ERGA

European Reference Genome Atlas

Open Data Policy

ERGA Open Data Policy (v0.0.5 - 20231121)

Authors

Christian de Guttry, Tom Brown, Peter Harrison, Matthieu Muffato, Jennifer Leonard, Joana Pauperio, Katja Reichel, Nick Juty, Alice Mouton, Sadye Paez, Rebekah Oomen, Ciara Stauton, Ann Mc Cartney, Elena Buzan, Chiara Bortoluzzi, Robert Waterhouse.

This document is based on the Biodiversity Genomics Europe (BGE) Data Management Plan.

Also for reference: https://www.earthbiogenome.org/data-sharing-management-best-practices
https://www.earthbiogenome.org/it-and-informatics-standards

Executive Summary

This document outlines the Open Data Policy (ODP) for the European Reference Genome Atlas (ERGA). This includes all data and metadata produced as part of the activities within and/or associated with ERGA. The ODP will be regularly reviewed and updated to adapt to changing circumstances, legal requirements, and best practices. Different aspects regarding data, their processing, and their publishing are addressed here.

- The standards of metadata to be collected and disseminated.
- Summary of common data types, formats, purposes, and storage.
- ERGA's commitment to FAIR and CARE principles for data.
- ERGA's approach to research outputs, e.g. genomes, software, surveys.

Keywords

FAIR, CARE, Open Data Policy, Metadata, ENA, INSDC, EBP, ERGA, BGE.

V0.0.5 updates

- Implementation of Council members comments
- Grammar

Table of Contents

List of Abbreviations

CARE - Collective Benefit, Authority to Control, Responsibility, Ethics
COPO - Collaborative OPen Omics
DOI - Digital Object Identifier
EBP - Earth BioGenome Project
ENA - European Nucleotide Archive
ERGA - European Reference Genome Atlas
FAIR - Findable, Accessible, Interoperable, Reusable
GAL - Genome Acquisition Laboratory
GDPR - General Data Protection Regulation
INSDC - International Nucleotide Sequence Database Collaboration
ITIC - IT and Infrastructure Committee
Mb, Gb, Tb - Megabase, Gigabase, Terabase
MB, GB, TB - Megabyte, Gigabyte, Terabyte
NCBI - National Centre for Biotechnology Information
ODP - Open Data Policy
SRA - Sequence Read Archive
SSP - Sample and Sample Processing Committee

1. Introduction

The European Reference Genome Atlas (ERGA) aims to create reference-quality genome assemblies for all eukaryotic life in Europe. As the European node of the Earth BioGenome Project (EBP) the genomes produced as part of the ERGA initiative should align with the standards set by the EBP regarding the quality of reference genomes produced, but also in the correct collection and publishing of metadata and data for each corresponding species.

This document outlines ERGA's policy regarding the collection, processing, storage, and publishing of metadata and data related to the production of high-quality reference genomes. This Open Data Policy (ODP) is treated as a living document, and as such is expected to be regularly reviewed and updated with the current best practices regarding handling of biological metadata and data. The aim of this ODP is to inform ERGA Members of the current expectations and best practices regarding data storage, open access research, and to ensure that all ERGA Members follow a compliant, current, robust, and reproducible approach. This ODP covers the following topics:

- Description of data sources, types of data, and formats of data and metadata collected as part of the ERGA initiative.
- Utility and purpose of collected data.
- Expected storage requirements for short- and long-term projects.
- Recommendations for how to align with the FAIR and CARE principles, in particular regarding persistent identifiers and the use of open repositories.
- Management of other research outputs, such as software and workflows.
- Information regarding allocation of resources, data security, ethics, intellectual property policies, and other relevant issues.

The growth of ERGA and consequently of EBP requires that collection, sampling, processing, sequencing, assembling, annotating, analysing, engaging, and dissemination happens efficiently and at scale. With this in mind, this document is written with the vision of all steps ultimately being automated as much as possible, with only minimal necessary manual interventions. We are actively pursuing the goal of achieving this through consistency across all stages of genome generation projects.

This document has been written by members of the European node (ERGA) with the aim of aligning ERGA with the global efforts of the EBP as well as the Convention on Biological Diversity, it's Nagoya Protocol and the EU Access and Benefit Sharing (ABS) Regulation (Regulation (EU) No 511/2014). However, a number of issues, particularly regarding Permits and Personal Data, relate specifically to Europe and so may be regionally specific. While the

ERGA Open Data Policy aims to align with the policies developed by the EBP for global coordination, it is important to note that European regulations take precedence to ensure that ERGA complies with all regional legislation.

2. Data and Metadata Summary

The primary purpose of the sequence data produced as part of the ERGA initiative is the production of reference-quality genomes and annotations and their use within the framework of genomic research. The primary sources of data produced to these ends include, but are not limited to:

- Genomic DNA-based sequencing technologies (e.g. short- and long-read DNA sequencing, restriction-based optical maps, linked-read sequencing, long-range conformation capture sequencing).
- Transcriptomic RNA-based sequencing technologies (e.g. long-read isoform sequencing, either directly from RNA or from transcribed cDNA, short-read sequencing of cDNA from RNA).

These data will be accompanied with metadata relating to the entire project, including: sampling, handling and processing of a sample, sequencing, assembly and annotation, through to final publication. The metadata associated with the data produced includes, but is not limited to:

- Origin and provenance of specimen/sample.
- Associated ethical and legal procedures and documents.
- Individuals involved in sampling, storage, preservation, and shipping of sample(s).
- Techniques and materials used in sample storage and preservation.
- Protocols and practices followed in the full process from sampling to preparing genetic material for sequencing.
- Bioinformatic workflows, software, and versions used in the generation of the genome sequence and annotation.

ERGA recommends that raw data and corresponding metadata be submitted to public repositories that are part of the INSDC, and in particular ENA as the European partner, as soon as possible after creation and quality control (QC).

2.1 Collection of Metadata

In order for a project to be considered part of the ERGA initiative (an ERGA-affiliated project), sufficient sample metadata must be provided with the project. The Sample and Sample Processing (SSP) committee, in collaboration with COPO have produced an ERGA Sample Manifest with a supporting Standard of Practice document. Once filled, either COPO acts as a broker to submit information to a public sequencing archive (ENA), or the research team can independently submit these data directly to a public repository. Submitters should always use the latest version of the Sample Manifest and Standard Operating Procedure (SOP) for completing their metadata obligations. If the metadata and data are being submitted directly to ENA by the research team, the metadata should be submitted in a format consistent with the COPO minimal sample information as outlined in the Sample Manifest and the restrictions set out by ENA (ERGA sample submission manifest).

The metadata to be entered into the Sample Manifest fall into the following categories:

- Sample submission information including specimen identifier and tube/well identifiers, as well as sample ambassador personal information.
- Taxonomic information including species name, family, and common name.
- Biological information of the sample including lifestage, sex, and organism part.
- Details of the submitting GAL and the associated organisational codes.
- Personal data of the collector, collection event, and collection localities.
- Information on taxonomic identification, taxonomic uncertainty, and risks.
- Details of the tissue preservation event.
- Information on DNA barcoding.
- Information on Biobanking and Vouchering.
- Information on regulatory compliances, indigenous rights, traditional knowledge, and permits.
- Additional information including a free text field to capture other important sample notes.

2.2 Data Types and Formats

The data produced to create a reference genome and perform analysis thereon come in a number of different formats, depending on the sequencing platform used and on any pre-processing steps. Below are listed a number of examples of expected data types to be produced as part of an ERGA project. In particular, the data formats for raw sequencing data, genome assembly, and annotation are taken from the Accepted Read Data Formats for ENA.

Data Type	Format
Sequencing data	
PacBio Long-read sequencing	.fasta(.gz), .fastq(.gz), .subreads.bam, .ccs.bam
Oxford Nanopore Long-read sequencing	.fastq(.gz), .fast5
Illumina short-read sequencing	.fastq(.gz), .bam, .cram
Bionano optical maps	.cmap
Sample metadata	
Environmental, logistic and laboratory data	.txt, .xlsx, .csv, .png, .jpg, .mp3, .mp4
Sampling/transport/ethical permit	.pdf
Personal data	.txt, .pdf, .csv, .xlsx,.mp3, .mp4.
Software metadata	
Pipeline	Snakemake, Nextflow, Galaxy workflow, docker or singularity container
Config/parameter file	.yaml
Research outputs	
Genome assembly	.fasta(.gz)
Genome annotation	.gff3(.gz), .gtf(.gz)
Genome reports	.pdf, .csv, .txt

2.3 Expected Storage Requirements

The production of a genome assembly and any subsequent analysis thereon can create large volumes of data for each project. The exact numbers, however, can vary greatly, depending on the technologies used, type of organism/genome, and the analysis performed. Typically the raw data for a diploid species with a 1 Gb genome would be expected to take up around 1 TB of space, with another 1 TB required for temporary data as part of the assembly process, with required space increasing with genome size. For a genome annotation, if only short-read RNAseq data are used to build transcript models, the raw data footprint is expected to be much smaller, below 100 GB. However, any raw long-read flowcells can take up to another 1 TB in storage space before processing. The intermediate files in producing annotations can require multiple TB of storage space.

3. FAIR Data

A cornerstone of ERGA's commitment to excellence is the adherence to the FAIR[1] Principles, stating that research data should be Findable, Accessible, Interoperable, and Reusable. The FAIR Principles are designed to ensure that research data and other digital research objects are able to be reused by other researchers in an open, clear, and understandable manner. For research projects associated with ERGA, this means that sample metadata, raw data, research outputs, and relevant software and/or workflows should be held in a permanent repository which is accessible and versioned. This allows objects to be reused by any other individual. ERGA makes the following recommendations and requirements for ensuring the FAIR Principles are adhered to by genome teams.

3.1 Data and Metadata Publishing on ENA

Any sample metadata, along with the raw sequencing data and resulting research objects, such as the genome assembly and annotation, should be submitted to a public repository. Within ERGA, we recommend utilising the metadata brokering system offered by COPO to upload sample metadata found in the ERGA Sample Manifest to the ENA. Following the FAIR principles, completing metadata to the standards set out in the ERGA Sample Manifest will ensure that rich metadata, written in approachable and accepted terms is associated with unique identifiers. Furthermore, once published in a public repository, it is guaranteed that the metadata will be hosted in a permanent, accessible location for future reference and use.

In order to be considered an ERGA-affiliated project, the submitted data, metadata, and assembly should be published in INSDC, with the minimal sample metadata information outlined in the above ERGA Sample Manifest. Note however the special considerations regarding Secure Data (Section 5 below). Furthermore, the BioSample, Sequence Read Archive (SRA) object, and genome assemblies/annotations must be associated with a BioProject, or in the case of a genome with multiple haplotype assemblies, an Umbrella BioProject. Once the relevant BioProject and BioSample have been created and raw sequencing data uploaded for the submitted sample, these objects will then be linked to ERGA's Umbrella BioProject and sub-project, where relevant. At current writing, the following Umbrella BioProjects exist under the ERGA Umbrella:

- 25 Genomes Project: Genome Data and Assemblies
- ATLASea : An Atlas of eukaryotic marine genomes
- Catalan Initiative for the Earth BioGenome Project (CBP)
- Darwin Tree of Life Project: Genome Data and Assemblies
- ENDEMIXIT

- European Reference Genome Atlas (ERGA): Biodiversity Genomics Europe (BGE) Project Genome Data and Assemblies
- The European Reference Genome Atlas Pilot Project
- The European Reference Genome Atlas (ERGA) Satellite Genomes
- The French node of the European Reference Genome Atlas (ERGA) initiative
- The Swiss node of the European Reference Genome Atlas (ERGA) initiative

Having all research data and metadata from an ERGA-associated genome project published on ENA and within an ERGA BioProject ensures that the data are held in a secure, archived storage location and have increased visibility through the association with an EBP-related initiative.

3.2 Protocol Publishing

Publishing the protocols used to collect the data and metadata in a genome assembly project assists in the accessibility of the generated information. ERGA recommends using a site such as WorkflowHub and Protocols.io to store protocols and refer to in the publishing of data and research outputs. Not only do these protocols give further context to the data and other metadata produced, but they also provide invaluable resources for researchers and furthermore allow for more efficient meta-analysis in the context of understanding best protocols for different projects.

3.3 Software Publishing

Wherever possible, any software (including workflows and pipelines) used to create research outputs as part of an ERGA project should be published alongside the metadata, raw data, and any associated publications. Within ERGA, we recommend using github to host any novel software and/or workflow and WorkflowHub for publishing releases of software and pipelines. Similar to BioProjects hosted on ENA, ERGA has a WorkflowHub Space which is used to host and publish any relevant software used to create research outputs as part of an ERGA-affiliated project. Once published, any workflows hosted here will be given a Digital Object Identifier, allowing for permanent look-up and referencing of workflows that should be cited in publications.

Where genome assemblies or annotations have been produced using existing tools or workflows, the relevant Persistent Identifier (e.g. DOI) should be published as part of the metadata associated with the project. This Identifier should be listed within the published metadata associated with the BioProject on INSDC. Similar to the data and metadata collected,

any software and pipelines should be written in a language that can not only be machine-read, but also understandable to future scientists. In line with the FAIR principles, the protocols used to generate research outputs should be clearly documented, understandable, accessible, findable in a public repository and ready to be redeployed by others.

4. CARE Principles

As an integral part of its commitment to excellence, ERGA actively incorporates the CARE (Collective Benefit, Authority to Control, Responsibility, Ethics) Principles into its data governance practices. These principles serve as the foundation of our data governance approach, with a particular emphasis on the critical aspect of 'Authority to Control'. ERGA considers the CARE Principles a framework for action and a living part of our research processes. To uphold this, we commit to the following:

1. **Meaningful Engagement:** ERGA is committed to meaningfully engaging with partners to align with the principle of Collective Benefit. ERGA has a Citizen Science Committee dedicated to engaging with external parties, and a Training and Knowledge Transfer Committee are committed to outreach, capacity development and knowledge transfer. In addition, ERGA has placed an “Open to Collaborate” Notice on the public website to foster collaborations and partnerships with Indigenous Peoples.
2. **Community-Led Decision-Making:** Members should actively engage with Indigenous Peoples and Local Communities (IPLCs) to ensure that data governance decisions are made collectively and respect community values and protocols. If partnering with an Indigenous Peoples. ERGA researchers are encouraged to use the Local Contexts Traditional Knowledge and Biocultural Labels and Notices. To support researchers with the implementation of these Labels and Notices, ERGA SSP and ELSI in partnership with Global Indigenous Data Alliance have prepared a supporting document “Guidance Documentation for Implementing the Traditional Knowledge and Biocultural Labels in the European Reference Genome Atlas”.
3. **Data Ownership and Autonomy:** Members should safeguard IPLC autonomy by respecting their ownership of data and seeking their informed consent for data collection, use, and sharing.
4. **Respect for Indigenous Knowledge:** Members should acknowledge and respect indigenous knowledge and cultural context, ensuring research activities honour this heritage.
5. **Data Access Protocols:** Members should establish clear and transparent data access protocols that respect the 'Authority to Control' by allowing communities to manage who accesses their data and under what conditions.

By following these actionable steps, ERGA aims to adhere to the CARE Principles but also to demonstrate its dedication to the ethical, legal, and social considerations that underpin the research process. Our commitment extends to promote clear attribution, equitable access to knowledge, and contributing to the broader well-being of society while respecting the rights and interests of the ERGA community members and partners.

4.1 Access and Benefit Sharing - The Nagoya Protocol

ERGA places paramount importance on adherence to the Convention on Biological Diversity (CBD) and its Nagoya Protocol, recognizing the significance of transparent and equitable access and utilisation of genetic resources, and traditional knowledge. The Nagoya Protocol codifies a bilateral mechanism for access and benefit-sharing ABS). It ensures that the benefits arising from the access and utilisation of genetic resources, are shared fairly and equitably with the countries providing those resources and any associated Indigenous People and Local Communities (IP and LC). ERGA will also align with the provisions in item 11 on Digital Sequence Information (DSI) codified in the CBD Kunming Montreal Global Biodiversity Framework (GBF) and will adhere with the multilateral system of ABS decided upon in COP16.

If an ERGA member is based in a country signatory to the CBD and the Nagoya Protocol, it is the responsibility of researchers undertaking an ERGA-affiliated project to ensure that all members of its research team are aware of their legal obligations regarding ABS, as well as the evolving discussions surrounding DSI and its relevance to genetic resources. ERGA is committed to promoting collaboration, respect, and ethical engagement through its initiatives. It fosters a global environment where genetic resources and DSI are used sustainably and in accordance with the GBF. This framework sets the stage for ‘open and responsible’ data management and benefit-sharing, ensuring that these valuable resources respect the rights of all community members while supporting research, innovation, and sustainable development.

5. Acknowledgment of Contributors

In accordance with our Open Data Policy, we place great emphasis on transparency and collaboration. We acknowledge and appreciate the invaluable contributions of all individuals and organisations involved in the collection and production of data under the ERGA umbrella. Appropriately recognizing all efforts fosters a culture of openness but also ensures that the data we share represents the collective dedication and expertise. We are committed to giving credit where it is due and promoting a spirit of collaboration in the pursuit of our shared data-driven objectives.

6. Secure Data

It is not necessarily always possible for all data produced as part of a project to be publicly available and searchable for a number of legal, ethical, and practical reasons. Here, we suggest following the mantra data being “As open as possible and as closed as necessary”. Below, we outline a number of situations where data and metadata should be securely stored, always accessible when necessary to those with access authority.

6.1 Permits

Permits created or requested in the process of sampling, transferring, or working on a biological sample should be stored in a secure and backed-up storage location. ERGA recommends a secure solution, such as EUDAT’s B2DROP Nextcloud solution, allowing collaborators to deposit and share files in a secure location, where they are stored and archived/backed-up. Once stored, all information should be indexed and searchable to ensure ease of accessing the relevant information when required in the future. In instances where researchers have reservations about publicly disclosing certain permits, it is essential to maintain confidentiality. ERGA provides secure data storage that is only available for uploading. Only designated members will have access to this directory. ERGA Executive Board approval is required for any considerations pertaining to the non-disclosure of permits.

6.2 Sensitive Data

Sensitive data typically include genomic sequences, particularly those from rare or endangered species, as their disclosure could potentially harm conservation efforts or lead to biopiracy. Protected species location data (and other similar metadata), essential for ecological studies,

must be handled confidentially to prevent poaching and ecological disruption. In addition, we recognize the significance of addressing Indigenous sensitivities around data, as Indigenous knowledge and cultural data are valuable and must be treated with the utmost respect. Personal information of researchers and collaborators involved in the consortium are also considered sensitive data.

6.2.1 Rare or Protected Species Metadata and Data

ERGA prioritises the responsible management of sensitive data, particularly protected species metadata and genomic information, in strict adherence to EU laws and regulations (Council Regulation (EC) No 338/97; Directive 2009/147/EC; Directive 92/43/EEC). The approach we suggest toward protected species data is characterised by utmost confidentiality. Researchers working on such species should carefully anonymize or aggregate location, sampling, and habitat data to prevent potential harm to these species or ecosystems. The researcher should meticulously adhere to EU and national laws governing the protection of endangered or protected species, ensuring that data disclosures do not facilitate poaching or exploitation.

6.2.2 Personal Data

ERGA is committed to upholding the principles set by the General Data Protection Regulation (GDPR), recognizing the importance of safeguarding personal data and privacy rights. GDPR serves as a framework for ensuring the fair, transparent, and lawful processing of personal data. It is focused on the protection of personal data and the use of personal data in a manner that respects individuals' rights and gives them control over their own information. For ERGA research projects, this entails a rigorous commitment to obtaining informed consent, protecting individuals' privacy, and implementing robust data security measures. Our dedication to GDPR compliance also extends to data storage practices, where we ensure that all data are stored in compliance with GDPR regulations. To enhance data security, privacy, and conform with GDPR guidelines, ERGA utilises a European-based server EUDAT. EUDAT reinforces our commitment to protecting individuals' rights and interests in the context of data privacy as genomics and data-driven research advance. In alignment with these regulations, ERGA mandates that researchers ensure that their research activities adhere to the principles and standards outlined by GDPR, fostering an environment of responsible and privacy-conscious genomic research.

7. Ethics

Ethics constitute a cornerstone of ERGA's mission and is central to upholding the highest standards of research integrity as well as responsible data management. ERGA acknowledges potential scenarios that warrant ethical considerations, e.g. ERGA recognizes the balance between data accessibility and protecting endangered species.

ERGA partners adhere to EU, international, and national laws, ensuring fundamental rights, privacy, data protection, and environmental preservation. Furthermore, ERGA is dedicated to responsible sampling practices (ERGA sampling code of practices), obtaining explicit permission for collecting samples from protected or endangered species. ERGA extends these ethical considerations to encompass the ever-important task of meaningful external engagement thatmay require Prior Informed Consent to be obtained in compliance with GDPR regulations and established principles for Citizen Scientists (ECSA 2015; https://www.ecsa.ngo/2016/05/17/10-principles-of-citizen-science/)

It is worth noting that a current ERGA-affiliated Project (ERGA-BGE) is funded by the European Union's Horizon Europe Research and Innovation Action. This funding reflects ERGA's commitment to conducting ethical and responsible genomics research, aligning with the rigorous ethical and scientific standards required by the Horizon Europe program. This dedication ensures that the research contributes positively to both science and society, meeting high ethical and quality criteria.

ERGA has the right to remove a sample/genome from the list of ERGA projects if it is later found not to have been collected and processed in a legal and ethical manner. We conduct routine control of data and metadata within the ERGA umbrella to maintain legal and ethical standards. In the event of a breach, the matter undergoes a thorough evaluation by the ERGA Executive Board. The evaluation considers the nature and severity of the breach, and potential measures range from requesting modifications that address the issue to the removal of the sample/genome or project from the ERGA umbrella.

References

1. Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
2. Carroll, SR, et al. 2020. The CARE Principles for Indigenous Data Governance. Data Science Journal, 19: 43, pp. 1–12. DOI: https://doi.org/10.5334/dsj-2020-043
3. European Union. (1997). Council Regulation (EC) No 338/97 of 9 December 1996 on the protection of species of wild fauna and flora by regulating trade therein. Official Journal of the European Union, L 61, 1–69
4. European Union. (2009). Directive 2009/147/EC of the European Parliament and of the Council of 30 November 2009 on the conservation of wild birds. Official Journal of the European Union, L 20, 7–25.
5. European Union. (1992). Council Directive 92/43/EEC of 21 May 1992 on the conservation of natural habitats and of wild fauna and flora. Official Journal of the European Union, L 206, 7–50.