Letter of Intent within the Call for National Research Data Infrastructures (NFDI)

Renewal Proposal 2024

1 Binding letter of intent as advance notification of a full renewal proposal

This is a binding letter of intent, required as advance notification for renewal proposals in 2024.

2 Formal details

Name of the consortium: **German Human Genome-Phenome Archive**

Acronym of the consortium: **GHGA**

Applicant institution: **German Cancer Research Center (DKFZ)**
Heads: Prof. Dr. med. Dr. h.c. Michael Baumann and Ursula Weyrich

Spokesperson: **Prof. Dr. Oliver Stegle**
o.stegle@dkfz-heidelberg.de, German Cancer Research Center (DKFZ), Heidelberg

Co-applicant institutions

Institutions in bold have been newly added as Co-applicant institutions.

Co-applicant institutions	City	Head of Institution
Berlin Institute of Health @Charité (BIH)	**Berlin**	**Prof. Dr. Christopher Baum (Chairman of the Executive Board)** **Dr. Doris Meder (Acting Administrative Director)**
Eberhard-Karls-Universität Tübingen (EKUT)	Tübingen	Prof. Dr. Dr. h.c. (Dōshisha) Karla Pollmann (President and Vice-Chancellor)
European Molecular Biology Laboratory (EMBL)	Heidelberg	Prof. Edith Heard (Director General)
Helmholtz Munich (HMGU)	München	Prof. Dr. med. Dr. h.c. Matthias H. Tschöp (Chief Executive Officer)
Klinikum rechts der Isar der Technischen Universität München (MRI)	**München**	**Dr. med. Martin Siess (Medical Director)** **Marie le Claire (Administrative Director)**
Max Delbrück Center for Molecular Medicine (MDC)	Berlin	Prof. Dr. Maike Sander (Chairman of the Board of Directors and Scientific Director) Prof. Dr. Heike Graßmann (Administrative Director)
Technical University of Munich (TUM)	München	Prof. Dr. Thomas F. Hofmann (President) Dr. Emanuel Schreiner (Senior Executive Vice President)
Technische Universität Dresden (TUD)	Dresden	Prof. Dr. Ursula M. Staudinger (Rector) Dipl.-Ök. Jan Gerken (Chancellor)
Universität zu Köln (UzK)	Köln	Prof. Dr. Joybrato Mukherjee (Rector) Karsten Gerlof (Chancellor)
University Hospital Heidelberg (UHH)	Heidelberg	Prof. Dr. Ingo Autenrieth (Chief Medical Director and Chairman of the Executive Board) Katrin Erk (Administrative Director)
University Hospital Tübingen (UKT)	Tübingen	Prof. Dr. med. Jens Maschmann (Chief Medical Director and Chairman of the Executive Board), Dr. Daniela Harsch (Commercial Director and Vice-Chairwoman of the Executive Board)
University of Heidelberg (UHD)	Heidelberg	Prof. Dr. Frauke Melchior (Rector) Holger Schroeter (Chancellor)

Co-spokespersons

- Dieter Beule (dieter.beule@bih-charite.de) - MDC and BIH, Berlin
- Ivo Buchhalter (i.buchhalter@dkfz-heidelberg.de) - DKFZ, Heidelberg
- Andreas Dahl (andreas.dahl@tu-dresden.de) - TU Dresden
- Julien Gagneur (julien.gagneur@tum.de) - TUM, München
- **Holm Graessner** (holm.graessner@med.uni-tuebingen.de), EKUT and UKT, Tübingen
- Daniel Hübschmann (d.huebschmann@dkfz-heidelberg.de) - DKFZ, Heidelberg
- Oliver Kohlbacher (oliver.kohlbacher@uni-tuebingen.de) - EKUT, Tübingen & de.NBI e.V.
- Jan Korbel (jan.korbel@embl.org) - EMBL, Heidelberg
- Fruzsina Molnár-Gábor (fruzsina.molnar-gabor@uni-heidelberg.de) - Uni Heidelberg
- Susanne Motameny (susanne.motameny@uni-koeln.de) - UzK, Köln
- Sven Nahnsen (sven.nahnsen@uni-tuebingen.de) - EKUT, Tübingen
- Stefan Wesner (swesner@uni-koeln.de) - UzK, Köln
- Juliane Winkelmann (juliane.winkelmann@helmholtz-munich.de) - MRI, TUM & HMGU, München
- Eva Winkler (eva.winkler@nct-heidelberg.de) - UHH, Heidelberg

Participant institutions

Institutions in bold have been newly added.

Participant Institutions	City
Charité - Universitätsmedizin Berlin (Charité)	Berlin
de.NBI e.V.	**Heidelberg**
Deutsches Zentrum für Neurodegenerative Erkrankungen e.V. (DZNE)	Bonn
EMBL-EBI Cambridge, UK	Hinxton, UK
German National Cohort (GNC / NAKO) e.V.	Heidelberg
Helmholtz-Zentrum für Infektionsforschung (HZI)	Braunschweig
Helmholtz-Zentrum für Informationssicherheit (CISPA)	Saarbrücken
Leibniz-Rechenzentrum (LRZ) der Bayerischen Akademie der Wissenschaft	Garching near Munich
Medizinische Hochschule Hannover (MHH)	**Hannover**
National Center for Tumor Diseases (NCT) Dresden	Dresden
National Center for Tumor Diseases (NCT) Heidelberg	Heidelberg
Universität des Saarlandes (UdS)	Saarbrücken
Universität Freiburg (U FR), Freiburg	**Freiburg**
Universitätsklinikum Schleswig-Holstein, Kiel (UKI)	Kiel
ZB MED – Informationszentrum Lebenswissenschaften (ZBMED)	Cologne

Participant individual

Individuals in bold have been newly added, individuals marked with an # changed from co-spokesperson to participant status.

- Viktor Achter, UzK, Köln
- Peer Bork, EMBL, Heidelberg#
- Benedikt Brors, DKFZ, Heidelberg
- **Nataliya Di Donato, MHH, Hannover**
- **Juliane Fluck, ZB MED, Cologne**
- Mario Fritz, CISPA Saarbrücken
- Stefan Fröhling, DKFZ, UHH, NCT Heidelberg, Heidelberg
- Hanno Glimm, National Center for Tumor Diseases (NCT) Dresden
- **Björn Grüning, U FR, Freiburg**
- Stephan Hachinger, LRZ München
- **Karsten Häcker, MDC, Berlin**
- Wolfgang Huber, EMBL, Heidelberg#
- Michael Hummel, Charité, Berlin
- Dirk Jäger, UHH, Heidelberg
- Thomas Keane, EMBL-EBI Cambridge, UK
- Jens Krüger, EKUT, Tübingen
- Martin Lablans, DKFZ, Heidelberg#
- Peter Lichter, DKFZ, Heidelberg#
- Nisar Malek, UKT, Tübingen
- Ninja Marnau, CISPA Saarbrücken
- Alice McHardy, HZI Braunschweig
- **Christian Mertes, MRI, München**
- Wolfgang E. Nagel, TU Dresden
- **Ralph Müller-Pfefferkorn, TU Dresden**
- Uwe Ohler, MDC, Berlin#
- Stephan Ossowski, UKT, Tübingen#
- **Leo Panreck, NAKO e.V., Heidelberg**
- Annette Peters, HMGU & NAKO, München#
- **Tobias Pischon, MDC, Berlin**
- **Stefan Pfister, DKFZ, Heidelberg**
- Olaf Rieß, UKT, Tübingen#
- **Peter Robinson, BIH, Berlin**
- Philip Rosenstiel, UKI, Kiel#
- Julio Saez-Rodriguez, UHH, Heidelberg
- Christoph Schickhardt, UHH, Heidelberg
- Thorsten Schlomm, Charité, Berlin#
- Joachim Schultze, DZNE, Bonn#
- **Julia Schulze-Hentrich, UdS, Saarbrücken**
- Thomas Ulas, DZNE, Bonn
- Thomas Walter, EKUT, Tübingen#
- Jörn Walter, UdS, Saarbrücken#

3 Objectives, work programme and research environment in the second funding period

3.1 Research area of the proposed consortium

- 21 Biologie / 2.11 Basic Research in Biology and Medicine
- 22 Medizin / 2.22 Medicine

3.2 Concise summary of the consortium's main objectives and task areas

With the German Human Genome-Phenome Archive (GHGA) we have established a national infrastructure for sensitive human omics data. GHGA provides a unique service to the German science community by making human genomic and related omics data accessible for research. Previously, this data has often been kept private due to high barriers to address data protection and liability issues, limiting the potential for secondary research use offered by these data. The GHGA service portfolio is designed to open up this potential, and is embedded in key national and international initiatives: Within Germany, GHGA is connected to other NFDI consortia and major national projects. For example, GHGA will store all genome data for the recently started German Model Project Genome Sequencing (MV GenomSeq), a national initiative to generate 100,000+ human genomes (exomes/whole genomes) from oncological and rare disease cases. Internationally, GHGA is the designated German node both within the federated European Genome-Phenome Archive (fEGA) and the European Genomics Data Infrastructure (GDI), thereby connecting Germany to major European research data spaces. Together with these initiatives, GHGA has established interoperable metadata standards to FAIR-ify omics data for research. While enabling international discoverability, GHGA protects the sensitive human research data it is entrusted through strict access control, high security standards, and strict enforcement of the relevant data protection regulation.

During the first funding period, GHGA has established the legal, ethical, organizational, and technical framework for its service portfolio, allowing it to manage and process omics data on a petabyte scale. To accommodate national data protection requirements, GHGA was designed as a federated infrastructure, building on established HPC centers and the nationally largest academic cloud infrastructures (de.NBI/ELIXIR-DE Cloud). Early in the project, we launched the 'GHGA Catalog' - the first service for the discovery of omics in Germany and beyond. More recently, the development of the second phase of GHGA services, enabling full management of the research data, has been completed and entered test operation. GHGA and the archive as its core service consolidate a wide range of life science communities that generate, process, or interpret human omics data. It thus serves researchers from medicine, medical data science, biomedical informatics, genomics, molecular biology, and related disciplines.

Within the second funding period, GHGA will expand its service portfolio and take the next steps towards a sustainable business model, thereby acting as a reliable, internationally connected national infrastructure for human omics data. We will also expand our scope and users through a series of community-driven measures, thereby fostering the uptake of data and creating new secondary use scenarios by linking human omics to other data modalities managed by other NFDI consortia and elsewhere. The project will be structured in three major task area blocks (TAs A-C), which by themselves consist of several individual task areas.

TAs A (Operations) will consolidate structures to ensure the continued and sustainable operation of the infrastructure. This includes the continued maintenance, and improvement of GHGA software and the technical infrastructure (e.g., expanding storage and analysis capacities, information security) as well as the continued improvement of processes and adaptation to changing regulatory requirements. Operational improvements will include refined access controls, continued development of the metadata model, and metadata exchange within the fEGA and GDI. The next operational phases beyond GHGA Archive will be GHGA Cloud and GHGA Atlas. GHGA Cloud establishes a scalable Trusted Research Environment (TRE) compatible with related genomics TREs across Europe. GHGA Cloud will improve security and democratize omics data analysis as data download to (typically less secure) infrastructures will no longer be required. GHGA Atlas will build on GHGA Cloud, providing an infrastructure to process, curate, and publish community-specific derived datasets. In addition to expanding the service offerings, we will also establish a sustainable business model, which includes additional funding mechanisms and income for the consortium. A key partnership is the involvement as core infrastructure provider for the national MV GenomSeq, which will support GHGA data hubs financially. TA block A also includes further core services necessary for the sustainable operation of a data infrastructure, including legal and project management, user training and communication. Technical developments will be aligned to related activities (e.g., NFDIs, Base4NFDI, and fEGA/GDI). To this end, we will ensure (meta-)data interoperability, but also enter joint open-source code development to enable a sustainable and open joint data space.

TAs B (Communities) will address the needs of GHGA's core communities. Measures were selected from community-driven proposals, followed by a prioritization by the scientific advisory board together with the board of directors of GHGA. The individual measures aim to (a) mobilize additional existing and forthcoming data sets from within our communities (e.g., oncological networks, rare disease networks, common disease), (b) support ongoing data generation projects with data depositions to GHGA (e.g., MV GenomSeq, NAKO), (c) promote the adoption of the archive by increasing outreach, training and deepen interactions with funders, (d) work with the communities to develop community-tailored analysis use cases and data resources for GHGA Atlas, (e) ensure the readiness of GHGA for evolving national and international legislation, most notably the EHDS, and enable the use of federated learning/AI within GHGA as well as across fEGA/GDI, and (f) develop new transparency mechanisms to enable data controllers to engage with patients. A major difference compared to the first funding period is that several community engagement activities are now co-funded by contributions from the communities themselves (e.g., through complementary external grants with contributions to GHGA). GHGA will continue its successful **flexible funding** program to adapt to changes and needs arising within the consortium but also within the communities GHGA is serving **(TA C)**.

3.3 Brief description of the proposed use of existing infrastructures, tools and services that are essential in order to fulfill the planned consortium's objectives

GHGA is building on established services offered by national (c.f. Section 3.4) and international infrastructures (cf. Section 4). A core infrastructure for the operation of core GHGA services are computational and storage capacities provided by the GHGA Data Hubs. This federated network provides access - and personnel support - for large-scale IT infrastructure for the operation of GHGA. All Data Hubs have committed to support GHGA operations and are contractually bound partners of the GHGA Operations Consortium, a separate cooperation that governs decision making and ensures the interoperability of the GHGA Data Hubs. Additional regulations are being defined via bilateral contracts between each data hub institution and DKFZ as GHGA Central, ensuring clear responsibilities for the operations of GHGA. The services offered by the data hubs are funded by a mixed funding model, which includes federal, state and project-based funding elements. Prominent partners include members of the Gauß-Allianz/HPC network, the de.NBI Cloud operated by the long-term funded German Network for Bioinformatics Infrastructure, as well as project funding from the MV GenomSeq. The use of cloud technologies for the GHGA software stack provided by de.NBI ensures re-usability of our developments and simple roll-out to other NFDI consortia, several of which build on the same cloud technology. To ensure GHGA services are developed according to community needs, we will collaborate with and employ standards from major biomedical research networks and data providers such as the MV GenomSeq, the German National Cohort (NAKO), the NCT network, the German Biobank Alliance, the German Centers for Health Research (DZG), Research4Rare as well as our NFDI and international partners listed below.

3.4 Interfaces to other NFDI consortia: brief description of existing agreements for collaboration and/or plans for future collaboration

Besides the integral embedding of GHGA into key national activities in the biomedical sector (see ch. 4 below), GHGA has multiple lines of engagement in the NFDI community. First, we are exploiting and will further strengthen the natural synergies with consortia in the life and biomedical sciences such as **4Health**, **4Bioimage**, **4Immuno**, **4Biodiversity,** or **4Microbiota**, all of which consider omics data as a data modality, thus creating natural synergies. GHGA will continue to work closely with **4Health** to create collaborative and secure data analysis solutions for omics using TREs, especially in the context of the German National Cohort (NAKO) as a common use case. The harmonization of metadata schemes is an ongoing activity with **4Health**, **4Immuno** and **4Bioimage,** where GHGA brings its special expertise in the handling of sensitive health data also in the context of metadata schemes. This activity will also create opportunities for joint activities together with the TS4NFDI base service. Beyond synergies on technology, metadata and use cases, we will also join forces with **4Health** to impact policy-making efforts by representing NFDI in the ministry-endorsed coordination group on Health Research Data Infrastructures (GFDI), which will play a leading role e.g. in the process of integrating NFDI data into the European Health Data Space.

As a continuous effort, GHGA will contribute its strong expertise in ELSA and data protection/GDPR matters within the NFDI consortia and communities. In this context, GHGA is driving and co-developing several collaborations within NFDI: GHGA is leading the **Task Force Ethics** (TFE) as part of the NFDI **ELSA Section**. The TFE aims to identify ethical issues that arise across multiple NFDI consortia and to develop solutions and resources (such as guidelines, publications, or training activities) to address them. Connected to ELSA, GHGA is also driving the development of fact sheets on FAIR research and data protection within the **Task Force Datenschutz**.

GHGA is leading the Assured project, a pan-NFDI consortium together with **KonsortSWD**, and **BERD@NFDI**, aiming to develop a training and accreditation scheme for scientists working with sensitive research data. Together with **KonsortSWD** and **4Health**, GHGA is also developing concepts for linking and accessing personal data while maintaining data privacy. Furthermore, GHGA is actively engaging with **IAM4NFDI** base service to develop their legal and data protection concept for the usage of highly sensitive health data.

4 International and national networking

Nationally, GHGA will continue to consolidate its role as the major research infrastructure for human omics data in Germany. As the infrastructure provider for the MV GenomSeq (**German Model Project Genome Sequencing**), GHGA will operate the genome analysis centers ("Genomrechenzentren"), which will create new synergies and datasets feeding into the archive and additional income. Together with the BfARM as the designated federal node of the model project and other partners, we will develop a versatile and secure TRE to make sure the generated genome data can be used for innovative ground-breaking research. This activity is complemented by multiple other collaborations we have established and will extend in the next funding phase: Among others, GHGA will work together with NAKO to set up an analysis platform for large-scale sequencing of the cohort and will establish analysis platforms for various ongoing research projects in the context of the NCT-network, the national prevention center, the Bavarian Health Cloud and others.

From the beginning, GHGA has emphasized **international embedding and compatibility** in all its design concepts. In the next funding phase, we will strengthen those activities to embed NFDI efforts in European initiatives and beyond. First, we will continue our work as a founding member of the **federated European Genome-phenome Archive (fEGA),** which makes GHGA data findable and compatible with other national data providers in the fEGA. With the GHGA infrastructure in place, we also plan to strengthen our engagement on a technical level to develop fEGA into a powerful platform for international data exchange and analysis. Second, these activities will be complemented by our engagement in the **European Genomic Data Infrastructure (GDI) project,** which is developing the data infrastructure for the **1+ Million Genomes initiative (1+MG).** GHGA will also act as a bridge between the research community and national policymakers. Working together with the relevant ministries in Germany, we will ensure that solutions developed in GDI fit the needs of the research communities and, via this, also open up new avenues within the **European Health Data Space (EHDS)**. Furthermore, GHGA is closely collaborating with ELIXIR and uses the infrastructure provided through the compute platform of ELIXIR-DE to operate parts of its infrastructure.

Beyond Europe, we aim to foster our activities within the **Global Alliance for Genomics and Health (GA4GH)** both on the policy-making level and by proposing GHGA-developed technologies for the advancement of GA4GH standards and tools.

These top-level activities are supplemented by community-specific engagement with international activities such as the **nf-core community** (workflows), **UNCAN.eu**, the **EOSC-A Health Data Task Force (human health data)**, **EOSC4Cancer** (Cancer) or the European Rare Disease Research Alliance - “**ERDERA**”.