate how the world has been changing, and especially how we have  
390 been changing the world. Reference genomes are a new benchmark. Vouchering is critical to  
391 genomics because it provides a permanent, verifiable, and accessioned record of the identity of  
392 the organism being sequenced and, in some cases, a sample of its genetic material (biobanking).  
393 When determining which of the many available vouchering methods is most appropriate,  
394 consideration should be given to e.g., the taxon, its size, its conservation status (Table 3). The  
395 SSP determined that a sample voucher helps contextualise the biology of the organism and thus  
396 increases the probability that the sequencing data generated will be [aligned with FAIR principles](#)  
397 [aligned](#) and useful into perpetuity.

398 A driving rationale for vouchering is the fluid nature of taxonomy, as new scientific insights lead  
399 to changes in the classification of species. As this happens, the prescribed identity assigned to  
400 a sequenced individual could be questioned. In such cases, the presence of a voucher can be  
401 used to re-examine the species to confirm, or alternatively revise and update, its identity.  
402 Furthermore, vouchers can improve data quality assurance, reduce the risk of data corruption,  
403 and eliminate the propagation of confusion when a taxonomic revision has taken place.

404 Even for taxonomically stable groups, a voucher specimen provides the possibility to join  
405 morphological and genome sequence information and verifies the specimen/ species from which  
406 the genome was produced. A physical voucher can also be used for other analyses, including  
407 photographic, x-ray, CT imaging, and/or chemical analyses such as stable isotopes. A  
408 biobanked sample could unlock opportunities for future exploration (e.g., RNA, secondary  
409 genetic marker analyses such as methylation).

410  
411

**Table 3** Vouchering methods available to specimens destined for genome sequencing. Note that multiple voucher types may be made for a single genome.

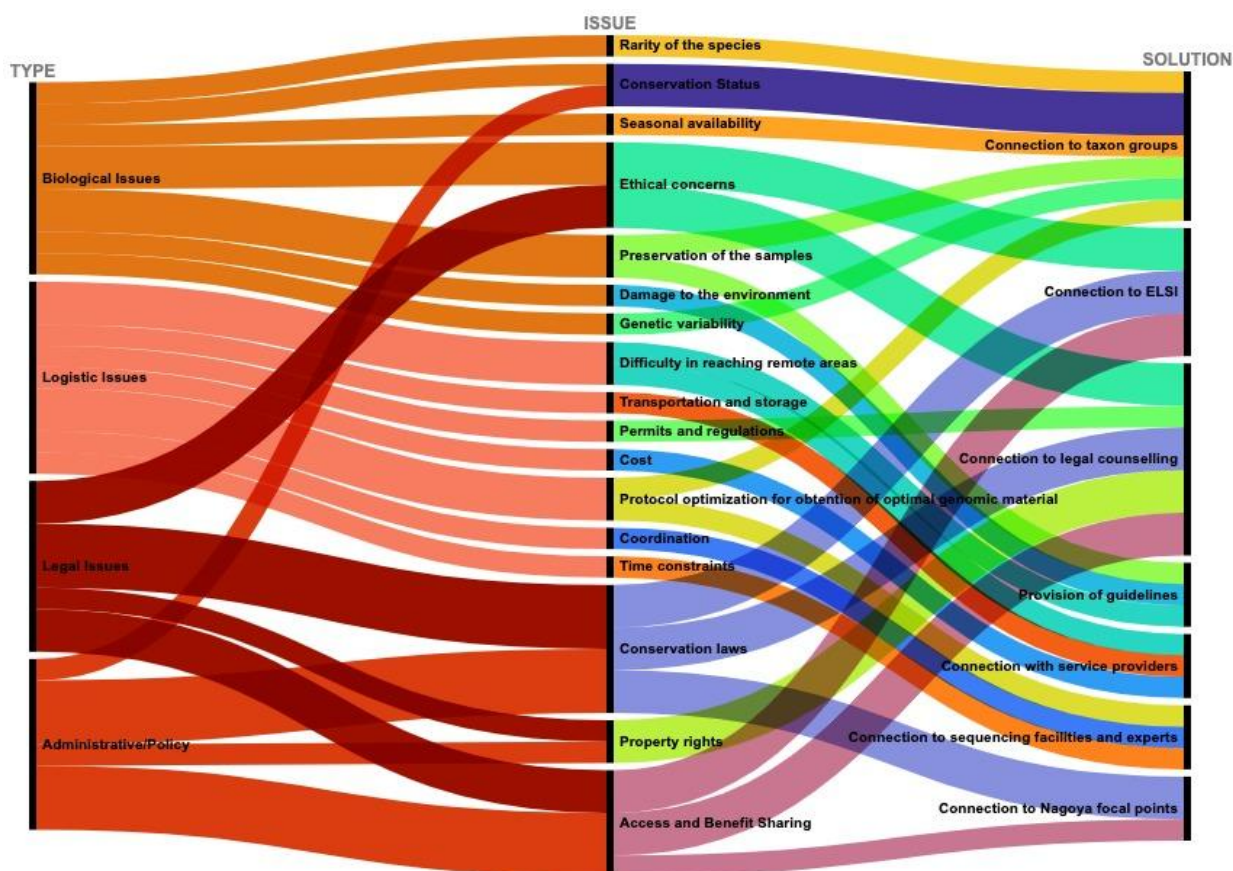
Desirability	Voucher type	Description	Suitable for	Potential Issues
<p style="text-align: center;"><b>High</b></p>  <p style="text-align: center;"><b>Low</b></p>	Primary voucher	Whole organism is preserved and deposited in a permanent collection. Vouchers can be dried, in a preservation liquid (ethanol), or frozen (e.g., biobanked tissue or cell culture vouchers).	Species that are of a suitable size for a permanent collection (taxon-specific), and can be legally and ethically collected	<ul style="list-style-type: none"> <li>• Not possible for very large/small species.</li> <li>• Species might be too rare to sacrifice for a voucher.</li> <li>• Preservation method determines possible additional future uses.</li> </ul>
	Secondary voucher: to complement - not replace - whole organism vouchering	E-voucher: digital image taken of whole organism and of diagnostic characteristics	Small species requiring destructive sampling to obtain sufficient genetic material for a high-quality genome assembly (e.g., single-cell protist)	<ul style="list-style-type: none"> <li>• Can require specialist equipment and expertise (e.g., microscope imaging of insect genitalia).</li> <li>• May have limited use in taxonomic identification.</li> <li>• Diagnostic characteristics may not be known.</li> </ul>
	Partial Voucher: tissue samples are taken, preserved, curated and stored in permanent collections.	For very large organisms (e.g., a whale), or very small (e.g., small insects), where preservation of the whole organism is not feasible.	<ul style="list-style-type: none"> <li>• Body part/tissue taken may not represent diagnostic taxonomic characteristics</li> </ul>	
	Proxy voucher: a sample that identified as the same species to be sequenced, and was collected from the same time and location	Species that are too small for direct or partial vouchering (e.g., bryophyte)	<ul style="list-style-type: none"> <li>• May not be the same as the sequenced species</li> </ul>	



412  
413  
414  
415  
416  
417  
418  
419  
420

## IV. Sample provision: connecting genome teams with sequencing centres

Sampling and sample transfer can be a complicated endeavour with its multilayer complexity arising from ~~three~~ four main categories: biological, logistic, administrative/policy and legal issues. These challenges can strongly influence the outcome of the project and impede the proper transfer of the samples to a sequencing centre (Box 1). The role of SSP is key to overcoming these issues and ensuring the legal, ethical, and timely flow of samples from sample collectors to sequencing centres (Figure 4).



421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432

**Figure 4** The role of SSP supporting critical issues prior to and after sample collection. Type of issue affecting sample provision, description of issues and solutions are indicated.

The distributed genomic infrastructure developed by ERGA promoted and supported the decentralisation of sequencing efforts across Europe. While many sampled species were sequenced within their country of origin, others were shipped to an international sequencing centre. Regardless of the length and duration of shipment involved, ERGA recommended cold-chain shipment, which is necessary to preserve the integrity of nucleic acids. [Since this can be a challenge for sample providers, ERGA tried to connect sample providers with sequencing centres that were geographically close in proximity and aided in sample transportation within the ERGA network. Maintaining the integrity of nucleic acids](#) This is a prerequisite to meet the EBP

433 standards of genome assembly utilising the current sequencing technology (Dahn et al. 2022).  
434 However, samples are often collected in remote locations, where access to appropriate courier  
435 service is financially not feasible or simply not available, a challenge that the ERGA Pilot also  
436 faced. Further, there is a series of legal procedures that require consideration to ensure  
437 compliance with regulations and safety standards, including, among others, chain of custody  
438 forms (to document the movement of the samples from collection to sequencing), material  
439 transfer agreements (a legal contract between two parties that governs the physical transfer of  
440 the biological samples between them, and which establishes the terms and conditions under  
441 which the materials will be transferred), import/ export permits (that may be required depending  
442 on the country of origin and destination), health certificates (required by some countries to  
443 ensure that the samples do not pose a risk to human or animal health), and/or CITES permits  
444 (required if the samples are from a species protected under the Convention on International  
445 Trade in Endangered Species of Wild Fauna and Flora), as well as ABS/ Nagoya relevant  
446 national implementations, among others. The ERGA Pilot project served as an opportunity to  
447 understand the magnitude and complexity of these needs and actions in a collective manner,  
448 with everyone implicated learning about pieces of information that could make an impact in the  
449 success of the full logistics chain. For instance, we learned that different shipping companies  
450 operate better in certain geographical regions, and that sometimes it is important to ask them  
451 explicitly to refill the dry ice during the transit. We also collectively learned about the bureaucratic  
452 idiosyncrasy of each country with respect to export and import permits and Nagoya protocol,  
453 with some countries being more flexible and others being more restrictive. All these pieces of  
454 information have been shared with SSP and are being leveraged to develop SOPs to facilitate  
455 the transit from species collectors to sequencing centres, and will have a strong impact in the  
456 implementation of larger projects such as Biodiversity Genomics Europe (see below).

## 457 Future taxon-specific best-practice guidelines

458 The biological diversity being sampled by large genome initiatives like ERGA necessitates the  
459 development of targeted best-practice sampling guidelines. [The approach of having different  
460 sampling procedures for different taxa is very commendable as it would—eliminates  
461 complications arising from structural and functional variations between the taxa.](#)

462 Such guidelines are imperative [to ensure](#) ~~so~~ that sampling efforts minimise the number of  
463 samples taken, maximise the data quality, and increase the scientific utility of the sample. To  
464 this end, the SSP will take a taxonomic approach that seeks to balance providing a set of  
465 guidelines that are comprehensive, with enough specificity to support fit-for-purpose sampling,  
466 while simultaneously not providing too much information and materials that may overwhelm field  
467 biologists.

468 To develop these guidelines, separate working groups have been set up for each of the following  
469 broad taxon groups: vascular plants, bryophytes and macroalgae, macroinvertebrates, protists,  
470 soft bodied invertebrates, fungi and lichens, chordates, and arthropods. The goal of each group  
471 is to create a working protocol for the sampling of specimens within that taxonomic group, and  
472 those will follow a set structure to ensure consistency and readability. There is a strong

473 foundation for these protocols (e.g. [dx.doi.org/10.17504/protocols.io.261gennyog47/v1](https://doi.org/10.17504/protocols.io.261gennyog47/v1)). ERGA  
474 has the intention of publishing these guidelines in open access over protocols.io

475 A key challenge in developing these guidelines will be to identify and include experts -taxonomic,  
476 field, and wet lab biologists- who are willing to voluntarily contribute their time and knowledge to  
477 the wider community. [The SSP has reached out to the ERGA repeatedly to gain insight into ERGA members' expertise and connect those to SSP. Based on this effort, SSP establishes communication with sample providers and ERGA member institutions that can provide expertise in e.g. sample handling, storing and species identification. This help is provided over the SSP email contact as well as a dedicated channel in the communication platform keybase \(<https://keybase.io/team/erga.listserv>\).](#) Vice versa, a future challenge will be to work towards an  
483 adoption of these guidelines by the biodiversity community at large. Integrating, documenting,  
484 and distributing this knowledge and 'know-how' is fundamental to ERGA and its umbrella  
485 organisation, the EBP. [Based on experiences in the ERGA pilot, members of the SSP and the ERGA BGE project consult with the EBP samples committee and the EBP executive board in areas where ERGA sees a need for larger adoption of processes and standards.](#)

488

## 489 V. Critical Bias Assessment

490 The biodiversity genomics community is subject to systematic biases that affect the accuracy  
491 and completeness of the produced data, and may limit the meaningfulness of the conclusions  
492 obtained. Bias comes in many forms, which have different impacts. The ELSI/ JEDI committee  
493 is more focused on the human dimension, and the SSP committee focused on country  
494 representation and taxonomic biases described here. ERGA as a consortium of European  
495 researchers is at its foundation intentionally geographically biased, while at the same time  
496 promoting and extending representation and participation of researchers across Europe. In the  
497 Pilot, prioritising this aim over the taxonomic breadth of the generated reference genomes  
498 resulted in the manifestation of taxonomic biases (see above).

499 Unbalanced representation of genomes being sequenced across the tree of life is  
500 common in biodiversity genomics initiatives, causing over-representation of some taxa with data  
501 available in public repositories. Non-model organisms and more "difficult" samples remain under-  
502 investigated because there are few standardised sampling collection, preservation, HMW-DNA  
503 extraction, and library preparation protocols available to manage non-optimal situations (e.g.,  
504 small size, existence of exoskeleton or spicules, presence of substances that impair adequate  
505 DNA extraction or sequencing, etc.). This lack of knowledge on certain taxa reflects the available  
506 taxonomic expertise. For example, experts in vertebrates, certain arthropod and plant groups  
507 are vastly more abundant than for other large taxonomic groups like mollusks, nematodes or  
508 annelids (Capa & Hutchings 2021; Engel et al. 2021), which SSP quickly realised while forming  
509 taxon expert groups (see above).

510 Beyond taxonomy, other sources of representation bias exist in reference genome  
511 projects. Sample bias can occur when samples do not accurately represent the known or  
512 unknown heterogeneity of the taxon being studied. SSP encourages sampling from the type  
513 locality. Habitat bias occurs when samples are more often collected in certain types of habitats  
514 that are more common or more easily accessible, under-representing knowledge about habitat-  
515 specific species (e.g., caves, deep-sea). ERGA aims to target this bias with [calls for funded field](#)  
516 [expeditions](#) to understudied hotspots of biodiversity in Europe. Historical bias can have strong  
517 impacts, as samples collected based on prior knowledge or historical information may not  
518 accurately reflect the current state of diversity.

519 A prime goal of SSP is to raise awareness of the importance of taxonomic representation for  
520 genomics, and biodiversity research more generally, and the study of research deserving  
521 groups, species, populations and habitats. SSP has played a key role in creating a bridge  
522 between taxonomy- and taxon-specific experts with sequencing centres, and aims to create the  
523 conditions to explore the feasibility of genome sequencing for all eukaryotes. Biodiversity  
524 genomics benefits the most when it is inclusive in all aspects. Many hotspots of biodiversity exist  
525 in Europe, and many are positioned in nations and regions that are deserving of additional  
526 support. By creating a European-wide network, SSP aims to support such regions through  
527 capacity and capability building for genomics.  
528

## 529 VI. Where do we head?

530 We believe that overall, sequencing and assembling the initial cohort of species that entered into  
531 ERGA's process was a success story. To a large extent this is thanks to collaboration and  
532 alignment with preexisting, well established biodiversity consortia e.g., DTOL. Similarly, we hope  
533 that our prioritisation efforts, the ERGA metadata manifest, as well as the stewardship of legal,  
534 FAIR and CARE information, can be utilised, improved, or adopted by other biodiversity  
535 genomics projects, national or international, irrespective of the project size. An immediate  
536 example of this is the EU-funded project [BGE - Biodiversity Genomics Europe](#), for which the  
537 ERGA initial phase has set the ground for key procedures of the sampling and sample  
538 processing process. The BGE consortium unites ERGA with the DNA barcoding community  
539 ([BIOSCAN Europe](#)) to promote the use of genomics to study and monitor biodiversity and create  
540 tools to tackle its decline. BGE will establish ERGA as the European node of the [Earth](#)  
541 [Biogenome Project](#) and formalise coordinated efforts, infrastructures and workflows to generate  
542 reference genomes of European species.

## 543 Towards a balanced and strategic prioritisation of species

544 As ERGA moves forward, the biases identified are being reflected upon to iteratively improve  
545 sampling and prioritisation. As dedicated projects are established, such as BGE, the selection  
546 and prioritisation of species for reference genome generation can better approximate governing  
547 principles (see above "Selecting species for biodiversity genomics projects"), and be less  
548 dependent on circumstantial feasibility aspects and funding availability for particular taxa. These

549 governing principles can be explicitly and objectively included into the species prioritisation  
550 process and with a more prominent role, while feasibility will likely remain an important aspect  
551 of species selection. Once priorities are established and weighted, the species selection process  
552 can be fully automated. Building on the first experiences of ERGA, such a process is being  
553 implemented in BGE. This process, which is developed with the larger ERGA community, gives  
554 after a check for technical feasibility more weight to taxonomic diversity, country of sample origin,  
555 countries with little representation in ERGA and involves researchers using including JEDI  
556 criteria (favouring favoring novel sample providers, as well as underrepresented groups, and  
557 involvement of non-scientific communities countries with little representation in ERGA) and  
558 applicability of produced genome resource, followed by a check for technical feasibility. ERGA  
559 is displaying its target species over the platform Genomes on a Tree  
560 (<https://goat.genomehubs.org/projects/ERGA>), in agreement with other nodes of the EBP.  
561 ERGA members as well as SSP sample managers engage with other genome initiatives when  
562 overlaps are detected and facilitate collaboration in order to prevent parallel efforts.

## 563 A live and comprehensive sampling metadata manifest

564 The ERGA metadata manifest and its SOP are living documents, which are regularly revised  
565 under strict version control (<https://github.com/ERGA-consortium/ERGA-sample-manifest>).  
566 During the Pilot phase, it became clear that the metadata core was not entirely comprehensive.  
567 For example, the first version could not capture sampling depth and only allowed inputting a  
568 precise location. This information is important in the marine context as it was not possible to  
569 correctly represent samples from trawls or transects. Updated releases of the manifest have  
570 acknowledged these gaps and now comprise fields for e.g., depth and latitudinal and longitudinal  
571 coordinates for two points instead of one for sampling transects, extended vocabulary for  
572 sampled tissues, etc. As ERGA progresses, adding more extensions might be necessary during  
573 the planned regular updates.

574 The question that is often raised in regard to metadata collection is what is the trade-off between  
575 comprehensiveness *versus* feasibility. Sampling for reference genome generation has many  
576 logistical steps that are important to document in the metadata record. Such extensive collection  
577 of metadata appears doable when the emphasis is on single (or a few) representative samples  
578 per species while we acknowledge that feasibility and applicability might be different for e.g.,  
579 population data or already collected material that cannot be obtained again. Yet, as the field of  
580 genomics moves forward and technological advances allow extracting more data at higher  
581 quality from material with varying quality samples, extending the high ERGA standards to any  
582 sample collected for genetic analyses appears as an appropriate perspective. In this light, the  
583 increase in frozen archives that ERGA supports will be a treasure trove for genome initiatives.

## 584 Streamlining legal compliance procedures

585 Biodiversity knows no boundaries and it is blissfully unaware of its traversal distribution across  
586 many national, political, and cultural borders that may have varying legal systems. However,  
587 ERGA is obligated to respect these borders and the legal systems within, and so a harmonisation

588 of procedures will be a crucial aspect of building a streamlined European sampling infrastructure  
589 for reference genome generation. ERGA's network provides cross-country communication,  
590 which should be extended to local authorities, and ensure efficient flow of information about  
591 specific legal requirements of sampling. Streamlining the steps required to ensure legal  
592 compliance therefore is an important way to increase the efficiency of the reference genome  
593 generation pipeline.

## 594 **A continued concerted effort**

595 Under the umbrella of the EBP and in the light of the progress that sequencing technology and  
596 data processing offer, there is a need to scale up the genome generation process. While ERGA  
597 has pioneered the establishment of a collaborative transnational effort for reference genome  
598 generation in Europe, other regional initiatives advance and face similar challenges. We here  
599 call for the establishment of collaborative concerted efforts among different consortia under the  
600 EBP flag, unifying standards across the whole workflow, starting with sampling and sampling  
601 processing and ending with making data available via open repositories.

## Glossary

Acronym	Explanation	Ressource
ABS	Access and Benefit-Sharing	<a href="https://absch.cbd.int/">https://absch.cbd.int/</a>
BGE	Biodiversity Genomics Europe	<a href="https://biodiversitygenomics.eu/">https://biodiversitygenomics.eu/</a>
BIOSCAN EUROPE	part of the International Barcode of Life Consortium (iBOL)	<a href="https://www.bioscaneurope.org/">https://www.bioscaneurope.org/</a>
CARE	Collective benefit, Authority to control, Responsibility and Ethics	<a href="https://www.gida-global.org/care">https://www.gida-global.org/care</a>
CITES	Convention on International Trade in Endangered Species of Wild Fauna and Flora	<a href="https://cites.org">https://cites.org</a>
COPO	Collaborative OPen Omics	<a href="https://copo-project.org/">https://copo-project.org/</a>
DToL	Darwin Tree of Life	<a href="https://www.darwintreeoflife.org/">https://www.darwintreeoflife.org/</a>
EBP	Earth Biogenome Project	<a href="https://www.earthbiogenome.org/">https://www.earthbiogenome.org/</a>
DAC	Data Analysis Committee	<a href="https://www.erga-biodiversity.eu/team-1/dac---data-analysis-committee">https://www.erga-biodiversity.eu/team-1/dac---data-analysis-committee</a>
ELSI	Ethical, Legal, and Social Issues	<a href="https://www.erga-biodiversity.eu/team-1/elsi---ethical%2C-legal%2C-and-social-issues">https://www.erga-biodiversity.eu/team-1/elsi---ethical%2C-legal%2C-and-social-issues</a>
ENA	European Nucleotide Archive	<a href="https://www.ebi.ac.uk/ena/browser/home">https://www.ebi.ac.uk/ena/browser/home</a>
ERGA	European Reference Genome Atlas	<a href="https://www.erga-biodiversity.eu/">https://www.erga-biodiversity.eu/</a>
FAIR	Findable, Accessible, Interoperable, and Reusable	<a href="https://www.go-fair.org/fair-principles/">https://www.go-fair.org/fair-principles/</a>
GoaT	Genomes on a Tree	<a href="https://goat.genomehubs.org/">https://goat.genomehubs.org/</a>
ITC	Inclusiveness Target Countries	-
JEDI	Justice, Equity, Diversity & Inclusion	<a href="https://jedicollaborative.com/">https://jedicollaborative.com/</a>
SOP	Standard Operating Procedure	-
SSP	Sampling & Sample Processing Committee	<a href="https://www.erga-biodiversity.eu/team-1/ssp---sampling-%26-sample-processing">https://www.erga-biodiversity.eu/team-1/ssp---sampling-%26-sample-processing</a>

## 603 References

- 604 Capa, M., & Hutchings, P. (2021). *Systematics and Diversity of Annelids*. MDPI, Basel.  
605 <https://doi.org/10.3390/books978-3-0365-1389-8>
- 606 Carroll, S. R., Herczog, E., Hudson, M., Russell, K., & Stall, S. (2021). Operationalizing the  
607 CARE and FAIR Principles for Indigenous data futures. *Scientific Data*, 8(1), Article 1.  
608 <https://doi.org/10.1038/s41597-021-00892-0>
- 609 Crandall, E. D., Toczydlowski, R. H., Liggins, L., Holmes, A. E., Ghoojarei, M., Gaither, M. R.,  
610 Wham, B. E., Pritt, A. L., Noble, C., Anderson, T. J., Barton, R. L., Berg, J. T., Beskid, S.  
611 G., Delgado, A., Farrell, E., Himmelsbach, N., Queeno, S. R., Trinh, T., Weyand, C., ...  
612 Toonen, R. J. (2023). Importance of timely metadata curation to the global surveillance  
613 of genetic diversity. *Conservation Biology*, 00, e14061.  
614 <https://doi.org/10.1111/cobi.14061>
- 615 Dahn, H. A., Mountcastle, J., Balacco, J., Winkler, S., Bista, I., Schmitt, A. D., Pettersson, O. V.,  
616 Formenti, G., Oliver, K., Smith, M., Tan, W., Kraus, A., Mac, S., Komoroske, L. M., Lama,  
617 T., Crawford, A. J., Murphy, R. W., Brown, S., Scott, A. F., ... Fedrigo, O. (2022).  
618 Benchmarking ultra-high molecular weight DNA preservation methods for long-read and  
619 long-range sequencing. *GigaScience*, 11, giac068.  
620 <https://doi.org/10.1093/gigascience/giac068>
- 621 Engel, M. S., Ceríaco, L. M. P., Daniel, G. M., Dellapé, P. M., Löbl, I., Marinov, M., Reis, R. E.,  
622 Young, M. T., Dubois, A., Agarwal, I., Lehmann A., P., Alvarado, M., Alvarez, N.,  
623 Andreone, F., Araujo-Vieira, K., Ascher, J. S., Baêta, D., Baldo, D., Bandeira, S. A., ...  
624 Zacharie, C. K. (2021). The taxonomic impediment: A shortage of taxonomists, not the  
625 lack of technical approaches. *Zoological Journal of the Linnean Society*, 193(2), 381–  
626 387. <https://doi.org/10.1093/zoolinlean/zlab072>
- 627 Formenti, G., Theissinger, K., Fernandes, C., Bista, I., Bombarely, A., Bleidorn, C., Ciofi, C.,  
628 Crottini, A., Godoy, J. A., Höglund, J., Malukiewicz, J., Mouton, A., Oomen, R. A., Paez,  
629 S., Palsbøll, P. J., Pampoulie, C., Ruiz-López, M. J., Svoldal, H., Theofanopoulou, C.,  
630 ... Zammit, G. (2022). The era of reference genomes in conservation genomics. *Trends*  
631 *in Ecology & Evolution*, 37(3), 197–202. <https://doi.org/10.1016/j.tree.2021.11.008>
- 632 Lawniczak, M., Davey, R., Rajan, J., Pereira-da-Conceicao, L., Kiliyas, E., Hollingsworth, P.,  
633 Barnes, I., Allen, H., Blaxter, M., Burgin, J., Broad, G., Crowley, L., Gaya, E., Holroyd,  
634 N., Lewis, O., McTaggart, S., Mieszkowska, N., Minotto, A., Shaw, F., & Sivess, L.  
635 (2022). Specimen and sample metadata standards for biodiversity genomics: A proposal



636 from the Darwin Tree of Life project. *Wellcome Open Research*, 7, 187.  
637 <https://doi.org/10.12688/wellcomeopenres.17605.1>

638 Lewin, H. A., Richards, S., Lieberman Aiden, E., Allende, M. L., Archibald, J. M., Bálint, M.,  
639 Barker, K. B., Baumgartner, B., Belov, K., Bertorelle, G., Blaxter, M. L., Cai, J., Caperello,  
640 N. D., Carlson, K., Castilla-Rubio, J. C., Chaw, S.-M., Chen, L., Childers, A. K.,  
641 Coddington, J. A., ... Zhang, G. (2022). The Earth BioGenome Project 2020: Starting the  
642 clock. *Proceedings of the National Academy of Sciences*, 119(4), e2115635118.  
643 <https://doi.org/10.1073/pnas.2115635118>

644 Mazzoni, C. J., Ciofi, C., Waterhouse, R. M. et al. (2023). [Biodiversity: an atlas of European  
645 reference genomes. \*Nature\* 619, 252. <https://doi.org/10.1038/d41586-023-02229-w>](https://doi.org/10.1038/d41586-023-02229-w)  
646 [McCartney, A. M., Formenti, G., Mouton, A. et al. \(2023\) The European Reference Genome  
647 Atlas: piloting a decentralised approach to equitable biodiversity genomics. \*bioRxiv\*  
648 2023.09.25.559365. <https://doi.org/10.1101/2023.09.25.559365>](https://doi.org/10.1101/2023.09.25.559365)

649 Shaw, F., Etuk, A., Minotto, A., González-Beltrán, A., Johnson, D., Rocca-Serra, P., Laporte,  
650 M.-A., Arnaud, E., Devare, M., Kersey, P., Sansone, S.-A., & Davey, R. (2020). COPO:  
651 A metadata platform for brokering FAIR data in the life sciences. *F1000Research*, 9, 495.  
652 <https://doi.org/10.12688/f1000research.23889.1>

653 Shaw, F., Minotto, A., McTaggart, S., Providence, A., Harrison, P., Pauperio, J., Rajan, J.,  
654 Burgin, J., Cochrane, G., Kiliyas, E., Lawniczak, M., & Davey, R. (2022). Managing sample  
655 metadata for biodiversity: Considerations from the Darwin Tree of Life project. *Wellcome*  
656 *Open Research*, 7(279). <https://doi.org/10.12688/wellcomeopenres.18499.1>

657 Theissinger, K., Fernandes, C., Formenti, G., Bista, I., Berg, P. R., Bleidorn, C., Bombarely, A.,  
658 Crottini, A., Gallo, G. R., Godoy, J. A., Jentoft, S., Malukiewicz, J., Mouton, A., Oomen,  
659 R. A., Paez, S., Palsbøll, P. J., Pampoulie, C., Ruiz-López, M. J., Secomandi, S., ...  
660 Zammit, G. (2023). How genomics can help biodiversity conservation. *Trends in*  
661 *Genetics*, 39(7), 545-559. <https://doi.org/10.1016/j.tig.2023.01.005>

662 Turner, H. (2022). *Cataloguing Culture: Legacies of Colonialism in Museum Documentation*.  
663 University of British Columbia Press.  
664 <https://press.uchicago.edu/ucp/books/book/distributed/C/bo70117236.html>