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

1. Binding letter of intent as advance notification or non-binding letter of intent

This is a binding letter of intent as advance notification for proposals in 2019.

2. Formal details

Planned name of the consortium German Human Genome-Phenome Archive

Acronym of the planned consortium GHGA

Applicant institution

The proposal is currently being jointly coordinated by two lead institutions. By the time of submission of the full proposal, both institutions will determine which institution will be responsible for the financial and administrative management of the proposed infrastructure.

Deutsches Krebsforschungszentrum
Im Neuenheimer Feld 280
69120 Heidelberg
Prof. Michael Baumann

Spokesperson
Oliver Stegle,
o.stegle@dkfz.de,
Deutsches Krebsforschungszentrum (DKFZ) and European Molecular Biology Laboratory (EMBL) Heidelberg

Eberhard Karls University Tübingen
Geschwister-Scholl-Platz
72072 Tübingen
Prof. Bernd Engler

Spokesperson
Oliver Kohlbacher, oliver.kohlbacher@uni-tuebingen.de,
University of Tübingen, MPI for Developmental Biology, and University Hospital Tübingen

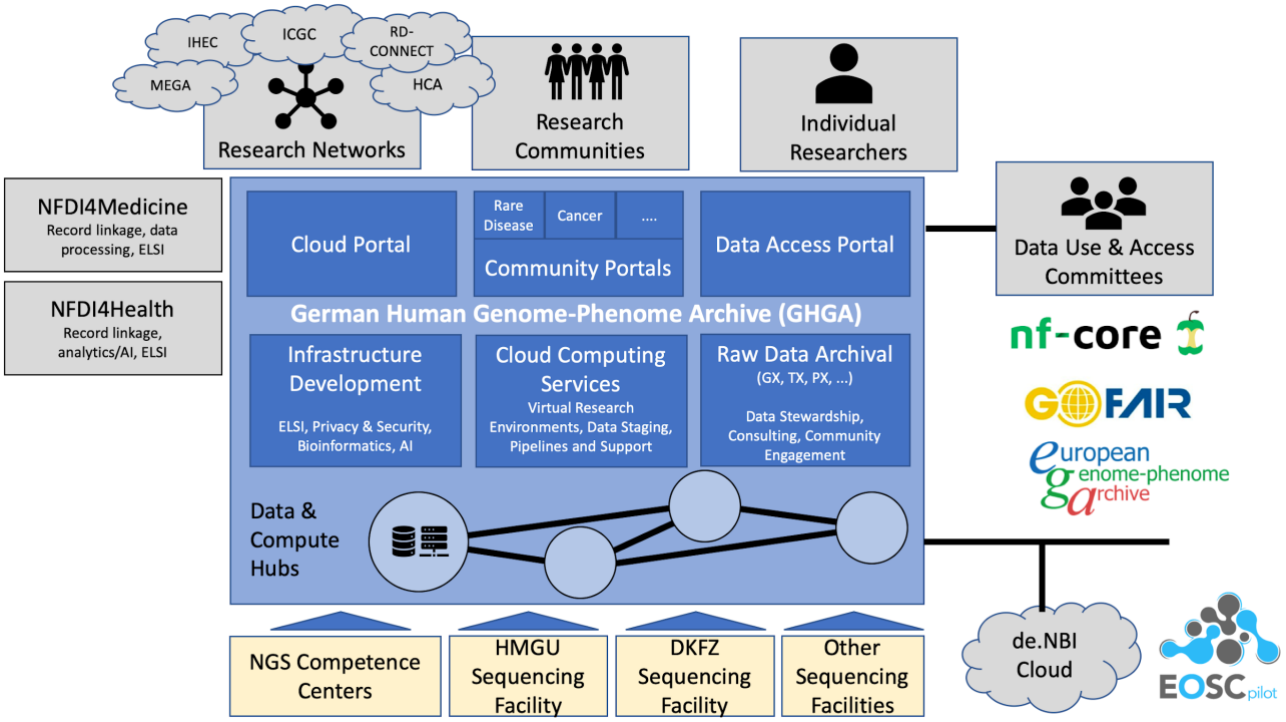
3. Objectives, work programme and research environment

Research area of the proposed consortium (according to the DFG classification system)

22 - Medicine, 21 - Biology

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

Genome sequencing and other omics technologies are among the most prominent and high-volume data sources in the life sciences, with major applications in basic biology, translational research and medicine. Clinical omics profiling of patients is expected to dominate large-scale data generation in the near future, providing unprecedented opportunities for use of these data in research. While initiatives exist to harmonize phenotypic data, in particular medical health records, there is a lack of infrastructure for high-volume omics data. While these data are already generated at scale by centers of excellence across Germany, legal, ethical and technical hurdles currently preclude managed access and data reuse for research at a national and international level. Such a national infrastructure could integrate existing and future omics data resources and link these to phenotype information. It will open up major scientific avenues and will deliver harmonized molecular profiles from large cohorts. Additionally, such an infrastructure will create an invaluable bridge between biomedical research and healthcare, opening the door for scientists in Germany to participate in key international research networks. This would tremendously boost genome science research in Germany and help to close the gap to European champions such as the United Kingdom, Denmark or Finland. Existing and forthcoming European infrastructure can complement national efforts, but cannot replace national infrastructures for financial, legal, and regulatory reasons. An overview of the structure of the planned infrastructure and its interactions is given in the following figure.



The core mission of this NFDI consortium is to address this need by establishing a national archive together with an analytics platform for human genome and phenome data. With an initial focus on human omics data types, ranging from whole genome sequencing data, epigenetic, transcriptome profiling, single-cell sequencing, proteomics to microbiome readout, the consortium seeks to establish a platform for data ingest, access, management and archival of human omics data. Access

to data and community buy-in will be achieved by engaging with clinical partners and major data generation centers in Germany, including biomedical research hubs and the recently established NGS competence centers (NGS-CN) of DFG.

Using state-of-the-art cloud technologies, GHGA will enable distributed analytics of large-scale sequencing datasets, thereby providing a platform for harmonized data processing, analysis and data reuse. The GHGA consortium will work closely with ethics and legal experts to address data processing, particularly data security concerns, and to establish an ethico-legal framework for population-scale data sharing and research, including harmonized patient consent. Best practice guidelines will be developed in accordance with applicable national and international regulation.

On a technical level, GHGA activities will build on and extend existing, reliable and secure high-performance computing infrastructures established by members of the consortium. A network of data hubs directly connected to the major data generators will handle the data in a distributed manner. Using cloud technologies, we will make this distributed infrastructure accessible to researchers in an integrated and seamless manner. Based on the needs, researchers will have access to raw sequence data, as well as analysis results generated using harmonized, internationally recognized analysis workflows. The consortium will drive open science solutions that are fully aligned with ELIXIR, the existing European Genome-Phenome Archive (EGA) at EBI and CRG and its federation strategy. To ensure quality and comparability with international standards, we will engage in projects such as GA4GH to foster international data exchange in current and upcoming studies (from ICGC ARGO to IHEC, HCA to rare disease studies to MEGA).

The GHGA will be open to data submission and projects across all fields of human omics. The initial focus will be on seed communities that drive the national efforts for research centric as well as clinical sequencing at scale - rare diseases, oncology and (genetic) epidemiology. These communities are well represented by members of the consortium, and personalized omics-based patient management is expected to play a major role in these domains. Building from these seed communities, the center will expand into other communities in the future, as well as handling additional data types such as proteomics and eventually imaging data. In parallel to establishing infrastructure and data resources, the consortium will drive innovation projects, flagship use cases and community portals to foster the immediate scientific exploitation of the established data resources. In particular, the availability of large homogenized national datasets and federated computing will enable population-scale omics studies, interrogating genotype-phenotype relationships in rare disease, human cancers and large epidemiological cohorts. To this end, GHGA will create interfaces with epigenomic resources (i.e. IHEC) as well as phenotype-centric data resources and networks (e.g., data integration centers within the Medical Informatics Initiative, NAMSE, RD-CONNECT, “bridgeheads” present in DKTK and comprehensive cancer centers). The consortium will also expand into novel omics technologies, including single-cell genomics, and foster interfaces between data opportunities and novel analytical methods based on machine learning and artificial intelligence. Finally, the GHGA will act as a platform for novel patient-centric data sharing initiatives, which create feedback loops between patients, clinicians and researchers, providing incentives for open data sharing and the democratization of omics research in Germany.

3.2 Brief description of the proposed use of existing infrastructures, tools and services that are essential in order to fulfil the planned consortium’s objectives

GHGA closely interacts with **Germany’s major sequencing centers**, namely the NGS Competence Network of DFG (NGS-CN, speaker and GHGA co-spokesperson Joachim Schultze), including the West-German Genome Center (WGGC, Bonn/Köln/Düsseldorf), the NGS Competence Center Tübingen (NCCT), as well as the sequencing facilities of the German Center for Cancer Research

(DKFZ, Heidelberg) and the Munich Sequencing Alliance. With these centers, data deposition paths and standardized meta-data transfer will be agreed upon to facilitate seamless data deposition to GHGA with minimal effort for the user.

GHGA builds on - and connects with - the existing infrastructure of the **European Genome-Phenome Archive (EGA)**, with whom we will share and harmonize organizational principles (project management, access control management, data deposition), but also joint software developments (e.g., portals, data processing pipelines). For the analysis of genomics data, we will also work closely with the **de.NBI cloud infrastructure** and related European infrastructures (**ELIXIR**), the compute cloud of the German Bioinformatics Infrastructure Network funded by BMBF. Both Tübingen and Heidelberg are de.NBI cloud sites and will contribute compute power, expertise and training capacity to GHGA.

The necessary compute and storage infrastructure of GHGA will be provided by the existing infrastructure of local **compute centers** established at the HPC institutions of consortium members (München, Heidelberg, Tübingen, Köln, Dresden). Funding of these compute centers is mainly through federal and state funds from various programs.

We will also establish a close interaction with the **German Biobank Alliance (GBA)** and the **German Biobank Node (GBN)**, since high-quality biomaterials and associated metadata are of utmost importance for the generation of reliable and reproducible research data. GBA (coordinated by GBN) provides a biobank network of 20 university-based high-end biobanks which are prepared to establish record linking between high-quality biomaterials and associated metadata with GHGA data sets. In addition, the available biomaterials might be used to generate additional research data, thus enabling targeted extension of existing data sets.

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

The GHGA consortium has agreed on a close partnership with the planned **NFDI4Medicine** and **NFDI4Health** consortia. Together, these three consortia provide comprehensive and complementary infrastructure components: bridging storage and management of healthcare, medical research, and public health data (NFDI4Health & NFDI4Medicine) with omics-centric raw data archival, processing and analytics (GHGA). By linking the data modalities addressed by these consortia (ideally, in a privacy-preserving manner), it will be possible to integrate previously disjoint datasets in an unprecedented manner. In this context, GHGA will provide large-volume data storage for omics raw data and capacity to processing these data. The data processing infrastructure in GHGA will yield quantitative molecular readouts (e.g., genetic variants, gene expression quantification, epigenetic states), which can jointly analyzed together with healthcare data and medical research data. GHGA will work with NFDI4Medicine on standardization of data exchange formats (e.g., as part of the National Core Data Set of the Medical Informatics Initiative extension modules ‘genomics’, ‘oncology’, and ‘rare disease’) to facilitate analysis of processed omics data jointly with associated clinical phenotypes. Similarly, GHGA will cooperate with NFDI4Health to facilitate linkage of ‘omics’ data to information harvested in structured public health and clinical trial data. Ethical, legal, and societal impact are synergistic cross-sectional topics to which all three consortia can contribute. The consortium will also work together with other synergistic efforts. In particular, the **NFDI4Microbiota** consortium, which is focused on Microbiome tool development and analysis, plans to build on GHGA as archive for human microbiome data. We will also work closely with the umbrella consortium **NFDI4Life** to coordinate cross-sectional issues within the life sciences.

4. Cross-cutting topics

Even though the constellation and relative timing of the different NFDI consortia is not at this point, we foresee several cross-cutting topics to which we can contribute and/or benefit from.

Standardized phenotyping of subjects and privacy-preserving record linkage

Although the management of human phenotype data, such as medical records, is not part of the aims of GHGA, the consortia and the compute platform will require such cross-cutting developments. Access to phenotype data for the corresponding samples will enable integrating molecular data modalities in GHGA with phenotypic outcomes. We expect this infrastructure to be established by the NFDI4Medicine consortium.

Consent management and ethico-legal framework for patient data

The GHGA consortium will establish mechanisms to manage patient consent, as well as an ethico-legal framework for research using patient data. These efforts will benefit all NFDI consortia working with sensitive data and person-related healthcare data in particular.

Cloud platforms for scientific computing

The GHGA infrastructure will make extensive use of distributed local cloud technologies, both for compute and storage. These activities will build on the established expertise in the context of ongoing initiatives such as de.NBI. Consequently, the consortium can contribute expertise and reference solutions in this area, as well as benefiting from centralized technological developments.

Standardization for data processing and workflow management

GHGA will establish standardized infrastructure for processing large volumes of omics data in a consistent and reproducible manner. Theses technical developments of computational workflows that can be executed using cloud computing systems are a core component of GHGA. The underlying workflow management solutions will be generic and reference solutions and software implementations will be openly shared with other consortia, both inside and outside of the life sciences.

Integrative data analytics

As part of the analysis workflows and community services, GHGA will make extensive use of machine learning and statistical inference for integrating different data modalities. This includes multi-omics data analysis, spatio-temporal modelling and the integration between omics data and other data modalities, such as imaging and health records. The consortium will benefit from and can contribute expertise in this area by sharing methods, expertise and software.