German Human Genome Archive (GHGA)

Draft Proposal within NFDI, Extended Abstract

The GHGA Consortium - March 2019

1. Formal details

Planned title of the consortium and acronym

German Human Genome Archive (GHGA)

Lead institutions

DKFZ Heidelberg and University of Tübingen

Name and work address of a contact person

Prof. **Oliver Kohlbacher**, University of Tübingen, University Hospital Tübingen, and Max Planck Institute for Developmental Biology, Sand 14, 72076 Tübingen (Oliver.Kohlbacher@uni-tuebingen.de)

Members of the planned consortium (including institutional affiliation, without address - preliminary**)**

- Prof. Michael Backes, CISPA
- Dr. Ivo Buchhalter, DKFZ
- Prof. Benedikt Brors, DKFZ
- Dr. Andreas Dahl, TU Dresden
- Prof. Stefan Fröhling / Peter Lichter, NCT Heidelberg
- Prof. Julien Gagneur, TU Munich
- Prof. Dirk Jäger, University Hospital Heidelberg
- Prof. Oliver Kohlbacher, University of Tübingen, MPI for Developmental Biology, and University Hospital Tübingen
- Dr. Jan Korbel, EMBL Heidelberg
- Prof. Dieter A. Kranzlmüller, LRZ München/LMU München
- Prof. Mechthild Krause, NCT Dreseden
- Dr. Martin Lablans, DKFZ
- Prof. Ulrich Lang, University of Cologne
- Prof. Thomas Meitinger / Prof. Juliane Winkelmann / Dr. Tim Strom, Helmholtz-Zentrum München/TUM
- Dr. Fruzsina Molnár-Gábor, Heidelberg Academy of Sciences and Humanities
- Prof. Wolfgang E. Nagel, TU Dresden
- Dr. Sven Nahnsen, Quantitative Biology Center, Tübingen
- Prof. Peter Nürnberg, Cologne Center for Genomics
- Prof. Stephan Ossowski, University Hospital Tübingen
- Prof. Olaf Rieß, University Hospital Tübingen
- Prof. Philip Rosenstiel, UK Kiel
- Prof. Thorsten Schlomm, Charité Berlin
- Prof. Joachim Schultze, DZNE & University of Bonn
- Dr. Oliver Stegle, DKFZ & EMBL Heidelberg
- Prof. Julio Saez-Rodriguez, Uniklinik Heidelberg
- Dr. Matthias Schlesner, DKFZ
- Prof. Thomas Walter, University of Tübingen
- Prof. Jörn Walter, Saarland University
- Prof. Eva Winkler, University of Heidelberg

Participants in the NFDI conference

- Dr. Oliver **Stegle**, DKFZ Heidelberg, EMBL (o.stegle@dkfz.de)
- Prof. Oliver **Kohlbacher**, University of Tübingen, MPI for Developmental Biology & University Hospital Tübingen (Oliver.Kohlbacher@uni-tuebingen.de)
- Prof. Joachim **Schultze**, DZNE & University of Bonn (j.schultze@uni-bonn.de)

Research area of the planned consortium (research area according to the DFG classification system

22 - Medicine, 21 - Biology

Participating research institutions

- DKFZ
- NCT Heidelberg
- NCT Dresden
- EMBL Heidelberg
- DKTK
- University Hospital Heidelberg
- University of Tübingen
- University Hospital Tübingen
- German Center for Neurodegenerative Diseases (DZNE), Bonn site
- University of Bonn
- University of Cologne
- UK Kiel
- MPI CBG, Dresden
- Helmholtz Zentrum München
- TU Munich
- Charité
- Helmholtz Center for Information Security (CISPA)
- Saarland University
- Heidelberg Academy of Sciences and Humanities

Participating infrastructure facilities and/or potential information service providers

- NGS Competence Center Tübingen (NCCT)
- West-German Genome Center (WGGC)
- Competence Centre for Genome Analysis Kiel (CCGA)
- Dresden-Concept Genome Center (DGC)
- German Cancer Research Center (DKFZ)
- Munich Sequencing Alliance
- Quantitative Biology Center (QBiC), University of Tübingen
- Heidelberg Center for Human Bioinformatics (HD-HuB)
- Center of Integrative Bioinformatics (CIBI), Tübingen
- Zentrum für Datenverarbeitung, Universität Tübingen
- Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH)
- Regionales Rechenzentrum der Universität zu Köln
- DZNE Bonn
- de.NBI Cloud
- LRZ München

Planned proposal submission date (2019, 2020, 2021)

2019

Overview diagram or organisational chart for the planned consortium

See Section 2.1 below.

2. Subject-specific and infrastructural focus of the planned consortium

2.1 Key questions/objectives of the consortium

Genome sequencing and omics are among the most prominent and high-volume data modalities in the life sciences, with major applications in basic biology, translational research and medicine. Clinical omics profiling of patients will be a dominant mode of data acquisition, providing unprecedented opportunities for use of these data in research. While initiatives exist to harmonize phenotypic data, including medical health records, there is a lack of infrastructure for high-volume sequencing 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. A national genomics infrastructure to integrate existing and future omics data resources will open up major scientific avenues, delivering 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 the field of genome sciences in Germany, helping 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.

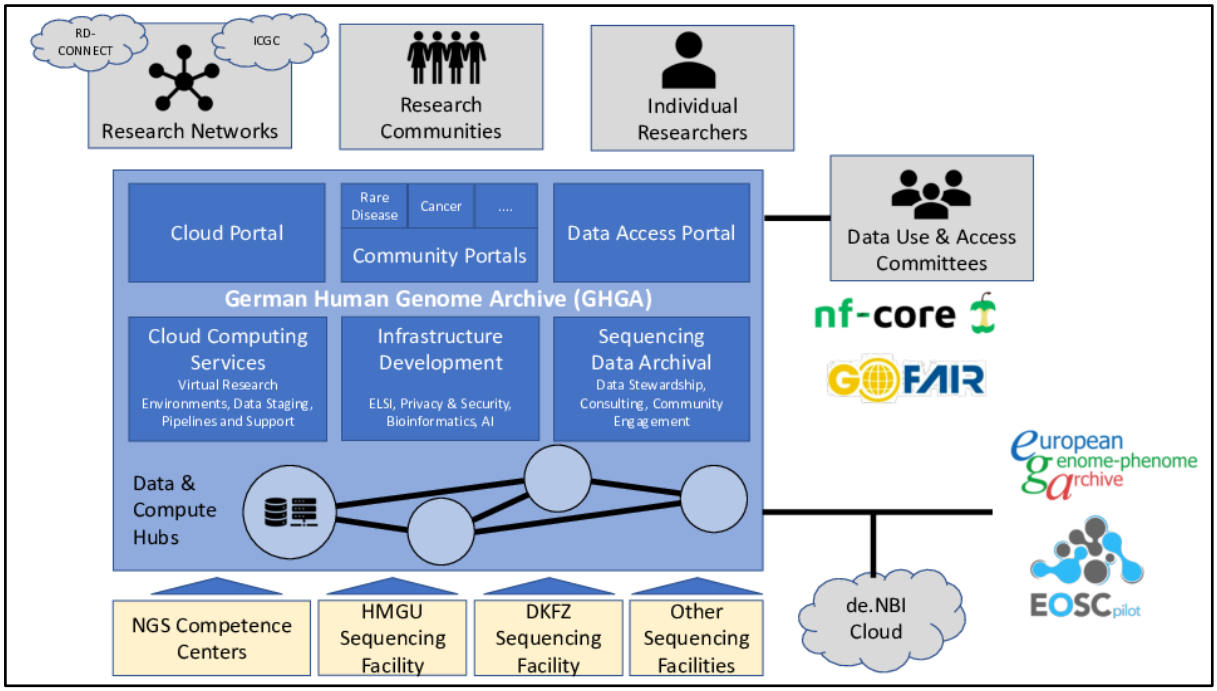

Fig. 1: Overall structure of the planned infrastructure and its interaction with the involved communities, and other national/international infrastructures/initiatives.

The core mission of this NFDI consortium is to address this need by establishing a **national genome archive for human genome data**. With an initial focus on human omics data types, the consortium seeks to establish a platform for **data ingest, access, management and archival** of sequence-based human 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 excellence centers. Using state-of-the-art cloud technologies, the GHGA will enable distributed analytics of large-scale sequencing datasets, thereby providing a platform for 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 infrastructure. A network of *data hubs* directly connected to the major data producers 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. The network will drive open science solutions that are fully aligned with ELIXIR, the existing European Genome-Phenome Archive (EGA) at EBI and its federation strategy. To ensure quality and comparability with international standards, the consortium will engage in projects such as GA4GH to foster international data exchange in current and upcoming studies (from ICGC ARGO to rare disease studies to "MEGA").

The GHGA will be open to data submission and projects across all fields of human genomics. The initial focus will be on two communities that drive the national efforts for clinical sequencing at scale - ***rare disease* biology** and ***oncology***. These communities are well represented by the partners of the consortium, and personalized omics-based patient management is expected to play a major role in both domains. Building from these seed communities, the center will expand into other areas in the future.

In parallel to establishing infrastructure and data resources, the consortium will drive **flagship use cases** and community portals to foster the scientific exploitation of the established data resources. The availability of large homogenized national datasets and federated computing will enable **population-scale omics studies,** interrogating genotype-phenotype relationships in rare disease and human cancers. To do so, the 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 foster interfaces between data opportunities and novel analytical methods, in particular methods based on **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.

2.2 Known needs/current status of research data management in the relevant discipline/subject-specific relevance of the planned consortium:

The only existing genome archives are international repositories such as the European Genome Archive. However, regulatory, legal, and technical hurdles limit data transfer and sharing within Germany and beyond national borders, and hence the vast majority of omics data that are generated in a clinical context are not centrally deposited in any archive. Instead, the predominant mode of data archival and management is via heterogeneous inhouse solutions developed by data generating institutions. Although case by case access can be arranged in principle, the scale of sequencing-based data in particular is a major limitation for collaborative research. A second major shortcoming of the *status quo* is the lack of consistent processing pipelines across centers. As a result, joint analysis of datasets

that span multiple centers is challenging and typically requires substantial data reprocessing efforts, which are only achievable by a small number of computationally well-equipped centers. The main aim of this proposal is to solve this challenge on a national level to then foster interfaces and data exchange with international data solutions.

2.3 Summary of the planned research data infrastructure that is specifically intended to address the needs of research users in their respective work processes

To address the needs of research users GHGA will provide the main services data archival, data access, and cloud-based analytics. As communities will have different needs for accessing/using the data and groups with different computational expertise will prefer different technical modes of access, GHGA will support alternative interoperable work processes. Access to these services will be obtained via a verified user account on the GHGA portal, which interfaces with existing authentication infrastructure (e.g., DFN AAI).

Data deposition - Data deposition requires the deposition of the raw data (e.g., FASTQ files) together with the associated technical metadata. Through the close interaction of GHGA with the DFG NGS competence centers and key clinical data generation centers, we are in a unique position that technical metadata and the primary sample metadata are already available in a digital format. The data deposition process will be streamlined to the point where it is fully automated, such that users are only required to check, expand, and confirm the metadata through the GHGA portal to deposit their data. Data deposition will build on automatically triggered annotation, quality control and integrity checks of the submission at one of the data hubs prior to deposition. The implementation of the ingest processes will be aligned and build on existing solutions, in particular the EGA.

Data access - The sensitive nature of the data stored in GHGA renders a controlled and secure access to the data critical. We intend to use the fundamental governance process for data access established at EGA as an orientation guide and to adapt this to the specific national and regional regulatory framework. The process involves an initial electronic application, review via a Data Access Committee (DAC), and subsequent data staging (if the application was granted). This process will be modelled fully electronically through the GHGA portal. Besides the generic, data set-centric, browsing of the available (meta-)data, community-centric science gateways (web-based & API) will provide more specific data exploration interfaces (see below). The data access portals will also support the generation of keys for linking samples in the GHGA with phenotypic data, such as those contained in medical informatics data processing systems.

Cloud-based data access - In order to avoid download/transfer of large-scale data, the GHGA Data Hubs offer HPC capacities that will be made available on a project basis through cloud-based technologies. Again, an electronic application review process (embedded in the data access process sketched above) grants access to data as well as to compute resources at one or more of the data hubs currently storing a replica of the required data. By providing a range of tested best-practice data processing workflows (containerized), GHGA data will be of consistent data quality and comparable across sites. These best practice workflows will respect compliance with relevant regulations for clouds.

Interfaces and portals for global genomics efforts - In order to connect to global efforts, such as reference variant information and cohorts available through ELIXIR and GA4GH, the GHGA will define appropriate interfaces. These will provide defined mechanisms for access to summary statistics and, where appropriate, search functionalities.

In order to address the specific needs of disease-related communities (initially, rare disease and cancer, but successively for other communities as well), we intend to build community gateways, i.e. science gateways tailor-made to cater to the specific needs and conventions of sufficiently large communities. The analysis of genomics data in the context of rare disease with its focus on few cases and deep phenotypical metadata (e.g., phenotype documentation with the Human Phenotype Ontology (HPO), orphanet codes) differs significantly from the large-scale statistical analysis typical for the cancer community. By building dedicated web-based access portals in close dialogue with the corresponding communities, we will ensure the usability of the infrastructure and maximize its impact on the communities. A clean and modular architecture of the underlying software stack will render future extensions to provide modified portals for different communities efficient and synergistic. More details on how these work processes will be integrated with the underlying infrastructure are given in the following section.

Ethico-legal framework - Underlying all activities will be an ethico-legal framework including emphasis on the above-listed normatively relevant issues of data processing, particularly data security, cloud computing and a unified consent structure that will be developed by the consortium as well as a policy on the handling of actionable incidental findings. When establishing this framework, compliance with both national and supranational regulatory provisions will be provided. Regarding consent, the legal basis for data storage and analysis will usually be an informed consent. Consents will be encoded and managed electronically. Consent modelling and encoding will be aligned with the current draft of a 'national broad consent' under development by the Medical Informatics Initiative and will be supplemented by specific information on genomics and research.

2.4 Description of data types, data processing, and data analysis methodologies

GHGA is designed to hold data from biological high-throughput technologies derived from human subjects and associated metadata. The expected data volume will be on the order of several petabytes per year during the initial phase, with further rapid growth. Data will be archived in open community standard formats (e.g., FASTQ, CRAM) wherever possible. Metadata associated with omics datasets chiefly falls into three categories: technical metadata (e.g., the instrumentation the data was generated and related information), sample metadata (e.g., tissue, organ, sample preparation), and patient phenotype data (e.g., HPO classification in rare diseases, oncological case documentation). Additional clinical phenotype data from existing external resources will be referenced and linked to GHGA data sets.

The **data archive** will require backed up file-system based storage of the underlying raw data, which will be kept on encrypted object storage, to be decrypted only upon access. Data ingest, quality control, and (technical) metadata extraction will be handled through reproducible containerized workflows. These workflows will be harmonized with other national and international initiatives and networks (e.g., ELIXIR, de.NBI, EGA, GA4GH). Metadata will be stored in relational database backends and will adhere to appropriate standards (e.g., aligned with core data sets developed within the medical informatics initiative). Data upload/transfer will be facilitated through established technologies (e.g., Aspera). Governance processes, role-based access management, electronic proposals, and user support will be handled through a central portal hosted at one of the data hubs.

A **cloud-based research environment** enables access to both access-controlled data and HPC resources. Together with an application for data access, users can apply for compute

resources. A review process with assigned priorities will prioritize these projects in case of insufficient resources. This process has been established at all HPC centers involved. We will unify this application process through the GHGA portal, building on experiences from previous projects such as de.NBI. Upon approval of projects, users are granted access to the relevant data as well as access to HPC resources (or virtual clusters) to perform the necessary computations. Dedicated staff will support users with the implementation of analysis pipelines and with technical issues related to HPC/cloud infrastructure.

2.5 Planned implementation of the FAIR principles and information about any existing policies or guidelines in the relevant discipline

We are working closely with the GO-FAIR initiative and each data hub will provide access to a federated metadata repository providing all relevant information on the omics data stored at any of the sites. These metadata repositories will act as FAIR Data Points, providing access to the metadata through REST APIs as well as through a user-friendly web-based UI. All data processing pipelines used in routine processing, quality control, data conversion, and annotation will be made available in a containerized form.
There are numerous international, national and state legislations governing the use of human genome data (e.g., GDPR, state data privacy laws, state hospital laws). Due to state-specific regulations, data hubs in different states within Germany will have to adhere to slightly different regulations. Within the proposed ethico-legal framework, we will develop best practice guidelines at each data hub and align these guidelines also being developed within related initiatives (e.g., Medical Informatics Initiative).

2.6 Planned measures for user participation and involvement

Sustainable infrastructures require the support of and close interaction with the communities they intend do serve. The GHGA addresses different communities on different levels. The three basic services it provides (archival, data access, cloud compute) cater to different, though overlapping, communities, which need to be engaged through different measures.
Archival is primarily of interest to researchers involved in data generation (data producers). Here, the direct interaction with the sequencing facilities (or other omics facilities) will be essential to engage this part of the community. For these communities we envision staying in contact through the website, social media, and an annual conference. Data consumers (researchers in the life sciences, bioinformaticians, AI researchers) will profit equally from the data access mechanisms (community portals, data staging, data download, FAIR data access APIs), but will typically choose different routes of access, depending on their IT proficiency and the type and extent of the questions asked. The community portals and an annual international symposium will bring together interested researchers with general interest in omics data analysis. Other parts of the communities need to be addressed more directly. The community portals will play a pivotal role there. We intend to evolve them into an indispensable tool at the heart of each of these communities. To this end, we will reach out with tutorials and presentations at the major conferences in the respective fields and direct communication with the established research networks.
All users will also be encouraged to provide feedback by automatically triggering questionnaires after data deposition or at defined periods after data access to obtain feedback on the user experience, but also to collect reliable information on the infrastructure's impact on the community (e.g., by asking for grants, publications supported by the GHGA).

2.7 Existing and intended degree of networking of the planned consortium

Substantial parts of the consortium are already working together to address key aims of the proposal, including cloud infrastructure, omics-based compute interfaces, workflow harmonization, or joint governance structures. GHGA is closely aligned with major national and international activities related to the research communities (initially cancer and rare disease) and on the infrastructure side (e.g., other genome archives, bioinformatics networks).

Nationally, GHGA will be closely integrated on the technical and governance side with the four established DFG-funded NGS competence centers and with established HPC infrastructures at the contributing compute centers. We intend to align our activities with the plans of related NFDI consortia (in particular, NFDI4Life, NFDI4Health, NFDI4Medicine, NFDI4Microbiome). Cross-referencing of data in GHGA with the phenotypic, clinical, or high-throughput data references within these consortia will be done through DOIs. We will be working closely with the Medical Informatics Initiative on data harmonization and interoperability (through its working group interoperability) to simplify linking of structured healthcare information to high-throughput data sets. We will harmonize our workflows with the corresponding activities within the de.NBI network (which the applicants are part of).

Internationally, we strongly involved in the Personal Health Train Implementation Network (PHT-IN) within the GO-FAIR initiative in order to facilitate distributed data analysis across sites (and countries). The infrastructure and technical development will be closely aligned with the development of GHGA's European counterpart EGA a the EBI. We will also synchronize with activities within Europe (ELIXIR) and beyond (GA4GH), for example, the GA4GH driver projects to which we intend to contribute and ELIXIR implementation studies. The developments of the federated cloud infrastructure will be aligned with the developments within the European Open Science Cloud (EOSC).

2.8 Additional information

None.