Dear Prof. Narayan,

On behalf of our co-authors we would like to thank you, and the two reviewers for providing such valuable feedback on our manuscript. With your contributions we believe that our manuscript is much improved both in terms of content and readability. We have submitted the following materials in response for your consideration:

1) A PDF that addresses both reviewers' comments in a point by point manner, noting that our co-developed responses are highlighted in purple font for reviewer 1 and blue font for reviewer 2. This PDF has also been updated to ensure it has all of the specified sections including a dedicated "Funding" and "Conflicts of Interest" section that includes co-authors that are PCI recommenders;
2) The URL to the revised manuscript which has now been resubmitted to BioRiv;
3) A .docx version of the resubmitted manuscript that includes all of the tracked changes that were implemented on the back of the reviewers comments for the resubmission;
4) The DOI for our revised Supplementary Materials (DOI: `10.5281/zenodo.10789421`)
5) The DOI for the scripts/code related to our manuscript (DOI: `10.5281/zenodo.10789421`)

We would again like to thank you and the reviewers for your time in considering our response.

Kind Regards,
Ann Mc Cartney

Link to the working manuscript draft where changes will be implemented:
ERGA_Flagship_ForReviewersComments

Handling Editor Comments by Jitendra Narayan, 04 Jan 2024 11:34

I strongly urge the author to carefully consider the constructive criticisms and comments made by the discerning reviewers. When writing responses, please explain the changes made in response to each critique, elaborate on any additional data or analyses performed, and provide thorough clarifications where necessary.

Handling Editor Response: We would like to thank the handling editor for their comments. We have responded below in a point by point fashion to each review comment.

Reviewer 1

The article details the procedures and challenges encountered while developing a pilot infrastructure for the production of reference genome resources. The authors mentioned that the results and insights gained from the pilot lay a strong foundation for ERGA and offer valuable knowledge to other national and transnational genomic resource initiatives.

Reviewer 1 Comment 1: Overall, the manuscript was well-written, with nice figures and rich references. However, the structure could use some improvement to enhance readability. One

way to achieve this, if it aligns with the ERGA implemented workflows, would be to reorganize the sections into four parts: 1) Background, 2) Development of a Decentralized Infrastructure, 3) Challenges, and 4) Future Directions.
Reviewer 1 Response 1: We would like to thank the reviewer for this recommendation. We have implemented the suggested changes into the text accordingly and agree that it facilitates the readability of the manuscript.

Reviewer 1 Comment 2: Section 2, can also be restructured into five subsections; 1. Genome Team Establishment , 2. Building a Representative Species List, 3.Developing a Communications and Coordination Strategy, 4.Developing a Capacity Building and Knowledge Transfer Strategy and 5. Technical Workflows.
Reviewer 1 Response 2: Again we would like to thank the reviewer for this structural recommendation. We have implemented the proposed changes.

Reviewer 1 Comment 3: Section 2.5, Technical Workflows: These are well described in Steps 2-9, and should be reassigned accordingly. Step 5, should be changed to Sample Preparation or similar since it describes not only HMW DNA isolation but also library prep considerations for each of the platforms.
Reviewer 1 Response 3: We agree, and have updated the text accordingly.

Reviewer 1 Comment 4: Section 3, Challenges, authors can assign the challenges into broad themes/subsections; For example, authors can assign the already described challenges into Social, Administrative and Technical Challenges or other relevant titles. Authors could also restructure this section to decribe challenges encoutered in specific sections of Section 2. Authors should avoid repeating titles in subsections. For example, Training and Knowledge Transfer appeared twice.
Reviewer 1 Response 4: We would like to thank the reviewer for this suggestion. To address it, we have restructured the decentralisation challenges section into three broad categories, namely 1) Technical, 2) Ethical and Legal, and 3) Social Justice. We have also edited subheaders to ensure that they are unique and are not found elsewhere in the manuscript.

Reviewer 2

Summary

The authors describe, at length, the pilot program for the European Reference Genome Atlas, which is the European node of the Earth Biogenome Project (EBP). EBP aspires to sequence the genomes of every eukaryotic species on our planet. The authors describe in detail the selection of species, development of infrastructure, and then nine steps toward the eventual sharing of completed reference genomes, from selection of genome teams, through sample collection and storage, DNA extraction, sequencing, assembly, annotation analysis and sharing of the data. They conclude with a discussion of the challenges of creating a decentralized network and ways to address these in the future.

Major Comments

Reviewer 2 Comment 1: This is a well-written description of the pilot version of gigantic undertaking, which is itself large in scope. While 98 reference genomes is nothing to sneeze at, the larger importance is that the authors have provided a template, which can be modified and applied around the world, towards the "moon-shot" goal of the EBP. I'm therefore glad that the authors have gone with Peer Community In, and I would suggest that they resist pressure from reviewers or editors to shorten this methods paper. It is full of important details that will be useful to others who try to replicate their success! The authors also clearly appreciate that it is at least as important *who* is doing science as *how* the science is being done, and have taken major steps to be inclusive in their science.

Reviewer 2 Comment 2: I did want to raise one important issue. The authors clearly understand the importance of ERGA's role in the global biodiversity community, as indicated by Case Study 4. For this reason, I strongly suggest that they use the relevant, established metadata standards and definitions whenever possible, to ensure that ERGA's hard won data are findable, accessible, interoperable (especially) and reusable (FAIR). Reviewing the ERGA Sample Manifest v2.4.3 that was linked in the article, the terms used are not from either Darwin Core (DwC), which is the relevant standard for biodiversity data, or MIxS, which is the relevant genomic metadata standard. This will be important if ERGA wants to share their metadata into GBIF, which uses Darwin Core, and I'd be surprised if they haven't already had issues with uploading to INSDC. Thoughtful people have put a lot of time into developing MIxS and DwC terms and definitions, and even if they are imperfect (for example neither has a term for permit information), the principles of precedence and standardization should be operative here. I don't know that addressing this issue should be a condition for PCI recommendation, as it will probably take some work and time to make changes in COPO's code. But that is also why it is important to address this issue now, rather than later.
Reviewer 2 Response 2: We would like to thank the reviewer for this comment, we have responded in Reviewer 2 Response 8.

Specific Comments
Reviewer 2 Comment 3:P3 Incorrect quantifier "Biodiversity and ecosystem decline, loss and degradation raise the prospect that *MUCH*, if not most, of the Earth's biodiversity will be lost forever before they can be genomically explored..." Also, I fully understand the intention of this sentence but it could be construed to mean that the only value in a species is found in its genomic resources. I know this is not the authors' intent but I suggest rewording.
Reviewer 2 Response 3: We would like to thank the reviewer for this suggestion. Although it was not our intent, we agree that it could be construed in this way and so we have updated the text accordingly.

Reviewer 2 Comment 4:P4 "However, the scientific enquiries that can be actualised from reference resources15 are limited in scope due in large to a current lack of standardisation across the multitude of actors involved throughout the production of complete reference resources.". Great sentence! I suggest replacing "actualised" with "realised"
Reviewer 2 Response 4: We have updated the text with the suggested change.

Reviewer 2 Comment 5:P6 "In other cases, *partnering sequencing* contributed their own grant funds" - sequencing partners?
Reviewer 2 Response 5: We would like to thank the reviewer for picking up on this error. We have implemented the suggested change.

Reviewer 2 Comment 6a:P7 "Building a representative species list" - I have wondered how to go about prioritization of species. This seems like a reasonable process, but surely phylogenetic representation could be considered. I'm curious about how target categories were selected though. (I note from page 26 that phylogenetic representation will be considered going forward) Figure 2b.
Reviewer 2 Response 6a: We agree with the reviewer that phylogenetic representation should be considered in the selection process. When the selection process was designed for this pilot, the main goal was to show that a decentralized approach across several European countries was possible and could be successfully conducted. To this end, ERGA decided to emphasize feasibility in the pilot to increase the chances of successfully sequencing genomes at the reference level across a wide geographic distribution. Developing and learning from such a process was crucial for scaling reference genome production for the ongoing Biodiversity Genomic Europe (BGE) project (https://biodiversitygenomics.eu) that resulted from the Pilot selection process that was described in this manuscript.

The larger ERGA community was involved in developing the prioritization process, it is based on explicit and objective criteria and the process is semi-automated based on sample providers' answers to the questionnaire used for the species proposal. By doing so, human intervention was only necessary in the curation of the initial database to ensure the automatization of the process.

In brief, the species selection process of BGE task is a four-stage process including (1) an exclusion stage (e.g., to avoid redundancy with other projects), (2) a prioritization stage employing a decision-tree model with phylogenetic representation at the highest decision level and additional ranking to ensure country and researcher representation, (3) a feasibility check with additional adjustment for genera with multiple species suggestions and (4) a final check of legal compliance. The entire species selection process is based on a total of 28 different criteria. However, this process was developed out of the pilot study and not a part of it, and so it is out of the scope of this manuscript. However, it will be published in a separate manuscript as part of the BGE project that details a comprehensive overview of the process and its establishment of.

Reviewer 2 Comment 6b: I am having trouble interpreting the "International Genome Team Composition". Are the bins the number of countries represented on a genome team? The text on page 8 clarifies that this is the number of international members, where "international" is defined as coming from a separate country than the sample. But the figure legend should be clearer. Or even expressing it in terms of number of countries would be clearer still.

Reviewer 2 Response 6b: We would like to thank the reviewer for this comment. We agree that the figure legend was vague in this regard and so we have elaborated the text in order to provide more clarification.

"*Figure 2: Sample, country and partnering institution distribution across Europe. a) Taxonomic distribution of the species included into infrastructure testing. b) Top: Distribution of sample ambassadors per participating country. Bottom-left: self identified sex distribution across sample ambassadors, Bottom-right: frequency of genome teams that have international collaborators i.e., collaborators that are outside of the country of origin that the sample was obtained fromions within genome teams. c) Map illustrating the distribution of sampling localities, cryopreserved specimens, collections holding vouchered specimens, sequencing library preparation hubs and sequencing facilities across Europe*[32]*."*

Reviewer 2 Comment 7:P9 GDPR should be added to the glossary. As a US citizen, I'm aware of GDPR, but other readers might not be.

Reviewer 2 Response 7: We would like to thank the reviewer for highlighting that the GDPR required additional description within the glossary in order for the manuscript to be more inclusive to readers beyond the European Union. We have updated the glossary with the following definition, with a link to the regulation website.

"GDPR - The General Data Protection Regulation (GDPR) was issued by the European Union and became applicable in 2018. The act aims to harmonise the data privacy regulations relating to personal information across Europe. The regulation protects the fundamental rights and freedoms of natural persons and their right to the protection of their data."

Reviewer 2 Comment 8: P10 Thanks for making the sample manifest publicly available. Great that you are using validation rules. I would strongly urge ERGA and COPO to adopt the Darwin Core and/or MIxS metadata standards for their metadata to ensure their FAIRness. See major comments above.

Reviewer 2 Response 8: We thank the reviewer for this observation and agree. The COPO developers have over the last few weeks been mapping the ERGA metadata fields to both Darwin Core and MIxS fields. The COPO API has been modified so that samples can be obtained mapped to either of these standards. This will be live on the COPO website in the due course of their release cycle.

Reviewer 2 Comment 9: P10 "Unique to ERGA, fields were developed to mandate important information disclosure..."These fields are not unique to ERGA. At GEOME we have developed similar fields to accept globally unique and persistent identifiers (EZIDs), as well as information about permits and TK/BC notices and labels. See Riginos et al. 2021. These fields are not covered by MIxS or Darwin Core - it might be a good time to meet to discuss standardization of this information.

Reviewer 2 Response 9: We would like to apologise to the reviewer for the misuse of this word. We agree that in fact the development of these fields were in large part informed by the work of the GEOME team. We have adjusted the text accordingly and included the reference supplied by the reviewer. We also agree that there is a need for this information to be standardised and would love to discuss this further with the reviewer.

"Inspired by the Genomic Observatories Metadatabase[25], ERGA alsoUnique to ERGA, fields were developed fields to mandate important information disclosure e.g., permanent unique identifiers (PUID) associated with ex-situ specimens, permits, and Indigenous rights and interests (TK and BC Labels and Notices)[22,26–28]."

Reviewer 2 Comment 10: P12 "All 98 of genome teams" -- All 98 genome teams
Reviewer 2 Response 10: We have updated the text accordingly to reflect this suggested change.

Reviewer 2 Comment 11: P13 "To initialise these partnerships, a sequencing platform landscape assessment was conducted across all of the countries that ERGA had council representation" -- across all of the countries that had ERGA council representation.
Reviewer 2 Response 11: We would like to thank the reviewer for catching this typographical error. We have updated the text to correct this.

Reviewer 2 Comment 12: P14 "Here, we recommended the following data-type volumes for assembly generation: 30X HiFi or 60X ONT, 25X Hi-C (per haplotype) and 25X (per haplotype) Illumina (in cases where ONT data was used), and the following data-type volumes for annotation: total of 100 million reads if >five tissue types are available, or 30 million reads if tissue samples are pooled." I am not an expert in genome sequencing as I work more at the population level, so I can't comment on the suitability of these recommendations. However, if these are official ERGA recommendations, there is a lot of room for misunderstanding here. I would spend the space to make them more clear, either in a table, or using several very clear sentences, with AND and OR statements.
Reviewer 2 Response 12: Thanks for the feedback. We extended this paragraph to better capture these recommendations, including proper citations:
"For the ERGA Pilot, based on a growing consensus in the genome sequencing community (Lariviere et al. 2024, Lawniczak et al. 2021), we recommended the following data-type volumes for assembly generation: 30X HiFi or 60X ONT for contigging, combined with 25X Hi-C (per haplotype). To improve the consensus accuracy for ONT-based assemblies, we also recommended 25X (per haplotype) of Illumina data. For annotation, recommendations were: total of 100 million reads if >five tissue types are available, or 30 million reads if tissue samples are pooled."
We note that these recommendations are rapidly becoming global standards for reference genome sequencing projects, particularly as part of the broader EBP (https://www.earthbiogenome.org/report-on-assembly-standards).

Reviewer 2 Comment 13: P16 I've done a little work to try to understand figure 3A, but haven't made much progress. How can annotation data be at the permitting stage?
Reviewer 2 Response 13: We would like to thank the author for this feedback on Fig.3a. The permitting is in reference to the permits for the sample collection that needs to take place for RNA/IsoSeq sequencing to be conducted. We have split this into categories as sometimes the sample collection done for annotation data is separate to that done for

long read and proximity ligation data. We have now clarified this also in the figure caption.

Reviewer 2 Comment 14: P18 I quite like this figure, with lots of information content. While I understand the utility of ToLIDs, I wonder if they are helpful here as I'd have to go to the supplemental table to decipher them. Just flagging a potential issue - handle as you see fit.
Reviewer 2 Response 14:
Thank you for highlighting this. We like to use ToLIDs because they are unambiguous, however to ease the work of the readers we now specify the full latin name and the acronym for broad taxa also in the figure caption.