Progress Report

German Human Genome-Phenome Archive (GHGA)

Consortium Progress Report

National Research Data Infrastructure (NFDI)

September 2023

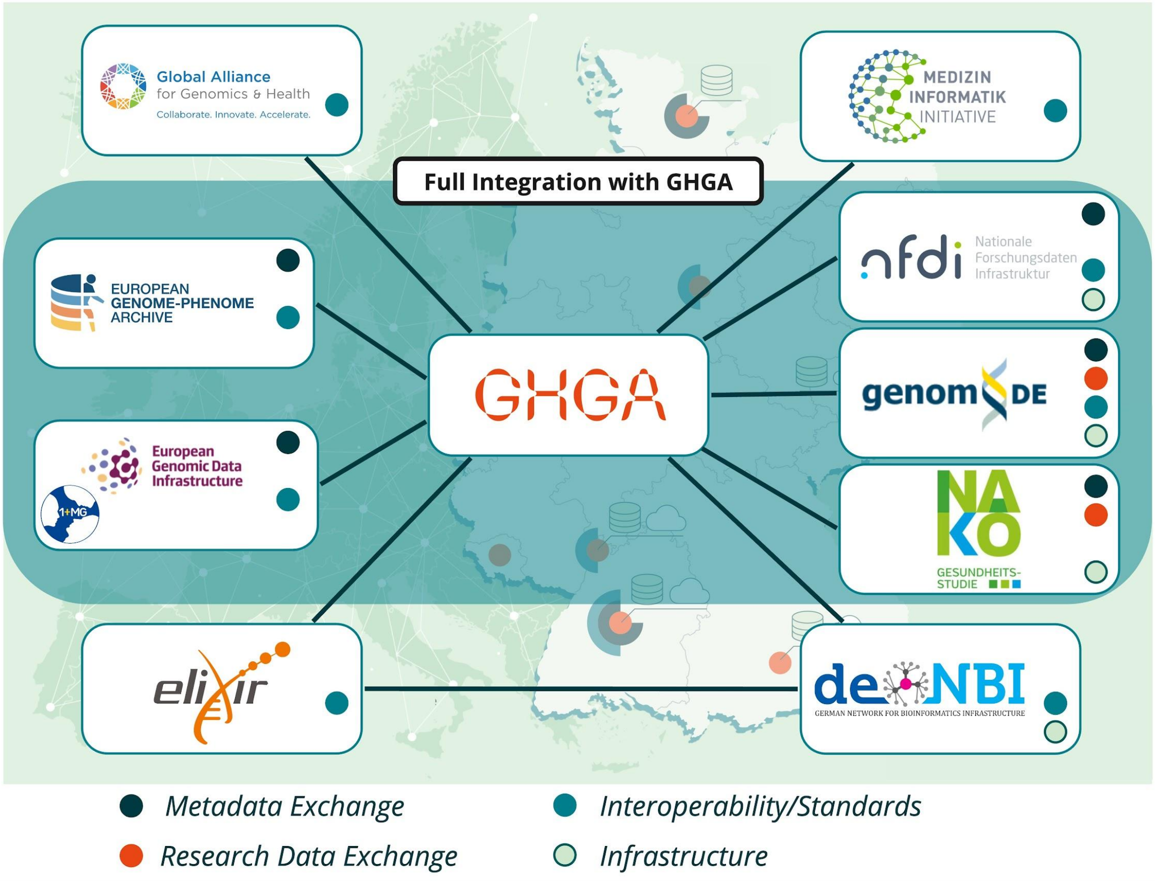
Full Integration with GHGA
GHGA
Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.
MEDIZIN INFORMATIK INITIATIVE
EUROPEAN GENOME-PHENOME ARCHIVE
nfdi Nationale Forschungsdaten Infrastruktur
European Genomic Data Infrastructure
1+MG
genomDE
NAKO GESUNDHEITS-STUDIE
elixir
de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE
Metadata Exchange
Interoperability/Standards
Research Data Exchange
Infrastructure

Figure 1: Alignment of GHGA with related international and national activities. Coloured dots indicate main areas of interaction.

Contents

B-1 Progress Report Template Part 1, for publication

B-1.1 General Information

Name of the consortium

German Human Genome-Phenome Archive (GHGA)

Research domains or research methods addressed by the consortium

GHGA is building a secure national omics data infrastructure that provides a framework for the use of human genome data for research purposes while preventing data misuse. GHGA is designed to bridge the current gap between research and medical care, and creates the opportunity to effectively use both data and technology to develop new therapies and diagnostic tools. This will bring together different fields, in particular biomedical and genomics research, data science and bioinformatics as well as healthcare.

Main DFG Review Boards covered

- Basic Research in Biology and Medicine (201)
- Medicine (205)
- Microbiology, Virology and Immunology (204)
- Neurosciences (206)

URL of the consortium website and repositories used for publishing output

Website:	www.ghga.de
Repositories:	catalog.ghga.de github.com/ghga-de github.com/GHGA-Training
Zenodo:	zenodo.org/communities/ghga/

B-1.2 Summary

GHGA, the German Human Genome-Phenome Archive, is a national infrastructure providing services for the secure archival, sharing, and processing of access-controlled human omics data. GHGA provides a national interface that connects national data providers and the scientific community to European genomic data resources and initiatives, such as the European Genome-Phenome Archive (EGA), the European Genomic Data Infrastructure (GDI) and 1+Million Genomes. In this first reporting period, GHGA has set up an integrated multi-institutional team, established a versatile legal framework and technologies for sharing of access controlled data, and established GHGA on the national and international landscape as the infrastructure for human genomic and omics data in Germany.

As part of our service portfolio, GHGA launched the first version of a national omics data platform - the GHGA Metadata Catalog - a platform that allows scientists to find and apply for access to human omics data. Currently, GHGA Metadata Catalog holds metadata from over 80 different datasets submitted via three data hubs with thousands of sharable genomes. In addition to significantly enhancing the *findability* of omics data, the launch of this service has delivered foundational technologies and tools for data sharing, including the establishment of a federated IT infrastructure, the technical and legal framework for human omics data sharing, a dedicated metadata model for human omics data, international data linkage and a sustainable and agile open source software development approach. GHGA also invested considerable resources in the development and implementation of legal, data protection, and information security concepts. In addition to underpinning the operations of GHGA services, the results from these activities have been made available to the community in the form of white papers, opinion articles and peer reviewed articles.

Building on the GHGA Metadata Catalog, GHGA is in the process of releasing the first major update to its services, which will complete the suite of services and tools to enable data providers to implement the full data life cycle for FAIR data sharing within GHGA. As part of this extended service portfolio, data deposited in GHGA will also be findable and accessible in major European archival efforts GHGA is connected to, most notably the federated EGA and GDI.

Together with the technical implementation measures, GHGA has established seven GHGA Data Hubs, which provide the necessary physical infrastructure to operate GHGA, and which are connected to major sequencing centres and local data providers. In parallel, GHGA has implemented various outreach measures to create close ties with the German bioinformatics and genomic research community. As part of these measures, GHGA contributed to the development, improvement and standardisation of important bioinformatics workflows, which have been developed together with global community efforts such as nf-core and ELIXIR.

With ethical and legal considerations being a key topic for GHGA, we have established a comprehensive ethical and legal framework, which on the one hand is closely integrated with the technical implementation of GHGA and connects to the needs of patients to ensure that their

interests and expectations for the secondary use of their genomic data are met. As part of these activities, we have contributed to the development of informed consent modules for genomic data, which feed into major transnational studies and the national broad consent. As part of these measures, GHGA has initiated a close engagement with patient representatives, supported medical ethics research and developed a new legal and ethical framework, which is forming the foundation of GHGA data services. To raise awareness for the need for FAIR data sharing more broadly, we have developed dedicated training programs for human omics data, and we have set up outreach measures that target the interested public. The conceptual work carried out in GHGA has resulted in a much-improved recognition of the needs of the human omics research community, both on a national and an international level. GHGA is embedded in genomDE, a national strategy program to establish genome sequencing in routine clinical care throughout Germany, where GHGA and its associated PIs take leading roles in the design of the underlying data infrastructures. The engagement in national strategic infrastructure activities have created valuable connections and extended mandate beyond the NFDI: GHGA has been recognized as the German node for the federated European Genome-Phenome Archive in 2022 and since end of 2022 is also charged with the creation of the German node within the European flagship project GDI. These activities will also naturally feed into and contribute towards upcoming developments such as the European Health Data Space.

B-1.3 Composition of the consortium

Applicant institution	Location	Duration
German Cancer Research Center (DKFZ)	Heidelberg	10/2020 -

Spokesperson	ORCID	Institution, location	Duration
Oliver Stegle	0000-0002-8818-7193	DKFZ & EMBL, Heidelberg	10/2020 -

Co-applicant institutions	Location	Duration
Eberhard-Karls-Universität Tübingen (EKUT)	Tübingen	10/2020 -
University Hospital Tübingen (UKT)	Tübingen	10/2020 -
Charité - Universitätsmedizin Berlin (Charité)	Berlin	10/2020 -
Technische Universität München (TUM)	München	10/2020 -
Europäisches Laboratorium für Molekularbiologie (EMBL)	Heidelberg	10/2020 -
Max Delbrück Center for Molecular Medicine (MDC)	Berlin	10/2020 -
Technische Universität Dresden (TU Dresden)	Dresden	10/2020 -
University Hospital Heidelberg (UHH)	Heidelberg	10/2020 -
Heidelberger Akademie der Wissenschaften (HAdW)	Heidelberg	10/2020 - 03/2022
University of Heidelberg (UHD)	Heidelberg	04/2022 -
Universität zu Köln (UzK)	Köln	10/2020 -
Universitätsklinikum Schleswig-Holstein, Kiel (UKI)	Kiel	10/2020 -
Helmholtz Zentrum München (HMGU)	München	10/2020 -
Dt. Zentrum für Neurodeg. Erkrankungen e.V. (DZNE)	Bonn	10/2020 -
Universität des Saarlandes (UdS)	Saarbrücken	10/2020 -
German National Cohort (GNC)	Heidelberg	10/2020 -
Helmholtz-Zentrum für Infektionsforschung (HZI)	Braunschweig	07/2021-12/2022

Co-spokespersons	ORCID	Institution, location	Task area(s) / Workstreams	Duration
Peer Bork	0000-0002-2627-833X	EMBL, Heidelberg	B3 / Metadata	10/2020 -
Ivo Buchhalter	0000-0003-0764-5832	DKFZ, Heidelberg	A3, C1, C2, C3, C5, D1, D2 / Data Hubs	10/2020 -
Andreas Dahl	0000-0002-2668-8371	TU Dresden	C1, D2 / Data Hubs	10/2020 -
Julien Gagneur	0000-0002-8924-8365	TU München	A1, C1, C2, C3, D2 / Workflows, Data Hubs	10/2020 -
Wolfgang Huber	0000-0002-0474-2218	EMBL, Heidelberg	A3 / Training	10/2020 -
Daniel Hübschmann	0000-0002-6041-7049	DKFZ, Heidelberg	C2, C3 / Workflows	10/2020 -
Oliver Kohlbacher	0000-0003-1739-4598	EKUT, Tübingen	A1, A3, C3, C4, D1, E1, E2 / All	10/2020 -

Jan Korbel	0000-0002-2798-3794	EMBL, Heidelberg	C2, C4, E2 / Workflows	10/2020 -
Martin Lablans	0000-0003-1880-5555	DKFZ, Heidelberg	A1, B2, C4 / Metadata	10/2020 -
Ulrich Lang (succeeded by S. Wesner)	0000-0001-7166-0805	UzK, Köln	C5, D1, D2 / Data Hubs	10/2020 - 10/2022
Peter Lichter	0000-0002-2960-5279	DKFZ, Heidelberg	A1 / Outreach	10/2020 -
Fruzsina Molnár-Gábor	0000-0002-9406-2776	HAdW (until 12/2021) U Heidelberg (since 01/2022)	B1 / ELSI	10/2020 -
Susanne Motameny	0000-0003-1186-1108	UzK, Köln	C1, D2 / Data Hubs	10/2020 -
Sven Nahnsen	0000-0002-4375-0691	EKUT, Tübingen	B2, B3, C1, C2, D1, D2 / Metadata	10/2020 -
Uwe Ohler	0000-0002-0881-3116	MDC, Berlin	A2, C2, C4 / Workflows	10/2020 -
Stephan Ossowski	0000-0002-7416-9568	UKT, Tübingen	A1, C2, C3 / Workflows	10/2020 -
Annette Peters	0000-0001-6645-0985	HMGU, München	A2 / Outreach	10/2020 -
Olaf Rieß	0000-0002-7011-1369	UKT, Tübingen	A1 / Outreach	10/2020 -
Philip Rosenstiel	0000-0002-9692-8828	UKI, Kiel	C1, D2 / Data Hubs	10/2020 -
Thorsten Schlomm	0000-0001-9557-4653	CHARITE, Berlin	A1 / Outreach	10/2020 -
Joachim Schultze	0000-0003-2812-9853	DZNE, Bonn	B3, D2 / Metadata	10/2020 -
Jörn Walter	0000-0003-0563-7417	UdS, Saarbrücken	A2 / Outreach	10/2020 -
Thomas Walter	-	EKUT, Tübingen	C5, D2 / Data Hubs	10/2020 -
Stefan Wesner	0000-0002-7270-7959	UzK, Köln	C5, D1, D2 / Data Hubs	10/2022 -
Juliane Winkelmann	0000-0002-3074-599X	HMGU & TUM, München	A1 / Outreach	10/2020 -
Eva C. Winkler	0000-0001-7460-0154	UHH, Heidelberg	B1, E2 / ELSI	10/2020 -

Participating institutions	Location	Duration
University of Cologne (UzK)	Cologne	10/2020 -
Max-Delbrück-Zentrum Berlin (MDC)	Berlin	10/2020 -
Charité, Berlin	Berlin	10/2020 -
German Cancer Research Center (DKFZ)	Heidelberg	10/2020 -
University Hospital Heidelberg	Heidelberg	10/2020 -
National Center for Tumor Diseases (NCT), Heidelberg	Heidelberg	10/2020 -
University Hospital Tübingen (UKT)	Tübingen	10/2020 -
Eberhard Karls University Tübingen (EKUT)	Tübingen	10/2020 -
Technical University Dresden (TUD)	Dresden	10/2020 -
National Center for Tumor Diseases (NCT) Dresden	Dresden	10/2020 -

Helmholtz Zentrum Munich (HMGU)	Munich	10/2020 -
Technical University Munich (TUM)	Munich	10/2020 -
Landesrechenzentrum München (LRZ)	Munich	10/2020 -
EMBL-EBI Cambridge, UK	Hinxton, UK	10/2020 -
Helmholtz-Zentrum für Informationssicherheit (CISPA)	Saarbrücken	10/2020 -
Helmholtz-Zentrum für Infektionsforschung (HZI)	Braunschweig	10/2020 -

Participating individuals	ORCID	Institution, location	Duration
Viktor Achter	0000-0002-3813-0746	UzK, Köln	10/2020 -
Dieter Beule	0000-0002-3284-0632	MDC, Berlin	10/2020 -
Benedikt Brors	0000-0001-5940-3101	DKFZ, Heidelberg	10/2020 -
Holm Graessner	0000-0001-9803-7183	UKT, Tübingen	10/2020 -
Michael Hummel	0000-0001-6717-605X	Charité, Berlin	10/2020 -
Dirk Jäger	-	UHH, Heidelberg	10/2020 -
Jens Krüger	0000-0002-2636-3163	EKUT, Tübingen	10/2020 -
Nisar Malek	0000-0002-9916-4608	UKT, Tübingen	10/2020 -
Thomas Meitinger	0000-0002-8838-8403	HMGU & TUM, München	10/2020 -
Wolfgang E. Nagel	-	TU Dresden	10/2020 -
Julio Saez-Rodriguez	0000-0002-8552-8976	UHH, Heidelberg	10/2020 -
Christoph Schickhardt	0000-0003-2038-1456	UHH, Heidelberg	10/2020 -
Thomas Keane	0000-0001-7532-6898	EMBL-EBI Cambridge, UK	10/2020 -
Mario Fritz	0000-0001-8949-9896	CISPA Saarbrücken	10/2020 -
Ninja Marnau	-	CISPA Saarbrücken	10/2020 -
Stephan Hachinger	0000-0001-8341-1478	LRZ München	10/2020 -
Alice McHardy	0000-0003-2370-3430	HZI Braunschweig	10/2020 -
Stefan Fröhling	0000-0001-7907-4595	National Center for Tumor Diseases (NCT) Heidelberg	10/2020 -
Hanno Glimm	0000-0003-4104-1135	National Center for Tumor Diseases (NCT) Dresden	10/2020 -