1st Conference on Research Data Infrastructure

Connecting RDM

https://doi.org/10.52825/CoRDI.v1i.394

Published: 07 Sept. 2023

The German Human Genome-Phenome Archive in an International Context: Toward a Federated Infrastructure for Managing and Analyzing Genomics and Health Data

Luiz Gadelha[https://orcid.org/0000-0002-8122-9522] and Jan Eufinger[https://orcid.org/0000-0002-3439-1674] on behalf of the GHGA-Consortium

German Human Genome-Phenome Archive, Germany

Abstract. With increasing numbers of human omics data, there is an urgent need for adequate resources for data sharing while also standardizing and harmonizing data processing. As part of the National Research Data Infrastructure (NFDI), the German Human Genome-Phenome Archive (GHGA) strives to connect the data from German researchers and their institutions to the international landscape of genome research. To achieve this, GHGA partners up with international activities such as the federated European Genome-Phenome Archive (EGA) [1] and the recently funded European Genomic Data Infrastructure (GDI) project to enable participation in international studies while ensuring at the same time the proper protection of the sensitive patient data included in GHGA.

Keywords: Genomics and Health Data, International Data Sharing, Federated Computing

1. Aims of GHGA and the need for international data sharing

To create a versatile and secure data infrastructure for genomics research, GHGA strives to provide (i) the necessary secure IT-infrastructure for Germany, (ii) an ethico-legal framework to handle omics data in a data-protection-compliant but open and FAIR [2] manner, (iii) harmonized metadata schemas, and (iv) standardized workflows to process the incoming omics data uniformly.

Genomic data is increasingly important in healthcare, allowing for clinical omics profiling of patients and enabling precision medicine, with tailored treatments having optimized efficiency for particular groups of patients. However, with many causal genetic alterations being typically very rare in given populations, successful knowledge generation in genome research critically depends on the availability of large cohorts of well curated reference datasets [3]. To exploit this potential it is therefore critical to share data at large scale, integrating data from different countries and their populations, incorporating diversity to the aggregated data set. To enable this, challenges bot on the technical side, caused by the need to manage and analyze the amounts of genomic data being collected, but also on the legal side, introduced by the inherent need for protection of individual's genome data, need to be met in a globally coordinated effort.

2. Federated Infrastructure to enable international science

Different projects are implementing federated infrastructures across national and continental scales in order to facilitate access to genomics and health data. The European Genome-Phenome Archive (EGA), run by the EMBL-EBI in the UK and the CRG in Spain, is one of the main repositories for genomics and health data started in 2008 [1]. There is ongoing work to expand EGA into a federated data infrastructure, the Federated EGA. In this model, countries will implement their own nodes following interoperability standards that will allow for sensitive data under controlled access to be stored in the national Federated EGA nodes. This avoids the need for further legal regulations and the exchange of person-related data with other nodes or the Central EGA (which corresponds to the original EGA). Similarly, the European Genomic Data Infrastructure (GDI) project is implementing a federated data infrastructure to allow access to genomics and health data across Europe. GDI will additionally allow for omics analysis workflows to be executed in the federated infrastructure. Jointly funded by the European Commission and the EU member states, GDI is an outcome of the Beyond 1 Million Genomes (B1MG) project and the 1+Million Genomes (1+MG) [4] initiative.

Positioned as the German national node both within the Federated EGA and the new GDI project, GHGA will enable German research projects to not only connect to those large scale international activities but also to contribute to the shaping of new standards and infrastructures for advancing research across Europe.

3. Global efforts to support genome research

Beyond these European activities, the Global Alliance for Genomics and Health (GA4GH) [5] provides many of the standards through which interoperability can be achieved in federated infrastructures. Consequently, both GHGA and its European partner projects include GA4GH developments into their mode of operation.

Data discovery is a key component of federated data infrastructures, and Beacon [6] is an example of data discovery protocol that enables researchers to search for and discover genomic and phenotypic data across different repositories and platforms without the need to expose sensitive information on individuals publicly. The Phenopackets [7] standard can be used for sharing phenotype and disease information along with associated data, such as genomics, diagnostics, and treatments. GA4GH Authentication and Authorization Infrastructure (AAI) and Passport standards [8] allow researchers to access data and resources across different platforms using a single set of credentials. The Data Use Ontology (DUO) [9] is used to annotate data sets with restrictions about their usage, standardizing the process of data access and use. Crypt4GH [10] provides a protocol for securely storing and sharing genomic data using public-key cryptography. GHGA closely follows and uses many of these standards to achieve interoperability for sharing genomics and health data within Germany and in Federated EGA and GDI. The GHGA Metadata Catalog was recently launched and consists of a public frontend for the discovery of study data from German research institutions, enabling the search of non-personal metadata. It aims to create a resource to collect information on human omics datasets available from German institutions for secondary research under controlled access conditions. The first data sets discoverable are 62 whole exomes, whole genomes and RNA sequencing data sets of 1310 patients suffering from 20 different rare cancer types from the NCT, DKFZ, and DKTK MASTER program. In the upcoming GHGA Archive, omics data will be stored and made available to other researchers after approval by the Data Access Committee of the data set controller.

GHGA is aiming to be more than an archive and consequently also contributes to the development and standardization of bioinformatics workflows for data analysis, benchmarking, statistical analysis, and visualizations . Here we are working with the global nf-core [11] community and also integrate GA4GH standards, such as the Workflow Execution Service (WES) and Tool Execution Service (TES) to also enable the application of the FAIR principles [12] in workflow development.

By delivering a national IT infrastructure for data sharing and analysis, an ethico-legal framework, metadata schemas, and standardized and reproducible workflows, GHGA will enable cross-project analysis and promote new collaborations and research projects in the international context of genome research. This will also be of high importance for new national projects such as e.g. the upcoming genomDE project within Germany.

Author contributions

Luiz Gadelha and Jan Eufinger wrote the initial draft. Members of the GHGA-Consortium revised and supervised the manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

The research project GHGA - German Human Genome-Phenome Archive is funded by the German Research Foundation (DFG) within the framework of the National Research Data Infrastructure (NFDI).

References

1. M. A. Freeberg *et al.*, “The European Genome-phenome Archive in 2021,” *Nucleic Acids Res.*, vol. 50, no. D1, pp. D980–D987, Jan. 2022, doi: 10.1093/nar/gkab1059.
2. M. D. Wilkinson *et al.*, “The FAIR Guiding Principles for scientific data management and stewardship,” *Sci. Data*, vol. 3, p. 160018, Mar. 2016, doi: 10.1038/sdata.2016.18.
3. Z. Stark *et al.*, “Integrating Genomics into Healthcare: A Global Responsibility,” *Am. J. Hum. Genet.*, vol. 104, no. 1, pp. 13–20, Jan. 2019, doi: 10.1016/j.ajhg.2018.11.014.
4. G. Saunders *et al.*, “Leveraging European infrastructures to access 1 million human genomes by 2022,” *Nat. Rev. Genet.*, vol. 20, no. 11, pp. 693–701, Nov. 2019, doi: 10.1038/s41576-019-0156-9.
5. H. L. Rehm *et al.*, “GA4GH: International policies and standards for data sharing across genomic research and healthcare,” *Cell Genomics*, vol. 1, no. 2, p. 100029, Nov. 2021, doi: 10.1016/j.xgen.2021.100029.
6. J. Rambla *et al.*, “Beacon v2 and Beacon networks: A ‘lingua franca’ for federated data discovery in biomedical genomics, and beyond,” *Hum. Mutat.*, p. humu.24369, Apr. 2022, doi: 10.1002/humu.24369.
7. J. O. B. Jacobsen *et al.*, “The GA4GH Phenopacket schema defines a computable representation of clinical data,” *Nat. Biotechnol.*, vol. 40, no. 6, pp. 817–820, Jun. 2022, doi: 10.1038/s41587-022-01357-4.
8. C. Voisin *et al.*, “GA4GH Passport standard for digital identity and access permissions,” *Cell Genomics*, vol. 1, no. 2, p. 100030, Nov. 2021, doi: 10.1016/j.xgen.2021.100030.
9. J. Lawson *et al.*, “The Data Use Ontology to streamline responsible access to human biomedical datasets,” *Cell Genomics*, vol. 1, no. 2, p. 100028, Nov. 2021, doi: 10.1016/j.xgen.2021.100028.
10. A. Senf *et al.*, “Crypt4GH: a file format standard enabling native access to encrypted data,” *Bioinformatics*, vol. 37, no. 17, pp. 2753–2754, Sep. 2021, doi: 10.1093/bioinformatics/btab087.

11. P. A. Ewels *et al.*, “The nf-core framework for community-curated bioinformatics pipelines,” *Nat. Biotechnol. 2020 383*, vol. 38, no. 3, pp. 276–278, Feb. 2020, doi: 10.1038/s41587-020-0439-x.
12. C. Goble *et al.*, “FAIR Computational Workflows,” *Data Intell.*, vol. 2, no. 1–2, pp. 108–121, Jan. 2020, doi: 10.1162/dint_a_00033.
13. M. Herschel, R. Diestelkämper, and H. Ben Lahmar, “A survey on provenance: What for? What form? What from?,” *VLDB J.*, vol. 26, no. 6, pp. 881–906, Dec. 2017, doi: 10.1007/s00778-017-0486-1.
14. S. Cohen-Boulakia *et al.*, “Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities,” *Future Gener. Comput. Syst.*, vol. 75, pp. 284–298, Oct. 2017, doi: 10.1016/j.future.2017.01.012.
15. J. Ison *et al.*, “EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats,” *Bioinformatics*, vol. 29, no. 10, pp. 1325–1332, May 2013, doi: 10.1093/bioinformatics/btt113.
16. A. Gray, C. Goble, and R. Jimenez, “Bioschemas: From Potato Salad to Protein Annotation,” in *Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks co-located with 16th International Semantic Web Conference (ISWC 2017)*, 2017.