





Peer Community In Genomics

Informed Choices, Cohesive Future: Decisions and Recommendations for ERGA

Jitendra Narayan  based on peer reviews by *Eric Crandall*  and *Justin Ideozu*

Ann M Mc Cartney, Giulio Formenti, Alice Mouton, Claudio Ciofi, Robert M Waterhouse, Camila J Mazzoni, Diego De Panis, Luisa S Schlude Marins, Henrique G Leitao, Genevieve Diedericks, Joseph Kirangwa, Marco Morselli, Judit Salces, Nuria Escudero, Alessio Iannucci, Chiara Natali, Hannes Svandal, Rosa Fernandez, Tim De Pooter, Geert Joris, Mojca Strazisar, Jo Wood, Katie E Herron, Ole Seehausen, Phillip C Watts, Felix Shaw, Robert P Davey, Alice Minotto, Jose Maria Fernandez Gonzalez, Astrid Bohne, Carla Alegria, Tyler Alioto, Paulo C Alves, Isabel R Amorim, Jean-Marc Aury, Niclas Backstrom, Petr Baldrian, Lorian Ballarin, Laima Baltrunaite, Endre Barta, Bertrand BedHom, Caroline Belser, Johannes Bergsten, Laurie Bertrand, Helena Bilandija, Mahesh Binzer-Panchal, Iliana Bista, Mark Blaxter, Paulo AV Borges, Guilherme Borges Dias, Mirte Bosse, Tom Brown, Remy Bruggmann, Elena Buena-Atienza, Josephine Burgin, Elena Buzan, Nicolas Casadei, Matteo Chiara, Sergio Chozas, Fedor F Ciampor, Angelica Crottini, Corinne Cruaud, Fernando Cruz, Love Dalen, Alessio De Biase, Javier del Campo, Teo Delic, Alice B Dennis, Martijn FL Derks, Maria Angela Diroma, Mihajla Djan, Simone Duprat, Klara Eleftheriadi, Philine GD Feulner, Jean-Francois Flot, Giobbe Forni, Bruno Fosso, Pascal Fournier, Christine Fournier-Chambrillon, Toni Gabaldon, Shilpa Garg, Carmela Gissi, Luca Giupponi, Jessica Gomez-Garrido, Josefa Gonzalez, Miguel L Grilo, Bjoern Gruening, Thomas Guerin, Nadege Guiglielmoni, Marta Gut, Marcel P Haesler, Christoph Hahn, Balint Halpern, Peter Harrison, Julia Heintz, Maris Hindrikson, Jacob Hoglund, Kerstin Howe, Graham Hughes, Benjamin Istace, Mark J. Cock, Franc Jancekovic, Zophonias O Jonsson, Sagane Joye-Dind, Janne J. Koskimaki, Boris Krystufek, Justyna Kubacka, Heiner Kuhl, Szilvia Kusza, Karine Labadie, Meri Lahteenaro, Henrik Lantz, Anton Lavrinienko, Lucas Leclere, Ricardo Jorge Lopes, Ole Madsen, Ghislaine Magdelenat, Giulia Magoga, Tereza Manousaki, Tapio Mappes, Joao Pedro Marques, Gemma I Martinez Redondo, Florian Maumus, Hendrik-Jan Megens, Shane A McCarthy, Jose Melo-Ferreira, Sofia L Mendes, Matteo Montagna, Joao Moreno, Mai-Britt Mosbech, Monica Moura, Zuzana Musilova, Eugene Myers, Will J. Nash, Alexander Nater, Pamela Nicholson, Manuel Niell, Reindert Nijland, Benjamin Noel, Karin Noren, Pedro H Oliveira, Remi-Andre Olsen, Lino Ometto, Stephan Ossowski, Vaidas Palinauskas, Snaebjorn Palsson, Jerome P Panibe, Joana Pauperio, Martina Pavlek, Emilie Payen, Julia Pawlowska, Jaume Pellicer, Graziano Pesole, Joao Pimenta, Martin Pippel, Anna Maria Pirttila, Nikos Poulakakis, Jeena Rajan, Ruben MC Rego, Roberto Resendes, Philipp Resl, Ana Riesgo, Patrik Rodin-Morch, Andre ER Soares, Carlos Rodriguez Fernandes, Maria M. Romeiras, Guilherme Roxo, Lukas Ruber, Maria Jose Ruiz-Lopez, Urmas Saarma, Luis P Silva, Manuela Sim-Sim, Lucile Soler, Vitor C Sousa, Carla Sousa Santos, Alberto Spada, Milomir Stefanovic, Viktor Steger, Josefin Stiller, Matthias Stock, Torsten Hugo H Struck, Hiranya Sudasinghe, Riikka Tapanainen, Christian Tellgren-Roth, Helena Trindade, Yevhen Tukalenko, Ilenia Urso, Benoit Vacherie, Steven M Van Belleghem, Kees van Oers, Carlos Vargas-Chavez, Nevena Velickovic, Noel Vella, Adriana Vella, Cristiano Vernesi, Sara Vicente, Sara Villa, Olga Vinnere Pettersson, Filip AM Volckaert, Judit Voros, Patrick Wincker, Sylke Winkler (2024) The European Reference Genome Atlas: piloting a decentralised approach to equitable biodiversity genomics. bioRxiv, ver. 4, peer-reviewed and recommended by Peer Community in Genomics. <https://doi.org/10.1101/2023.09.25.559365>

Submitted: 03 October 2023, Recommended: 11 May 2024

Cite this recommendation as:

Narayan, J. (2024) Informed Choices, Cohesive Future: Decisions and Recommendations for ERGA. *Peer Community in Genomics*, 100298. [10.24072/pci.genomics.100298](https://doi.org/10.24072/pci.genomics.100298)

Published: 11 May 2024

Copyright: This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

The European Reference Genome Atlas (ERGA) (Mc Cartney et al, 2024, Mazzoni et al, 2023) demonstrates the collaborative spirit and intellectual abilities of researchers from 33 European countries. This ambitious project, which is part of the [Earth BioGenome Project](#) (Lewin et al., 2018) Phase II, has embarked on an unprecedented mission: to decipher the genetic makeup of 150,000 species over a span of four years. At the heart of ERGA is a decentralized pilot infrastructure specifically built to assist the production of high-quality reference genomes. This structure acts as a scaffold for the massive task of genome sequencing, giving the necessary framework to manage the complexity of genomic research. The research paper under consideration offers a comprehensive narrative of ERGA's evolution, outlining both successes and challenges encountered along the road.

One of the most significant issues addressed in the manuscript is the equitable distribution of resources and expertise among participating laboratories and countries. In a project of this magnitude, it is critical to leverage the pooled talents and capacities of researchers from across Europe. [ERGA's pan-European network](#) promotes communications and collaboration, creating an environment in which knowledge flows freely and barriers are overcome. This adoption of strong coordination and communication tactics will be essential to ERGA's success. Scientific collaboration depends on efficient communication channels because they allow researchers to share resources, collaborate on new initiatives, and exchange ideas. Through a diverse range of gatherings, courses, and virtual discussion boards, ERGA fosters an environment of transparency and cooperation among members, enabling scientists to overcome challenges and make significant discoveries. The importance ERGA places on training and information transfer programmes is a pillar of its strategy. Understanding the importance of capacity development, ERGA invests in providing researchers with the knowledge and abilities necessary for effectively navigating the complicated terrain of genomic research. A wide range of subjects are covered in training programmes (Larivière et al. 2023), from sample preparation and collection to data processing methods and sequencing technology. Through the development of a group of highly qualified experts, ERGA creates the foundation for continued advancement and creativity in the genomics sector.

This manuscript also covers in detail the technological workflows and sequencing techniques used in ERGA's pilot infrastructure. With the aid of cutting-edge sequencing technologies based on both long-read and short-read sequencing, they are working to unravel the complex structure of the genetic code with a level of accuracy and precision never before possible. To guarantee the accuracy of genetic data and prevent mistakes and flaws that can jeopardize the findings' integrity, quality control methods are put in place. Despite having a focus on genome sequencing due to its technological complexities, ERGA also remains firm in its dedication to metadata collection and sample validation. Metadata serves as a critical link between raw genetic data and useful scientific insights, giving necessary context and allowing researchers to draw practical findings from their investigations. Sample validation approaches improve the reliability and reproducibility of the results, providing users confidence in the quality of the genetic data provided by ERGA.

Looking ahead, ERGA envisions its decentralized infrastructure serving as a model for global collaborative research efforts. By embracing diversity, encouraging cooperation, and pushing for open access to data and resources, ERGA hopes to catalyze scientific discovery and generate positive change in the field of biodiversity genomics. ERGA aims to promote a more equitable and sustainable future for all by ongoing interaction with stakeholders, intensive outreach and education activities, and policy change advocacy. In addition to its immediate goals, ERGA considers the long-term implications of its work. As genomic technology progresses,

the potential application of high-quality reference genomes will continue to grow. From informing conservation efforts and illuminating evolutionary histories to revolutionizing healthcare and agriculture, it is likely that ERGA's contributions will have far-reaching consequences for people and the planet as a whole.

Furthermore, ERGA understands the importance of interdisciplinary collaboration in addressing the difficult challenges of the twenty-first century. ERGA aims to integrate genetic research into larger initiatives to promote sustainability and biodiversity conservation by forming relationships with stakeholders from other areas, such as policymakers, conservationists, and indigenous groups. Through shared knowledge and community action, ERGA seeks to create a future in which mankind coexists peacefully with the natural world, guided by a thorough grasp of its genetic legacy and ecological interconnectivity.

Finally, the manuscript exemplifies ERGA's collaborative ambitions and achievements, capturing the spirit of creativity and collaboration that defines this ground-breaking effort. As ERGA continues to push the boundaries of genetic research, it remains dedicated to scientific excellence, inclusivity, and the quest of knowledge for the benefit of society. I wholeheartedly recommend the publication of this groundbreaking initiative, offering my enthusiastic endorsement for its valuable contribution to the scientific community.

References

Larivière, D., Abueg, L., Brajuka, N. et al. (2024). Scalable, accessible and reproducible reference genome assembly and evaluation in Galaxy. *Nature Biotechnology* 42, 367-370. <https://doi.org/10.1038/s41587-023-02100-3>

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., ... Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, 115(17), 4325-4333. <https://doi.org/10.1073/pnas.1720115115>

Mazzoni, C. J., Claudio, C. i, Waterhouse, R. M. (2023). Biodiversity: an atlas of European reference genomes. *Nature* 619 : 252-252. <https://doi.org/10.1038/d41586-023-02229-w>

Mc Cartney, A. M., Formenti, G., Mouton, A., Panis, D. de, Marins, L. S., Leitão, H. G., Diedericks, G., Kirangwa, J., Morselli, M., Salces-Ortiz, J., Escudero, N., Iannucci, A., Natali, C., Svoldal, H., Fernández, R., Pooter, T. de, Joris, G., Strazisar, M., Wood, J., ... Mazzoni, C. J. (2024). The European Reference Genome Atlas: piloting a decentralised approach to equitable biodiversity genomics. *bioRxiv*, ver. 4 peer-reviewed and recommended by Peer Community in Genomics. <https://doi.org/10.1101/2023.09.25.559365>

Reviews

Evaluation round #2

Reviewed by **Eric Crandall** , 01 April 2024

The authors have done well to address both reviewers' comments. Just a few very minor comments below. Congratulations!

Specific comments:

P4 "are limited in scope due in large PART to a current lack of standardisation"

P8 the percentages for self-reported gender are reversed and differ from what is in Figure 2b.

P13 "To support this RECOMMENDATION we issued supporting guidance for bioanking"

Finally, a lot of text was moved around, so I may have missed something, but I can't find the text that was reported to have been added in Review 2 Response 12 to clarify recommendations. I've looked in the tracked-changes Word document as well as the posted preprint. I felt that the text was a good addition, and would like to see that it makes it into the final version.

Evaluation round #1

DOI or URL of the preprint: <https://doi.org/10.1101/2023.09.25.559365>

Version of the preprint: 2

Authors' reply, 07 March 2024

[Download author's reply](#)

[Download tracked changes file](#)

Decision by Jitendra Narayan , posted 04 January 2024, validated 08 January 2024

Revision !

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.

Reviewed by Justin Ideozu, 11 December 2023

Title: The European Reference Genome Atlas: piloting a decentralized approach to equitable biodiversity genomics

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.

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.

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
5. Technical Workflows

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.

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 describe challenges encountered in specific sections of Section 2. Authors should avoid repeating titles in subsections. For example, Training and Knowledge Transfer appeared twice.

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

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.

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.

Specific Comments

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.

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"

P6 "In other cases, **partnering sequencing** contributed their own grant funds" - sequencing partners?

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. 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.

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

P9 Step 2: Pre-sampling requirements: Taking all of this into account requires a lot of effort and I congratulate the authors for making it a part of their infrastructure from the beginning.

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 MlxS metadata standards for their metadata to ensure their FAIRness. See major comments above.

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 MlxS or Darwin Core - it might be a good time to meet to discuss standardization of this information.

P12 "All 98 of genome teams" – All 98 genome teams

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.

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.

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?

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.

Literature Cited

Riginos C, Crandall ED, Liggins L, Gaither MR, Ewing RB, Meyer C, Andrews KR, Euclide PT, Titus BM, Therkildsen NO, Salces-Castellano A, Stewart LC, Toonen RJ, Deck J. 2020. Building a global genomics observatory: Using GEOME (the Genomic Observatories Metadatabase) to expedite and improve deposition and retrieval of genetic data and metadata for biodiversity research. *Molecular Ecology Resources* 20:1458–1469. DOI: 10.1111/1755-0998.13269.