German Human Genome-Phenome Archive

(GHGA)

Renewal Proposal 2024

A Consortium within the NFDI

In cooperation with

Table of Contents

List of Abbreviations

Abbreviation	Definition
1+MG	1+ Million Genomes Project
AAI	Authentication and Authorisation Infrastructure
ACHSE	Allianz Chronischer Seltener Erkrankungen e.V.
AI	Artificial Intelligence
API	Application Programming Interface
BfArM	Federal Institute for Drugs and Medical Devices (Bundesinstitut für Arzneimittel und Medizinprodukte)
BIH	Berlin Institute for Health
BMBF	Bundesministerium für Bildung und Forschung / Federal Ministry of Education and Research
BMG	Bundesministerium für Gesundheit / Federal Ministry of Health
BoD	Board of Directors - Main governing body of GHGA
CDS	Central Data Stewards - Data steward team located at GHGA Central
cf.	See also
CISPA	Helmholtz Centre for Information Security
cSPE	community Secure Processing Environment
DAC, eDAC	Data Access Committee, electronic Data Access Committee
DMG	Data Mobilisation grant
de.KCD	German Competence Center for Cloud Technologies for Data Management and Processing
de.NBI	German Network for Bioinformatics Infrastructure (ELIXIR-DE)
DFG	Deutsche Forschungsgemeinschaft/German Research Foundation
DFN	German Research Network
DKFZ	German Cancer Research Center
DKTK	German Consortium for Translational Cancer Research DNA Deoxyribonucleic acid
DS	Data Steward
DZIF	German centre for infection research
DZNE	Deutsches Zentrum für Neurodegenerative Erkrankungen e.V.
EBI	European Bioinformatics Institute
EGA, FEGA, cEGA	European Genome-Phenome Archive, federated EGA, central EGA
EHDS	European Health Data Space
EKUT	Eberhard-Karls-Universität Tübingen
ELSI	ethical, legal, and societal impact
EMBL	Europäisches Laboratorium für Molekularbiologie
EOSC, EOSC4Cancer	European Open Science Cloud, EOSC for cancer
ERDERA	European Rare Diseases Research Alliance
ESHG	European Society of Human Genetics
ETL	Extract-Transform-Load - Standard data loading process
EU	European Union

Abbreviation	Definition
FAIR	Findable, Accessible, Interoperable, and Re-usable
FDG	Research Data Act (Forschungsdatengesetz)
FEGA	Federated European Genome-Phenome Archive
GA4GH	Global Alliance for Genomics and Health
GDC	Genome Data Centre (Genomrechenzentrum) within MV GenomSeq
GBN	German Biobanking Node
GDI	Genomic Data Infrastructure Initiative
GDNG	Health Data Utilisation Act (Gesundheitsdatennutzungsgesetz)
GDPR	General Data Protection Regulation
GFH	Deutsche Gesellschaft für Humangenetik
GHGA	German Human Genome-Phenome Archive
GNC	German National Cohort / NAKO
gnomAD	Genome Aggregation Database
GSC	Genomic Standards Consortium
HPC	High-Performance Computing
HPO	Human Phenotype Ontology
HZI	Helmholtz Centre for Infection Research
IAM4NFDI	Base4NFDI project for Identity & Access Management
ICGC	International Cancer Genome Consortium
IHEC	International Human Epigenome Consortium
IFF	Internal Flex Fund
IIP	Innovation & Implementation Project
ISMS	Information Security Management System
IT	Information Technology
ITCF	Information Technology Core Facility of DKFZ
KGI4NFDI	Base4NFDI project for a Knowledge Graph Infrastructure Service
KI	Künstliche Intelligenz
KiTZ	Hopp Children's Cancer Center Heidelberg
KMS	Key Milestones
KPI	Key Performance Indicator
LDS	Lead Data Steward
LIMS	Laboratory information management system
LoC	Letter of Commitment
LRZ	Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften
MDC	Max Delbrück Center for Molecular Medicine
MDM	Metadata Model
MHH	Medizinische Hochschule Hannover
MII/MI-I	Medical Informatics Initiative
MoU	Memorandum of Understanding

Abbreviation	Definition
MPI	Max Planck Institute
MTB	Molecular Tumor Board
MV GenomSeq	Modellvorhaben Genomsequenzierung, Model Project Genome Sequencing (according to §64e SGB V)
NAKO e.V.	Nationale Kohorte / German National Cohort (GNC)
NCCT	NGS Competence Center Tübingen
NCT	Nationales Centrum für Tumorerkrankungen
NFDI	Nationale Forschungsdateninfrastrukturen / National Research Data Infrastructures
NGS	Next-Generation Sequencing
NGS-CN	NGS Competence Network
OCB	Data Hub Operations Consortium Board
PaGODA Study	“Patients‘ perspectives on Governance of an Omics Database” Study
PI	Principal Investigator
PM	Person Month(s) or Project Management
QBiC	Quantitative Biology Center at the University of Tübingen QC Quality Control
RD	Rare Disease(s)
RDC	Research Data centre
RDM	Research Data Management
RNA	Ribonucleic acid
RRZK	Regionales Rechenzentrum der Universität zu Köln / Regional Computing Center of UzK
rSPE	Restricted Secure Processing Environment
SAB	Scientific Advisory Board
SC	Steering Committee
Seq	Sequencing
SOP	Standard Operating Procedure
SPE	Secure Processing Environment
TA	Task Area
TL / TLC	Team Lead / Team Lead Committee
ToU	GHGA Terms of Use (cf. https://docs.ghga.de/)
TRE	Trusted Research Environment
TS4NFDI	Base4NFDI project for Terminology Services
TUD	Technische Universität Dresden
TUM	Technische Universität München
TUMUH	University Hospital of the Technical University of Munich
UAB	User Advisory Board
UdS	Universität des Saarlandes
UHH	University Hospital Heidelberg
UK	United Kingdom
UKI	Universitätsklinikum Schleswig-Holstein

Abbreviation	Definition
UKT	University Hospital Tübingen
UzK	Universität zu Köln / Cologne University
WES	Whole-Exome Sequencing
WGS	Whole Genome Sequencing
ZDV	Zentrum für Datenverarbeitung / Center for Data Processing
ZIH	Zentrum für Informationsdienste und Hochleistungsrechnen / Center for Information Services and High-Performance Computing

B-1 Proposal Part 1

1 General Information

Name of the consortium in English and German

German Human Genome-Phenome Archive / Deutsches Humangenom-Phänomarchiv

Summary of the proposal in English

GHGA, the German Human Genome-Phenome Archive, is a national infrastructure that facilitates the secure archival, sharing, and processing of access-controlled human omics data. It is connected to national data providers and scientific communities using omics technologies, and it collaborates with European resources and initiatives such as the European Genome-Phenome Archive (EGA), the European Genomic Data Infrastructure (GDI), and the 1+ Million Genomes project (1+MG).

During its initial project phase, GHGA established a multi-disciplinary team, created a legal framework, and set up services for managing omics data. We launched the GHGA Metadata Catalog, Germany's first national omics data repository, recently expanded by the GHGA Archive, a full-fledged archive for access-controlled human omics data. GHGA has established omics metadata standards, unified data workflows, and connected national stakeholders to key international initiatives.

GHGA has developed an integrated legal, ethical, and technical framework to address human omics data sharing, establishing a foundation for future growth. GHGA is positioned as a prominent infrastructure for omics data in Germany, facilitating strategic connections nationally and internationally. GHGA is the German node for the federated European Genome-Phenome Archive (fEGA) and is a member of the national initiatives forum of the Global Alliance for Genomics and Health (GA4GH). These activities will naturally also contribute to upcoming developments such as the European Health Data Space.

The engagement in national strategic initiatives has created key connections and enabled an extended mandate that is fully aligned with, but extends beyond, the scope of the NFDI. In particular, GHGA has been a major driver of strategic national developments in genome medicine, and the consortium has been mandated by national ministries to operate core infrastructure components for the Model Project Genome Sequencing (MV GenomSeq), connecting this major programme to the European Genomic Data Infrastructure Initiative.

In the next funding phase, we will consolidate our activities, and build on the commitments of key data providers to grow the archive. With the core infrastructure operational, we will reinforce our efforts to enable the community to deposit data. A second focus area is the establishment of secure processing environments in order to create new secondary use

opportunities aligned with national needs. We have initiated driver projects with our core communities to pilot new services and showcase the added value of GHGA. Finally, building a sustainable long-term business model based on income from GHGA-associated research and infrastructure projects will ensure the future of GHGA as a national research data infrastructure.

Summary of the proposal in German

GHGA, das Deutsche Humangenom-Phänomarchiv, ist eine nationale Forschungsdateninfrastruktur, die das sichere Archivieren, Teilen und Verarbeiten von zugangsbeschränkten menschlichen Omics-Daten ermöglicht. Es ist mit nationalen Datenanbietern und der wissenschaftlichen Gemeinschaft verbunden und arbeitet eng mit europäischen Ressourcen und Initiativen wie dem European Genome-Phenome Archive (EGA), der European Genomic Data Infrastructure (GDI) und dem 1+ Million Genomes (1+MG) Projekt zusammen.

In der ersten Förderphase haben wir einen rechtlichen Rahmen geschaffen und die technisch-organisatorische Plattform für das Management von Omics-Daten etabliert. Mit dem GHGA Metadatenkatalog und dem GHGA Archiv wurde eine umfangreiche FAIRe Dateninfrastruktur in Betrieb gebracht. Des Weiteren hat GHGA an der Etablierung von Metadatenstandards, der Vereinheitlichung von Omics-Daten-Workflows und der (inter-)nationalen Vernetzung nationaler Akteure maßgeblich mitgewirkt. GHGA wurde zum deutschen Knoten für das föderierte European Genome-Phenome Archive ernannt und entwickelt sich zu einem nationalen Leitprojekt innerhalb der Global Alliance for Genomic and Health. GHGA ist eng in strategische nationale Entwicklungen eingebunden und wurde von den Bundesministerien für Bildung und Forschung sowie Gesundheit beauftragt, zentrale Infrastrukturkomponenten für das nationale Modellvorhaben Genomsequenzierung zu betreiben und Deutschland mit GDI zu verbinden. In diesem Kontext haben die GHGA-Datenknoten auch die Zulassung als Genomrechenzentren durch das BfArM erhalten.

In der nächsten Finanzierungsphase werden wir unsere Aktivitäten weiter konsolidieren und das Archiv, basierend auf den Beiträgen wichtiger Datenanbieter, weiter ausbauen. Aufbauend auf dem verlässlichen Betrieb unserer Kerninfrastruktur werden wir unsere Bemühungen verstärken, die Community bei der Hinterlegung ihrer Daten zu unterstützen, was zu erheblichen genomischen Datenressourcen führen wird. Ein zweiter zentraler Bereich der zukünftigen Entwicklung besteht darin, neue Werkzeuge zu liefern, um die Daten für die Sekundärnutzung zugänglicher zu machen. Technisch wird dies durch den Aufbau einer sicheren Verarbeitungsumgebung auf Basis von Cloud-Technologien erreicht, die die Rahmenbedingungen in Deutschland berücksichtigt. Wissenschaftlich haben wir

mehrere Leitprojekte in den Bereichen Onkologie, seltene Krankheiten und allgemeine Krankheiten/Prävention definiert, die unsere Lösungen testen und gleichzeitig den Mehrwert des Datenaustauschs demonstrieren werden. Dies wird zu einem dringend benötigten Kulturwandel beitragen. GHGA-assoziierte Schwesterprojekte und insbesondere das nationale Modellvorhaben Genomsequenzierung werden es uns ermöglichen, ein nachhaltiges Finanzierungs- und Geschäftsmodell zu entwickeln, um das Konsortium auf eine zukünftige Institutionalisierung vorzubereiten.

Applicant institution

Applicant institution	Location
German Cancer Research Center (DKFZ)	Heidelberg

Spokesperson

Spokesperson	Institution, location
Oliver Stegle	DKFZ and EMBL, Heidelberg

Co-applicant institutions

Co-applicant institutions	Location
Berlin Institute of Health @Charité (BIH)	Berlin
Eberhard-Karls-Universität Tübingen (EKUT)	Tübingen
European Molecular Biology Laboratory (EMBL)	Heidelberg
Helmholtz Munich (HMGU)	Munich
University Hospital of the Technical University of Munich (TUMUH)	Munich
Max Delbrück Center for Molecular Medicine (MDC)	Berlin
Technical University of Munich (TUM)	Munich
Technische Universität Dresden (TUD)	Dresden
Universität zu Köln (UzK)	Cologne
University Hospital Heidelberg (UHH)	Heidelberg
University Hospital Tübingen (UKT)	Tübingen
University of Heidelberg (UHD)	Heidelberg

Co-spokespersons

Co-spokespersons	Institution, location	Task area(s)
Dieter Beule	MDC and BIH, Berlin	A2, A4, B2
Ivo Buchhalter	DKFZ, Heidelberg	A1, A2 A4
Andreas Dahl	TUD, Dresden	A4
Julien Gagneur	TUM, Munich	A4, B1, B2
Holm Graessner	UKT, Tübingen	B1, B2, B3
Daniel Hübschmann	DKFZ, Heidelberg	B1, B2
Oliver Kohlbacher	EKUT, Tübingen	A2, A4, B3, C1, C2
Jan Korbel	EMBL, Heidelberg	B4, C2

Co-spokespersons	Institution, location	Task area(s)
Fruzsina Molnár-Gábor	UHD, Heidelberg	B5
Susanne Motameny	UzK, Cologne	A4
Sven Nahnsen	EKUT, Tübingen	B2, B4
Stefan Wesner	UzK, Cologne	A2
Juliane Winkelmann	TUMUH, TUM, & HMGU, Munich	B3
Eva Winkler	UHH, Heidelberg	B5, C2

Participants

Participating Institutions

Participating institutions	Location
Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM)	Bonn
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)	Brunswick
Helmholtz-Zentrum für Informationssicherheit (CISPA)	Saarbrücken
Medizinische Hochschule Hannover (MHH)	Hanover
National Center for Tumor Diseases (NCT DD) Dresden	Dresden
National Center for Tumor Diseases (NCT HD) Heidelberg	Heidelberg
Universität des Saarlandes (UdS)	Saarbrücken
Universität Freiburg (UFR)	Freiburg
Universitätsklinikum Schleswig-Holstein, Kiel (UKI)	Kiel
ZB MED – Informationszentrum Lebenswissenschaften (ZBMED)	Cologne

Participating individuals

Participating individuals	Institution, location
Viktor Achter	UzK, Cologne
Peer Bork	EMBL, Heidelberg
Benedikt Brors	DKFZ, Heidelberg
Nataliya Di Donato	MHH, Hanover
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	UFR, Freiburg
Karsten Häcker	MDC, Berlin
Britta Hänisch	BfArM, Bonn
Wolfgang Huber	EMBL, Heidelberg
Dirk Jäger	NCT Heidelberg/UHH/DKFZ, Heidelberg

Participating individuals	Institution, location
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, Brunswick
Christian Mertes	TUMUH, Munich
Ralph Müller-Pfefferkorn	TUD, Dresden
Wolfgang E. Nagel	TUD, Dresden
Uwe Ohler	MDC, Berlin
Stephan Ossowski	UKT, Tübingen
Leo Panreck	NAKO e.V., Heidelberg
Annette Peters	HMGU & NAKO, Munich
Tobias Pischon	MDC, Berlin
Stefan Pfister	DKFZ and KiTZ, Heidelberg
Peter Robinson	BIH, Berlin
Philip Rosenstiel	UKI, Kiel
Christoph Schickhardt	NCT-HD / DKFZ Heidelberg
Thorsten Schlomm	Charité, Berlin
Joachim Schultze	DZNE, Bonn
Julia Schulze-Hentrich	UdS, Saarbrücken
Cornelia Specht	GBN and Charité, Berlin
Thomas Ulas	DZNE, Bonn
Thomas Walter	EKUT, Tübingen
Jörn Walter	UdS, Saarbrücken

Contributions of the participants

de.NBI e.V., represented by Oliver Kohlbacher, will work together on training aspects (TA A4) and enable GHGA access to the biomatics infrastructure, services, and other resources provided via de.NBI/ELXIR-DE. • **Viktor Achter (UzK)** is heading the HPC centre at UzK and will be responsible for the operations of the data hub in Cologne (TA A2). • **Peer Bork (EMBL)** will support the efforts for interoperability of the GHGA platform, especially with respect to alignment with NFDI4Microbiota (TA B4). • **Benedikt Brors (DKFZ)** will support GHGA use cases in cancer bioinformatics (TA B1 and B2) and will connect GHGA to the PM⁴Onco project. • **Nataliya Di Donato (MHH)** will lead the prospective data hub at the MHH (TA A2 and A4). • **Juliane Fluck (ZBMED)** will further support the interaction with

NFDI4Health along the joint use cases (TA B4) and will foster NFDI interactions in the biomedical field. • **Mario Fritz (CISPA)** will provide expertise on privacy-preserving analysis and data processing (TA A3). • **Stefan Fröhling (NCT HD)** will support the operation of GHGA by providing clinically annotated genome, transcriptome, and methylome data from MASTER and additional clinical networks (TA B1&B2) as well as long-standing expertise in FAIR (Findable, Accessible, Interoperable, Reusable) data sharing. • **Hanno Glimm (NCT DD)** will support the operation of GHGA by providing clinically annotated genome, transcriptome, and methylome data from MASTER and additional clinical networks (TA B1&B2). • **Björn Grüning (UFR)** will contribute his experience with various GA4GH standards by engaging with the Galaxy Public Health community (TA B2). • **Karsten Häcker (MDC)** will, together with Dieter Beule, operate the GHGA data hub in Berlin at MDC (TA A2). • **Britta Hänisch (BfArM)** will work together with GHGA and the GHGA data hubs (TA A2) to enable the secondary use of the research data generated within the MV GenomSeq. • **Wolfgang Huber (EMBL)** will continue to support GHGA by engaging in training activities (TA B3) and to support GHGA by connecting to long-standing community efforts such as Bioconductor. • **Dirk Jäger (NCT HD, UHH, DKFZ Heidelberg)** is supporting GHGA by connecting GHGA to oncological communities active within the NCT Heidelberg (TA B1). • **Thomas Keane (EMBL-EBI)** is leading the EGA and is working together with GHGA within the FEGA ensuring international connectivity of GHGA (TA B4). • **Jens Krüger (EKUT)** will provide expertise on Trusted Research Environments (TREs) and related technical issues and legal implications (TA A3) and will support the operation of the Tübingen data hub (TA A2 and A4). • **Martin Lablans (DKFZ)** will contribute expertise on record linkage and connect GHGA to efforts in the medical informatics initiative (TA B4). • **Peter Lichter (DKFZ and NCT HD)** will work with GHGA to connect to communities around cancer genomics (TA B1) and will use GHGA as a data platform for clinical programmes at the NCT HD (CATCH and COGNITION). • **Nisar Malek (UKT)** will contribute his experience in personalised medicine and as speaker of the German Network for Personalized Medicine (DNPM) support the mobilisation of omics data from this network (TA B1). • **Ninja Marnau (CISPA)** will support GHGA with her expertise on IT security and privacy law (TA C2). • **Alice McHardy (HZI)** will support the implementation of strategies for handling sensitive data such as biomedical data from patient cohorts, as well as for data tied to both the GHGA and the NFDI4Microbiota consortium, such as human microbiome data sets (TA B1). • **Christian Mertes (TUMUH and TUM)** is leading the diagnostic platform at the Institute of Human Genetics at the TUMUH and contributes with his expertise in processing large-scale NGS datasets (TA B2) as well as in running secure IT Infrastructure at scale (TA A2). • **Wolfgang E. Nagel (TUD)** will support the operation of the Dresden data hub (TA A2 and A4). • **Ralph Müller-Pfefferkorn (TUD)** will provide expertise on data analytics platforms (TA A3) and will

support the operation of the Dresden data hub (TA A2 and A4). • **Uwe Ohler (MDC)** will facilitate the operational embedding of GHGA at the Berlin data hub (TA A2) and contribute expertise in omics workflows (TA B2). • **Stephan Ossowski** will support the community engagement efforts and provide expertise on scalable methodologies and infrastructure for research on rare diseases as well as bioinformatic development of methods for personalised medicine and NGS-based diagnostics (TA B1 and B2). • **Leo Panreck** will bridge efforts in GHGA with data management efforts in the German National Cohort (NAKO e.V.) to make NAKO Omics data available via the GHGA platform and to enable interoperability of GHGA with further data sources around cohort studies (TA B1). • **Annette Peters (HMGU)** will support GHGA by deepening the connection to NAKO's efforts around genotyping and omics data and help connecting to the epidemiological research communities (TA B1). • **Tobias Pischon** will enable close interaction with NFDI4Health along the joint use cases (TA B1) and will support NFDI interactions in the biomedical and epidemiological communities (TA B2). • **Stefan Pfister** will provide access to datasets of the ITCC-P4 platform, the world's largest repertoire of patient-derived xenograft (PDX) models of paediatric tumours, and use GHGA as the primary archive for archive for the ITCC PedCanPortal developing a global platform for the integration and sharing of data from paediatric precision oncology studies (TA B1 and B2). • **Peter Robinson (BIH)** will contribute his experience in structured clinical data, in particular the phenopackets standard, to standardise phenotypic data deposition. Peter has led the Human Phenotype Ontology (HPO) project since its inception in 2008 and will extend HPO content as needed to represent GHGA clinical data (B4). • **Philip Rosenstiel (UKI)** will contribute his experience in genomic medicine in general and genomics of chronic inflammatory disease to our community outreach efforts (TA B1&B4). • **Christoph Schickhardt (NCT-HD, DKFZ)** will contribute his expertise in the fields of ethics of genomics and ethical governance of genetic & clinical data sharing, and connect the GHGA consortium to the Working Group Consent of the German Medical Informatics Initiative (TA B5). • **Thorsten Schlomm** will enable the connection of DNA-Med Connect, a clinical data hub within the MV GenomSeq, to the GHGA Data Infrastructure, ensuring close alignment on metadata schemata and on patient outreach measures. (TA B1, B2). • **Joachim Schultze** will work engage in TA B1, B2 and B4 to support the metadata framework development and to develop FAIR scoring metrics. • **Julia Schulze-Hentrich** will engage in public outreach measures to promote the GHGA brand and core functionalities and to widen the spectrum of GHGA data usage (e.g., within the functional (epi)genomics community (TA B3)). • **Cornelia Specht (GBN/Charité)** will support the interoperability of GHGA with the German Biobanking Node (GBN) and BBMRI-ERIC (TA B1). • **Thomas Ulas** will engage in TA B1, B2 and B4 to support the metadata framework development and to develop FAIR scoring metrics. • **Thomas Walter** will provide expertise on TREs and related

technical issues and legal implications (TA A3) and will support the operation of the Tübingen data hub (TA A2 and A4). • **Jörn Walter (UdS)** will engage in public outreach measures to promote the GHGA brand and core functionalities and to widen the spectrum of GHGA data usage (e.g., within the functional (epi)genomics community (TA B3)).

Names and numbers of the DFG review boards (*DFG-Fachkollegien*) that reflect the subject orientation of the proposed consortium

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

2 Scope and Objectives

2.1 Research domains or research methods addressed by the consortium

GHGA addresses the secure and safe data storage, management, sharing and analysis of the full spectrum of human omics data, following the well-established controlled-access principles [1,2].

2.1.1. Research domains

GHGA serves diverse scientific communities and stakeholders sharing an interest in human omics data. The largest community for human omics is biomedical researchers interested in molecular aetiology and potential therapies for various diseases. On the one hand, these users require a secure, trustworthy, and convenient repository for their omics data, which GHGA has established and launched in the first project period. On the other hand, the same researchers also require access to large community reference datasets generated in other labs and/or in the context of international consortia, and will require utilising external datasets to enable, replicate, and verify discoveries. Whilst the data resources that are suitable for deposition in GHGA are still building up, GHGA will increasingly become an important resource to epidemiological communities (i.e., when linked to external clinical data, for example, via record linkage to the Medical Informatics Initiative (MII), the German Centers for Health Research (DZG), the Model Project Genome Sequencing (MV GenomSeq (via the national trust centre at the Robert Koch Institute)), biobanks (via German Biobank Node) and NAKO). Patients, patient networks, but also citizens and citizen scientists are additional key stakeholders of GHGA.

Entirely different approaches are required to communicate to these distinct communities effectively, on how their valuable data will be used and how GHGA complies with changing ethical and legal requirements for access to their data over time. GHGA has conducted significant efforts to understand the needs and expectations of patients [3], including patient outreach campaigns, information sessions, and research in this field. While initially, the majority of the human data generated will represent genomic data [4], the integration of

additional omics types such as epigenomes, transcriptomics, or proteomes (which are increasingly collected from the same sample donors) will turn GHGA into a unique resource where such multi-omics data will be accessible in a transparent and reproducible manner, fostering research in systems medicine. In the first project period, the ingest of such data modalities has been piloted with oncology data, of which we host transcriptome and epigenome data, and we will further provide access to upcoming data types such as spatial and single-cell transcriptomics.

GHGA is also foreseen to be a key resource and major driver for the development of new computational methods. In addition to classical bioinformatics, we observe a rise of artificial intelligence (AI) in omics-related applications. This is motivated by growing data volumes on the one hand, and by the specific structure of genome sequences on the other hand, rendering AI-based approaches particularly applicable. The AI community is increasingly becoming interested in GHGA as a resource and we have taken the first steps to make tools accessible on our infrastructure. As such, we have developed workflows and analysis methods in order to equip the community with the tools needed to exploit GHGA as a resource.

In order to serve our communities as efficiently as possible, we have employed a step-by-step approach to grow our user community across all research domains that generate or use human omics data and derived data types. Initially, our community engagement measures were focused on rare disease (RD) and cancer. We have since made significant progress in incorporating common disease genetics, and single-cell biology. In the second funding period, we will continue this trajectory. In addition to consolidating our current target communities, we will also expand into new communities, including prevention.

2.1.2 Impact

In its first funding period, GHGA has already had a major impact on its communities and stakeholders. We would like to highlight four core areas that affected all our communities visibly:

Establishing a national genome initiative bridging genome science and translation in patient care: A large structural achievement has been GHGA's support to help realise a national genome initiative and making it visible to decision-makers. GHGA has been one of the major drivers of the MV GenomSeq as several GHGA (co-)spokespersons and participants have been strategically involved in the conception of this project (cf. 4.1.3). Our role has also been to make sure everything is designed to foster secondary use of data and ensure a viable and scalable data infrastructure for the project. The launch of this project in July 2024 is also a major milestone to integrate genome research and clinical care.

Creating a legally sound national archive for genome data: GHGA has overcome two-decades-old issues related to legal uncertainties on the deposition of human genome

data in Germany. We now provide *the* national genome data service. We have been designated as national representatives in European key initiatives, including fEGA and GDI. Nationally, the GHGA data hubs are the only infrastructures approved as genome data centres in the MV GenomSeq, and contractual connection to GHGA is one of the criteria required to operate a genome data centre [5]. The ELSI task areas in GHGA have established general-purpose templates for patient consent forms, ensuring the secondary use of the data for research [6].

Training & enabling biomedical scientists to use genomic data for research. We have run a multitude of courses, conferences, webinars, online lectures, and a podcast series to reach out to various stakeholders and the general public about genomic medicine (cf. nfdi-ID 2001-2003 in Appendix 5). The training components also include practical guidelines on how to handle sensitive data and the role of data protection, increasing the awareness of data protection in the community - a topic essential to building and maintaining trust in our infrastructure. We also provide workflows and best practices for analysing omics data and thus help train the next generation of scientists working at the intersection of research and genomic healthcare.

Fostering culture change: GHGA is the designated data-sharing platform for omics data across all major initiatives in Germany (MV GenomSeq, NCT, NAKO, DZHK, ERDERA, Solve RD etc.). By convincing these initiatives to share their data through GHGA, we have made a major step forward in avoiding the siloisation of omics data and towards a culture of data sharing, which (traditionally) has been limited in scope and has been delayed due to publication embargoes. This change is reflected in both the willingness to share data, the implementation of firm data-sharing commitments, and joint efforts to build infrastructures as evident from external funding going into the consortium. Multiple of these data controllers are in the process of establishing internal policies that will permit data sharing via GHGA also prior to publication. Much of that cultural change has been driven by our interaction with patient representatives, who unanimously reaffirm the overwhelming desire to utilise the data for research.

2.1.3 Aims for the next funding period

Building on these past achievements, we have defined three overarching aims for the second funding period:

Consolidation and sustainability: A major aim for the upcoming funding period is to consolidate our service portfolio, and to ensure the long-term viability of GHGA as the national archive for genome data. Building on our national mandate, we will consolidate metadata standards, consolidate our technical infrastructure, reshape best practices for data sharing, and further support the FAIR-ification of human omics data. Making GHGA sustainable entails a streamlined governance structure, the roll-out of a solid business

model, acquisition of external funding, and mobilising financial commitments from major stakeholders. We will also identify services that require long-term support and develop outsourcing models where appropriate, for example using commercial cloud services. To achieve this, we will work closely also with the NFDI e.V. and other NFDI consortia.

Enhancing community value: We strive to make the archive as useful as possible by populating it with primary data, as well as with data derived from that. Our partnerships with NAKO, MV GenomSeq, and other data controllers ensure the steady growth in data (10+k genomes per year; c.f. Table 4). Automated data analysis workflows that GHGA will execute on behalf of (and driven by feedback from) our communities will make derived data and results directly accessible and comparable (e.g., beacon services, variant databases, annotation services). This will help to grow our user community and provide significant added value, extending the established data archival and sharing functionality of GHGA. Our strategy for including new communities focuses on technological innovation (e.g., multi-omics and single-cell technologies), as well as expanding the range of indications beyond cancer and RD (e.g., common disease, healthy controls). The new community driver projects outlined in the work programme (B1) are a key mechanism to achieve this.

Expanding our national and international embedding: With GHGA firmly established as the German national human genome archive, we will strengthen our embedding within the NFDI and other national stakeholders. GHGA is established as the national node for genomics across a growing portfolio of projects, including federated EGA, the European Genomic Data Infrastructure (GDI), ELIXIR, and will eventually be part of the European Health Data Space (EHDS). By consolidating these roles, and leveraging the synergies between them, GHGA will act as a one-stop-shop for genomic data for Germany. Beyond our communities in genomics, the consortium is well-positioned to act as a driver project in EOSC, being an early mover in the transition from classical HPC solutions to cloud technologies. GHGA will bring the European Health Data Space to the NFDI, the first of multiple European data spaces to come, providing opportunities for synergies and joint national positions. Finally, we will deepen our connections with individual NFDI consortia, most notably NFDI4Health and help develop the services of Base4NFDI and implement them within GHGA.

2.2 Objectives and measuring success

As a national research data infrastructure, GHGA has been designed with a series of core objectives in mind. Primarily, these objectives are based on the requirements of our target communities, data controllers, and associated networks and communities.

2.2.1 Core Objectives

The following **core objectives** (CO-1 through CO-10) have been identified:

- **CO-1 - National Archive:** Operate a national secure and trustworthy long-term archive of human omics data federated within Europe
- **CO-2 - Business Model:** Consolidate GHGA services and develop a business model for data archival
- **CO-3 - MV GenomSeq:** Coordinate and operate the genome data infrastructure backbone for the national genome initiative Model Project Genome Sequencing
- **CO-4 - SPE:** Develop concepts, develop and operate a secure processing environment for GHGA users, thereby democratising access to data
- **CO-5 - Enable communities:** Facilitate responsible data sharing and provide best practices for data sharing via community driver projects
- **CO-6 - National gateway:** Act as a national gateway for relevant European initiatives and data spaces and research data infrastructures
- **CO-7 - ELSI Framework:** Adapt to changes of the (inter)national legal and ethical framework
- **CO-8 - Training:** Train the next generation of scientists on the efficient and responsible use of omics data in research
- **CO-9 - Increase value of data:** Increase value of the research data by integrating multiple omics modalities and connecting omics data to phenotype data
- **CO-10 - Increase FAIRness:** Increase FAIRness of omics data by establishing record linkage, metadata references, and linkages with other infrastructures

These individual objectives collectively support the major aims for the next funding period. The aim to further **enhance the community value** of GHGA will be underpinned by the engagement with the MV GenomSeq, novel technical solutions, the role as international gateway, targeted community measures, training and the deposition of multi-omics datasets (CO 3-6, 8 & 9). As above, the **expansion of the national and international embedding** of GHGA is supported via our MV GenomSeq and international engagement but will be further strengthened via measures to address upcoming legal challenges, the focus on multi-model data and measures to further develop metadata linkages (CO 3,6,7 & 9-10). Finally, the transition to a **sustainable infrastructure** will be enabled by the objectives to deliver a long-term archive, develop a viable business model for the infrastructure and use the role in MV GenomSeq to deliver services of national relevance (CO 1-3).

Activities related to infrastructure services (TA A1-A4), as well as project management, legal and sustainability (TA C1-C2) are long-term activities of GHGA, thus **requiring sustainable funding**.

2.2.2 Measuring Success: KPIs and Key Milestones

The fulfilment of the core objectives will be assessed either by achieving some of the **key milestones (KMSs)** from the work programme as qualitative indicators of implementation progress. Additional quantitative **key performance indicators (KPIs)** are defined to

continuously track progress in key areas, for which we specify target numbers for the end of the upcoming funding period**.** The following KPIs and KMS will measure the success of our core objectives:

- **CO-1: National Archive:** KPI 1: number of samples with deposited omics data (target: 100,000); KPI 2: data requests received (target: 200); KPI 3: number of samples from which data has been staged (target: 5,000)
- **CO-2: Business Model:** KPI 4: amount of external funding received that supports GHGA (target: 10 M€); Milestone C2.M4.T1 - White paper on business model published
- **CO-3 - MV GenomSeq:** Milestone C2.M2.T3: Legal framework negotiated and executed; Milestone A4.M3.T1: First data set deposited
- **CO-4 - SPE:** Milestone A3.M2.T1: Published white paper on GHGA cSPE; Milestones A4.M3.T4 First access granted for used on cSPE
- **CO-5: Enable communities:** KPI 5: formalised interactions (e.g., MoU, joint grant) with target communities (target: 5); KPI 6: organised 10+ community events; Milestone B4.M1.T2: Data from GHGA findable in NFDI4Health Portal
- **CO-6 - National Gateway:** Milestone B4.M2.T5 Metadata exchange with GDI established; Milestone A4.M1.T2: First data access request received via EGA fulfilled
- **CO-7 - ELSI Framework:** Milestone B5.M1.T2 Concept for integration with EHDS; Milestone B5.M2.T1 Compliance measure developed for align with emerging national legislation
- **CO-8: Training:** KPI 7: number of participants in GHGA training events (target: 2,500); KPI 8: number of GHGA-organised training events (target: 30)
- **CO-9 - Increase Value:** Milestone B4.M3.T3: First multi-omics metadata model operational; Milestone A4.M3.T5: First multi-omics data set deposited in GHGA
- **CO-10: Increase FAIRness:** KPI 9: number of infrastructures GHGA exchanges (meta)data with (target: 5)

By achieving these KMSs we will be able to prove that the corresponding core objectives have been achieved. The selected KPIs will help us track progress for those core objectives that are easier to quantify. Both will be reported regularly.

3 Consortium

Table 1: Members participating in other NFDI consortia.

Name	Also participating in consortium/consortia
Bork, Peer	NFDI4Microbiota
Fluck, Juliane	NFDI4Health
Grüning, Björn	DataPlant, NFDI4Bioimage
Krüger, Jens	DataPlant
McHardy, Alice	NFDI4Microbiota

Name	Also participating in consortium/consortia
Müller-Pfefferkorn, Ralph	NFDI4Earth, Text+, FAIRmat, NFDI4xCS, NFDI4DS, NFDI4Chem
Nagel, Wolfgang	NFDI4Earth, Text+, FAIRmat, NFDI4xCS, NFDI4DS, NFDI4Chem
Nahnsen, Sven	DataPlant, NFDI4Immuno, NFDI4Bioimage
Peters, Annette	NFDI4Health
Pischon, Tobias	NFDI4Health
Stegle, Oliver	NFDI4Health
Walter, Thomas	DataPlant
Wesner, Stefan	FAIRmat

3.1 Composition of the consortium and its embedding in the community of interest

3.1.1 Structure of the consortium and changes in its composition

Consortium Composition

The consortium for the German Human-Genome-Phenome Archive (GHGA) has been carefully assembled to align with the project's objectives, ensuring comprehensive coverage of expertise in research data management. The (co-)applicant institutions and participants were selected to include:

- **Clinicians:** These members maintain close ties with the communities GHGA serves, ensuring the relevance and applicability of data resources and services.
- **Ethical, Legal, and Social Implications (ELSI) Experts:** These specialists provide a robust ethical and legal foundation for data access and processing, addressing crucial aspects of data governance, consent, and compliance.
- **Omics Centres:** These centres are responsible for generating the majority of academic omics data in Germany, contributing critical data resources to GHGA.
- **Bioinformaticians:** Experienced in large-scale omics projects, these professionals bring expertise in data analysis, integration, and interpretation.
- **High-performance Computing Centres:** These centres offer the necessary infrastructure to operate a scalable and robust data management system, supporting the computational needs of GHGA.

Changes to the Consortium

The composition of the consortium has evolved in response to changing priorities, funding, and capacity, ensuring that GHGA is well-placed to continue meeting its goals. Key changes include the evolution of data hubs and membership changes detailed in the following.

Evolution of Data Hubs

While the number of data hubs has been perceived as a problem as the federated structure of GHGA adds to its complexity, we have been encouraged at a political level that a number between five and ten data hubs would be desirable to appropriately represent the federated structure of Germany. This is also aligned with the assessment of the Federal Office for Data

Protection and Informational Freedom that infrastructures such as national genomics data should not be stored in a single centralised location [7]. Besides these external drivers, the required capacity of 50 PB for the projected data volume by 2030 also requires us to rely on multiple partners to provide these resources as part of their own contributions, as the NFDI funding scheme currently does not fund the creation of physical data infrastructures or pay for comparable services. We thus strive to work with a small number of potent data hubs. Since the original proposal, this setup has been evolving as follows:

Munich: The TUMUH has taken over the operations of the Munich data hub from TUM. **Kiel:** Previously a data hub, will now no longer be operating in this capacity, reflecting a shift in focus and resources. **Berlin:** Has been onboarded as a prospective data hub during the first funding period and has now been established as a core data hub. This addition enhances GHGA's capacity to handle large volumes of omics data and supports a more distributed and resilient infrastructure. **Hannover** (MHH, Hannover Medical School): Identified as a prospective data hub, MHH's inclusion aims to expand GHGA's data management capabilities, particularly in regions critical for German academic and clinical research. All core data hubs have committed to providing a harmonised service level to the consortium via contractual arrangements according to our tailor-made legal concept [8].

New and leaving members and impact of changes

The careful reduction in the number of co-spokespersons (15, down from 27) allows for a new consolidated and streamlined governance now that the key data archiving infrastructure is operational. Individuals who are no longer co-spokespersons will continue to contribute their expertise as participants. For the GHGA members retiring in the previous or upcoming funding period, corresponding successors have been identified and included (Stefan Wesner for Ulrich Lang, Julia Schulze-Henrich for Jörn Walter, Holm Graessner for Olaf Rieß). Julio Saez-Rodriguez has been appointed Head of Research at EMBL-EBI and is thus no longer available for the consortium, but will continue to support GHGA through his new role at EMBL-EBI. Stefan Hachinger left the consortium as a participant as LRZ now provides its services as a computing centre through TUMUH. Michael Hummel retired as director of GBN/GBA and was succeeded by Cornelia Specht. As indicated in our LoI [9], Dieter Beule (BIH) and Christian Mertes (TUMUH) have already been added as Co-Spokesperson and Participant, respectively, in the current funding period.

The inclusion of new members improves GHGA's connections within NFDI and to our user communities:

- **Nataliya Di Donato** (MHH) joins as a participant for considering options on the integration of MHH - a major national centre for diagnostic sequencing - as a potential GHGA data hub.
- **Juliane Fluck** (ZBMED) joins as a participant to coordinate the strategic and scientific

alignment with NFDI4Health on metadata standards, record linkage, and the joint activities of health infrastructures in the NFDI [10].

- **Björn Grüning** (University of Freiburg) joins as a participant: An expert in Galaxy, which is integral to the European Genomic Data Infrastructure (GDI) and the EOSC4Cancer project, will contribute to the creation of execution environments and leverage synergies to position Germany strategically within the European context.
- **Karsten Häcker** (MDC) joins as a participant to provide IT support for the operation of the Berlin data hub.
- **Britta Hänisch** (BfArM): joins as participant to create a bridge between GHGA and the BfArM as legally responsible operator of the data infrastructure for the Model Project Genome Sequencing (MV GenomSeq).
- **Leo Panreck:** (NAKO): joins as participant to setup joint use cases and interactions with the NAKO.
- **Tobias Pischon** (MDC Berlin): joins as participant to jointly develop use cases and federated analysis strategies in the context of NAKO and other cohorts with data in both NFDI4Health and GHGA.
- **Stefan Pfister** (DKFZ) joins as a participant to support data mobilisation from within the DKFZ.
- **Peter Robinson** (Charité) joins as a participant: Recently relocated to Germany, Robinson brings expertise in Phenopackets, a crucial service for representing phenotypes in the rare disease community [11]. His involvement will help standardise and evolve data and metadata representation, enhancing data utility and interoperability.
- The inclusion of **de.NBI e.V.** as a participating institution (also represented by co-spokesperson Kohlbacher), will allow closer technical alignment on cloud and computing standards between GHGA and the de.NBI/ELIXIR-DE Cloud.

3.1.2 Co-applicant institutions & participants

Co-Applicant Institutions

The **German Cancer Research Center (DKFZ)** operates one of the largest sequencing facilities in continental Europe and has been at the forefront of translating omics technologies to the clinics. Through the National Center of Tumor Diseases (NCT), the Hopp Children's Cancer Center Heidelberg (KiTZ), and the German Cancer Consortium (DKTK), genomes from more than 1,000 cancer patients per year are sequenced with the results routinely feeding into molecular tumour boards. DKFZ will act as GHGA Central, while at the same time bring institutional expertise and scientific networks to establish the federated GHGA Data Infrastructure. **Stegle (Spokesperson)** heads the division for Computational Genomics and Systems Genetics and is also a group leader at EMBL Heidelberg. He is a member of the special group of 1+Million Genomes (1+MG) for Germany, pillar-co-lead of

the Genomic Data Infrastructure Initiative (GDI), and steering committee member of MV GenomSeq. His research is focused on statistical and computational methods for tying together large population-variation data resources. He has led, and contributed to, major international studies, including in cancer (PCAWG), population genetics and single-cell genomics (Human Cell Atlas). **Buchhalter (Co-Spokesperson)** heads the Omics IT and Data Management core facility, which performs data management, curation, and processes thousands of omics data sets per year, from both research and clinical projects. **Hübschmann (Co-Spokesperson)** heads the Unit for Bioinformatics and Precision Medicine and is the deputy head of the Molecular Precision Oncology Program at the NCT HD. He is responsible for translational analysis of omics data for molecular tumour boards, and will be responsible for developing infrastructure of a genomic newborn screening project in Heidelberg. **Brors (Participant)** heads the division of Applied Bioinformatics and has been leading bioinformatics analysis in three German networks within the International Cancer Genome Consortium (ICGC) as well as leading the data coordination centre of the German contribution to the International Human Epigenome Consortium (IHEC). **Lablans (Participant)** is head of the division of Federated Information Systems, and develops infrastructure solutions for data protection-compliant federated data management, including the *Mainzelliste* record linkage solution and networks such as the DKTK and the German Biobank Alliance, where his bridgehead software connects 21 national partners. **Lichter (Participant)** is head of the division of Molecular Genetics as well as co-director of NCT HD. His research interests are to decipher the genetic basis of human cancers. He has led numerous national and international consortia, including the German contribution to ICGC. **Pfister** (Participant) is head of the Department of Pediatric Neurooncology (DKFZ), PI of the international paediatric precision oncology programme INFORM, and the director of KiTZ. He will provide expertise and access to datasets collected in paediatric cancer also as part of the involvement of GHGA In the ITCC PedCanPortal project. **Schickhardt (Participant)** is Senior Scientist in the Section of Translational Medical Ethics and a member of Data Use and Access Boards within the MII. In the area of bioethics, his research focuses on ethical, social, and data protection aspects of clinical and omics data sharing.

University Hospital Heidelberg (UHH) is among the largest and most prestigious medical centres in Europe where in more than 40 clinics 59,700 inpatients are seen every year, 40% of these contacts are cancer-related. The **NCT Heidelberg (NCT HD)** was founded as an exceptional alliance between DKFZ and UHH together with the Heidelberg Medical Faculty and German Cancer Aid. **Winkler (Co-Spokesperson)** is heading the Section of Translational Medical Ethics and was awarded a Heisenberg professorship for her research. She is a member of the Data Sharing Workgroup of the MII as well as of the Regulatory and Ethics Working Group of the Global Alliance for Genomic and Health. **Jäger (Participant)** is

Managing Director of NCT HD, Director of the Medical Oncology Department at UHH and Head of the DKFZ Clinical Cooperation Unit Applied Tumour Immunity.

The **University of Heidelberg (UHD**) is the oldest university in Germany and possesses a unique research portfolio. **Molnár-Gábor (Co-Spokesperson)** has extensive experience in advising international and national consortia and organisations on legal solutions for data protection and sharing. Among others, she is a member of the ELSI working group in the 1+MG, forum editor of the GA4GH Data Protection and International Health Data Sharing Forum and member of the European Group on Ethics in Science and New Technologies.

The **European Molecular Biology Laboratory** (EMBL) is Europe's flagship institution for molecular biology and biodata. **Huber, Korbel**, and **Stegle** have joint appointments with the EMBL-European Bioinformatics Institute (EMBL-EBI), a subsidiary of EMBL, which hosts and operates the EGA and where it was launched in 2008. These direct links, together with the agreement of EMBL-EBI to provide code and support to GHGA as an additional participant (cf. LoC from **Thomas Keane (Participant)**, **EMBL-EBI Hinxton**, UK) will secure the necessary expertise for transferring technology from the EGA to our consortium. **Korbel (Co-Spokesperson)** is EMBL senior scientist and EMBL's Head of Data Science. He is a member of the EOSC-A Health Data Task Force, previously led the EOSCpilot Demonstrator Project 'Pan-Cancer', and is a WP co-lead of the EOSC4Cancer project. He is also acting as a mirror group member in Germany for 1+MG. The research focus of his group is to unravel determinants and consequences of human genetic variation. He co-initiated PCAWG, an initiative for the international sharing of cancer genomes, and a forerunner project in international omics data reprocessing. **Bork (Participant)** is head of the Structural and Computational Biology Research Unit at EMBL. He is a leading expert on research of the microbiome and its connection to human phenotypes and diseases, and coordinates the Heidelberg service centre within de.NBI/ELIXIR-DE network. **Huber (Participant)** is an EMBL senior scientist and expert in statistical data integration of various omics data types. He is a co-founder and board member of the Bioconductor project, which provides open-source tools for the analysis and comprehension of high-throughput omics data.

Eberhard Karls University Tübingen (EKUT) is one of the leading German research universities and one of the eleven German universities of excellence. EKUT has been centralising its omics data generation and management with the establishment of the **Quantitative Biology Center** (**QBiC** - founding director Kohlbacher, current director Nahnsen), a DFG-co-funded core facility. **Kohlbacher (Co-Spokesperson; EKUT, de.NBI)** is professor for applied bioinformatics at EKUT and director of the Institute for Bioinformatics and Medical Informatics at EKUT. His research focuses on methods for the analysis of large-scale omics data (genomics, epigenomics, proteomics, metabolomics). He is also the (co-)scientific speaker of the German Network for Bioinformatics Infrastructure (de.NBI) and

(co-)Head of Node of ELIXIR Germany. He is a member of the National Steering Boards of the Medical Informatics Initiative (MII) and MV GenomSeq. **Nahnsen (Co-Spokesperson)** is the director of QBiC and professor for biomedical data science. His research focuses on FAIR data management and reproducible omics data processing, and he has initiated the internationally renowned nf-core project for scalable, automated, and fully reproducible data analytics of omics data. **Krüger (Participant)** heads the high-performance and cloud computing group at the ZDV and is responsible for the de.NBI/ELIXIR-DE Cloud site in Tübingen. His research focuses on sustainable science gateways and workflows. **T. Walter (Participant)** is a Professor for Information Services and the director of the university's HPC centre (ZDV). He is responsible for the operations and strategic planning of the university's IT infrastructure.

The **University Hospital Tübingen (UKT)** is a highly specialised clinical centre with more than 420,000 patients treated per year. The Institute of Medical Genetics and Applied Genomics (IMGAG) is part of several European Reference Networks (ERNs) for rare diseases, is leading the European Solve-RD consortium and the Clinical Research Network of the European Rare Disease Research Alliance (ERDERA), and is one of four German competence centres for Next Generation Sequencing (NGS Competence Center Tübingen, NCCT) funded by the DFG. **Graessner (Co-Spokesperson)** is managing director of the Rare Disease Centre Tübingen and Coordinator of the European Reference Network for Rare Neurological Diseases. He is also coordinator of the H2020 European flagship diagnostic Rare Disease project Solve-RD and co-lead of the Clinical Research Network of ERDERA. **Ossowski (Participant)** is professor for computational biomedical genomics and vice director of the Institute for Bioinformatics and Medical Informatics at EKUT, leads the Bioinformatics for Diagnostics group at UKT, and develops methods for personalised medicine and NGS-based diagnostics. **Malek (Participant)** is Director and Chair of the Dept. of Internal Medicine I at UKT and EKUT. He has been the driving force behind establishing the molecular tumour boards in Tübingen and establishing a state-wide infrastructure (Centres for Personalized Medicine) for personalised oncology.

The **TUM University Hospital (TUMUH),** formerly known as Klinikum rechts der Isar (MRI), is one of the leading medical centres in Europe and treats more than 65,000 inpatients and 265,000 outpatients a year. In the German MV GenomSeq, the TUMUH participates for both indications, cancer and rare disease with its accredited diagnostic labs. **Mertes (Participant)** is leading the diagnostic platform at the Institute of Human Genetics and is the coordinator of the GHGA workflow workstream. He has expertise in processing large-scale NGS datasets and omics-based diagnostics as well as administering secure IT Infrastructure.

The **Helmholtz Munich (HMGU), Computational Health Center** aims to develop and apply cutting edge computational methods to promote personalised health. **Winkelmann (Co-Spokesperson) (HMGU, TUMUH, TUM)** is director of the Institute of Neurogenomics at HMGU and director of the Institute of Human Genetics at TUMUH. She heads the university's diagnostic programme as well as research programmes on rare diseases such as an unsolved rare disease programme based on trio genome and transcriptome sequencing funded by the Bavarian Ministry of Research. Her lab has extensive experience in leveraging large-scale sequencing datasets to identify driver genes of neurological disorders.

Technical University of Munich (TUM) was one of the first universities in Germany to be named a University of Excellence and is consistently ranked as one of the top European universities. **Gagneur (Co-Spokesperson)** is Professor for Computational Molecular Medicine at the faculty of informatics of the TUM. The focus of his research is to improve our understanding of the genetic basis of gene regulation and its implication in diseases. His expertise includes machine learning methods for sequence-based modelling, variant interpretation, and multi omics-based diagnostics of rare diseases.

The **Berlin Institute of Health (BIH)** is transferring research findings into novel approaches towards personalised prediction, prevention, diagnostics and therapy and, conversely, using clinical observations to develop new research ideas. It is integrated with Charité and tightly connected to MDC. **Beule (Co-Spokesperson)** leads the BIH Core Unit Bioinformatics and is affiliated with both MDC and Charité. The unit has designed and built a cross-institutional omics data management system for routine clinical applications. **Robinson (Participant)** is a Humboldt Professor for Artificial Intelligence at FU and HU Berlin and at BIH. He is an expert in structured clinical data, in particular the Phenopackets standard and has led the Human Phenotype Ontology (HPO) project since its inception in 2008.

The **Max Delbrück Center for Molecular Medicine (MDC)** was founded with the goal of understanding the molecular basis of health and disease by bringing together researchers from different disciplines. Together with the BIH, the MDC also drives translational efforts, creating infrastructure, large-scale studies, and connections between basic science and clinical research. **Häcker** (Participant) has been the CIO of the MDC since 2017. Prior to this, he was the Head of Corporate IT at the Forschungsverbund Berlin (FVB) research association. He is a board member of the VOICE - Federal Association for IT Users (Bundesverband der IT-Anwender e.V.) and a Member of the Presidential Committee FOCUS.ICT at Deutsches Institut für Normung e.V. (DIN). **Ohler (Participant)** is a full professor at Humboldt University and Senior Research Group Leader at the MDC where he coordinates the cross-cutting Data Science initiative. **Pischon (Participant)** heads the Molecular Epidemiology Research Group studying the relationship between lifestyle, genetic,

metabolic, and environmental factors with risks and outcomes of chronic diseases in human populations at the molecular level. He coordinates the Cluster Berlin-Brandenburg at NAKO.
As a University of Excellence **TU Dresden (TUD)** belongs to the top group of universities in Germany. **Dahl (Co-Spokesperson, DcGC)** has comprehensive expertise in NGS based technologies and is heading the DRESDEN-concept Genome Center, one of four DFG NGS competence centres. **Nagel (Participant, ZIH)** is director of the Center for Information Services and High-Performance Computing (ZIH). He holds a professorship in computer architecture and his expertise is centred on High Performance Computing and data-intensive computing. **Müller-Pfefferkorn (Participant, ZIH)** heads the department for distributed and data intensive computing at ZIH. His expertise is on (meta)data management, data infrastructures and data analytics.
The Cologne Center for Genomics (CCG) and the Computing Center (ITCC/RRZK) both at the **University of Cologne (UzK)** have an established close cooperation in operating the sequencing processing and storage infrastructure for NGS. **Motameny (Co-Spokesperson)** heads the CCG NGS data analysis group and leads the Next Generation Sequencing Competence Network Special Interest Group “Data Management & Protection”. She has comprehensive expertise in NGS data analysis with a focus on data management and storage integration. **Wesner (Co-Spokesperson)** is the Director of the IT centre and full Professor of Parallel and Distributed Systems and has been Coordinator of several European Cloud Computing projects. He has years of experience in Cloud and High-Performance Computing as well as operating large scale scientific data and compute infrastructure. **Achter (Participant)** heads the department for HPC and visualisation and his department provides IT services for excellence clusters in life sciences.

Participant Institutions and Participating Individuals[1]

The **Federal Institute for Drugs and Medical Devices (in German: Bundesinstitut für Arzneimittel und Medizinprodukte – BfArM)** is the medical regulatory body in Germany. It operates under the Federal Ministry of Health (BMG) and is also the platform operator for MV GenomSeq. **Hänisch (Participant)** is a Professor of Pharmacoepidemiology (Uni Bonn) and head of the Research division at BfArM. She is overseeing the installation of the data platform within MV GenomSeq.
Charité is the largest university hospital in Europe, seeing and treating over 900.000 patients per year. The Charité Comprehensive Cancer Center (CCCC) is one of the most active study centres providing the central infrastructure for translational precision oncology programmes embedded in the German Consortium for Translational Cancer Research (DKTK). **Schlomm (Participant)** is Director of the Department of Urology and carries

[1] Sorted alphabetically according to the institutions

research concerned with the clinical validation of molecular markers and patient-centred value-based healthcare. He is the founder of the German Network for Applied Precision Medicine (DNA-Med) which improves access to precision medicine and supports clinical decision-making based on real-world data. Since 2014, **Specht (Participant)** is managing director at the German Biobank Node (GBN) where she also manages the German Biobank Alliance (both BMBF-funded) which currently consists of 36 qualified large biobanks.

The non-profit **de.NBI e.V.** was established to represent and support German bioinformatics researchers, advocate for funding, and provide essential tools, resources, and training for the scientific community. Kohlbacher (Co-Spokesperson) is a board member of de.NBI e.V. de.NBI e.V. is closely aligned with de.NBI/ELIXIR-DE supported through federal funds at the Jülich Research Centre.

Schultze (Participant) is Professor for Genomics & Immunoregulation at Uni Bonn and director systems medicine at the **German Center for Neurodegenerative Diseases (DZNE).** His research focuses on the regulation of the immune system, systems medicine and the development of federated machine learning approaches together with industrial partners. **Ulas** (Participant) is a trained bioinformatician. His primary research interest is focused on leveraging computational methods to gain insights into complex biological systems, with a specific focus on the immune system and genomics.

NAKO e.V. is the legal entity implementing and owning the data of the **German National Cohort (GNC)**. It is a prospective population-based cohort study, which recruited more than 200,000 men and women in 18 study centres around the country. **Panreck (Participant)** is involved in the project management of NAKO since 2018 and since 2024 acts as Team Lead Research Data Management. **Peters (Participant)** is the NAKO PI for **HMGU** and former chairwoman of the board of directors of NAKO and the director of the Institute of Epidemiology at the HMGU with decades of experience in collecting and providing human cohort data including genetic and omics data under changing legal and ethical requirements. She also leads the expert group on omics within the NAKO and is responsible for building up the central biorepository of the NAKO at HMGU.

McHardy (Participant) is an expert in computational microbiome and pathogen research. Her lab at the **Helmholtz Centre for Infection Research (HZI)** works on computational techniques, oftentimes using machine learning methods, for the analyses of the biomolecular data, such as metagenomics data, and translating their application into medical and biological research questions. She is on the board of directors of NFDI4Microbiota, which advances the generation of high quality, FAIR microbiome research data.

The **Helmholtz Center for Information Security (CISPA)** explores all aspects of information security. **Fritz** (Participant) is a faculty at CISPA, an honorary professor at Saarland University, and a fellow of the European Laboratory for Learning and Intelligent

Systems (ELLIS). His research focuses on trustworthy artificial intelligence, especially at the intersection of information security and machine learning. **Marnau** (Participant) is a research group leader and legal scholar specialising in IT security and privacy law.

Hannover Medical School (MHH) is one of the most research-intensive medical universities and the largest transplant centre in Germany with the additional focus on stem cell research/regenerative medicine, infection and biomedical engineering and implant research (DFG Cluster of Excellence Hearing4All). **Di Donato** (Participant) is a Professor for Human Genetics, chair of the Department of Human Genetics and recognised expert in genetics of neurodevelopmental disorders and rare paediatric syndromes.

Schulze-Hentrich (Participant) is Professor of Genetics and Epigenetics at **Saarland University (UdS)** with a research focus on the analysis of gene expression and epigenetic modifications in health and disease in particular neurodegenerative disorders. She has coordinated the transnational BMBF-ANR-CIHR project decipherPD (epigenomics of Parkinson's disease) and leads the epigenomics data analysis group of solveRD. **J. Walter (Participant)** is Professor of Genetics and Epigenetics with a long-standing expertise in epigenetics and epigenomics. He has coordinated the German Epigenome Network DEEP, serves as deputy scientific coordinator of the International Human Epigenome Consortium, IHEC, and is co-founder of Single-Cell Omics Germany Initiative.

At **Kiel University (UKI)**, omics-based medical life science research is a long-standing focus, which is documented by leading roles in international scientific consortia and publications. **Rosenstiel (Participant)** has contributed to major international genomics consortia (IHEC, ICGC, IIIBD), and coordinates an H2020 project on epigenetic maps of inflammation and inflammation-associated carcinogenesis.

The **National Center for Tumor Diseases (NCT)** is a long-term cooperation between the DKFZ, excellent partners in university medicine, and other outstanding research partners at various locations in Germany. **Fröhling (Participant, DKFZ, UHH, NCT HD)** is Managing Director of NCT Heidelberg, Head of the Division of Translational Medical Oncology at DKFZ, and Head of the Division of Translational Precision Medicine (UHH). His work is at the interface of cancer research and clinical care centres on multidimensional tumour characterisation as a basis for clinical trials investigating novel cancer therapies. **Glimm** (Participant, DKFZ, NCT DD) heads the department for Translational Medical Oncology at NCT Dresden and TUD. He investigates the molecular and cellular mechanisms for cancer development, proliferation and evolution, and engages in clinical as well as experimental activities to ensure a rapid turnaround of scientific results into clinical application and clinical outcome into new hypotheses.

Grüning (Participant) leads the European Galaxy team at **Albert-Ludwigs University of Freiburg (UFR)**. He is part of de.NBI/ELIXIR-DE, the ELIXIR tools platform, and manages

the Freiburg part of the de.NBI/ELIXIR-DE Cloud. With experience running the pan-European Galaxy server, he is adept at managing virtualised, cloud environments. As a core member of conda-forge, Bioconda, and BioContainers, he promotes sustainable software deployments.

The **German National Library of Medicine (ZBMED)** – Information Centre for Life Sciences in Cologne is the central specialist library for medicine, public health, nutrition, environmental and agricultural sciences in Germany. **Fluck (Participant)** is the head of Knowledge Management and Deputy Scientific Director (provisional) at ZBMED. She specialises in the fields of text and data mining and is the Spokesperson of **NFDI4Health**.

3.1.3 Embedding of the consortium in our communities

Overarching strategy for community integration

A prominent embedding of GHGA in our target communities is ensured by the selection of co-spokespersons and participants, who are cornerstone members and leading figures of key national networks, and represent major stakeholders for genomic medicine. This setup has facilitated the development of close ties with major data generators, legal experts, and genomic medicine scientists - the key user groups of GHGA. Similarly important, GHGA has taken substantial efforts to establish a dialog with patient communities in order to align our services with their expectations and needs, in particular with respect to balancing high standards for data protection with retaining utility of deposited data for human omics research. As part of these activities, we have established a comprehensive communication strategy and platform, which provides tailored communication channels for the respective stakeholders.

Embedding in the scientific community

Facilitated by the role of our member institutions as major omics data providers, we have established a community-tailored metadata model [12,13], thereby addressing the needs of the German omics community, ensuring interoperability with relevant national and international standards. We have further established best practices concerning the formulation of patient consents, and we have actively engaged with the German Medical Informatics Initiative (MII) to ensure that these activities are fully aligned with the MII broad consent, which has been rolled out on a national scale.

Engaging research and data communities

As part of our communication strategy, we use a range of communication channels to reach our different target groups. Briefly, the online representation of GHGA conveys our mission and services, complemented by tailored social media channels, and a newsletter to address GHGA users and communities. To foster direct interactions, GHGA has (co)-organised more than a dozen community workshops, symposia and conferences, and our members have been and continue to be well-represented at key national and international events of our

core communities - from bioinformatics to cancer and rare disease to patient care. The organisation of strategic workshops also serves as a tool to develop and extend our communities, such as joint meetings with the single cell omics community, and strategy workshops to unite national and international genome data to tackle hurdles in deploying multi-cloud solutions. At events, informational materials, such as the recently finished GHGA brochure [14], are handed out to interested parties. As an additional outreach and community measure, GHGA offers a scientific lecture and webinar series, and we run courses that address challenges and opportunities related to omics research and data sharing principles. Beyond data sharing and access, we also promote FAIR concepts in the downstream bioinformatics analysis. The Workflow workstream has joined forces with the nf-core community to develop and release reference omics workflows, which are compatible with the GHGA Data Infrastructure. Collectively, these engagements have resulted in five white papers and 180 public-facing events (cf. nfdi-ID 1007-1009 in Appendix 5) for information and training in public and scientific communities.

Addressing patients and the general public

GHGA is actively engaging with the general public to increase the understanding and awareness around genomic research in general and the benefits of national and international data sharing in particular. We have developed a diversified portfolio of measures to address this, ranging from local events at GHGA sites, such as a science slam and science pop up store, to our German-language podcast 'Der Code des Lebens', which has been streamed more than 16,000 times with hundreds of regular subscribers. The podcast, which has recently been complemented by a shorter format entitled "Genomhäppchen (Genome bites), adresses a major niche as German language materials on genomics research and data sharing are scarce. Given the success of these efforts, we will extend these activities and develop additional materials in the next funding phase.

Patient engagement activities

As the subjects of the data deposited in GHGA, patients are a pivotal stakeholder group for GHGA. The GHGA consortium has therefore put great emphasis on engaging with patients both directly and via our associated networks. To this end, the ELSI workstream has conducted a qualitative study (the 'PaGODA' study) with patients from the cancer and rare diseases communities to learn more about patients' views and needs regarding data infrastructures such as GHGA and, in collaboration with the study participants, developed a patient engagement concept for GHGA. In response to the needs identified, GHGA is in the process of establishing a structured interaction with existing or a dedicated patient advisory board (cf. 3.4.1). These activities have resulted in a concept paper [15], and will inform our forward-looking strategy. We have also started to utilise our network of users, and in

particular data controllers, as multipliers to reach a larger patient collective. For example, MV GenomSeq (new participant BfArM, see 4.1.3) will be a driver project to directly impact and communicate with patients on a nation-wide scale.

Management of community expansions

Flexible funding instruments have proven valuable to quickly onboard new communities, initiating targeted measures to interact with our stakeholders. A prime example of a new community we added to the consortium was GHGA's COVID-19 response measures, contributing to making viral host genome data available for research (cogdat.de). More recently, we have also used this flexibility to contribute to the establishment of the MV GenomSeq, the onboarding of new data hubs, and we have initiated the Assured project (B3.M5) - a new training measure developed together with other NFDI consortia to address communities' needs in working with sensitive data.

Forward-looking enhancement of community value

The now operational core Archive service (cf. Service1: GHGA Archive) will enable new dimensions of community interactions. We strive to make the archive as useful as possible by populating it with primary data, as well as with data derived from that. Our partnerships with the National Cohort, the Model Project Genome Sequencing, and other data controllers assure the steady growth in data (10+k genomes per year, cf. Table 4). The continued mobilisation of community datasets will be a forward-looking priority. To this end, we will establish a distributed network of data stewards, tasked to identify new datasets and users within their local networks, and facilitate data ingestion (A4.M3).

We will empower our communities by providing data analysis workflows that can be executed on behalf of (and driven by feedback from) our communities. We will also make derived data and results directly accessible (e.g., beacon services, variant databases, annotation services cf. B2). These measures will help to grow our user base, providing significant added value to the communities. We will continue to monitor opportunities to include new communities where opportunities emerge (B1), for example additional omics modalities, single-cell technologies, and using flexible funding to initiate corresponding measures (C1).

The GHGA Archive, as the cornerstone of our services, will also be leveraged to develop new modes of interaction with our communities, by developing new use cases and embedding the archive service into the scientific data management processes. We have developed driver projects in strategic areas (B1), with the aim to fast track the development, demonstration and application of our services, in a community-centric manner. These driver projects will in particular be major use cases for the GHGA secure processing environment (A3.M2).

In response to the expected growth in data and the expected uptake of our services, we will widen the engagement of external advisors into our governance model (cf. 3.4.1), to ensure

that community needs are taken into account when refining our service portfolio. This includes both the implementation of a patient engagement concept as well as a GHGA User Advisory Board (UAB), representing data controllers and users who wish to access data in GHGA.

3.1.4 Training activities

The embedding of key stakeholders of our communities has enabled GHGA to (co-)organise in person training events in topics that are relevant to GHGA together with our communities, covering topics such as bioinformatics analysis of human omics data, processing of omics data, translational oncology and beyond, entailing a total of 36 training sessions with approx. 2,000 participants (cf. nfdi-ID 1008 in Appendix 5 - the total number of 65 events includes the number of webinars (29)). The embedding with these networks will remain a major facilitator to connect to the most relevant communities (cf. B1, B2, B3). Complementary to these, we have established a successful webinar series, averaging around 86 live attendees per session and a substantial number of streaming participants (124 on average). Covering topics related to GHGA such as data protection, FAIR data sharing and metadata, but also methods for data visualisation and statistical genomics. Both training arms have primarily attracted young scientists, from MSc students to PhD candidates and early career researchers. Several of our co-applicant institutions accept attendance certificates, and while these events are open to the entire community in Germany, we also continuously attract a steady stream of local participants who profit.

In addition to the organisation or training events together with our communities, we have taken first steps towards addressing training gaps in a targeted manner, which has led to bespoke and GHGA-specific training offerings. The most relevant example is the Assured project, which is further described under 3.2.2 and B3.M5. Finally, to further ensure that the platform and online experience of GHGA matches the community needs, we plan to establish a dedicated measure within the TA Outreach and Training (B3.M6) looking into user experience. Using mixed method approaches, we will continuously assess the needs of our user communities and evolve our services accordingly.

3.2 The consortium within the NFDI and the national academic research system

3.2.1 Contributions to the NFDI and commitment

As one of nine first-round NFDI consortia, GHGA played an active role in developing and shaping the NFDI and its governance structures. GHGA and its associated co-spokespersons have contributed to the installed governance bodies on all levels, both within the association (consortia assembly, sections) but also within the more organisational circles (speaker, management, finances, communications).

3.2.2 Multidisciplinary use cases within NFDI

Data infrastructures must meet diverse and evolving multidisciplinary research needs. NFDI in particular was established precisely to make a "significant contribution to addressing new interdisciplinary research questions of high societal relevance" [16]. Multidisciplinary use cases address the RDM-challenges related to interdisciplinary research questions [17]. GHGA has already been engaged in multi-disciplinary use cases in the context of ELSI, where GHGA is prominently represented, within the framework of the NFDI, to the cross-disciplinary alignment of ELSI, data protection and common (base) infrastructure components. In this context, we actively contribute expertise through dedicated training courses on means of GDPR-compliant processing of sensitive data.

An example of a multi-disciplinary activity is the **Assured project** (also see B3.M5 in the work programme), a joint effort GHGA tackles in collaboration with KonsortSWD and BERD@NFDI, aiming to develop a cross-disciplinary training and accreditation service for users of sensitive research data. This training programme, initially funded from GHGA flex funds and organised together with KonsortSWD and BERD@NFDI, is now part of the core work programme for the second funding phase (cf. B3.M5) and is also embedded in the work programmes of the collaborating consortia. In addition to conveying skills, this programme will also hand out accreditation to participants, which in the future could be made mandatory for users requesting data from GHGA and other NFDI initiatives hosting sensitive data. Our co-applicant institutions have expressed strong interest in using the accreditation internally to address the skills gap in this domain.

The Assured project is an excellent example of an activity we did not foresee at the time applying for the first funding period. More generally, we appreciate that multi-disciplinary use cases in a dynamic research environment cannot be anticipated ahead of time. We therefore commit to:

1. Collaboratively issuing open calls for multidisciplinary use cases with other consortia, incorporating a joint selection process. GHGA would, for example, be interested in use-cases pertaining to privacy, data protection or ethical considerations of data sharing and reuse.
2. Assist researchers in finding other consortia that complement our expertise. At the same time help researchers from other research contexts with use cases brought to our attention by other consortia.
3. To support both the calls and ad-hoc cases, we pledge to use resources from our flex funds (cf. C1). We will coordinate with other participating consortia to share costs for multidisciplinary use cases as appropriate.

To support these commitments, we are currently in the process of organising a "Biomedicine@NFDI" workshop together with the NFDI life science consortia (NFDI4Health, NFDI4Immuno, NFDI4BIOIMAGE, NFDI4Microbiota) in November 2024.

3.2.3 Scientific interactions within the NFDI

Naturally, the closest ties emerged with other NFDI consortia in the life and biomedical sciences, i.e. **NFDI4Health** and **NFDI4Biodiversity** in the first round, **NFDI4Microbiota** in the 2nd and **NFDI4Immuno** and **NFDI4Bioimage** in the 3rd round. For the latter two consortia, collaborations were already initiated during their conception, which was facilitated by the co-location of coordination centres (GHGA and **NFDI4Immuno** are coordinated at the DKFZ and there is also a **NFDI4Bioimage** coordination hub at DKFZ). As a result, these NFDI networks are closely aligned concerning their services, metadata, and on a legal level. The harmonisation of metadata schemas is an ongoing activity with **NFDI4Health**, **NFDI4Immuno** and **NFDI4Bioimage**, with GHGA contributing its expertise in the handling of sensitive health data, also in the context of metadata. We expect this activity will also create opportunities to contribute to and build on the **TS4NFDI** base service. GHGA is also engaging with **IAM4NFDI** base service, to develop their legal and data protection concept to ensure it provides the required safeguards to be used in the context of sensitive health data.

As consolidation and provisioning of data to the biomedical research communities are a common mission of **NFDI4Health** and GHGA, the consortia have committed to intensify their collaboration in a joint Memorandum of Understanding [10]. As part of this alliance, the consortia will engage in joint use cases aiming to create linked data resources and develop new analysis, exploration, and federated query tools for personal health data that have the potential to eventually improve population health. The consortia will also engage in overarching metadata standards, as well as legal and ethical frameworks for NFDI and international communities.

GHGA is also part of the FAIR Data Spaces (FAIR DS) project, where GHGA PIs contribute to the development of an ethical legal framework for data exchange between public and private stakeholders. FAIR DS aims to link activities of the NFDI and the Gaia-X community at a European level. GHGA PIs Buchhalter (DKFZ), Kohlbacher (EKUT) and Korbel (EMBL) are PIs in the de.KCD project ("German Competence Center for Cloud Technologies for Data Management and Processing", as part of the BMBF call for Data Competence Centers), which will provide ample opportunities to reinforce the bridge between de.NBI and NFDI. Building upon the technology stack, services and use cases in GHGA, we will contribute to pan-NFDI activities to leverage synergies and develop and use joint services. In addition to the existing interactions, we will also contribute to, and leverage, opportunities for technological collaboration. Through the NFDI sections and Base4NFDI, we are actively exploring joint authentication services, as well as joining pan-NFDI metadata standards and efforts to foster interactions and interoperability between NFDI structures and European initiatives such as the EHDS.

3.2.4 Relevance to the German science system

In parallel to the embedding of GHGA within the NFDI community, the consortium is broadly integrated and connected to a number of key infrastructural activities, initiatives and networks within the field of biomedical sciences.

GHGA's role in the German science system is unique: it is the only national archive for human omics data and thus indispensable for many national initiatives. No other infrastructure has the legal, technical, or organisational capabilities to take its role.

As mentioned above (3.1.3), GHGA has established collaborations with the Medical Informatics Initiative and the MV GenomSeq project, where GHGA fills a gap in expertise in the management, handling, and the analysis of human omics data. GHGA also contributes to, and builds upon, the de.NBI and ELIXIR, for example via the establishment of shared resources and using the physical infrastructure of the de.NBI/ELIXIR-DE Cloud. These activities and engagement are primarily at the national-level, however GHGA is uniquely positioned to bridge national initiatives to key international networks; as a result GHGA occupies an important strategic position within the national research system. Most importantly, GHGA as a consortium holds critical roles within the MV GenomSeq (cf. 2.1.2 Impact & 4.1.3), a strategic initiative launched by the German Ministries of Health (BMG) that aims to foster innovation in genomic medicine and establish genomic sequencing as part of the routine diagnostic in Germany. GHGA has been a major contributor to the conceptualisation of the data infrastructure of the MV GenomSeq, and GHGA PIs contribute expertise at all levels. The GHGA data hubs also have an exclusive role as being approved by BfArM - the responsible institution for the overall infrastructure of MV GenomSeq - as genome data centres. Without GHGA, the implementation of the Model Project Genome Sequencing in Germany would be impossible or at least delayed by several years. As a national infrastructure to represent Germany within the European Genomic Data Infrastructure Initiative (GDI) (cf. 3.3, 4.1.3), GHGA also has the mandate to coordinate metadata exchange with European networks. Going forward, the deep involvement of GHGA in the MV GenomSeq will be a major driver and opportunity for future developments. In addition to being a key use case, data provider and customer for GHGA, the MV GenomSeq will also facilitate the consolidation of our roles and community interactions. Finally, we expect income from the MV GenomSeq, providing a unique opportunity to establish a long-term business model for GHGA (cf. C2.M4).

3.3 International networking

A deep international embedding has been a cornerstone principle of the GHGA Consortium. Human omics research cannot thrive within national borders as it must integrate the overwhelming complexity and diversity of human biology by aligning efforts for enabling data

sharing, as has also been noted by the reviewers of our initial project application. Consequently, GHGA has taken, and will continue to take (cf. B4) considerable efforts to align with a series of relevant international endeavours, in particular infrastructures with a focus on human omics data (see Table 2). In fact, many of our communities and stakeholders nationally have natural international counterparts, with whom we interact. Specifically, GHGA has been a founding member of the **federated European Genome-phenome Archive (FEGA),** a consortium of national genomic data infrastructures with the goal to enable cross-border data findability and access (collaboration agreement with EGA signed in June 2022 as the second country to join the federation). GHGA members are represented and contribute to all governance bodies of the FEGA, and together with the four remaining founding members (Sweden, Finland, Norway, Spain) is shaping the technological and legal foundations of the FEGA. Closely connected to FEGA, GHGA is playing a major role in contributing to the EC-led **1+ Million Genomes initiative (1+MG)**, of which Germany is an official member [18]. GHGA members contribute to 1+MG working groups, the special group, and we have been mandated by the BMG and BMBF to act as national implementation partners within the **European Genomic Data Infrastructure (GDI) project**. As part of GDI, GHGA is a key stakeholder in aligning and designing the architecture of a cross-border genome data sharing infrastructure that will be deployed in over 20 participating countries. The engagement of GHGA in GDI represents the endorsement of BMBF and BMG and the consortium also received additional financial support from the EU and the BMBF for this task. As part of the first activities, GDI released a "Starter Kit" to support federated data access within the national nodes of GDI. Technologies included are based on standards developed within large international activities such as the **Global Alliance for Genomics and Health (GA4GH)**, which also form the basis of key GHGA services, thus providing the basis for a successful technical and legal alignment with the emerging GDI network. At the end of 2023, GHGA has also been invited to join GA4GH's "National Initiatives Forum (NIF)", which includes national programmes focused on advancing genomics and to translate genomics into health systems. Through this, GHGA will further strengthen its international engagement, and its alignment with related global programmes on goals and practices. GHGA members are also contributing to legal and ethical aspects of international data sharing, also within GA4GH, which has resulted in several publications in this field [19,20]. Further, GHGA members are contributing to the development of and actively maintaining NGS workflows within the **nf-core community**, which is an international community with the aim to standardise and harmonise NGS data processing. A forward-looking opportunity and goal is to strengthen our interaction with the **European Open Science cloud (EOSC)**. Given the strong dependency on and technological alignment with cross-border data sharing using cloud solutions, GHGA is well

positioned to act as a national driver project and use case for EOSC, an opportunity we plan to build on in the next funding period (B4.M2).

Table 2: Areas of alignment with key national / international initiatives

Networks	Metadata Exchange	Data Exchange	Interoperability	Standards	Infrastructure
NFDIs	**x**		**x**	**x**	**x**
MII/NUM	**x**		**x**	**x**	
MV GenomSeq	**x**	**x**	**x**	**x**	**x**
FEGA	**x**		**x**	**x**	
GDI/1+MG	**x**		**x**	**x**	
GA4GH			**x**	**x**	
ELIXIR-DE			**x**	**x**	**x**
NAKO	**x**	**x**			**x**
de.NBI			**x**	**x**	**x**
nf-core			**x**	**x**	**x**
Galaxy			**x**	**x**	**x**

Besides infrastructure-focussed activities, GHGA members have contributed to key international research projects within our focus research communities:

1. In the cancer field, GHGA members are represented in recently funded EU projects such as UNCAN.eu or EOSC4Cancer, ensuring that GHGA developments are aligned with these overarching coordination efforts. Members of GHGA are also contributing to the ambition of the Hopp Children's Cancer Center Heidelberg (KiTZ), which has the goal to establish a clinical data repository for the collection of detailed datasets from paediatric cancer patients [21]. Since the network (ITCC PedCanPortal) includes world-wide partners who will contribute with their datasets, including children's hospitals in Canada and Australia, this engagement will further strengthen the international activities of GHGA with a prominent use case.
2. Within the field of rare diseases, GHGA members have leading roles in the European Rare Diseases Research Alliance (ERDERA). It builds on Solve-RD [22], which is coordinated by GHGA members and the European Joint Programme on Rare Diseases (EJP-RD) in which GHGA members participate and are WP leads (cf. B1.M2).

Future strategy & perspective

GHGA will continue to actively engage with European networks and initiatives, aiming to establish and act as a strong bridge between national and international initiatives. New directions will come from growing ties with GA4GH, which has accepted GHGA as a member of the National Initiatives Forum. Strategically, we will retain our current strategy, which is to operate a self-sustained national infrastructure that can be linked and is interoperable with different standards. We have already established metadata standards that allow joint embedding in GDI and FEGA, further details are described in B4. Together with

partners in the NFDI, we hope to underline the role of the NFDI also on the European level.

3.4 Organisational structure and viability

3.4.1 Governance Structure

Governance Bodies

While the governance structure drafted in the original application proved to work well, we made certain adjustments to consolidate the consortium and to meet the changing needs as we moved from development phases to the production phase of the infrastructure (cf. Fig. 1). We will describe the main governing bodies, but also how the decisions made in the governance structure are implemented in a distributed team.

Based on the changes in the task areas in the upcoming funding period, we have consolidated the number of Co-Spokespersons (SPs) responsible for the coordination to now 15 individuals, including the spokesperson of GHGA. In brief, the governance structure consists of the following bodies:

- **Board of Directors (BoD)**: The BoD consists of four members elected by the SC and chaired by the current spokesperson of GHGA, Oliver Stegle, with a term of three years. The current members are Oliver Kohlbacher (Deputy spokesperson), Eva Winkler and Jan Korbel. Guest status is used to foster smooth transitions as the BoD changes composition in the future (currently Julien Gagneur). The main tasks of the BoD are the overall coordination of the infrastructure and project execution as well as the interaction with the funding (DFG) and umbrella (i.e. NFDI) organisations of GHGA. It reports to the SC and the external advisory boards (SAB, UAB) and appoints the Team Leads (TLs). The BoD meets monthly and is supported by the **GHGA Office**.
- **Steering Committee (SC):** The main steering board of GHGA is the SC, which is formed by the SPs and TLs. The SC meets bi-monthly and is chaired by the spokesperson and its deputy. Through the SC, all TAs come together to ensure overall alignment and decision about new strategies. SPs and TLs have voting rights on all matters except financial decisions and approval of Flex Funds projects (C1), where the voting rights are limited to the SPs.
- **General Assembly (GA):** Since the SC now already includes all Co-Spokespersons and GHGA Team Leads, the Members Assembly (which in the initial governance structure was the assembly of all 27 Co-Spokespersons) is no longer necessary. Instead, all SPs, Participants and GHGA Team Members (staff funded via GHGA) meet bi-annually in the GA. The BoD and SC report on the overall progress and the whole consortium can discuss and develop the overall strategy.

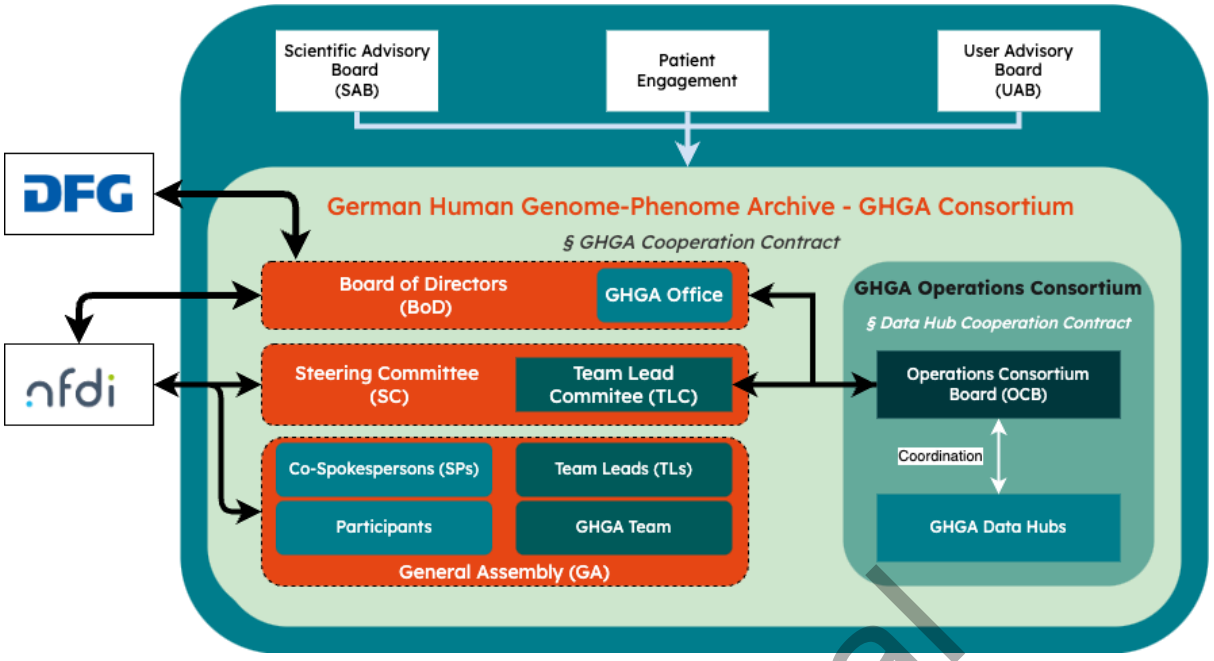

Fig. 1: Governance Structure for the next funding phase. All bodies within the light green area will again be contractually bound by the revised GHGA Cooperation Contract. As a subset to this, all data hubs form the GHGA Operations Consortium regulated by the GHGA Data Hub Cooperation Contract, which is coordinated via the Operations Consortium Board (OCB). The main decision-making body within GHGA is the Steering Committee (SC), which decides i.a. on the uptake of new members, and elects the Board of Directors (BoD). The main tasks of the BoD, supported by the GHGA Office, are the overall coordination and execution of the GHGA project and the interaction with the funding (DFG) and umbrella (i.e. NFDI) organisations of GHGA. The Team Leads (TLs) form the Team Lead Committee (TLC), which is responsible for the implementation of the GHGA TAs aims and alignment with the Co-Spokespersons (SPs) coordinating the TAs. Together with the SPs, the TLs form the Steering Committee (SC) which supports the BoD with strategic planning decisions and joint reporting to other governance bodies. The GA (all participating members and staff) ensures overall strategy alignment, while the Scientific Advisory Board (SAB), the User Advisory Board (UAB) and the patient engagement activities provide external advice on the functionalities and directions of GHGA.

- **Team Lead Committee (TLC):** The Team Leads form the TLC, which meets weekly and is chaired by the Technical and Administrative Leads (cf. below). Besides the Team Leads, other Team Members can be regular members of the TLC (e.g., to represent partner projects such as GDI).
- **Data Hub Operations Consortium Board (OCB):** To support the operations of the GHGA Data Infrastructure, the GHGA Central and the GHGA data hubs form the GHGA Operations Consortium, which is based on an additional collaboration contract. The OCB is responsible for maintenance, development, security and sustainability of the (externally funded) physical infrastructure of the data hubs (storage, compute), load balancing across the data hubs (w.r.t. storage, compute and personnel capacities) and for supporting data stewardship processes across the consortium (A2 & A4).

External Advisory Boards

On questions of overall strategy, GHGA is advised by its **Scientific Advisory Board (SAB)** consisting of internationally renowned scientists involved in research data infrastructures and international initiatives with direct relevance to GHGA. Continuing the interactions in the

previous period, the SAB will provide strategic input on an annual basis, with bi-annual invitations to the GHGA Annual Meetings. The SAB is especially relevant with respect to international embedding, ELSI matters, and the selection and adoption of emerging technologies. SAB members are proposed by the SC and selected by the BoD. The office term of SAB members is three years, with the possibility of re-appointment. Currently the SAB consists of the following members:

- Soren Brunak (Prof. for Disease Systems Biology, University of Copenhagen)
- Ruth Horn (Prof. for Ethics in Medicine, University of Augsburg)
- Janet Kelso (Prof. for Genetics, MPI Leipzig)
- Bartha M. Knoppers (Law Prof., Centre of Genomics and Policy, McGill University, Montreal)
- Ilka Lappalainen (Dep. Head of ELIXIR Finland at CSC - IT Center for Science, Espoo, Finland)
- Christine Mundlos (Deputy Director, ACHSE e.V.)
- Augusto Rendon (Head of Bioinformatics, Genomics England)

All efforts of GHGA are based on the willingness of patients and citizens to make their personal data available for research. Based on the results of our study on patient involvement in the governance of GHGA of the ELSI workstream with patient experts as co-researchers we have a strategy for impactful and sustainable **Patient Engagement** (B5.M4). In a next step an alignment with the patient engagement strategy of the MV GenomSeq is key, as well as those of other partner initiatives of GHGA, which have already implemented a close exchange with patients (e.g., OneNCT and ERDERA). To institutionalise patient engagement within GHGA, we will collaborate with patient representatives, established patient advisory boards, initiatives such as ACHSE e.V. and the projects mentioned previously and then decide for the most efficient mechanisms (cf. B5.M4 for details). This will be supported by our activities to create suitable communication materials on data sharing specifically for patients (B3.M3) We are confident that involving patient organisations and patient representatives in the governance of GHGA will help to define quality standards for meaningful and sustainable patient engagement in the governance of omics data and data infrastructures and will also increase the acceptance of a platform such as GHGA by the wider public.

The **User Advisory Board (UAB)** will consist of five representatives of the users of GHGA that are recruited from users of GHGA (i.e., mainly data depositors and data requesters) via an online call for nominations, followed by an online vote of the GHGA user community. Members should not be directly involved with the consortium (co-spokespersons, participants, funded team members), however due to the strong embedding of GHGA in the national community this might not always be fully avoidable. The UAB is convened online twice per year and receives early insights into the GHGA roadmap and new features before

release. The advice of the UAB will help prioritise the developments of the GHGA services and to adjust them to the needs of individual communities and across communities.

Team Structure within GHGA

The success of GHGA relies above all on the strong GHGA Team and its commitment to the project. During the initial period, several team members have taken on leadership roles, taking responsibility to coordinate the development of core GHGA products of the TAs. To ensure GHGA can support professional development of the recruited staff, we will appoint **GHGA Team Leads (TLs)** to formalise these roles and responsibilities in the next funding phase. TLs are senior members of the GHGA Team, responsible for the coordination of one or multiple Task Areas (cf. Table 3). The new team lead structure will enable agile decision making with effective distribution of responsibilities on the one hand and will foster career development of currently recruited staff on the other hand.

Together with the SPs of the TAs, the Team Leads are responsible for:

- Overall coordination of TA aims and strategy
- Supervision of the involved GHGA Team Members together with the PIs at the respective institutions
- Resource management of the TA, design of and contribution to new Flex Funds projects
- Alignment with other TAs and the overall GHGA strategy
- Embedding of GHGA in the NFDI and with other core partners
- Fulfilment of reporting duties towards the GHGA governance bodies (BoD, SAB), NFDI and funders

Table 3: Overview of GHGA Team Leads

	Title	Areas of Responsibility	Team member in this role
TAs A	Team Lead Technology / Product Management	Overall Coordination of TA A and Alignment with TA-B and TA-C including National and International Networking with TA-B4	Koray Kirli
	Team Lead Operations	Lead of DevOps team in TA-A1/Coordination of Data Hub Operations in A2	N.N.
	Information Security Coordinator (ISC)	Responsible for the ISMS of GHGA Central	Pascal Kraft
	Team Lead Architecture	Leading the Development Team / TA-A3	Leon Kuchenbecker
	Team Lead Data Steward	Leading the Data Steward Team / TA-A4	Paul Menges
TAs B	Team Lead Community Engagement	Coordinating Community Engagement and Data Service Projects / TA-B1 and TA-B2	Andrew Behrens
	Team Lead Communications and Training	Overall Coordination of Outreach and Training / TA-B3	Ulrike Träger
	Ethics Coordinator	Coordination of Ethics Activities	Andreas Bruns
	Team Lead Data Protection and Legal	Overall coordination of all legal and DP processes / TA-B5 and TA-C2	Simon Parker
TAs C	Team Lead Administration	Coordination of PM, Finances, HR and Governance, including National and International Networking with TA-A4 / TA-C2	Jan Eufinger

3.4.2 Financial Management

As described in the progress report in 2023, at the start of the project, we have taken great care to adapt the funding and adjust the goals of the first funding period of GHGA to match the incurred budgetary cuts (26.5% cut of the original proposal). All institutions could be kept within the project although with reduced budget, a fact that also influenced the timeline of the project development. Through three rounds of Flex Funds, we have supported relevant areas and new directions. Reduced cash flow in the first 2.5 years of the project, due to well-known challenges in the recruitment of IT experts combined with the impact of the pandemics, have led to more funds being available in 2023 and 2024. Those additional funds were used to strengthen strategically relevant areas especially in the development of the GHGA Data Infrastructure by additional recruitments but also through focussed involvement of external experts and support. Especially helpful in the context of budget adjustments were the availability of funds from own contributions, which were not only used for personnel but especially for necessary outsourcing contracts and investments.

For the next phase of GHGA, we have developed a robust financial management plan ensuring transparency, accountability, and efficient use of awarded funds. As detailed in the Work Programme, each TA is equipped with a core budget allowing the maintenance of a robust and sustainable work force carrying out the described tasks. We have again budgeted up to the equivalent of five positions for Flex Funds, which will be assigned for key measures for the success of GHGA (cf. C1). Based on the experiences gathered before, C2 will ensure efficient financial reporting allowing efficient use of the available funds, including efficient processes and contractual arrangements to reinforce critical parts of the project by rerouting unspent funds between the institutions. We have also started to acquire additional funding, beyond the NFDI support, which will be important steps towards a sustainable business model (C2.M4).

3.4.3 Consolidation and Demand-Driven Adaptations of the GHGA Service Portfolio

The GHGA consortium is committed to enhancing the efficiency and effectiveness of its services through strategic consolidation of structures and services.

Structurally, we have adapted the task areas and governance structure to the needs of an operational physical infrastructure and will continue to do so. The governance structure also includes the consortium's two advisory boards (SAB, UAB) who will assist the BoD and SC to effectively incorporate community feedback and prioritise structural change appropriately and balance the interests of different stakeholders in the process. Structural changes concerning the physical data infrastructure will be advised on by the GHGA Operations Consortium Board (OCB), which will suggest such changes based on the existing capacity and the demand. Measures could e.g. be the increase or decrease of the number of data hubs based on the capacity and performance or the adaptation of SPE capacities based on

changing demand. These structural changes will, however, be constrained by the available external funding (own contributions or additional grants), which would be prerequisite to an expansion here. Part of our structural consolidation efforts are also close collaboration with other NFDI consortia. Integration will be closest with NFDI4Health, where we are able to host the omics data for the studies referenced in NFDI4Health [10]. With other NFDI consortia we will also be working closely on harmonising metadata standards. Another focus of the upcoming funding period will be the integration of Base4NFDI services as soon as they are production-ready. Our highest priorities are IAM4NFDI for AAI services and PID4NFDI. GHGA nevertheless faces some issues arising from the sensitivity and visibility of the project that results in higher demands on IT security and a legal basis than many other NFDI consortia. There are thus still several points to be discussed (e.g., a reliable legal/contractual basis for the authentication of persons) before we can achieve a full integration. We are confident that we will be able to solve these open issues with Base4NFDI in the coming year.

Service consolidation based on demand requires close interaction with our stakeholders and target communities. Due to their diversity, we need to collect their input through different channels: our advisory boards (SAB, UAB) are clearly one route for feedback to the consortium, but it is mostly suited to strategic decisions. Many decisions on the modification of the services, however, need to be based on the requests and demand for a broader user base. There are different feedback channels that have been established. The GHGA Helpdesk plays a crucial role in providing direct assistance to users and collecting feedback (cf. nfdi-ID 5007 in Appendix 5). This service ensures that we can respond promptly to user enquiries and continuously refine our offerings. Community engagement is further enhanced through organising workshops at relevant conferences and hosting GHGA User Meetings (cf. B3.M4). These events provide valuable opportunities for knowledge exchange, networking, and gathering feedback. GHGA also supports seed funding for community projects via Flex Funds (C1) to foster innovation and collaboration within the research community. Through these efforts, GHGA aims to maintain a service portfolio that is not only relevant and user-centric but also sustainable, thereby supporting the advancement of scientific research and the development of new healthcare solutions. To ensure that our training services are demand-oriented, GHGA is actively engaged in community outreach and user integration. This involves gathering feedback through patient engagement, the User Advisory Board, workshops at conferences, and social media campaigns. Other feedback mechanisms include questionnaires distributed during lectures and courses. GHGA also monitors key performance indicators, such as participation rates in training events, to better understand the demand for our services. This data-driven approach helps

us tailor our offerings to meet the specific needs of our users. Based on this feedback, the TLC will draft suggestions that will be decided by the BoD.

3.5 Operating model

As with most data infrastructures, GHGA relies on a mixed funding model. A long-term base funding is indispensable for GHGA to be able to fulfil its mission as a national infrastructure for long-term archival of research data. On top of this base funding, additional - usual project or service-related, funding is required for its expansion, adaptation to new technologies, communities, and roles within other strategic initiatives.

GHGA relies on long-term **base funding** from the NFDI and on significant contributions by the institutions hosting data hubs as part of the GHGA Operations Consortium to deliver its services (cf. 8 for a detailed breakdown of the requested funding and the own contributions - which are roughly equivalent). The operating model of GHGA **does not currently foresee fees for data deposition** - as these would deter potential data depositors, in particular for smaller-scale projects (several of the co-applicants have had this experience in smaller-scale local RDM infrastructures). Large-scale projects (1,000+ genomes - e.g., NAKO, MV GenomSeq, genomic newborn screening) are encouraged - and are expected to - support GHGA via staff or financial contributions to permit the scaling of the infrastructure. This **external community funding** has already created significant income and NAKO is the first community to contribute in a substantial and sustainable way. Going forward, we expect the MV GenomSeq to be the second major project to contribute financially, thereby enabling GHGA to operate a cost-neutral operation model, providing a major leap towards sustainability (C2.M4). **External research funding** will be used to finance additional data services, usually together with members of the community these services will be aimed at. These could be large-scale data processing projects, AI-based tools based on GHGA data, or community-specific workflows which will then become available through GHGA as a sustainable platform. We have been able to procure significant external funding through European and German funders (e.g., NAKO, GDI, FAIR-DS projects, BMBF Decade Against Cancer, genomic newborn screening) and expect the amount of external funding to increase significantly in the upcoming funding period due to the growing availability of data and the growing interest in human omics data. Alignment will be achieved by GHGA co-spokespersons and participants acting as Co-PIs on the respective research grants, thereby also reinforcing community interests represented by these grants.

Commercial data access fees will be explored in the upcoming funding period in two directions. First, we would like to enable very large-scale data analysis projects that require the elasticity of a commercial hyperscaler to be accessible to external users. On the one hand, this will require a robust solution to ensure appropriate security models, motivating the

development of a dedicated GHGA community SPE (A3.M2) and on the other hand a fee-based access model. Second, commercial access to the data will also necessitate a fee-based model, as there are legal implications (competition and tax law) that would prohibit us from providing significant services for free to such users. We will develop and deploy reliable business models for this in the upcoming funding period, working closely together with major data providers like NAKO and MV GenomSeq, taking inspiration from existing solutions such as the UK Biobank Research Analysis Platform.

One major remaining open issue is a national strategy for **hardware infrastructure funding**. The current funding scheme does not permit hardware funding through the NFDI. Hence, data infrastructures like GHGA that are dealing with data on the petabyte scale need to make considerable effort to procure the required storage and compute through other funding lines. Currently, GHGA is profiting from significant contributions of its co-applicant institutions (more than 20 M€ in investments, external funding, and operating costs estimated for the upcoming funding period), much of which has been given to these institutions through external funding lines (e.g., the de.NBI/ELIXIR-DE Cloud infrastructure through BMBF).

A more reliable **long-term funding strategy for national research data infrastructures** is thus one of the political goals GHGA will be working for in the upcoming funding period. The second difficulty for the GHGA operating model has been DFG's strict adherence to annual budgets. The uncertainty of whether the whole budget can be carried over to the next calendar year - this has to be decided by DFG on a yearly basis - made it impossible to give staff longer-term contracts and in the end made it very difficult, or even impossible, to fill certain positions.

4 Research Data Management Strategy

4.1 Scientific relevance and quality of the measures

4.1.1 Overview of the project and changes from the initial proposal

In our initial proposal, we mapped out and initiated a roadmap to set up a comprehensive human omics data infrastructure in Germany from scratch. This roadmap included components to address the standardisation and capture of metadata of omics data, making data accessible in an archive connected to EGA, curation of community reference data sets, and finally the provision of data in cloud-based research analysis platform. Over the five years since we conceived this idea, we had to both implement substantial components, and adjust it to shifting legal boundary conditions, developments within Europe, and to substantial changes in national requirements. We would therefore briefly sketch an updated vision for GHGA before laying out both the successes and challenges during the first funding period, as well as the perspective for the upcoming funding period.

It was clear from the start of the project that we should mobilise the metadata first, then deliver the research (raw) data to our communities, and build additional services based on the data in the archive. Over the course of the projects, we thus developed a four-stage model for the development and rollout of the core data infrastructure of GHGA (Fig. 2).

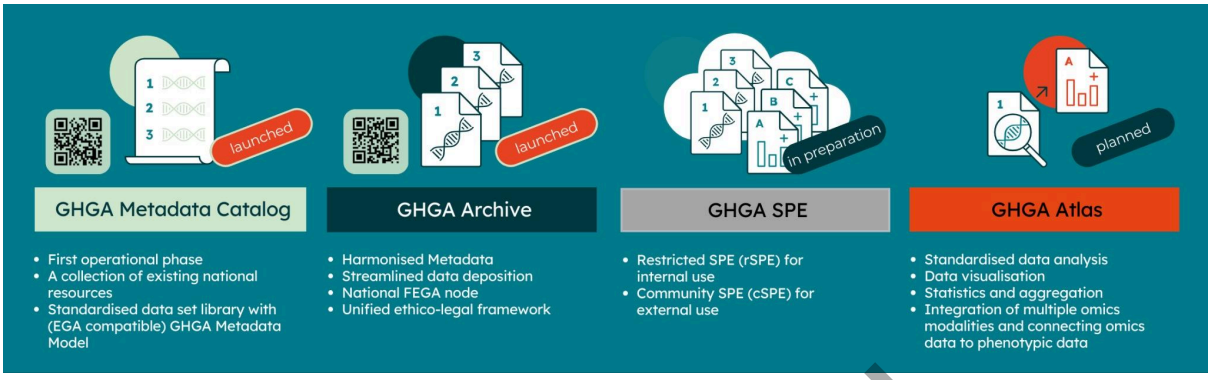

Fig. 2: Phases of the GHGA core data infrastructure and associated services. Overview of the core data archive and usage services in GHGA, ordered by phases of development and launch (from left to right).

Since 2023, the **GHGA Metadata Catalog** (catalog.ghga.de) has enabled data use by making the data findable and accessible, although the responsibility of sharing the data upon request did still reside with the data controllers. With the start of the **Archive phase** (data.ghga.de) in 2024, GHGA now provides archive services for the raw research data as well. In the initial proposal, we did foresee prioritisation of community-curated reference data sets (such as variant frequency databases) over cloud-based data access and usage.

A general shift in Germany and Europe away from downloadable data towards trusted research environments, most notably the legal mandate for several national SPEs for healthcare data in Germany and the end of download access to data from the UK Biobank, prompted us to shift these priorities. As a result, we are now working on the **SPE Phase** of GHGA which provides a cloud-based analytics platform thus avoiding the data download (currently still possible) in favour of more secure SPEs. Initially, this will consist of a **restricted SPE (rSPE)**, primarily aimed at internal use (quality control, dedicated processing by GHGA on behalf of the data controller). In the next funding phase, the rSPE will be complemented by a **community SPE (cSPE)**, which will be accessible to all researchers and offer improved scalability. The community reference data sets, statistics, and integrated visualisation, will thus be an important goal of the upcoming funding phase only. Of course, the four phases of the infrastructure development have been accompanied and supported by numerous measures (e.g., outreach, legal groundwork, technical developments) required to achieve these phases.

4.1.2 Challenges and successes of the first funding period

GHGA has managed to establish a national archive for human omics data with a full range of services around the ingestion, archival, and release of human omics data for research. While

not every single task laid out in the original proposal could be implemented, and the scope had to be reduced owing to budget cuts, the NFDI consortium has delivered a robust research data infrastructure that is widely accepted by the scientific community and has become a foundational infrastructure for many human genomics projects in Germany as well as a national data hub in many European initiatives.

Successes

Overall, of the 125 tasks we had planned in the original proposal, we have completed 62 tasks, we had to modify and conclude 17 tasks, and only 2 tasks were failed. The remaining tasks are not yet due (based on their milestones) or had to be removed (based on the budget cuts).

One of our major successes was the successful establishment of a young, distributed, **motivated team** tackling the multitude of challenges jointly and with great enthusiasm. This team consists of currently around 100 people (including PIs), is highly international (all GHGA internal communication is in English), and has allowed us to address all challenges in a timely manner. Several key members of the team took on leadership roles (cf. 3.4 Team Structure), allowing for the infrastructure to develop its own dynamics and agility. With this team, we were able to establish the **ethico-legal framework** of GHGA [8], which currently entails: the contractual framework joining all data hubs and GHGA Central; templates for all contracts between GHGA, data controllers, and data requesters; the consent toolkit to formulate GHGA-compatible informed consent forms; extensive data protection documentation; an ISO 27001-prepared ISMS. In parallel with the development of the ethico-legal framework - and always closely aligned - we have developed an extensive **architecture and open-source implementation** of the data management infrastructure for GHGA Catalog and Archive (github.com/ghga-de). The modular and modern architecture ensures sustainability, extensibility, and scalability. While the most visible part is the GHGA data portal, there is a complex service-oriented infrastructure underneath that provides secure management of the metadata and research data (including key management), provides data stewards with a toolkit for data submission and provisioning, and enables the execution of analysis workflows on the data. The infrastructure supports a fully distributed storage of the data in federated data hubs. Our **outreach and training activities** were very successful and reached a large audience - GHGA has become a brand in Germany with wide recognition in the scientific community. While the training activities have not yet reached the extent we had envisioned (due to the delays mentioned above) we were still able to register more than 1,100 people for the GHGA Lecture Series and Webinars alone with more than 8,100 views online.

Challenges

We had to face significant **challenges** during the implementation: The initial **funding cuts** of around 27% of the proposed budget meant that we had to reduce our work programme. Primarily, the cuts led to a delay in implementing the core infrastructure and reduced the scope of later phases of the implementation, the SPE phase (referred to as GHGA cloud in the initial proposal), as well as the Atlas phase with its community-driven reference data sets. The delay in implementing the archive infrastructure obviously delayed data ingest and thus the analysis of these data sets, which is only now starting. An initial prototype of the SPE phase (restricted SPE - cf. A3) is still achievable in the current funding phase, but the large-scale cSPE and the Atlas phase will only be implemented in the upcoming funding phase. While we initially implemented even cuts to all members to maintain a collaborative atmosphere in the consortium, we still re-allocated underspent funds and flex funds very strategically to ensure that we could implement critical services that would otherwise not be funded. The implementation of a **reliable ethico-legal framework** also turned out to be more challenging than expected. In part, this was due to the changing legal landscape, including the challenges of negotiating data protection and IT security issues across all contributing entities of the federated structure. Obviously the **COVID-19 pandemic** severely impacted GHGA's productivity, primarily by impacting its ability to recruit additional staff and by redirecting (after approval by DFG) some of its resources for pandemic research. For example, GHGA was instrumental in setting up the SAR-CoV-2 Genomics Data Platform (CoGDat). While these were certainly worthwhile efforts, the resources were missing for GHGA's main mission and thus induced delays. A challenge, as well as an opportunity, was the introduction of the **MV GenomSeq** (cf. below), which was an unforeseen opportunity that GHGA had to take, but which still detracted from implementing the main mission. The increased visibility and responsibility of GHGA due to its role in MV GenomSeq also motivated the decision to implement an Information Security Management System compliant with (though not yet audited) ISO 27001. A related development was the launch of the **Genome Data Infrastructure Initiative (GDI)**, which required changes to our operational model and the data exchange mechanisms with European networks. GHGA had to develop robust solutions on how GHGA data can be exposed in multiple different metadata exchange networks, currently FEGA and GDI, and potentially the EHDS in the future.

4.1.3 Model Project Genome Sequencing (MV GenomSeq)

The major unexpected challenge and opportunity was the introduction of the Model Project Genome Sequencing (MV GenomSeq), an initiative by the German Genome Medicine community and the German Ministry of Health (BMG). Through a change in legislation and extensive negotiations with statutory health insurance, German patients with certain oncological indications, unsolved rare diseases, and hereditary tumour predisposition

syndrome have become eligible for genome sequencing (whole genome/WGS, whole exome/WES, and - for the first two years - panel sequencing). Health insurance companies have set aside a budget of 700 M€ for this project for the next five years (after which the project will be evaluated), which will result in 80,000 - 100,000 genomes being sequenced.

The preparation of this project started three years ago with the initiation of a project on the National Genome Strategy (genomDE). GHGA contributed significantly here: 5 out of 16 members of the national steering board of genomDE are also members of GHGA and dedicated significant efforts to the conception of the project. We were able to position GHGA not only as the national research data infrastructure holding the genome data for research access, but also our data hubs as genome data centres (GDCs). GDCs will receive additional federal funding to maintain the data infrastructure for the model project. Four of GHGA's data hubs have been unconditionally approved by BfArM (the federal institute tasked with legal responsibility for the data platform of MV GenomSeq), and two additional data hubs have been conditionally approved [23]. This official seal of approval and the additional funding for the data hubs illustrate the importance of GHGA in the German genomics community.

Significant efforts were required to ensure that MV GenomSeq built on the existing contractual concepts developed by GHGA, uses a metadata model compatible with the GHGA model, uses the existing infrastructure of the data hubs, and GHGA-compatible informed consent. GHGA was also tasked to ensure that MV GenomSeq data can in the future also be shared within the GDI network and 1+MG. Members of GHGA were also instrumental in ensuring the research compatibility as well as the interoperability (through HL7/FHIR) of the clinical data and technical metadata associated with the sequencing data. Our success here also illustrates that there is no alternative data infrastructure available in Germany that would compete with GHGA in this area.

We expect the 'first data in' from MV GenomSeq towards the end of 2024 and see it as one of the major drivers in genomic medicine in Germany in the coming years.

4.1.4 Further development and professionalisation of community RDM

Through our input for MV GenomSeq we could establish national standards that will have to be implemented at the sequencing centres of most German university hospitals (24 out of 38 university hospitals have - so far - been approved for participation). We intend to build upon this and suggest these standards also for research projects, by streamlining the data deposition process into GHGA. Currently, we perceive that the community is much more willing to share data for research through a national FEGA node than through EGA itself. With the availability of GHGA Archive, we now expect a significant increase in data deposition, which will certainly benefit the community as a whole. We would consider the number of data depositions, the completeness of the metadata, and the number of data use

requests to be the core KPIs toward assessing the success of the data management strategy. These KPIs - along with many others - are being monitored and reported.

Much of our outreach and training activities in the upcoming funding phase (B3) as well as our community activities - in particular the driver projects (B1) and data services (B2) will be our main routes through which we will provide the community with best practices for omics RDM. And vice versa, it will also provide us with direct feedback on how to prioritise our developments to meet the community needs.

4.2 Metadata standards

GHGA has developed a metadata model (MDM) to facilitate scientific communities to enrich the context of their submitted genomic data as well as to retrieve data of interest. The corresponding whitepaper [12] not only describes the basic and extended metadata modules, but also covers the importance of ontologies and terminologies traversing into the details of the different ontologies used to build our MDM. The model is based on the existing MDM of EGA and has been extended to meet our specific needs and discussed with the European partners (FEGA hubs). The selected ontologies and vocabularies are evaluated based on their quality and sustainability using fairsharing.org. The LinkML framework has been used to build the MDM, which is now available at the GHGA GitHub Repository [13].

To enhance the software integration of the MDM, GHGA has recently developed *schemapack*, a linked data modelling framework built on top of the industry standard JSON schema and will transition to express the MDM in *schemapack* instead of LinkML within the next six months (see B4.M3). Schemapack itself is open-source software and available through a GHGA GitHub repository[2]. Data harmonisation and standardisation facilitates cross-project analyses and is therefore seen as one of the pillars of continuous development of the MDM (see also B4.M1). The MDM is aligned with the relevant (inter)national standards, is interoperable with the metadata of (F)EGA, builds on GA4GH recommendations, and is aligned with the national standards for clinical data (HL7/FHIR and national core data set of the Medical Informatics Initiative). In the upcoming funding period, this alignment will be maintained and expanded to support interoperability with a wider range of infrastructures and support metadata from other additional omics modalities (see TA B4 for details). The GHGA data dictionary lists and describes all the aspects of the data model as well as the application domains. The most up-to-date data dictionary publicly available as part of the GHGA User Documentation. In addition, we have developed a quick submission guide that provides a user-friendly and easy walkthrough to help the data submitters to prepare, format, and submit their metadata to the GHGA portal. Within the FEGA and GDI projects, we are working on a convergence of the different MDMs in genomics. The

[2] https://github.com/ghga-de/schemapack

metadata team is and will also be involved in training and outreach activities.

4.3 Implementation of the FAIR principles and data quality assurance

One of the biggest hurdles towards **FAIR-ification of human omics data** in the past has been a reliable legal basis - many of the discussions with potential data submitters were centred around the frustration of not being able to deposit data due to uncertainty and fear related to data protection issues. By providing a clear legal basis and detailed data protection documentation that submitters can share with their respective data protection officers as well as the assurances that come from a national archive within national boundaries, we have resolved the biggest issue - the lack of submissions to public archives.

The FAIR principles [24] are furthermore the drivers for us developing common metadata standards, standardised workflows to increase interoperability and reusability of research data. In addition, as the FAIR principles are crucial to achieving a cultural change towards fairer and more equal conditions in biomedical research, we continuously work on highlighting this to our communities and provide information and training on FAIR, e.g. with webinars, and courses (in collaboration with e.g. EMBL and ELIXIR-DE). Crucially, GHGA builds on the architecture development of the GHGA service, to make research data findable and accessible.

Key **quality assurances** for the consortium include completeness and accuracy of metadata, but also specific metrics on the research data itself. The metadata white paper [12] sets out our references and internal standards for minimum viable and target specifications on metadata. In parallel to metadata standards, the consortium has, and will, continue to develop solutions to automatically score and quantify data quality, which includes dedicated measures on data quality and curation (B2.M1). While GHGA's focus is on the management of omics raw data, the value of the data in the archive depends on the extent to which the data can be linked and integrated with other data modalities, including clinical data. Assurance of accurate data linkage will be addressed in the context with our communities, including in the context of driver projects (B1) such as MV GenomSeq, but also uses cases on common disease and prevention.

4.3.1 GHGA Data Management Infrastructure

From the very start of the project, it was evident that neither the software artefacts, nor the metadata standards available from the FEGA at the time were suitable for the (nationally) federated infrastructure and heightened data privacy requirements GHGA was facing. While most European countries decided to implement centralised national genomics data infrastructures, the assessment of the Federal Office for Data Protection and Informational Freedom was that the national genomics data should not be stored in a single centralised location [7] and additional secure measures, such as two-factor authentication and identity

management were deemed essential. GHGA thus had to implement a different, federated data management architecture.

Together with the complex national legal requirements, the construction of this federated infrastructure was perceived as one of the key challenges of the project. Consequently, architectural development and legal development were parallelised, although there were frequent alignment steps necessary.

Besides the technical-legal alignment, GHGA architecture development had to make sure to integrate existing standards and tools, most notably those from the GA4GH, and be interoperable with the European developments (GDI, ELIXIR, FEGA). The architecture had to be modular and extensible in order to support a stepwise implementation according to the four development phases of GHGA (cf. Fig. 2). These design requirements led to a microservice-based architecture employing established cloud technologies. The overall architecture implements a hub-and-spoke model, where a number of central services, including the central metadata management, are hosted centrally (at DKFZ) and the genomic raw data is kept distributed at the data hubs (cf. Fig. 3). The typical data path starts at either one of the local sequencing centres or an external submitter who, after accepting the necessary data processing contract, provides the omics raw data as well as the corresponding metadata. GHGA Central assigns the data hub storing this data based on available capacities (external submissions) and/or proximity to the sequencing centres (local submissions). The data is initially stored in a staging area and relocated to the archive storage after quality control. Metadata is then forwarded to the central metadata store, from where it can be forwarded and/or exposed to the GHGA's European counterpart EGA and eventually also in GDI. External data users can find and access the data in a seamless fashion, without being exposed to the underlying federation model. After signing a data access agreement with the controller of the data, GHGA can make the data accessible for download or (in the SPE Phase of GHGA) stage it to local Trusted Research Environments (TRE) or Secure Processing Environments (SPE)[3] and provide access to the data user.

[3] Note that we will use the terms secure processing environment (SPE) and trusted research environment (TRE) interchangeably in this proposal - within a research data infrastructure the secure processing occurs always in a research context.

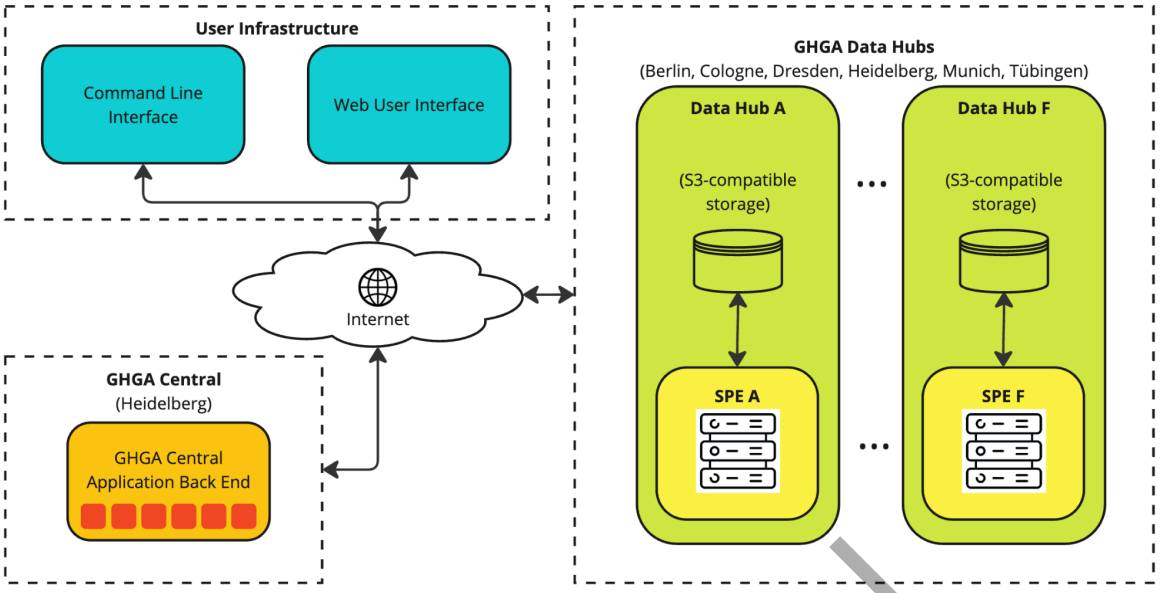

Fig. 3: High-level architectural view of the GHGA infrastructure. The GHGA Archive is accessible via command line and a web user interface. The connectivity and interaction to the federated data, distributed across GHGA data hubs, is transparently managed via the GHGA Central Application backend, operated at GHGA Central. Data exchange is managed via S3-compatible storage interface. Data can be either downloaded or accessed via one of several GHGA SPE instances.

Implementation

The GHGA software stack has been implemented as open-source software under a permissive Apache-2.0 licence. Artefacts are publicly accessible on the GHGA GitHub repository[4]. The complete software stack integrates existing tools (e.g., Crypt4GH [25]), adheres to international standards (in particular, those developed by GA4GH), and relies heavily on established open-source software components. The overall software architecture follows a few core foundational design principles: we employ an event-driven microservice architecture integrated into a single API. We adhere to a hexagonal architecture (ports and adapters architecture), which is particularly useful to separate concerns and helps to implement the zero-trust paradigm of GHGA as well as making the software infrastructure agnostic [26]. Standards of architecture and implementation are defined in a series of design documents, which GHGA also publishes once they reach the necessary maturity on a dedicated website. The microservice-based architecture is deployed centrally for production as well as for testing and development purposes in a Kubernetes cluster hosted on an OpenStack-based cloud environment providing robustness and scalability. One critical aspect of the secure management of research data is cryptography and key management. GHGA adheres to the GA4GH standards for cryptography and employs an industrial-strength standard solution for key management (HashiCorp Vault). Through the use of API-based key management (e.g., through the data stewardship toolkit) human

[4] https://github.com/ghga-de

intervention (and mistakes) in key management is minimised.

Operational Status

GHGA has been operational in the Catalog phase since May 2023 and the infrastructure demonstrates that the data managed within **GHGA Metadata Catalog** makes genomic data accessible and has been used in data usage requests. With the launch of **GHGA Archive** in August 2024, we now have a full implementation of a national genomics archive. The coming months will focus on polishing its components, improving the data management processes, and mobilising the data already at the data hubs (but not yet ingested). A detailed overview of the data sets is given in Table 4 below. We will also increase the number of data hubs connected step by step and we hope to have all of them fully operational by the end of 2024. By being the German node in the European FEGA infrastructure and the GDI project, we have now caught up to the genomic research data infrastructures elsewhere in Europe.

4.4 Services provided by the consortium

GHGA offers a comprehensive set of interconnected services, addressing specific needs of the communities ranging from the core services of Archive and Catalog to Training the communities (cf. Fig. 4). Certain services (such as SPEs and MV GenomSeq Genome Data Centres) are in preparation and will achieve readiness, at least partially, during the first funding period. The Atlas service for community reference data will only assume operation during the second funding period, due to the cuts and subsequent delays discussed above.

Service 1: Catalog - A National Catalog for Human Omics Data

The core service of GHGA is the development and operation of the GHGA Data Infrastructure for the secure and FAIR sharing of human omics data with controlled access mechanisms. This serves the need of the omics community in Germany to share and access human omics data for research and at the same time provides key connectivity and contributions to international omics networks such as FEGA and GDI.

The **GHGA Metadata Catalog**[5] was launched in 2023, addressing the needs to find and identify suitable human omics data for research. This service builds on key developments of GHGA, including the software architecture, the GHGA Metadata model [12] and the legal model [8]. The GHGA Metadata Catalog currently holds a total of 82 datasets with records from over 3,200 individuals (WGS, WES, transcriptomes) submitted by the GHGA data hubs, which can be requested for access via the corresponding Data Access Committees.

[5] catalog.ghga.de

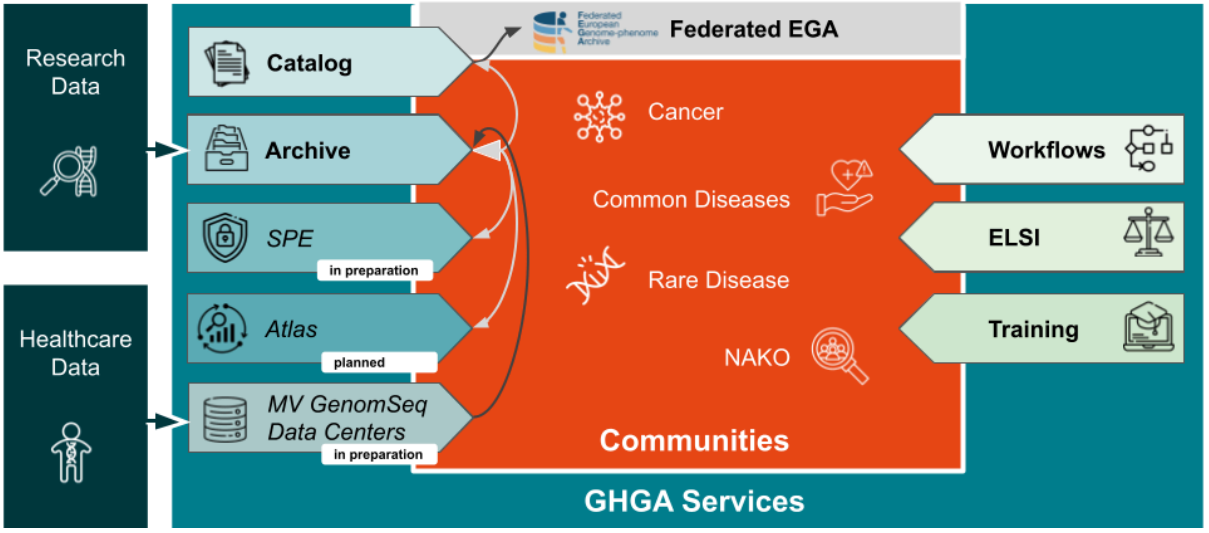

Fig. 4: Overview of the services that are currently operational (in bold), services that will assume (partial) operations within the current funding phase (marked as 'in preparation'), and services that will become operational only in the upcoming funding phase (marked as 'planned'). Data flow indicated with arrows with bidirectional links in black and unidirectional links in grey.

Service 2: Archive - Controlled-access human omics data archival

On August 1st, 2024, **GHGA Archive**[6] entered production and now offers the full range of data controlled-access data management (data ingest, annotation, archival, access management, and data provisioning, all based on GA4GH standards). Data ingest is now available via the Data Stewardship Toolkit (DSTK [27]), which facilitates the efficient ingest of data available in submission inboxes or locally at data hubs through a data steward. Our priority for the coming months is the ingest of the datasets committed to be deposited in GHGA (Table 4), which in part are already available at the data hubs, and preparing the data hubs for the upcoming submission to all data hubs for MV GenomSeq. Metadata from these datasets will be shared across the FEGA network to enable international findability of the data as well. Data requests are forwarded to the DAC of the data controllers associated with each project. After signing a data transfer agreement, the data controller approves the data access request and identifies the persons who have been permitted to access the data to GHGA data stewards, who then re-encrypt (Crypt4GH) and provision the data for download. The size of these committed data sets exceeds 50 PB, hence data transfer and full ingest of the data and metadata will be an ongoing process taking up to several months just for data transfer for individual studies.

[6] data.ghga.de

Table 4: Datasets committed to be deposited in GHGA Archive.

Dataset description and size	Partner Project	Data Hubs
An estimated 100,000 WES/WGS over five years (2024 - 2028) from the national model project 'genomic medicine' (§64e SGB V)	MV GenomSeq	All
200k imputed genotype panels as part of the current NAKO funding period (2026), 15k WGS from NAKO as part of a Helmholtz Sequencing effort (2025), 20k WGS from German National Cohort (NAKO) as part of the pilot project Genomes of Europe (GoE, 2026),	NAKO/GoE	Heidelberg, Munich, Tübingen
6,000+ samples (lcWGS/WES/RNA) datasets from 1,200+ patients with diverse childhood malignancies	INFORM	Heidelberg
Approximately 12,000 datasets (WGS/WES/RNA) from 5,000+ patients with rare cancers	DKFZ/NCT/DKTK MASTER	Heidelberg
500 WGS from two studies with rare diseases	Various	Tübingen
80 whole genomes (individuals) from hereditary breast cancer	MV GenomSeq (Pilot)	Cologne
Approximately 400 genomes (individuals) per year + 70 datasets with ca 24 individuals	Various	Dresden
100+ WGS/year CADS rare disease programme, 1,400+ NAMSE WES, 100+ multi-omics WGS/WGBS/RNAseq/SCseq TerminateNB/CRC1588, 200+ WGS limb cohort, BeyondTheExome, etc.	BIH-CADS, NAMSE, NCT-Berlin, IonGER, CRCs	Berlin
2,400 WGS samples and 300 RNA samples from the Bavarian Genomes project 1,200 WGS samples and 700 RNA samples from the DZHKomics resource	Bavarian Genomes and DZHK	Munich
1,000 WES samples from molecular diagnostics lab	Private lab	Munich
About 500 WGS based on 2-3 projects per year with WES and/or WGS data	Various	Kiel
About 1,000 whole genomes from COVID-19 patients	DeCOI	Cologne, Tübingen
Total number of individuals for which genomic data has a firm commitment for submission	**351,480**	

Service 3: Workflows - Standardised Community Workflows

GHGA aims to deliver harmonised and standardised community workflows for processing various types of omics data. Key developments in this area are (i) workflow development and maintenance, (ii) workflow benchmarking and evaluation, and (iii) the establishment of runtime configurations including associated reference files. GHGA has partnered with key communities and consortia (e.g., nf-core, NGS-CN, Solve-RD, DZHK and SATURN3). So far, six workflows for different data modalities have been released and entered continuous maintenance (cf. www.ghga.de/resources/data-analysis). Alignment with the nf-core framework ensures adherence to FAIR4RS best practices [28]. By integrating this continuous benchmarking tool with the existing CI/CD pipelines, the service provides a mechanism to continually assess the scientific utility and validity of workflows, providing a feedback loop for the community to assess workflow reliability. As part of this service, the Workflows Workstream has also contributed to the rare disease community by maintaining the Detection of RNA Outlier Pipeline (DROP [29]). The workflow development is aligned with, and prioritised according to, community needs; for example, the workflow team has

supported the Solve-RD's RNA working group in the analysis of RNA sequencing samples using DROP, conducted in multi-day hackathons to solve rare disease cases, now termed as Solvathons. In the next funding period, the development of workflows will be driven by community requirements for data services (B2), and in particular to underpin community driver projects (B1).

Service 4: ELSI - Ethico-legal tools for the communities

Besides other activities vital to the success of GHGA (e.g., patient involvement) and the legal model, the ELSI team has developed consent modules and recommendations as a service to data providers who would like to collect and share prospective data via GHGA. The consent modules can be used to update or create new consent forms to enable data sharing for scientific research use [6]. Our modules are not limited to GHGA and are compatible with the GA4GH Data Use Ontology (DUO), as such this is a service to the research community more broadly.

Service 5: Training - Training, Outreach, and Educational Activities

The training and educational activities across GHGA are bundled in a dedicated service that has the objective to support the users and associated communities on various levels, covering different aspects of data sharing in the field of biomedical research and health care. The service offers training in different formats (cf. training), including webinars, online as well as in-person courses, a lecture series, and additionally two podcasts. The latter specifically aim at educating the general public regarding genomics topics - all other formats are primarily aimed at our users and communities. For all activities, we draw from the expertise of internal and external experts in the respective areas. To address the needs of our communities we provide webinars on RDM and FAIR principles that are tailored to the omics research community, adding to existing - more general - training opportunities. Training materials on data protection and consent in a (bio)medical context, ranging from more general webinars for beginners to highly specialised courses for data stewards, uniquely fill a gap in the training space of researchers in this field in Germany. In addition, we provide training opportunities for more junior members of our communities on bioinformatics workflows as well as statistical and analytical methods. In addition to webinars, we have been involved in courses with collaborating institutions (e.g., EMBL) or hackathons for large projects (e.g., ELIXIR, de.NBI), and have presented GHGA by holding workshops at conferences. We offer a monthly GHGA Lecture streamed on the Internet, which sees leading international experts in the field of biomedical data sharing or similar (international) initiatives share their expertise and experience.

Service 6: SPE - Secure Processing Environments (in preparation)

The SPE Service (cf. TAs A3 and A4) will provide access to the data archived via GHGA without downloads through a secure processing environment. This will be established in two

steps. Initially, the data will be made accessible through limited compute resources at the data hubs to the GHGA data stewards only. This restricted SPE (rSPE) will not be accessible for external users, but will already support the quality control of the data, the processing required within MV GenomSeq, and the initial community analyses for the Atlas phase. This service will become available before the end of the current funding phase. The second step will be the community SPE (cSPE) that will provide increased compute capacities as well as additional checks and restrictions that permit its use by external researchers. It will increase the security of the data and its availability will initiate a transition away from downloads to an SPE-only access model. The rSPE service is expected to assume test operations in Q2/2025 and the cSPE service will follow in the upcoming funding period (A3/A4).

Service 7 - Genome Data Centres - Integration with MV GenomSeq (in preparation)

With the establishment of MV GenomSeq (cf. 4.1.3), a significant number of genome (WES/WGS) data sets are expected to become available for secondary use, as it is mandatory to ask the patients for their consent for research use. GHGA will coordinate the operations of the corresponding genome data centres, provide standardised data ingest pipelines for GHGA and thus make the data accessible as soon as it has been deposited with the genome data centres. Genome data centres are expected to assume their operations - subject to the timely conclusion of contractual negotiations of the GHGA data hubs and all University Hospitals with BfArM - in Q1/2025.

Service 8 - Atlas - Community reference data sets and derived data sets (planned)

The Atlas Service relies on the rSPE's availability to permit analyses across different data sets (always dependent on the data controller's permission). TAs B1 and B2 detail some of these services and their scientific use cases, mostly leveraging the availability of large-scale omics data in one central data infrastructure. The standardised workflows and harmonised metadata models of GHGA are critical to enabling these analyses. The aggregated - anonymised - data of relevance to the community (e.g., indication-specific variant databases) will support the communities and provide added value to the data submitters. This service will become available in the upcoming funding phase.

4.5 Impact of changes of external conditions/constraints

A series of developments in the data infrastructure landscape have impacted GHGA's development, providing important opportunities to further strengthen the positioning of GHGA at the interface between research and clinical care. Nationally, the formation of MV GenomSeq will boost data production with immediate relevance for GHGA users. GHGA is fully integrated with this initiative and seeks to mobilise these data for secondary research. In addition, GHGA has established close ties with NAKO, the German National Cohort, and will host NAKO's omics data. Internationally, GHGA has been mandated to develop and operate

the German node within the European Genomics Data Infrastructure - GDI project (cf. 3.3 International networking). Overall, these developments reflect the broad recognition of GHGA as a national data infrastructure and platform for human omics data, which in turn will open up further participation and contributions to upcoming developments, such as the European Health Data Space (EHDS).

With respect to constraints, it needs to be mentioned that GHGA, similar to other infrastructural projects, struggled to recruit talent in competitive areas, such as individuals with specialised expertise in software development, information security, or in the operation of large-scale IT infrastructures. The lack of long-term funding perspectives combined with the salary constraints of the public sector have made it difficult to attract individuals with these profiles. To compensate for these factors, we have broadened the search space for suitable personnel on the one hand but have also invested into training individuals to develop the necessary skills for specialised tasks.

GHGA has been significantly impacted by the COVID19 pandemic as some of its resources have been dedicated to managing SARS-CoV19-related data on the national level. For the sake of brevity, we will not reiterate these activities since they were detailed as part of the previous report and have little impact on the upcoming funding period.

5 Work Programme

The GHGA work programme is split into 11 task areas (TAs), which cluster into three blocks: TAs A1-A4 are concerned with the operations, maintenance, and further development of the core data management infrastructure. Task Areas B1-B5 are related to community interaction and aim to maximise the community impact of the infrastructure. Task Areas C1-C2 are concerned with project management. The following Table 5 gives an overview of the TAs described below and the responsible (co-)spokespersons.

Table 5: Overview of Task Areas, Measures and Responsible Co-Spokespersons

Task Area	Measures	Responsible Co-spokesperson(s)
Block A: Operations and Development		
5.1 A1. Operations - Central	A1.M1: Operation and Maintenance of Central GHGA Services A1.M2: Operation and Maintenance of the Central Infrastructure A1.M3: Information and IT Security A1.M4: Coordination of Operation of Data Hubs	I. Buchhalter
5.2 A2. Operations - Data Hubs	A2.M1: Data Hub Coordination A2.M2: Data Hub Operations and Maintenance A2.M3: Data Hub IT Security A2.M4: Implementation of MV GenomSeq Genome Data Centres	O. Kohlbacher, S. Wesner
5.3 A3. Architecture & Development	A3.M1: GHGA Archive Software Maintenance A3.M2: GHGA SPE Requirement Engineering and Development	O. Kohlbacher, O. Stegle

Task Area	Measures	Responsible Co-spokesperson(s)
	A3.M3: Requirement Engineering and Developments of Data Services and Interactive Tools	
5.4 A4. Data Stewardship - Central and Data Hubs	A4.M1: Central Data Stewardship Coordination A4.M2: Central User Support A4.M3: Federated Data Mobilisation and Local User Support A4.M4: Connection to Sequencing Centres	I. Buchhalter; S. Motameny; A. Dahl; D. Beule; O. Kohlbacher, J. Gagneur
Block B: Communities		
5.5 B1. Community Driver Projects	B1.M1: Cancer Genomics B1.M2: Rare Disease Genomics B1.M3: Common Disease & Prevention	H. Graessner; O. Stegle
5.6 B2. Community Data Services	B2.M1: Data Quality Control & Curation Tools B2.M2: Community Interfacing and Portals B2.M3: Integrated Data Processing Tools	J. Gagneur; D. Hübschmann
5.7 B3. Outreach & Training	B3.M1: Platform Communication B3.M2: Scientific and Clinical Outreach B3.M3: Patient and Public Communication B3.M4: User Training B3.M5: Assured Training Scheme B3.M6: User Experience	J. Winkelmann, O. Kohlbacher
5.8 B4. National and International Connectivity and Metadata Alignment	B4.M1: National Alignment within the NFDI and Beyond B4.M2: International Alignment B4.M3: Metadata Model Maintenance, Development and Alignment	J. Korbel; S. Nahnsen
5.9 B5. Legal and Ethical Issues	B5.M1: Legal Advice / EHDS and GHGA B5.M2: Legal Embedding in the National Infrastructure (GDNG) B5.M3: Integrated Ethics B5.M4: Patient Engagement	E. Winkler; F. Molnár-Gábor
Block C: Management		
5.10 C1. Flex Funds	C1.M1: Innovation & Implementation Projects (IIP) C1.M2: Data Mobilisation grants (DMG) C1.M3: Internal Flex Funds (IFF)	O. Stegle; O. Kohlbacher
5.11 C2. Project Management, Legal, Sustainability	C2.M1: Project Management and Governance C2.M2: Legal and Data Protection Affairs C2.M3: Finances, Human Resources and Reporting C2.M4: Sustainability and Strategic Development	O. Stegle; O. Kohlbacher

5.1 TA A1: Operations - Central

Overview of the Task Area

The central operations of GHGA are fundamental to managing core services that ensure the secure handling of human omics data. These operations include managing user accounts, operating the GHGA data portal, coordinating data upload/download services, and managing encryption keys. Central operations ensure the smooth, secure functioning of GHGA's services, supporting scientific data access and utilisation. Additionally, they navigate regulatory and technical challenges, ensuring compliance with stringent standards. Coordinating closely with infrastructure providers, central operations support service maintenance and efficiency, reinforcing GHGA's commitment to security, compliance, and continuous improvement.

5.1.1 Measure A1.M1: Operation and Maintenance of Central GHGA Services

Consortium Member	Contribution
Buchhalter (DKFZ)	Co-Spokesperson, Coordination of TA
Stegle (DKFZ)	Spokesperson, technical and strategic advice
Kirli (DKFZ)	Team Lead Technology / Product Management
Kraft (DKFZ)	Information Security Coordination

Goals: Ensure the secure, reliable, and efficient operation of central GHGA services and maintain high standards of data governance through continuous logging, monitoring, regular audits and implementation of modern software operation practices. The operation, maintenance, and improvement of central GHGA services (GHGA Central) are critical to ensuring that users can securely and efficiently access and manage research data. In this task we will ensure the **continued operation of the central infrastructure and services**. These services include the web portal of GHGA including the request management dashboard that facilitates efficient applications for user permissions, and the data steward dashboard which supports the data steward team. Centralised **management of user permissions and cryptographic keys** ensures data security and integrity, crucial for maintaining the trust of data providers and requesters. The file registry service is essential for cataloguing and securing all files within the system, ensuring they are properly managed, restricted as necessary, and accessible when needed. Operational and maintenance tasks involve **continuous logging and monitoring** to promptly identify and resolve any issues, ensuring uninterrupted service availability. We will continuously improve our internal policies, keeping them in line with external frameworks. Regular audits will be conducted to assess compliance with these policies and regulatory standards, maintaining high standards of data governance. **Incident management protocols** will be expanded and continuously reviewed and adapted to swiftly address and mitigate any operational disruptions, minimising downtime and impact on users. Keeping these services operational requires a dedicated focus on modern software operation practices, ensuring that GHGA remains at the forefront of technological advancements and can efficiently handle large volumes of data. To fulfil these critical tasks, **highly trained personnel** are needed as they directly impact the usability, reliability, and security of GHGA's central services. Their responsibilities include the seamless operation and continuous enhancement of the service infrastructure, which is foundational to GHGA's mission of providing secure, accessible, and high-quality omics data. Their diligent work ensures that the infrastructure can support the scientific community's growing needs, thereby significantly contributing to the project's overarching goals of advancing genomic research and fostering collaboration both nationally and internationally. Through professional operation and proactive improvement, this task area helps maintain the GHGA infrastructure as a safe place for genomics data.

5.1.2 Measure A1.M2: Operation and Maintenance of the Central Infrastructure

Consortium Member	Contribution
Buchhalter (DKFZ)	Co-Spokesperson, Coordination of TA
Kirli (DKFZ)	Team Lead Technology / Product Management
Kraft (DKFZ)	Information Security Coordination

Goals: Collaborate closely with our dedicated infrastructure providers to ensure the reliable performance and scalability of our technical backbone, including hardware components, virtualisation, containerised applications, and key/secret management systems. Enhance the capacity of the infrastructure provider to offer essential compute and storage resources, network monitoring, and data backup services, ensuring robust and uninterrupted service delivery. The **operation and maintenance support for compute and storage infrastructure** are crucial for ensuring the reliable performance and scalability of GHGA's technical backbone. The software stack, maintained and operated in A1.M1, relies on hardware such as servers, compute resources, storage, and networking infrastructure as well as modern virtualisation and containerized application management. The infrastructure provided by DKFZ's ITCF and de.NBI/ELIXIR-DE in Heidelberg allows GHGA Central to run the software stack required to deploy the GHGA services. Currently GHGA uses IBM COS S3 technology provided by the DKFZ's ITCF and OpenStack/Kubernetes service provided by de.NBI/ELIXIR-DE. These services provide the virtualisation layers needed for deploying containerised applications and enable optimised scaling and secure networking. **Backup and disaster recovery:** Ensuring robust, and continuous operation is a priority for GHGA to fulfil its mission as a national data infrastructure. A critical component is backup and disaster recovery for which GHGA has developed detailed plans for GHGA Central (as well as for the data hubs), which need to be adapted to the constantly evolving infrastructure and changing technical infrastructure. They also need to be tested on a regular basis and aligned with the efforts of the data hubs (A2). These efforts ensure robust and uninterrupted service delivery, enabling GHGA to continue offering high-quality, secure, and scalable services to the scientific community. Disaster recovery will be tested regularly on the test infrastructure and, during scheduled maintenance periods, also on the production infrastructure.

5.1.3 Measure A1.M3: Information and IT Security

Consortium Member	Contribution
Buchhalter (DKFZ)	Co-Spokesperson, Coordination of TA
Marnau / Fritz (CISPA)	Advanced information security approaches
Kirli (DKFZ)	Team Lead Technology / Product Management
Kraft (DKFZ)	Information Security Coordinator

Goals: Establish, maintain, and continuously improve an ISO 27001-compliant Information Security Management System (ISMS). Ensuring the security of GHGA's infrastructure is paramount. This is achieved through a comprehensive **Information Security Management System** (ISMS) compliant with the ISO 27001 standard. The ISMS

of GHGA Central provides a structured approach to meeting the applications of managing sensitive data, ensuring its confidentiality, integrity, and availability. This system underpins all security measures and practices within GHGA Central, creating a robust foundation for protecting human omics data and connects to associated ISMS (e.g., the ISO 27001 certified ISMS of de.NBI/ELIXIR-DE Cloud or the ISMS of other data hubs). The current ISMS encompasses all central infrastructures and interfaces to external infrastructures as well as relevant SOPs. While it has been implemented in an ISO 27001-compliant way, it still requires a formal external audit, which will be conducted by external auditors in the first year of the upcoming funding phase. The ISMS will also require continuing monitoring, updating, and review. The ISMS defines interfaces with service providers (cf. A1.M2) and data hubs (A2.M3). Common policies, developed together with the OCB, will support the implementation of harmonised ISMS standards across GHGA Central and all data hubs. **Regular security audits** are a critical component of our security strategy. These include penetration testing to identify and address potential vulnerabilities, as well as supply chain attack assessments to secure our entire operational ecosystem. Vulnerability analysis is conducted to proactively detect and mitigate security threats. Logging and monitoring activities are thoroughly implemented to ensure any anomalies or incidents are promptly detected and addressed, maintaining the integrity of the GHGA infrastructure. In addition to these measures, coordination across GHGA Central and the data hubs is essential for a unified security posture. GHGA Central will participate in the internal security audits coordinated by the OCB (cf. A2.M3). By maintaining stringent security protocols and continuously improving our security practices, GHGA ensures the highest standards of information security and thereby data protection. This commitment not only safeguards the data but also reinforces trust with data providers and users, ensuring the secure and efficient handling of human omics data and a willingness from data controllers to deposit data with GHGA. Fulfilling these diverse tasks will require the expertise of both GHGA staff as well as the support of external providers. Therefore, we will further train and sensitise our staff regarding information security and additionally collaborate closely with companies supporting us with penetration testing and monitoring (see B3.M4).

5.1.4 Measure A1.M4: Coordination of Operation of Data Hubs

Consortium Member	Contribution
Buchhalter (DKFZ)	Co-Spokesperson, Coordination of TA
Kirli (DKFZ)	Team Lead Technology / Product Management
Kraft (DKFZ)	Information Security Coordinator

Goals: Ensure effective communication and collaboration between central operations and GHGA data hubs, develop technical standards and requirements. GHGA Central's operation plays an important role in the GHGA network, and GHGA data hubs depend on this interaction for providing GHGA services. GHGA Central and the data hubs need to

coordinate during operation on updates to the software, and the assessment of resources for incoming data. This task involves managing and synchronising central tasks with data hubs to ensure a unified approach to data handling and service delivery. Regular coordination within the OCB (cf. A2.M1) ensures that all hubs are aligned with central policies and procedures, fostering a collaborative approach to define GHGA-wide policies, procedures and concepts that are co-developed together with the data hubs. Legal obligations, service levels and the division of roles and responsibilities are formalised in the Central-to-Data Hub Bilateral Contracts (cf. [8]). Communication and collaboration between central operations and data hubs are essential for maintaining a cohesive and integrated system. This unified approach helps in addressing any operational challenges swiftly and efficiently.

Tasks and Deliverables

Tasks	Deliverables	Due Date
A1.M1.T1 Ensuring operation, maintenance, and improvement of Central GHGA Services	Core software stack update released	M12, M24, …, M60
A1.M2.T1 Collaborating with infrastructure providers	Regular meetings and updates	M12, M24, …, M60
A1.M2.T2 Enhancing the capacity of the infrastructure providers	Adequate compute and storage capacity for the project goals	M48
A1.M3.T1 Establishing a certified Information Security Management System	ISO 27001 certification of central ISMS	M18
A1.M3.T2 Security auditing and pentesting	Penetration testing concluded	M12, M24, M36, M48
A1.M4.T1 Coordinating the operation across the GHGA data hub network	Integration of new data hub (MHH)	M24

Dependencies, Interactions, Risks, and Mitigation Strategies

Dependencies and Interactions: Effective coordination with other task areas is crucial for GHGA's central operations. Dependencies include A2 for secure and efficient data transfer and alignment of security measures, A3 for supporting software development with necessary environments and resources, A4 for collaboration with data stewardship teams to ensure continuous service, and B1 and B2 which rely on robust central resources for bioinformatics workflows. Regular coordination meetings and integrated planning sessions are essential for maintaining alignment and addressing issues promptly. Shared documentation and tools facilitate easy access to information and collaborative work.

Risks and mitigation strategies: A1.M1: Risks include problems with the central software and load balancing. To mitigate this risk an efficient communication with A3 will be established to fix upcoming issues timely. A1.M2: The associated risks are general availability of services and infrastructure as well as skilled personnel. To minimise these risks, we will work closely with external professionals and perform for example regular data recovery tests. In A1.M3, risks include vulnerabilities, attacks, and a lack of compliance with policies. To mitigate these, we will regularly train our personnel, continuously update our documentation and risk management plans as well as incorporate results from penetration tests and security audits. The predominant risks in A1.M4 are a lack of, or inefficient,

communication with the data hubs. Therefore, we will make sure that the respective boards meet regularly and that well-structured minutes and concrete action plans are distributed.

Justification of Requested Funds

As outlined in 7.1, a total of five positions is requested for this TA. These include two senior positions for the "Team Lead Technology / Product Management" and "Team Lead Operations" (cf. Team Structure) at DKFZ, as well as two DevOps engineers at DKFZ, also covering the information security topics, and one position based at EKUT. In addition, 400 k€ are needed for outsourcing efforts to ensure sustainable and secure operations. Funds will be supplemented by DKFZ's own contribution financing an additional DevOps Engineer as well as 500 k€ for additional support by third-party.

5.2 TA A2: Operations - Data Hubs

Overview of the Task Area

This task area is concerned with the secure and resilient operation of the federated data hubs physically storing the research data and personal metadata within GHGA. The data hubs provide and further enhance the secure and reliable services developed in the previous phase of GHGA for accessing storage and (in the GHGA SPE phase, cf. 4.1) compute capacity for SPEs. The data hubs are connected to GHGA Central using a well-defined, standardised, and secure object storage interface. A2 is split into three measures: A2.M1 deals with coordination, evolution of the services, load-balancing across the hubs, and the definition of reporting templates and key performance indicators (KPI) for the data services. A2.M2 aligns the day-to-day operations and maintenance of the data hub infrastructure across sites. A2.M3 coordinates the activities maintaining, adapting, and evaluating IT security measures and ensures equivalent service levels across all data hubs. The data hubs, as well as GHGA Central, are reporting their activities towards the GHGA Data Hub Operations Consortium Board (OCB, cf. 3.4.1).

5.2.1 Measure A2.M1: Data Hub Coordination

Consortium Member	Contribution
Wesner (UzK)	Co-spokesperson TA A2, coordination of the TA
Kohlbacher (EKUT)	Co-spokesperson TA A2, coordination of the TA

Goal: Coordinate all data hubs, ensure joint standards for operations, balance the data hub capacities. **Coordination and Communication:** The OCB (cf. 3.4.1) will coordinate activities through regular meetings and communication channels among the data hubs, in close alignment with GHGA Central (cf. A1.M4). This includes monthly coordination meetings to discuss ongoing operations, load balancing issues, emerging challenges, and strategic planning for the further evolution of the services. The OCB will be supported by C2 and will build on the existing central GHGA communication platform and documentation resources. The OCB defines constraints for operational procedures to ensure security, consistency, and efficiency of services across all hubs. **Capacity Balancing:** The OCB will

ensure effective load balancing across the six federated data hubs within the GHGA Data Infrastructure. This involves the dynamic distribution of data storage and computational tasks to optimise resource utilisation, enhance performance, and maintain system resilience. Data hub resources are allocated based on the capacity reports and a technical assessment of the best-suited hubs for large scale data submission. Policies proven in practical context will be encoded into software for increased automation (A3.M1). Based on proposals by the Lead Data Steward (cf. A4) the OCB assigns data submitters to data hubs (A4.M1). As part of this measure, we will also develop concepts for geo-redundant replication of the data across data hubs. This replication will increase resilience and support load balancing, in particular during the GHGA SPE phase. **Reporting and Documentation:** The OCB will rely for its assessments on the comprehensive reporting system defined within this measure. This will include detailed logs of data load distributions, data access patterns, performance metrics, and incident reports. Quarterly and annual reports will be compiled to summarise the operational status and data capacity and will be reported to the GHGA BoD and SC. Documentation will be integrated into the central GHGA documentation and maintained there. This will not only ensure compliance with regulatory requirements, but also facilitate knowledge transfer and training for new staff or data hubs.

5.2.2 Measure A2.M2: Data Hub Operations and Maintenance

Consortium Member	Contribution
Krüger/Walter (EKUT)	Operations of the data hub Tübingen
Achter (UzK)	Operations of the data hub Köln
Mertes (TUMUH)	Operations of the data hub Munich
Buchhalter (DKFZ)	Operations of the data hub Heidelberg
Beule / Häcker (MDC)	Operations of the data hub Berlin
Müller-Pfefferkorn / Nagel (TUD)	Operations of the data hub Dresden
Di Donato (MHH, participant)	Stepwise establishment of an additional data hub at MHH

Goals: Ensure a secure continuous operation of storage and compute infrastructure across all data hubs. Operations, maintenance, and monitoring: Local DevOps staff will implement, maintain and continuously improve the secure and performant operation of their local data hub infrastructure. This includes routine maintenance tasks, such as software updates, hardware checks, and system optimisations. Continuous monitoring tools will be implemented to track the performance, health, and security of the data hubs, allowing for the early detection and resolution of potential issues along the reporting metrics defined in A2.M1. Regular maintenance schedules and emergency protocols will be established to minimise downtime and ensure uninterrupted service. Each data hub will provide secure and performant object storage services (currently based on the S3 protocol), along the agreed service definitions connecting all hubs to GHGA Central. **Backup and Data Life Cycle Support:** In alignment with the Central-to-Data Hub Bilateral Contract, and the GHGA ToUs, each data hub will implement policies and technical-organisational measures to meet its

obligations, ensuring that data is appropriately categorised, stored, and transitioned through its various stages. Geographically redundant storage systems and off-site backup strategies are used to enhance data protection and disaster recovery capabilities. Automated policy driven workflows will be established and monitored to handle regular tiered backup, long-term archival, and archive retrieval. **Operations of the rSPE:** Existing, local compute infrastructure at each data hub, consisting of dedicated hardened compute services using isolation approaches such as virtual machines (VMs), are deployed for GHGA processing tasks such as data encryption, re-encryption, download, and quality control. This restricted SPE (rSPE) infrastructure needs to be maintained and updated to provide a well-defined and standardised services supporting secure, allowing controlled access to sensitive research data by data stewards, for example to execute community data services (B2). **Operation of the cSPE:** The Data Hub Operations team will support the evaluation and operations of solutions to establish the community SPE (cSPE, cf. A3.M2), which will allow for interactive secure access to research data by third party users. This will involve setting up secure network links, access controls, provenance monitoring and security incident and event management systems to maintain the integrity and confidentiality of the data. Depending on the outcomes of the concept development (cf. A3.M2), the operations team will also consider connections to external dedicated TRE solutions, for example by establishing an additional data hub operated on commercial cloud solutions, thereby establishing the basis to use existing commercial solutions such as DNAnexus. Access control rights will be implemented based on emerging standards in GA4GH, FEGA, GDI as well as taking the development of the EHDS into account.

5.2.3 Measure A2.M3: Data Hub IT Security

Consortium Member	Contribution
Krüger/Walter (EKUT)	IT security coordination data hub Tübingen
Achter (UzK)	IT security coordination data hub Cologne
Mertes (TUMUH)	IT security coordination data hub Munich
Buchhalter (DKFZ)	IT security coordination data hub Heidelberg
Beule /Häcker (BIH)	IT security coordination data hub Berlin
Müller-Pfefferkorn / Nagel (TUD)	IT security coordination data hub Dresden
Di Donato (MHH, participant)	Stepwise establishment IT security coordination at MHH

Goals: Maintenance, harmonisation, and further development of IT security across all data hubs and with GHGA Central. Coordinated Implementation of IT Security Measures and ISMS: This measure will ensure the continued harmonisation and standardisation of IT security policies, protocols, and procedures across the data hubs and their alignment with the central IT security measures (cf. A1.M3). All of these measures are governed by an ISMS, which formally specifies the systematic approach to IT security at the data hub. Similar to the measures in A1, the data hubs will need to harmonise their existing ISMS to ensure the standards as specified in the Central-to-Data Hub Bilateral Contract with

GHGA Central, as well as requirements from ISO 27001 as a common framework (some data hubs are already operating ISO 27001-certified infrastructure) are met. Once these ISMS have reached sufficient maturity, they will be audited externally to be certified according to ISO 27001 (or BSI Grundschutz). The OCB will supervise the implementation of these measures, ensuring that each site maintains an equivalent level of security and might, if necessary, suspend a hub compromising overall security levels. To support the OCB, a dedicated IT Security Taskforce consisting of representatives from data hubs and GHGA Central has been established. This taskforce will review current security practices, share insights, evaluate emerging threat intelligence and risks, and report these findings along with recommended measures to the OCB. **Cross-Site Friendly Audits:** To promote continuous improvement and adherence to security standards, cross-site friendly audits by local security experts (e.g., SecOps, CERT or infrastructure experts) will be conducted regularly. These audits will have teams from different data hubs auditing each other's sites in a collaborative and constructive manner to identify vulnerabilities, share best practices, and foster a culture of mutual support and transparency. Findings from these audits will be documented and used to enhance the overall security posture of the GHGA data hubs. **Security Audits:** In addition to friendly audits, formal security audits will be conducted at each data hub site. These audits will be carried out by external security experts to provide an objective assessment of the security infrastructure and practices. The audits will cover all aspects of IT security, including access controls, data encryption, network security, and incident response procedures. Detailed audit reports will be generated, highlighting areas of compliance and recommendations for improvement. External Audits will be transitioned to ISO 27001 audits after accreditation. **Pen Testing:** Regular penetration testing (pen testing) will be conducted by external contractors who will attempt to breach the systems using advanced techniques and tools. The findings from these tests will provide valuable insights into the effectiveness of existing security measures and highlight areas that require strengthening.

5.2.4 Measure A2.M4: Implementation of MV GenomSeq Genome Data Centres

Consortium Member	Contribution
Krüger/Walter (EKUT)	MV GenomSeq operation hub Tübingen
Achter (UzK)	MV GenomSeq operation hub Cologne
Mertes (TUMUH)	MV GenomSeq operation hub Munich
Buchhalter (DKFZ)	MV GenomSeq operation hub Heidelberg
Beule / Häcker (BIH/MDC)	MV GenomSeq operation hub Berlin
Müller-Pfefferkorn / Nagel (TUD)	MV GenomSeq operation hub Dresden
Malek (UKT)	Connection to Clinical Data Nodes in MV GenomSeq

Goals: Implement services of MV GenomSeq Genome Data Centres (GDC) at GHGA data hubs. *Note: This measure is listed for completeness' sake - it will be funded externally by BfArM, and no DFG funding has been requested for this measure.*

Definition of GDC services: BfArM has approved the GHGA data hubs to operate genome data centres within the Model Project Genome Sequencing [23]. The data hubs will jointly negotiate the specific services to be provided to the BfArM as part of their dual role as GDC. The legal and organisational structure that is established within GHGA will be mirrored as closely as possible, with additional bilateral agreements between the GDC and the BfArM as platform operators governing individual service offerings. It is expected that the GDC services will be reimbursed by the platform operator for the additional services provided, thus allowing to implement and operate GDCs without additional cost to GHGA. **Operation of GDC services:** The data hubs will implement the GDC services locally. These are foreseen to include encrypted data transfer and ingestion from data-generating university medical centres, metadata management, sequence data quality control, and the generation of billing signals to trigger reimbursement. Dedicated GDC operations staff and data stewards will work side by side with the local GHGA team. **Ingest of research-consented data into GHGA:** Data from MV GenomSeq that has been consented for research will be periodically ingested into GHGA to make them accessible for secondary research. The alignment between the GDC and GHGA data hubs means this can be implemented as a simple metadata transformation without any need to copy data, potentially retaining encryption keys. **Making MV GenomSeq data available:** Research data that are part of GHGA will be shared and made available alongside other GHGA data with the BfArM acting as the data controller.

Tasks and Deliverables

Task	Deliverables	Due Date
A2.M1.T1 Establishing the governance model for second funding period	First GHGA2 OCB meeting held	M3
A2.M1.T2 Balancing of data hub capacities	SOP for capacity balancing drafted by OCB and approved by BoD	M12
A2.M2.T1 Establishing a GHGA Data Storage and Compute Service Interface	cSPE operational at two data hubs	M24
A2.M2.T2 Creating GHGA data hub utilisation report	First report of defined metrics to the OCB	M12
A2.M3.T1 Accomplishing certification	ISMS at all data hubs certified	M48
A2.M3.T2 Conducting IT security audits of data hubs	Accomplished security audits alternating between friendly and regular audits	M24, M36, M48, M60
A2.M3.T3 Penetration testing on data hubs	Pen testing concluded and reported	M24, M48
A2.M4.T1 Ingest of research-consented MV GenomSeq data into GHGA	First MV GenomSeq data release accessible via GHGA	M12
A2.M4.T2 Make research-consented MV GenomSeq data accessible via GHGA SPE	MV GenomSeq data accessible by SPE	M36

Dependencies, Interactions, Risks, and Mitigation Strategies

Dependencies and interactions: A2 will closely cooperate with its sister A1 to ensure close interaction between central and data hub operations. Similarly, there will be tight interactions with A4 as the operations of the local data hubs will be essential for the central and decentral data steward teams. As the key focus of this task area is the provision of reliable and

trustworthy data and compute services the major dependency is on the evolving requirements from the user community and the use cases. Furthermore, changing regulations and also technology evolution might make adaptations to the services mandatory in order to ensure continuous service levels. **Risks and mitigation strategies:** A2.M2: Risks include operational issues, maintenance challenges, and infrastructure establishment. The risk management plan includes regular risk assessments, contingency plans, and protocols for rapid response. Continuous monitoring will address potential challenges, ensuring stability and reliability. Dedicated SOPs and runbooks will minimise operational risks. A2.M3: IT security risks include vulnerabilities, security breaches, and compliance issues. Mitigation involves a comprehensive risk management plan with regular assessments, contingency planning, and protocols for rapid response to incidents as well as early certification of the ISMS. Regular pen testing and audits will ensure ongoing protection of sensitive data.

Justification of Requested Funds

As outlined in 7.2, for each of the data hubs, one position for operations at each of the current data hubs (DKFZ, EKUT, MDC, TUMUH, TUD, UzK) is requested for A2.M2 and A2.M3. MHH will build up its new data hub based on own contributions. For coordination tasks (A2.M1), EKUT and UzK apply for ½ position. In addition, each data hub needs 50 k€ for external consulting (e.g., for penetration tests, audits (A2.M3)). Activities of the data hubs are supported with significant own contributions by the individual institutions, cf. Description and Summary of Contributions by (Co-) Applicants for further details.

5.3 TA A3: Architecture & Development

Overview of the Task Area

This task area addresses the software development and maintenance required for the continuous operations of the existing GHGA Archive, subsequent feature extension (including a tighter integration into FEGA), as well as new developments required for the subsequent phases (GHGA SPE and Atlas). Requirement engineering and software development fall into the responsibility of this task area, together with the stakeholders involved. A3 is split into three measures: A3.M1 will deal with overarching software development and maintenance tasks to robustly operate the Archive as core services. A3.M2 will establish technical requirements and software solutions to operate a federated secure execution environment and take conceptual steps towards providing a fully featured SPE. A3.M3 will establish a portfolio of data services that can be executed by data controllers and users to facilitate exploiting and making use of datasets in GHGA.

5.3.1 Measure A3.M1: GHGA Archive Software Maintenance

Consortium Member	Contribution
Kohlbacher (EKUT)	Co-spokesperson TA A3, coordination of the TA
Stegle (DKFZ)	Co-spokesperson TA A3, coordination of the TA
Kuchenbecker (EKUT)	Team Lead Architecture
Kirli (DKFZ)	Team Lead Technology / Product Management
Kraft (DKFZ)	Information Security Coordinator

Goal: Continuous improvements and maintenance of the GHGA Archive software stack. **Maintenance of GHGA Archive:** Building on the operational GHGA Archive, this measure will deliver continued development and maintenance to ensure a secure and user-centric operation. Areas that are particularly in flux and require continuous effort include the development and adaptation of interfaces to integrate GHGA with FEGA and the GDI network. Maintenance tasks also include the development of software tools to facilitate manual and increasingly automated load balancing solutions (cf. A2). **New tools for data stewardship:** New tools and functionality as required will be developed. A significant area of development are refinements to the data stewardship toolkit, to simplify and streamline data ingest and data access. The toolkit will be adapted and extended to comply with the requirements of the MV GenomSeq, including support for automated quality control, and necessary logistics to comply with the changes to key management once the external trust centre (to be operated by RKI) is operational. **Refinement of the GHGA interface:** A third area of development are advances of the GHGA portal, based on user feedback together with B3 (cf. B3.M6). Maintaining a high level of usability and accessibility of the portal will be a priority. **Documentation & IT security**: Fourth, A3.M1 will coordinate the documentation of software products as required for IT security, e.g. implementing recommendations from audits and pen tests (cf. A1.M3 & A2.M3). All IT components will be documented in the GHGA Internal Documentation.

5.3.2 Measure A3.M2: GHGA SPE Requirement Engineering and Development

Consortium Member	Contribution
Kohlbacher (EKUT)	Co-spokesperson TA A3, coordination of the TA
Stegle (DKFZ)	Co-spokesperson TA A3, coordination of the TA
Marnau / Fritz (CISPA)	Advanced information security approaches
Kuchenbecker (EKUT)	Team Lead Architecture
Kirli (DKFZ)	Team Lead Technology / Product Management
Kraft (DKFZ)	Information Security Coordinator
Parker (DKFZ)	Team Lead Data Protection and Legal

Goal: Development of a federated, community Secure Processing Environment (cSPE) allowing the execution of GHGA data services across all data hubs. **Concept development:** While we will establish a prototype of a restricted SPE (rSPE) that is only available for internal use in the consortium (data stewards) already in the first funding phase, additional steps are required for a full-fledged community SPE (cSPE) that is secure and scalable enough to permit large-scale analyses by external researchers. We will assess existing solutions and systems, focusing on core workflows developed in B2. Solutions will

be synchronised with operational requirements (e.g., to guarantee I/O throughput), but also appropriate replication of datasets across hubs where compute is available. Suitable technical solutions will be developed, ideally based on an existing federated analysis platform (e.g., based on the FLAME platform developed as a national federated omics data analysis platform within the MII). Rollout will occur in a stepwise fashion, initially focusing on data hubs with de.NBI/ELIXIR-DE Cloud sites and existing kubernetes deployments (Tübingen, Heidelberg). **Definition of the technical requirements for a GHGA cSPE:** We will determine the technical requirements for establishing a cSPE within GHGA. Appropriate security measures will be determined, building on standards for SPEs that are being developed within FEGA, GDI, and EOSC. We will consider both commercial solutions (e.g., DNAnexus, Edgeless Systems, StackIT) and open-source SPE implementations (e.g., GWDG TRE) and test the various platforms. **cSPE implementation:** Selected candidate solutions for a GHGA cSPE will be tested and integrated into the GHGA architecture, including necessary changes to metadata, microservices, and the data portal for seamless access. **Business model and long-term support:** Building on the experience from the pilot phase, we will develop concepts for sustainable operations, including business model, documentation, SOP, and training (cf. C2.M4).

5.3.3 Measure A3.M3: Requirement Engineering and Developments of Data Services and Interactive Tools

Consortium Member	Contribution
Kohlbacher (EKUT)	Co-spokesperson TA A3, coordination of the TA
Stegle (DKFZ)	Co-spokesperson TA A3, coordination of the TA
Kuchenbecker (EKUT)	Team Lead Architecture
Kirli (DKFZ)	Team Lead Technology / Product Management
Kraft (DKFZ)	Information Security Coordinator
Parker (DKFZ)	Team Lead Data Protection and Legal
Behrens (TUM)	Team Lead Community Engagement

Goals: Contribute to the identification of the most suitable driver projects, prioritising data services, and piloting their application in the context of community-driven use cases in collaboration with B1. Identify driver projects for GHGA Atlas: Considerations for driver projects include data availability (including consent), community size, leverage to attract third party funding, and alignment with community-driven requirements. An initial set of driver projects include the national cohort NAKO and the model project genome sequencing (MV GenomSeq). **Define the requirements and priorities for GHGA data services:** Together with our target communities, and based on the selection of driver projects (B1), we will define the requirements for Atlas and prioritise the development of data services (B2), covering low-level tasks such as quality control, annotation, but also derived data generation and services for federated query systems (e.g., Beacon). **User-facing data management tools** will be essential to enable our users to interact with the platform effectively. Similar to the Data Stewardship Toolkit [27], we will create a GHGA User Toolkit,

which includes tools to readily submit, download, and check data as well as associated metadata. These tools will be Python-based to ensure they are accessible to a large community of bioinformaticians and life scientists alike. In those areas where there is overlap with existing FEGA tools (e.g., pyega3), compatibility and reuse of existing components will be ensured. Its scripting capabilities will also enable seamless integration with other infrastructures and its library character will simplify metadata transformations. **Implement data services and an execution environment:** We will implement the necessary orchestration infrastructure to execute data services within the secure execution environment (cf. A3.M2) across GHGA data hubs. Implementations will be synchronised with initiatives on a European level, most notably GDI where relevant software solutions (e.g., Beacon) are being developed. After reaching production status with GHGA Archive, the team will have capacity to more proactively engage with GDI and contribute additional software products. **Platform integration and production:** We will integrate the services into the platform, making them accessible to data controllers and (where applicable) users. Solutions will be documented and tested. Training / SOPs will be incorporated into the GHGA service portfolio.

Tasks and Deliverables

Task	Deliverables	Due Date
A3.M1.T1 Releasing of annual GHGA software updates	Update sent to all users	M12, M24, …, M60
A3.M1.T2 Improving and maintaining of data stewardship tools	First major update of data stewardship tools	M24
A3.M1.T3 Refining the GHGA interface	Updated GHGA interface based on user feedback	M24, M48
A3.M1.T4 Improving the security of the infrastructure based on audits and pen testing	Implementation of feedback from audits and pen testing	M26, M38, M50, M60
A3.M2.T1 Drafting a concept for a cSPE	Published Whitepaper on a GHGA cSPE	M12
A3.M2.T2 Implementing technical requirements for a cSPE	Launch of the GHGA cSPE	M36
A3.M2.T3 Delivering technical documentation for the development of a sustainable business model	First Task Force Meeting attended (cf. C2.M4)	M6
A3.M3.T1 Defining of requirements and complete concept for Research Analysis Platform in GHGA	Communication of developed concepts via newsletter and at Annual Meeting	M24
A3.M3.T2 Implementing execution environment	Beta testing of new services concluded	M40
A3.M3.T3 Integrating services into the platform	First community data service integrated and in production	M48

Dependencies, Interactions, Risks, and Mitigation Strategies

Dependencies and Interactions: A3 receives requirements from communities, in particular B1 & B2 where the requirements for community driver projects and data services will be defined, but also B5 which connects GHGA to international networks and standards. It also receives requirements from the data steward team (A4) who pass on feedback and suggest

improvements to the GHGA toolset from the users' perspective. A3 will deliver software solutions to accommodate these requirements in the core infrastructure, which are deployed by A1 & A2. **Risks and Mitigations Strategies:** The most important risks are related to unclear requirements from the communities. Within GHGA we will ensure regular coordination meetings to ensure efficient information flow. The early definition of driver projects (B1) and community data services (B2) will ensure that exemplary data services can be defined early, in fact several key directions are defined at the time of writing. More complex interactions are expected with external stakeholders, most importantly international standards to ensure interoperability of the GHGA software stack with FEGA or GDI. Both networks are in flux, with constant changes of metadata standards, and operational models. The key mitigation strategy is the decision to develop GHGA as self-sustained national services as outlined in 4, thereby ensuring a substantial level of independence and an integration model that is focused on the exchange of metadata. A second risk are employment related delays and technical challenges. In the previous funding period, we have developed effect strategies, including overarching working groups and a close interaction between development and operations. We will also ensure a high degree of alignment between the development environment and release so that a timely and regular release cycle (at least annually) can be achieved.

Justification of Requested Funds

As outlined in 7.3, we will be applying for three positions for software developers (at DKFZ, EKUT, and EMBL) plus one senior software developer at DKFZ. Funded by own contributions, the team will be complemented by an additional developer (DKFZ) and the Team Lead Architecture (EKUT).

5.4 TA A4: Data Stewardship - Central and Data Hubs

Overview of the Task Area

Data stewardship (DS) in GHGA is focused on supporting the scientific community in sharing human omics data in a safe and efficient manner according to the FAIR principles. The goal is to ensure a uniform experience for GHGA users while at the same time incorporating the various local communities connected to the GHGA data hubs. A4 is split into four measures: A4.M1 ensures the overall coordination across GHGA Central and the GHGA data hubs by a Lead Data Steward. A4.M2 organises an efficient and uniform user support at GHGA Central. In A4.M3, local data stewards at the data hubs connect to the networks and researchers that are linked to their respective institution. A4.M4 connects the co-located sequencing centres at the GHGA data hubs as major producers of omics data in Germany.

5.4.1 Measure A4.M1: Central Data Stewardship Coordination

Consortium Member	Contribution
Buchhalter (DKFZ)	Coordination of TA, Oversight of Lead Data Steward activities
Kohlbacher (EKUT)	Oversight of Lead Data Steward activities
Menges (DKFZ)	Team Lead Data Stewardship

Goals: Coordinate the overall team of data stewardship activities, maintain and update processes and SOPs, report on data usage to the BoD, manage legal aspects of data transfer processes, and coordinate data exchange with other (inter)national data infrastructures. Overall Coordination: The Lead Data Steward (LDS) is responsible for coordinating the distributed team of data stewards (centrally and at the data hubs). The central data steward team (CDS) supports the centrally managed processes of data submission and data access of GHGA, including the management of necessary legal processes (with C2). The LDS coordinates the data stewardship operations by maintaining SOPs for all relevant processes, assigning tasks to data stewards in an agile manner, and supervising the onboarding and training processes of the data stewardship team. The LDS serves as liaison between the central data stewards (A4.M2), the local data stewards at the data hubs (A4.M3 & A4.M4), and the Data Hub Operations Consortium Board (OCB). The LDS reports to the OCB about DS operations and recommends the assignment of projects to data hubs to balance the overall load. The LDS also serves as an escalation point for user complaints. Further, the LDS is responsible to collect feedback from the DS team and users about operations, processes, and interfacing to define new requirements for the Data Steward Toolkit, Metadata Model, Data Portal GUI and internal SOP catalogue. **(Inter)national Data Infrastructures:** Interconnectivity with central EGA is maintained by the CDS by aligning and forwarding metadata to the EGA and eventually GDI (c.f. B4). In turn, EGA will forward user requests concerning German data sets to GHGA. The LDS and CDS also actively support the connection of GHGA to other emerging European data infrastructures including the EHDS and GDI, as well as connect to national initiatives such as NFDI Helpdesk and the MV GenomSeq on the national level. **Product Ownership and Process Development**: The LDS - supported by the data stewardship team - acts as a product owner for the data portal and works together with A1 and A3 to ensure continuous development and improvement of the GHGA Data Infrastructure, especially with respect to user experience and further automation of processes. They also closely interact with B4 to maintain and develop the GHGA Metadata Model.

5.4.2. Measure A4.M2: Central User Support

Consortium Member	Contribution
Buchhalter (DKFZ)	Coordination of TA, Oversight of central data stewardship team
Menges (DKFZ)	Team Lead Data Stewardship

Goals: Provide user support by the central data stewards team (CDS) under the supervision of the LDS using the GHGA Helpdesk for the management of core user

processes in GHGA. Management of Data Submission to GHGA: The CDS assist in signing of Data Processing Contracts between users and GHGA Central, validate non-personal metadata of submissions, run data and metadata QC checks (cf. B2.M1**),** and perform checks to confirm that no personal metadata that can be used to identify a subject is submitted. For submissions that pass the checks, they load the metadata to the GHGA portal to make submissions publicly findable. **Management of Data Access Requests:** The CDS forwards and tracks incoming data access requests to DACs to guarantee a timely handling of the request. Once data access requests are granted to GHGA users, the CDS works together with the local data stewards who make the data that are stored at the respective hub available for download or, once available, stage them into GHGA SPEs (cf. A4.M3). **User Management:** The communication with users is orchestrated through the ticketing system of the GHGA Helpdesk. The CDS further handles the identity management in GHGA by validating Independent Verification Addresses to ensure that authorisations are assigned to the correct user accounts. The CDS team will work with the training and outreach team (B3) to enhance user training (e.g., submission guidelines, user documentation, training materials, webinars) and create materials for the onboarding of new data stewards.

5.4.3 Measure A4.M3: Federated Data Mobilisation and Local User Support

Consortium Member	Contribution
Buchhalter (DKFZ)	Data Stewardship at Data Hub Heidelberg
Motameny (UzK)	Coordination of TA, Data Stewardship at Data Hub Cologne
Dahl (TUD)	Data Stewardship at Data Hub Dresden
Gagneur (TUM)	Data Stewardship at Data Hub Munich
Beule (MDC and BIH)	Data Stewardship at Data Hub Berlin
Kohlbacher (EKUT)	Data Stewardship at Data Hub Tübingen

Goals: Local data ingest/egress processes, SPE data management & user support. At each data hub, GHGA stewards support users of the local sequencing centres (all data hubs are co-located at major academic sequencing centres) with data management and submission to GHGA. Their primary duty is to support the users of GHGA at their respective data hub's institutions. External submissions and requests will be reviewed by the LDS and then assigned by GHGA Central to an appropriate data hub for ingestion into local S3 storage. **Local Data Ingest:** The local data stewards help local GHGA users during the submission of data to GHGA by answering questions concerning the collection and submission of metadata, and managing the technical transfer of the research data and personal metadata into the data hub's storage infrastructure. As such they serve as local ambassadors for the usage and continuous development of GHGA. **Support for Community Driver Projects for Data Mobilisation:** The local data stewards will support the community use cases described in B1 to enable swift data deposition in GHGA and to enable further adaptations of the overall platform. In particular, data deposition from MV

GenomSeq, ERDERA, NAKO, the Bavarian Genomes project, DigiMed Bayern, KORA, and DZHKomics will be supported. As part of these activities, the deposition of complex datasets, including multi-omics profiles will be supported. **Local Data and Resource Management for Data Processing in GHGA SPEs:** Local data stewards are also responsible for data management and resource configuration at the local GHGA rSPE infrastructure. Once data access has been granted centrally to a data requester, data stewards decide on the best data hub to perform the calculations, estimate and configure the required computational resources of the local rSPE, and make the data available for processing. Local data stewards also ensure the proper handling of encryption and decryption keys and that rSPEs are closed and data are removed at the end of the project. They will also operate the necessary processes that make data that are stored at the local data hub available in the GHGA cSPE, once this is established.

5.4.4 Measure A4.M4: Connection to Sequencing Centres

Consortium Member	Contribution
Buchhalter (DKFZ)	Co-Spokesperson TA A4, Connection to Heidelberg sequencing centre
Motameny (UzK)	Co-Spokesperson TA A4, Coordination of TA, Connection to Cologne sequencing centre
Dahl (TUD)	Co-Spokesperson TA A4, Connection to Dresden sequencing centre
Gagneur (TUM)	Co-Spokesperson TA A4, Connection to Munich sequencing centre
Beule (MDC and BIH)	Co-Spokesperson TA A4, Connection to Berlin sequencing centre
Kohlbacher (EKUT)	Co-Spokesperson TA A4, Connection to Tübingen sequencing centre

Goals: Ingest data from local sequencing centres into GHGA. In order to continuously ingest high-quality data sets with rich metadata for GHGA, the local data stewards will build strong connections to the co-located sequencing centres at the data hubs. **ETL Processes:** Data stewards will help with the implementation of interfaces that allow seamless submission of human omics data and their corresponding metadata from the local sequencing centres to GHGA. Tools for standardised mapping of metadata from the local LIMS to the GHGA metadata have been established and will be further maintained, making it easy for researchers at the sequencing centres to submit their data into GHGA. Also, ETL processes will be maintained and adapted to changing metadata requirements to streamline submission of data to GHGA. **Outreach to Users:** The data stewards will, supported by B3, furthermore provide information material about FAIR data sharing and GHGA to the sequencing centres to be used during counselling of new sequencing projects. In this way, researchers are approached in the planning phase of a sequencing project and encouraged to make their data available for secondary research. In user meetings at the sequencing centres, the data stewards will inform about the services of GHGA, promote FAIR data sharing and collection of high-quality metadata.

Tasks and Deliverables

Task	Deliverables	Due Date
A4.M1.T1 Coordinating all GHGA data stewards	Documentation and report on regular Data Committee meetings to the OCB	M12, M24, ..., M60
A4.M1.T2 Aligning and forwarding metadata to the EGA	First data access request received via EGA fulfilled	M12
A4.M2.T1 Supporting users in submissions and data requests	Documentation and report on submissions and requests to the OCB	M12, M24, …, M60
A4.M2.T2 Proactively engaging with user queries	FAQs for users published	M12
A4.M3.T1 Ingesting MV GenomSeq data	First dataset deposited	M6
A4.M3.T2 Ingesting first data sets	1 PB data stored	M24
A4.M3.T2 Expanding number of ingested data sets	20 PB data stored	M48
A4.M3.T3 Ingesting data from community use cases	Ingestion of data from cancer and RD communities as described in B1.M1 and B1.M2	M20
A4.M3.T4 Managing resources for rSPE and cSPE compute	First access staged into cSPE	M36
A4.M3.T5 Fostering deposition of multi-omics datasets	First multi-omics dataset deposited in GHGA	M24
A4.M4.T1 Implementing interfaces that allow seamless data submission	ETL processes established at all sequencing centres	M24
A4.M4.T2 Providing information on FAIR data sharing and GHGA to sequencing centres	User meeting held at all sequencing centres	M48

Dependencies, Interactions, Risks, and Mitigation Strategies

Dependencies and Interactions: TA A4 depends on the success of nearly all other task areas. Data stewards will only have a user base to support if all aspects of central and local operations (A1 and A2), the legal framework (B5, C2), training and outreach (B3), connection to the scientific communities (B1, B2), international cooperations and metadata (B4) and the technology stack (A3) develop successfully. Internally, the central data steward team of A4.M1 and A4.M2 will have strong interactions with the central operations team of A1 for the operational tasks as well as with the project management team of C2 concerning legal documents. They will use the DS tools developed in A3 in their routine work and pass user feedback to the architecture team to further improve the user-facing GHGA software. Together with B3, outreach and training material for promoting GHGA towards its user community will be developed. The local data stewards in A4.M3 will have close interactions with the local operations teams of the data hubs (A2) to solve technical problems and streamline local data management processes. External interactions will be with NGS sequencing centres (A4.M4) and other international and national data infrastructures (B4 and A4.M1). An important task of the DS team is the support of the community driver projects through close interaction with B1. **Risks and Mitigations:** In the scenario of delayed technical solutions, for example for data ingest or other features of the data stewardship tool, A4 will revert to manual solutions and SOPs to implement the processes. In the scenario that GHGA will be well adopted and quickly build a large active user base,

the main risk for this task area are unsatisfied users due to longer than expected response times. These can occur when GHGA infrastructure and software development are delayed or if an insufficient number of personnel are available. The planned mitigation measure is to distribute tasks across the already existing team of data stewards from the first phase of GHGA in order to minimise response times. Also, transparent expectation management towards users will ensure high user satisfaction even in times of increased response times. Prioritising time consuming and uniformly structured DS tasks for automation will help to reduce the workload while considering restricted development capacities. In the unlikely scenario that GHGA is not well adopted by the scientific community, we will intensify our outreach efforts towards the scientific community via the network of local data stewards. The successful support of the community driver projects (B1) will also mitigate this risk.

Justification of Requested Funds

As outlined in 7.4, we will be applying on the one hand for three positions for the central data stewardship team, including the lead data steward position. This will be supported by a CDS position via own contributions from the DKFZ. On the other hand, the current six data hubs will receive each half a DS position to support local DS activities. This will be supplemented by own contributions as outlined in Description and Summary of Contributions by (Co-) Applicants.

5.5 TA B1: Community Driver Projects

Overview of the Task Area

Community driver projects will support and accelerate GHGA's overarching objective to connect to, and be embedded within, key genomics communities. Initial driver projects will be focused on the core communities cancer (B1.M1), rare diseases (B1.M2), and common diseases (B1.M3), aiming to embed GHGA in the scientific process beyond archival as an 'endpoint'. B1.M1 and B1.M2 will curate national reference datasets based on MV GenomSeq and other initiatives for cancer and rare diseases to enable the deployment of standardised analysis workflows (cf. B2). Both measures will focus on variant annotation and community-driven variant interpretation environments. B1.M3 will establish close ties between GHGA and NAKO, enabling the NAKO omics community to utilise the GHGA SPE and data services for population genetics and to develop new prevention use cases. In addition to connecting GHGA to these key communities based on the (added) value chain creating highly refined datasets by *curating and archiving data*, these projects will also test run technical advances, and support the national and European embedding (cf. B4), e.g. via shared use cases together with other NFDIs.

5.5.1 Measure B1.M1: Cancer Genomics

Consortium Member	Contribution
Hübschmann (DKFZ, NCT HD)	Establishment of diagnostic research infrastructure
Fröhling (DKFZ, UHH, NCT HD)	Community alignment, integration of cancer genomics data
Brors (DKFZ)	Cancer genomics alignment
Pfister (DKFZ/KiTZ)	Integration of GHGA with paediatric cancer community
Lichter (DKFZ)	Community integration and alignment
Glimm (NCT DD)	Integration of cancer genomics data
Jäger (NCT HD)	Cancer genomics and immunology alignment
Behrens (TUM)	Team Lead Community Engagement

Goals: Establish GHGA as a major diagnostic and research data hub for the German cancer genomics and precision oncology communities. We will achieve these goals by i) curating data from MV GenomSeq, ii) implementing analysis tools and APIs for genome-analysis and variant-interpretation services for somatic alterations and targetability, and iii) deploying visualisation and decision-support platforms to facilitate evidence-based treatment recommendations. This driver project will closely interact with the MII (in particular the project PM4Onco), Cancer Core Europe (CCE), 1+MG and GDI, EHDS, and resources like COSMIC or CIVIC. Key medical needs in precision oncology are (i) identification of targetable lesions, (ii) making evidence-based treatment recommendations, and (iii) allowing for longitudinal follow-up and re-assessment of a patient's molecular data for secondary prevention or for several rounds of molecularly guided treatment. To address the community needs **infrastructurally**, we will exploit synergies with Rare Disease Genomics (B1.M2), including alignment on common variant annotation services, germline variant calling and reporting, but also address cancer-specific aspects including calling and reporting of **somatic** variants, gene fusions, estimation of tumour purity & ploidy, etc. We will furthermore develop a backbone and interfaces to (i) allow for seamless interaction in the oncology domain, first and foremost the **MV GenomSeq** (record linkage with the clinical Data Nodes, in the oncology networks DNPM, nNGM, and MASTER upon approval by the trusted third party, the RKI), (ii) build large resources containing oncology world knowledge and blocks of evidence (oncoKB, CIVIC) depending on the licences available to the individual user and (iii) provide dashboards as well as visualisation and decision support systems, both for cohort analyses (e.g., cBioPortal, B2) and for processing of an individual patient's data for molecular tumour boards (e.g., cBioPortal individual patient view, the Knowledge Connector, or the molecular tumour board portal). We will aim at performing quality control, benchmarking and harmonisation at the level of MTB recommendations. To help address medical needs and to engage with the cancer genomics and precision oncology communities, we will build and curate a pan cancer reference **data resource** from scientific submissions and, in particular, from large case numbers of MV GenomSeq.

5.5.2 Measure B1.M2: Rare Disease Genomics

Consortium Member	Contribution
Graessner (UKT)	Coordination of TA, data ingestion logistics, Solvathons
Ossowski (UKT)	Establishment of diagnostic research infrastructure
Gagneur (TUM)	Organisation of Solvathons, link to TA B2
Behrens (TUM)	Team Lead Community Engagement

Goals: Make GHGA the diagnostic research data infrastructure for the German RD genomics community. We will achieve these goals by i) curating and staging data from MV GenomSeq, ii) implementing APIs for various genome-analysis and variant-interpretation services, and iii) enable deploying workflows and organising the recently established Solvathons to foster reanalysis across this national RD reference dataset. This driver project will connect the German RD genomics community with infrastructures and activities as implemented in the diagnostic research workstream of European Rare Disease Research Alliance (ERDERA) on European level. **National RD Reference Dataset:** We will **curate and archive a national RD reference dataset** based on data from MV GenomSeq. This dataset will be used in GHGA annotation services and community-driven variant interpretation Solvathon events and processes thereby engaging the RD genomics community. **Establishment of Interfaces:** We will support the RD community in these activities by establishing interfaces to 1) the clinical decision support system megSAP/GSvar, 2) HerediVar (new system for shared interpretation of variants developed by the HerediCaRe consortium), 3) clinical Data Node RD (also known as zKDK Rare Disease in MV GenomSeq), complementing the respective RD3 database used in ERDERA, 4) long-read analysis pipelines (wf-human-variation and megSAP) in accordance with the respective ERDERA plans. This new diagnostic research infrastructure will support a harmonised data reanalysis of the national RD reference dataset with modern and ERDERA-abiding variant annotation standards (B2), thereby ensuring consistency with other ERDERA datasets and improved variant calling. **Solvathons for Community Engagement**: We will organise **Solvathons** for the RD community, which are collaborative events/processes involving bioinformatic and clinical/genetic experts with the aim of solving diagnostically unsolved cases by joint data analysis and interpretation. Thereby, we will implement various diagnostic RD research use cases, aiming at community-driven interpretation and consensus classification of variants for different RD groups, variant types and omics technologies. The Solvathon concept has been developed, tested, and successfully applied in the EU programme Solve-RD. This driver project will thus integrate the German RD genomics community with the respective European community activities in ERDERA.

5.5.3 Measure B1.M3: Common Disease & Prevention

Consortium Member	Contribution
Stegle (DKFZ)	Coordination of technical implementation and genetic risk factors
Fluck (ZBMED)	Alignment with NFDI4Health on metadata and record linkage.
Panreck (NAKO)	Integration with NAKO data access and policies
Peters (HMGU/NAKO)	Definition of epidemiological use cases and scientific priorities
Pischon (MDC/NAKO)	Connecting GHGA and NAKO / NFDI4Health
Rosenstiel (UKI)	Genomic medicine and connection to clinical communities common disease
Specht (Charité / GBN)	Connection to German Biobanking Network
Behrens (TUM)	Team Lead Community Engagement

Goals: Establish GHGA as the genomic data infrastructure of the German National Cohort (NAKO) and enable common disease and molecular prevention initiatives. In order to address indication areas beyond cancer and RD, we will i) establish adapted legal and technical measures to foster interoperability with NAKO and NFDI4Health, ii) establish practical use cases and pilot applications of the GHGA Release Analysis Platform, and iii) develop new use cases in molecular prevention such as a genomic newborn screening. **Alignment:** First, we will establish technical and organisational alignment between the existing NAKO data infrastructure, which in large parts is operated at DKFZ, thus providing ample opportunities for alignment. This entails the development of the necessary contractual framework on the one hand, and ETL workflows (including record linkage) to facilitate data export for authorised NAKO participants on the other hand. Building on this new bridge, we will ingest all forthcoming NAKO omics data into GHGA. At the time of writing, the funded omics data include genotyping arrays for all 200k participants, as well as approx. 15,000 whole genome sequences funded from the Helmholtz Association (granted) and up to 20,000 additional WGS from the Genomes of Europe Initiative (final decision pending). In parallel to alignment of infrastructure solutions with NAKO, this use case will also foster the alignment with NFDI4Health on multiple levels (cf. MoU), with NAKO being a major driver of this interaction. We will ensure that archived data, and in particular, metadata in NFDI4Health and GHGA are linked to maximise findability.

Bioinformatics support: Second, we will support the NAKO omics community to conduct quality control, genomic data processing, and downstream analyses. Building on all NAKO omics data ingested, we will work with the NAKO community as a pilot use case to conduct quality control but also genetic analyses such as GWAS and rare variant studies, using the GHGA platform and cSPE concept (A3) - a major pillar of our technical roadmap. This use case is very suitable to test drive this system, as the NAKO omics community, while external to GHGA, is focused and motivated and also has expertise with using HPC and cloud technologies. **New use cases:** Third, we will establish new use cases based on NAKO data for molecular prevention. The communities that come together in GHGA will provide a strong opportunity to mobilise the NAKO reference data with the aim to foster stratified preventive measures, e.g. genomic risk prediction for common diseases, including cancer. This activity

will also exploit synergies with the National Cancer Prevention Center, coordinated at DKFZ. *These measures will be primarily supported by NAKO funding that is dedicated to achieving these aims.* **Molecular Prevention:** Fourth, emerging molecular prevention projects include a genomic newborn screening programme to be run jointly by Heidelberg University and DKFZ, in which, embedded into an ELSI framework, utilising elaborate consent management and based on careful selection of diseases to be screened, whole genome sequencing will be performed in 3,000 newborns over three years. In this context we will also implement record linkage with German biobanks (GBN/GBA).

Tasks and Deliverables

Task	Deliverables	Due Date
B1.M1.T1 Curating cancer datasets from MV GenomSeq	Handover of curated data to data stewards for data ingestion	M24
B1.M1.T2 Implementing analysis tools and APIs for genome-analysis for somatic alterations and targetability	Provision of resources containing oncology world knowledge and blocks of evidence	M36
B1.M1.T3 Deploying visualisation and decision support platforms for treatment recommendations	Accessible dashboard for decision support	M40
B1.M2.T1 Curating of RD datasets from MV GenomSeq	Handover of curated data to data stewards for data ingestion	M24
B1.M2.T2 Implementing APIs for various genome-analysis and variant-interpretation services	RD diagnostic research infrastructure established	M24
B1.M2.T3 Using deployed workflows for self-organised Solvathons for rare disease cases	First GHGA Solvathon organised	M30
B1.M3.T1 Establishing legal and technical measures for interoperability with NAKO and NFDI4Health	NAKO contractually and legally integrated & first data deposited	M6
B1.M3.T2 Running use cases and pilots with common disease data sets	First NAKO community analysis on GHGA cSPE	M24
B1.M3.T3 Developing strategies for new use cases in molecular prevention	Written joint agreement on two new use cases	M24, M48
B1.M3.T4 Participating in efforts for newborn screening project	Aligned ELSI framework for ingest of newborn data into GHGA	M18

Dependencies, Interactions, Risks, and Mitigation Strategies

Dependencies & interactions. B1 depends on the success of a large number of TA, including central and local operations (A1 and A2), the legal framework (B5), deployment of community data services (B2), outreach and training of the key genomics communities (B3) and international cooperations (B4) develop successfully. Internally, B1 will have strong interactions with the community data service team of B2 for the deployment of workflows as well as with the outreach and training team of B3 for engaging genomics communities. Together with A4, the respective specific data stewardship will be planned and deployed.

Risks & mitigation. The main risks for this task area are (i) the timely availability of needed workflows and SPE as well as (ii) the timely fitness of GHGA to accommodate the archival of analysis-enriched national reference datasets. The identified risks will be mitigated through a profound and proactive collaboration with task areas B2 and B3 for community engagement

and data services as well as A1 and A2 for respectively needed central and local operations adaptations. Additionally, redundancy and alternatives in data providers will ensure robustness to delayed data deposition or legal hurdles to use the data as intended.

Justification of Requested Funds

As outlined in 7.5, we will be applying for a total of 2 FTE for this task area, which includes 50% of the envisioned Team Lead Community Engagement (all measures, cf. Team Structure within GHGA), a 50% position for the integration of NAKO and GHGA (B1.M3) and a full position for B1.M2. This will be complemented by a full position based on DKFZ own contributions for B1.M1, and a half position from TUM own contributions for B1.M2. B1.M3 is also further supported via complementary funds from NAKO.

5.6 TA B2: Community Data Services

Overview of the Task Area

A key aim of GHGA is to provide value-added resources for communities through secondary use of research data. This requires continued deployment of workflows and data solutions in order to present harmonised and quality-controlled (QC) data across submissions via the GHGA data portal, on the one hand, and to support analysis for tasks in B1 on the other hand. In doing so, B2 primarily aims to achieve the following objectives: B2.M1: implementing data and metadata QC on submission, ensuring FAIRness. B2.M2: leveraging unique large national datasets and community-driven collaborations to provide novel and innovative products and portals such as a federated Beacon service across GHGA data hubs and a variant frequency database for German cohorts. B2.M3: implementing and running state-of-the-art workflows to service community-driven projects and portals (B2.M2), and to enable harmonised data analysis.

5.6.1 Measure B2.M1: Data Quality Control & Curation Tools

Consortium Member	Contribution
Nahnsen (EKUT)	Implementation of BioQC workflow
Ulas / Schultze (DZNE)	Development of metadata scoring metric
Bork (EMBL)	Collecting user requirements and feedback for improving the scoring metric
Behrens (TUM)	Team Lead Community Engagement

Goals: Ensure data quality, integrity and FAIRness with standardised technical and metadata quality control reports. B2.M1 aims to integrate processes ensuring high-quality data submissions to GHGA. Evaluation of datasets and metadata with scoring metrics will help the community derive maximum value from well-managed, high-quality FAIR data and metadata. **Validation metrics for improving data quality (BioQC):** Human errors, including those during metadata annotation, are pervasive. B2.M1 will align with, and drive forward, the FEGA BioQC Working Group, engaging the community in the supporting validation of submissions. These efforts aim to ensure robust validation of submissions. The validation process will build upon pipelines ensuring correct sex annotations, familial

relationships, and sample matching across donors (sexdetermine and GRAF for DNA or DROP for RNA-seq [29–31]). These tools will help identify discrepancies and ensure the integrity of submitted biological data. **FAIR scoring metric:** Metadata harmonisation is crucial to ensure data findability, accessibility, and reusability. B2.M1 will develop a scoring metric, which will be calculated based on the completeness and quality of the submitted metadata, and the use of controlled vocabularies and ontologies. Positive feedback mechanisms (e.g., badges) will incentivise and encourage data submitters to provide comprehensive metadata that adheres to FAIR principles. The system will be transparent, providing clear indications of progress and areas for improvement.

5.6.2 Measure B2.M2: Community Interfacing and Portals

Consortium Member	Contribution
Gagneur (TUM)	Co-Spokesperson TA B2, TA coordination, gnomAD integration, variant frequency database
Mertes (TUMUH)	Variant frequency database
Hübschmann (DKFZ, NCT)	Co-Spokesperson TA B2, Variant frequency database, User Empowerment in Cancer Genomics
Beule (BIH)	User Empowerment in Cancer Genomics: cBioPortal toolchain
Ohler (MDC)	Services definition and alignment
Grüning (UFR)	Connecting GHGA to Galaxy community
Behrens (TUM)	Team Lead Community Engagement

Goals: Integrate workflows and data services with community and driver project resources to produce data products and portals with added value and secondary use potential. To support secondary use for all GHGA-hosted data, and in particular the driver projects (B1) including MV GenomSeq, B2.M2 will coordinate the development of added-value data products, such as federated variant querying and population-specific variant frequency statistics for genetic disorders, aggregated data portals for cohort analyses and hypothesis generation like federated cBioPortal resources, interactive workflow solutions using Galaxy, and patient-level dashboards as well as decision support systems for identification of targetable lesions or formulation of individual evidence-based treatment recommendations. **Community projects and data services coordination:** Management and alignment with members of B1 and B2 as well as data controllers, subsequent data ingestion and integration into GHGA with Data Stewardship (A4) will need to be coordinated. **German variant frequency database and federated gnomAD partnership:** In RD diagnostics, variant frequency within a population is a crucial filter criterion. Therefore, multiple driver projects (B1), including the MV GenomSeq, require a variant frequency database. Building upon the DZHKomics resource comprising 1,200 whole genomes from healthy donors, B2.M3 will build a prototype for a frequency database within GHGA and automate the processing of incoming data. This resource will then be expanded with new healthy-donor cohorts from the NAKO, NAPKON, genomic newborn screening (B1.M3) and non-affected parents or relatives from MV GenomSeq from genomic newborn screening. A

primary outcome of this task is the integration of processed variant frequencies into the gnomAD browser as a federated service. As an initial step for this task, consortium and team members are already participants of the federated gnomAD working group. **Beacon implementation:** The GA4GH Beacon API supports the discovery of genomic variants and biomedical data across distributed resources. Querying genomic variants, frequencies, and related clinical information is essential in rare disease diagnostics. As such, B2.M2 will implement a federated Beacon service across datasets at all six GHGA data hubs, and integrate the German variant frequency database, supplying GHGA users and driver project communities with an invaluable resource of aggregated variants across datasets. The GHGA Beacon v2 instance will also be integrated into the GDI Beacon Network. **User Empowerment in Cancer Genomics:** cBioPortal (cf. B1.M1) allows access-controlled data sharing and cross-institution collaboration. While several well-populated but disjunct instances of cBioPortal are running in Germany, GHGA will link these resources in order to reduce redundancy, and to provide a single comprehensive view on available data. Furthermore, a toolchain will be developed such that access to GHGA-hosted datasets can directly be applied for from the cBioPortal instance. In addition to displaying cancer omics data at cohort level, a central task for precision oncology is to offer single patient dashboards for identification of targetable lesions and preparation of molecular tumour boards (MTBs). cBioPortal also has a functionality for this use case, but other tools (Knowledge Connector, molecular tumour board portal) also exist. GHGA will provide interfaces and APIs for various tools required by the precision oncology community.

5.6.3 Measure B2.M3: Integrated Data Processing Tools

Consortium Member	Contribution
Gagneur (TUM)	Co-Spokesperson TA B2, TA coordination, variant annotation and RNA-seq-based workflows
Hübschmann / Brors (DKFZ, NCT)	Co-Spokesperson TA B2, Variant calling and benchmarking
Stegle (DKFZ/EMBL)	Single-cell multi-omics and variant calling; spatial transcriptomics
Graessner (UKT)	Coordination with B1.M2
Behrens (TUM)	Team Lead Community Engagement

Goals: Deploy, test, and run scalable workflows and data services enabling the development of products in M2, and addressing the primary needs of GHGA Community Driver Projects (B1). B2.M3 will continue to provide multi omics data processing workflows that serve specific communities and use cases and support the direct development of products and portals described in B2.M2. While products from B2.M2 will directly serve users of GHGA, the services described here will not be user-facing but rather internally developed and deployed to achieve the overall aims of B1 and B3. To ensure full compatibility with existing and emerging international standards, we will engage with international communities (e.g., GA4GH, ELIXIR, gnomAD and de.NBI). Furthermore, since the consortium members leading this TA also produce leading-edge research on these

topics, modern state-of-the-art algorithms will easily be transferred and integrated into the GHGA data service catalogue. **Variant calling, benchmarking, and annotation:** Accurately calling and annotating genetic variants is crucial for identifying and interpreting those that may be responsible for disease and is therefore a basic service needed by all driver projects (B1). B2.M3 will improve variant benchmarking and reference call sets through extended ring trials performed with the DFG sequencing facilities. Additionally, we will evaluate and integrate new variant effect prediction tools, creating a scalable annotation service supporting the latest AI-based prediction methods. For this, we will leverage and build upon the expertise and research already being conducted by contributing consortium members (e.g., Gagneur and Stegle [32–34]). Together, these efforts will ensure that variants are both called and annotated with modern, accurate approaches across different datasets, supporting the goal of understanding, diagnosing and treating complex genetic disease. Regular cycles of data freeze, processing, and release will provide state-of-the-art annotations back to key communities and will support events such as Solvathons (B1.M2) to help in variant prioritisation and interpretation. **Readiness for new omics modalities:** To continue supporting driver projects and communities as multiple omics data modalities are routinely generated, B2.M3 will additionally implement and maintain data services to support this diverse research data. The Detection of Outliers Pipeline (DROP), developed in the group of Gagneur, aids rare disease diagnostics using RNA-seq data. B2.M3 will implement and utilise this tool to call expression and splicing outliers potentially underlying rare disease. In addition, GHGA will ensure readiness to support single-cell multi-omics and spatial transcriptomics data (in collaboration with the BMBF-funded SATURN3 project), as well as single cell variant calling analysis to aid understanding of tissue/tumour heterogeneity in hard-to-treat cancer types.

Tasks and Deliverables

Task	Deliverables	Due Date
B2.M1.T1 Establishing validation metrics for improving data quality pipeline	BioQC pipeline implementation	M24
B2.M1.T2 Developing metadata FAIR scoring metric	Rollout of the scoring metric system across GHGA submissions	M24
B2.M2.T1 Building prototype for a frequency database within GHGA	First release of the German variant frequency database	M24
B2.M2.T2 Expanding frequency database with healthy-donor cohorts	Second release including non-rare disease cases and disease-stratified frequencies	M60
B2.M2.T3 Implementing federated Beacon service across GHGA datasets	Beacon v2 deployment, connection to GDI Beacon network, and implementation of a GHGA Beacon v2 data ingestion pipeline	M36
B2.M2.T4 Linking and integrating cBioportal into GHGA	cBioportal toolchain developed for GHGA datasets	M36
B2.M3.T1 Improving variant benchmarking and reference call sets	Development of somatic benchmarking datasets	M24

Task	Deliverables	Due Date
B2.M3.T2 Evaluating new variant effect prediction tools	Benchmarking of variant callers with developed dataset	M60
B2.M3.T3 Integrating new variant effect prediction tools	Implementation and deployment of (AI-based) variant effect prediction methods	M60
B2.M3.T4 Supporting research data usage with services	Implementation of DROP workflow	M24

Dependencies, Interactions, Risks, and Mitigation Strategies

Dependencies: TA B2 depends primarily on Community Driver Projects (B1) for establishing and maintaining connection to relevant communities and aligning on common goals and required services from these communities, and on Data Stewardship (A4) and legal regulation (C2) for acquiring data access and processing rights to this data. Additionally, data services within GHGA can be successfully deployed only if all aspects of central and local operations (A1 and A2), outreach and training of the key genomics communities (B3) and international cooperations (B4) are sufficiently developed. **Interactions:** B2 will continue to work with established national and international communities (B4) such as GA4GH, FEGA, nf-core, and Galaxy, specifically for alignment on technical specifications relating to workflows, data, metadata, QC, and portal implementation standards and use-cases.

Risks & mitigation: The main risks for this task area are (i) difficulties or delays in obtaining community datasets and the required access that would enable the production of planned products and portals (M2), and (ii) successful technical implementation of these solutions in the GHGA architecture. Strong coordination and collaboration with TA B2 and B3 will be required to engage these communities effectively and mitigate risks for data acquisition. For technical risk mitigation, coordination with A1 and A2 for central and local data hub operations, and especially A3 for architectural and research environment deployment support, will ensure that data services are duly supported through expertise in other task areas.

Justification of Requested Funds

As outlined in 7.6, we will apply for one half position each for BIH/Beule (for tasks in B2.M2), and DKFZ/Huebschmann (for tasks in B2.M3), and a full position for TUM/Gagneur, which includes the 2nd 50% of the envisioned Team Lead Community Engagement position (shared with B1, cf. Team Structure), as well as contributions to tasks in B2.M2 and B2.M3. In addition, a full position for 3 years is budgeted for DZNE (Ulas and Schultze) to carry out B2.M1.

5.7 TA B3: Outreach & Training

Overview of the Task Area

The Outreach and Training TA will support GHGA's central infrastructure by coordinating communication with GHGA stakeholders. This effort aims to foster a cultural change where

researchers, clinicians, as well as the public more actively support and engage in FAIR data sharing. We will provide information and training on new functionalities, best practice examples, and updates from the consortium, for associated communities and the wider research community. Events and publications will highlight the importance of FAIR data sharing for secondary research and healthcare. By actively engaging patients and the broader public, we aim to educate, empower and foster trust and support for genomic research. B3 will create multi-media training materials and focus on user experience to enable efficient and secure use of GHGA by users from various communities. Internal training will ensure team members process data safely. Work in B3 will be leveraged through existing partnerships within the NFDI, GDI, FEGA, and the bioinformatics communities (de.NBI, de.KCD, ELIXIR).

5.7.1 Measure B3.M1: Platform Communication

Consortium Member	Contribution
Stegle (DKFZ)	Coordination of central communication efforts, crisis communication
Walter (UdS)	Creation of content for communication
Schulze-Hentrich (UdS)	Creation of content for communication
Winkelmann (HMGU)	Creation of content for communication
Kohlbacher (EKUT)	Creation of content for communication
Träger (DKFZ)	Team Lead Communications and Training

Goals: Coordinated communication around the GHGA platform, its functions and development updates, streamlining both internal communication within the consortium and coordinating external communication to GHGA users and other stakeholders. Throughout the operations of the GHGA platform, services will continuously evolve with new features and functionalities (cf. A3). To keep both internal GHGA members and external users updated, several communication channels have been established to reach our key stakeholders in the first funding phase. The website ghga.de serves as the primary information source, featuring background information on the consortium, relevant events as well as news. In addition, a regular newsletter and social media channels such as LinkedIn are used to inform interested parties of our progress and developments. Internally, email lists and instant messaging (Slack) are used to communicate and discuss new developments. These information sources will be maintained. Fact sheets and other media, containing concise and up-to-date information on the GHGA platform, its functions and impact, will be created and distributed via the mentioned channels. Bringing the consortium members and GHGA users together, we will also organise in person meetings - in the form of an Annual Meeting and User Symposium, respectively. These will foster collaboration and exchange within the communities in addition to providing information about the GHGA platform to our users. We will continue to update and expand the established crisis communication plan. The crisis communication plan will use the established communication channels and align with institutions involved in GHGA to reach the affected community and

inform in a transparent way, while complying to legal requirements in the case of a data breach (together with the Data Protection and Management Team in C2).

5.7.2 Measure B3.M2: Scientific and Clinical Outreach

Consortium Member	Contribution
Walter (UdS)	Scientific outreach
Schulze-Hentrich (UdS)	Scientific outreach
Winkelmann (HMGU)	Clinical outreach
Kohlbacher (EKUT)	Link to MV GenomSeq and bioinformatics community
Stegle (DKFZ)	Link to MV GenomSeq, MII, and bioinformatics community
Graessner (UKT)	Link to medical societies and use cases (B1)
Schlomm (Charité)	Link to oncology communities via DNA-med
Träger (DKFZ)	Team Lead Communications and Training

Goals: Actively identify and engage (new) scientific and clinical communities to disseminate GHGA's comprehensive genomic resources and services to a wider audience, highlighting the role of FAIR data sharing for secondary research and health care, and encouraging the integration and sharing of genomic data into research and clinical practice. We will continue and expand our successful scientific outreach activities towards national and international scientific, as well as clinical, communities and stakeholders. Our main goal is to disseminate GHGA's core key resources, genomic services and their use in medical genomics, and functional genomic analysis to new fields of users. We will operate in close collaboration with B3.M1, building on a variety of online channels, print media as well as personal interactions in workshops and conferences booths. **Extend reach and build partnerships:** Scientists and clinicians will be addressed through scientific programmes, academic communities/associations and various medical societies, reaching users and communities beyond current GHGA partners and use-cases (detailed in B1), thereby increasing inclusivity and widening the data sets and data types available within GHGA. Our aim is to expand the reach and spectrum of scientific, clinical, and basic research users, building and fostering direct interaction, (new) scientific collaborations and partnerships with GHGA. We will **create awareness** of GHGA's rich resources and data use highlighting the impact of GHGA's policy of FAIR data sharing and its broad application for basic and applied clinical research and health care. The goal is to drive a cultural change, influencing the research culture to encourage FAIR data sharing via GHGA. By positioning GHGA as the central omics repository in Germany, we aim to increase both the amount and diversity of data deposited in GHGA, reaching out to new data resources encompassing various molecular data types (e.g. genomic, epigenomic, proteomic, and metabolomic data) and different disease states, thereby facilitating integrated functional interpretation. **Medical genomics:** Beyond the strategic interaction with MV GenomSeq (and the preparation of successor projects), we will be strengthening our interactions with clinical communities. We will i) approach the medical societies (e.g.,

German Society for Human Genetics) at their annual conferences to promote the GHGA consortium, its services and resources, and establish new partnerships and use cases, ii) approach clinicians at university clinics involved with the MV GenomSeq to encourage the consent of patients for secondary research use, providing material for their patients in collaboration with B3.M3, iii) promote hackathons and Solvathons across the rare disease and cancer communities (with B1), iv) hold workshops at ESHG together with ELIXIR and EGA, as well as v) feature our work in widely read popular media (e.g., Ärzteblatt) to gain visibility in the clinical community. **International scientific outreach:** We will continue to engage with national and international scientific communities (MV GenomSeq, MII, de.NBI, HCA, IHEC, ELIXIR, 4DNucleome) through personal contacts by i) promoting GHGA's scientific achievements on conferences via booths, posters and topical talks/panel discussions, and ii) organising panel discussions and targeted workshops on topics related to FAIR data sharing, GHGA services, and showcasing application examples (e.g. what one can do with annotated data in GHGA or the potential of functional integrated analysis etc). To enhance the national and international visibility of GHGA (and FEGA) in the scientific (clinical) community we will also i) aim to publish (with B3.M1) on GHGA and FAIR data sharing efforts in a high-ranking international scientific journal, and ii) extend contacts and collaborations to academic societies, working groups, and umbrella organisations (e.g., Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften, Leopoldina, Academia Europea, AG Gentechnology) via topical publications and joint lectures.

Expand educational activities: Other activities to involve communities via online measures include i) the continuation and broadening of our well-established lecture series 'Advances in Data-Driven Biomedicine' in the direction of medical genomics (coordinated with MV GenomSeq and NFDIs such as NFD4Health), ii) the continuation of our well-established webinar series on topics such as FAIR data sharing, metadata, and application examples and, in cooperation with B3.M1, B3.M4 and B3.M5, on data protection and showcasing of new GHGA tools and workflows, and iii) the contribution to summer schools, webinars and lecture series organised by other initiatives (EMBO, MPG) or local institutes (e.g., Helmholtz lecture series) to reach out in new communities, focusing on early career scientists. We will also evaluate the needs of new communities (e.g., via questionnaires (with B3.M6)). Engaging scientific and clinical communities will raise the scientific and clinical awareness of GHGA resources and services leading to an increased use of the GHGA Archive. Closer ties to these communities will also result in adaptations of and innovations by GHGA to better serve both the clinical and scientific clientele. Showcasing success stories and case studies will further illustrate the practical benefits of GHGA.

5.7.3 Measure B3.M3: Patient and Public Communication

Consortium Member	Contribution
Graessner (UKT)	Production lead for podcasts, link to RD community
Stegle (DKFZ)	Coordination of patient communication strategy and its execution, content creation
Kohlbacher (EKUT)	Content creation
Winkler (NCT)	Bioethical input on patient engagement in the communication
Walter (UdS)	Content creation
Schulze-Hentrich (UdS)	Content creation
Winkelmann (HMGU)	Content creation, link to RD community
Schlomm (Charité)	Content creation, link to cancer community
Träger (DKFZ)	Team Lead Communications and Training

Goals: Further development and implementation of GHGAs communication strategy to highlight the value of omics data sharing in research and healthcare to a wider audience. Public trust in genome research and health data sharing is lower in Germany than in other countries (YourDNAYourSay), partly due to a lack of public outreach and information available in German. We aim to bridge this gap by actively informing and engaging the public, and specifically patients, who have their genome sequenced, to gain a better understanding of their opinions regarding data sharing and genomic research. We aim to build trust and shift the public opinion toward FAIR data sharing. A key component of our activities is the continuous production of the podcast 'Der Code des Lebens' ('The Code of Life'), which features episodes on different aspects of genome research. With 32 episodes published to date, the podcast regularly charts in the top 15 of German life sciences podcasts (Podcast Charts). Taking feedback into account, a second, shorter podcast format ("Genomhäppchen") has been established and was first published in May 2024. In addition, GHGAs public outreach strategy involves participation in various events, for which we developed specific formats, with the goal to engage in dialogue with the public. Ranging from Science Slam, and interactive debates on data sharing to Lego Challenges, we aim to continue developing novel and appealing formats. Once established and tested, a portfolio of public outreach formats will be made available for all GHGA institutions (especially data hubs, cf. A4) to use at local open days and long nights of science. Similarly, GHGA pursues collaborations with partners like the EMBL exhibition 'The World of Molecular Biology' allowing GHGA to leverage the wide outreach of these events. Classical press work, with press releases featuring success stories from GHGA and active placement of content in media such as *Apotheken Umschau* or FAZ, will round off our public outreach efforts.

GHGA specifically focuses the needs and interests of patients with regard to the sharing and use of their omics data for research. The PaGODA study [15], conducted during the first funding period, explored patient views and tasked us to develop a patient communication strategy. This strategy allows patients to become experts, make informed decisions about their data, and engage actively in science. We plan to launch a website specifically for patients within the current funding period and conduct focus groups to develop relevant

information materials in close collaboration with patients. The content will be distributed via social media, but also by engaging with patient organisations. By training medical professionals and offering easily accessible information they can provide to their patients, we can utilise them as multiplicators to reach our core target group of patients. Moreover, we plan to develop training courses for interested patients to learn more about omics and data sharing. In collaboration with the Patient-Expert-Academy for Cancer Diseases (PEAK) and ACHSE, we will initially gear these courses towards cancer and rare diseases patients. We also want to improve and harmonise how experts talk about omics to patients and the public. While language guides are available in English, no such guidelines for sensitive and patient-preferred language are available in German. Working closely together with our patient engagement efforts (B5.M4) we will develop a guideline for sensitive and patient-preferred language around omics research.

5.7.4 Measure B3.M4: User Training

Consortium Member	Contribution
Kohlbacher (UKT)	Coordinate training activities and connection to ELIXIR Training
Huber (EMBL)	Connection to international open-source activities (e.g., BioConductor)
Träger (DKFZ)	Team Lead Communications and Training

Goals: Providing training opportunities for current and future GHGA users will allow communities to utilise GHGA optimally. Additionally, internal training will ensure safe data processing. Our user-centric training programme is designed to empower GHGA users by providing comprehensive and evolving education on key aspects of the GHGA infrastructure. The established GHGA User Documentation, which is being openly developed via GitHub, will be updated continuously to ensure the best possible user experience. We will offer multimedia material, linked to the User Documentation, and training sessions on how to effectively upload and download data, along with guidance on preparing data appropriately for submission to GHGA and 'Bring your own data' workshops. In later stages, new material on additional functions developed in A3 within the GHGA platform will be added. To ensure the optimal delivery of our training programme, we will implement a training platform (linked to ELIXIR TeSS) into the GHGA portal that contains all of our training materials, thereby improving their useability. The training sessions will encompass both webinars and in-person workshops, at the data hubs and at different community-specific conferences (e.g., GCB[7] (bioinformatics), GfH/ESHG[8] (human genetics), GCC/GCRC (Cancer)[9]). Using a Train-the-Trainer concept, this task area aims to enable staff at the data hubs and across all TAs to train users. To implement this, training courses will be designed and materials shared, and GHGA staff will be instructed in the best training

[7] German Conference on Bioinformatics
[8] European Human Genetics Conference / German Society for Human Genetics Annual Meeting
[9] German Cancer Congress (DKK) / German Cancer Research Congress (DKFK)

techniques. For the data hubs in particular, a roadshow will be designed with training materials that can be run by local data stewards and engage local GHGA users - both existing and future. This hands-on approach is intrinsically linked to B3.M6 since it will allow us to interact directly with our users and establish a knowledgebase of common questions. In doing so, we contribute to foster an organically growing user community. Regular user feedback will be incorporated via the User Advisory Board (cf. 3.4.1) and regular GHGA User Meetings.

To support the community in ensuring safe data processing, this task area aims to continue to develop internal training materials, which we also want to share with other NFDI consortia with similar demands (e.g., on sensitive data, omics data, SPEs, data protection, ethics). We already successfully established a course on data protection for internal and external participants, which will be continued in the Assured programme (B3.M5). While the training of new staff will continue in this area, additional courses on aspects such as GHGA information security measures and best practices in data sharing will be developed. We will also continue to educate researchers on topics such as FAIR, statistical analysis or data visualisation in our Webinar series and in various workshops organised in collaboration with de.NBI/ELIXIR-DE.

5.7.5 Measure B3.M5: Assured Training Scheme

Consortium Member	Contribution
Parker (DKFZ)	Team Lead, PI for Assured

Goals: Develop a nationally recognised training and accreditation scheme (entitled: 'Assured') for researchers and research data professionals handling sensitive research data in Germany. Assured will be developed in collaboration with other NFDI Consortia (KonsortSWD and BERD@NFDI). It will comprise of a core set of training modules and assessments with additional materials available for various specialisations. The development of Assured will have a number of benefits: it will support researchers to learn the skills required to handle sensitive research data more easily and facilitate their access to it. It will help Research Data centres (RDCs) train staff involved in the use of sensitive research data and offer them a career progression path, reduce the risk of misuse of the data they store, and encourage data producers to share the data they create by increasing trust in the RDC. It will also integrate aspects of ethical considerations, developed together with the ethics team (B5.M3). Assured will also be linked to an existing externally developed Authentication and Authorisation Infrastructure (AAI) so that those who have undergone training can easily 'carry' their accreditation alongside their identity.

This measure will support the development of the GHGA Data Infrastructure in two ways, primarily with regards to the GHGA SPE phase of the project. Providing training for researchers is an important part of a data security model as the provision of training will

reduce the likelihood that users of GHGA SPE will produce disclosive results. Secondly, through the use of accreditation and modules that relate to data access professionals, GHGA can train staff who can operate GHGA SPE safely and staff will be able to obtain transferable skills. In doing so, Assured will support B3.M4 by providing a standardised framework and materials for staff training in this area.

In addition, it may be beneficial to offer Assured training to data requesters before the introduction of GHGA SPE as a means to facilitate the safe use of human omics data and encourage data controllers to share data via GHGA. In the current funding period, the work of Assured is supported by flex funds which have been used to recruit a Training Coordinator for Data Protection who has contributed to the development and enhancement of Assured's core training materials, and the creation of a brand identity for the project. In the next funding period, Assured will launch the first version of its core training modules targeting researchers and research data access professionals, produce modules covering specific roles, RDCs, and data types, and develop those aimed at other stakeholders. In addition, the link to an AAI will be implemented. Assured will also be adopted by a number of RDCs as their required training for researchers. We will also explore achieving international recognition for Assured, such that researchers who have trained under Assured can have their accreditation recognised outside of Germany.

5.7.6 Measure B3.M6: User Experience

Consortium Member	Contribution
Kohlbacher (EKUT)	Coordination of the measure
Stegle (DKFZ)	Coordination of the measure, connection to GDI
Huber (EMBL)	Coordination of the measure with training
Kirli (DKFZ)	Team Lead, Product Management

Goals: Gaining a better understanding of what GHGA users need and expect from our infrastructure will enable GHGA to enhance its products and services (cf. A3.M1), supporting a safe and successful user experience resulting in increased trust in the GHGA infrastructure in the target communities. Assessing user needs and UX metrics: B3.M6 will focus on surveying users of GHGA in order to provide information for the improvement and further development of the services offered by GHGA, which will be implemented by A3, A4, B2, B3 and other TAs. The topics to be investigated include, but are not limited to, (meta)data preparation for submission, the data up- and download processes, training needs, materials and events, as well as public engagement activities. UX research makes it possible to involve the user community in the continuous development process of our services, ensuring they meet the needs of the communities. This will enable community-orientation of GHGA services and build-up of trust in our infrastructure. In an iterative process, throughout the course of the GHGA project, we will do extensive research into the needs and characteristics of our target audiences by collecting data on use,

expectation, perception, and evaluation of GHGA services and tools. We will apply a mixed-methods approach, combining quantitative and qualitative methods such as different types of surveys, expert interviews, and document analysis among others. By assessing our services and better knowing our users, we will obtain valuable information regarding the desired functioning of the GHGA infrastructure that we will use for (i) feature optimisation, (ii) development of new services, (iii) improvement of training and communication strategies, and (iv) addressing emerging challenges in our user community.

Providing improved user documentation: We will further develop an online GHGA User Forum connected to the existing User Documentation that enables an ongoing exchange between different GHGA users and GHGA. The forum, similar to the Bioconductor forum or Stackoverflow, will address users' needs to ask questions not easily answered by the training material and the online help system. The forum will provide added value compared to the Helpdesk for multiple reasons, especially in rapidly growing a collaborative user community: a diversity of experts will be able to contribute answers, users other than the one asking the question can also benefit from the discussion (complex questions e.g. about bioinformatics analyses benefit from a discussion with multiple users), and the question and answers are automatically publicly archived and become findable with search engines. The forum, together with a condensed FAQ section, will establish a continuously growing online help and reference system supporting users of GHGA enabling us to improve the overall user experience.

Tasks and Deliverables

Task	Deliverables	Due Date
B3.M1.T1 Creating fact sheets and further information material	Fact sheets and other media for internal and external use published	M24
B3.M1.T2 Connecting GHGA with its users	Annual Meetings and GHGA User symposium held	M12, M24, …, M60
B3.M2.T1 Engaging with young scientists	First lecture held	M3
B3.M2.T2 Engaging with disease communities	GHGA workshops or other event held for core communities, human genetics and new communities	M6, M12, …, M60
B3.M2.T3 Enhancing national and international visibility of GHGA	Publications targeting different communities (clinical, scientific and public)	M36
B3.M2.T4 Engaging with NFDI communities	Regular participation and active contribution to NFDI conferences	M6, M12, …, M60
B3.M3.T1 Informing and engaging with the public on data sharing	Published training material on data sharing for patients and medical staff	M36
B3.M3.T2 Increasing public trust in FAIR data sharing	First monthly podcast released	M1
B3.M3.T4 Fostering public attention by press releases	Two publications in print media with national reach	M48
B3.M3.T5 Updating of the patient specific website	Regular publication of new content on patient specific website	M6, M12, …, M60
B3.M3.T6 Developing of a language guide on personal data sharing	Publication of language guide	M12

Task	Deliverables	Due Date
B3.M4.T1 Distributing GHGA training material in a FAIR format	Release of training platform linked to ELIXIR TeSS	M10
B3.M4.T2 Training of users and staff for working with the GHGA Data Infrastructure	Updated User Documentation and regular workshops with collaborators (four-monthly) and internal training courses (biannually)	M6, M12, ..., M60
B3.M5.T1 Training staff for safe operation of GHGA SPE	Implementation of Assured within GHGA SPE (depending on development of GHGA SPE)	M12
B3.M5.T2 Rolling out of Assured training programme nationally	Launch of a national Assured network of supporting RDCs	M12
B3.M5.T3 Extending Assured modules to additional audiences	Materials related to additional audiences: service managers and data producers	M19
B3.M5.T4 Achieving international recognition of Assured	Mutual recognition between Assured and a similar training initiative outside of Germany (i.e. Safe Researcher Training in the UK) and mutual recognition across multiple international partners	M28-M36
B3.M6.T1 Researching user requirements	Concept published to better address the different target groups: an essential document for optimisation of the training strategies and for adaptation of current services or/and creation of new services.	M18
B3.M6.T2 Implementing user feedback	SOP developed to distribute user feedback to relevant workstreams and integrate into infrastructure development	M30
B3.M6.T3 Establishing feedback loop for user empowerment	User Forum launched	M48

Dependencies, Interactions, Risks, and Mitigation Strategies

Dependencies: Communication, Outreach and Training TAs are tightly interconnected, with a continuous exchange about activities and materials, and a coordinated use of communication platforms. The success of these projects strongly depends on the development of the communication platforms, and integrating and promoting activities contributed by other TAs. Contents and materials developed by other TAs will be used for communication, outreach and training, and materials will be continuously adjusted and aligned with new developments, particularly when approaching new communities. **Interactions**: The task area interacts with other NFDIs to ensure that the GHGA's communication strategy aligns with the global NFDI communication goals, and with the NFDI section EduTrain to align with other NFDIs on best training practices and standards. As part of the MV GenomSeq, GHGA is closely tied to the Genome Data Centres. In this role, GHGA's overall communication strategy and messages will be closely linked with, complementary to, and aligned with, BfArM and the BMG as necessary. In addition, GHGA's overall communication strategy and messaging is closely linked and complementary to MV GenomSeq (TMF, BfArM, RKI) given that we address the same target group on overlapping aspects of the topic of genomic research. **Risks & mitigation:** The main risks for this task area are (i) the timely availability of needed workflows and SPE as well as (ii) the timely fitness of GHGA to accommodate the archival of analysis-enriched national reference

datasets. The identified risks will be mitigated through a profound and proactive collaboration with task areas B2 and B3 for community engagement and data services as well as A1 and A2 for respectively needed central and local operations adaptations.

Justification of Requested Funds

As outlined in 7.7, activities in this TA need to be supported by several positions experienced in stakeholder communication and training (5.8 FTE from DFG, 1 FTE from own contributions). Activities will be coordinated by the Team Lead Communication and Training (cf. Team Structure, full position at DKFZ), who will coordinate the TA and lead the local team responsible for patient engagement (1 FTE 50% DFG/50% DKFZ own contribution, B3.M3) and the Assured programme (B3.M5, 80% position, together with the Team Lead Data Protection cf. C2). B3.M2 will be supported via two positions at HMGU (Winkelmann) and UDS (Walter / Schulze-Henrich), ensuring broad connectivity into pivotal clinical and scientific communities. UKT (Graessner) will continue to produce audio-visual materials supported by one FTE (50% DFG, 50% UKT own contribution). A dedicated training coordinator at EKUT will be responsible for B3.M4 and all other training related activities.

5.8 TA B4: National and International Connectivity and Metadata Alignment

Overview of the Task Area

This task area aims to strengthen GHGA's integration into the national and international research community and align its activities with related infrastructures. One key aspect of these interactions is the alignment of metadata models to improve interoperability and thus FAIR-ness of the data managed in GHGA. The GHGA metadata model will be extended to facilitate the deposition of multi-omics datasets, and to capture technical metadata for proteomics, epigenomics, as well as single-cell and spatial-omics data. We will engage with key stakeholders, both nationally and internationally, to exchange harmonised metadata and thus increase the reach of the data archived via GHGA. This comprehensive alignment effort will ultimately position GHGA as a hub for genomics data.

5.8.1 Measure B4.M1: National Alignment within the NFDI and Beyond

Consortium Member	Contribution
Nahnsen (EKUT)	Coordination of NFDI alignment
Korbel (EMBL)	Coordination beyond NFDIs
Ulas (DZNE)	Alignment of NFDI base services
Kohlbacher (EKUT)	Interaction with MII
Fluck (ZBMED)	Alignment with NFDI4Health
McHardy (HZI) & Korbel (EMBL)	Alignment with NFDI4Microbiota
Eufinger (DKFZ)	Team Lead Administration, NFDI Networking

Goals: Tighter integration of GHGA with NFDI consortia in the life sciences, metadata-related Base4NFDI services, and other national initiatives. Integration with NFDI4Health: NFDI4Health is the NFDI consortium most closely related to GHGA due to its focus on health science and clinical studies. It has established a metadata catalogue

FAIR-ifying study data. By harmonising the metadata model of GHGA and the metadata schema of NFDI4Health the automated exchange of metadata between GHGA and NFDI4Health will be enabled, thereby integrating GHGA studies into the NFDI4Health study portal. This will make GHGA data findable in a different context while simultaneously increasing the completeness of the data in NFDI4Health. The implementation of this integration has been agreed upon with NFDI4Health in a recently published Memorandum of Understanding (5). **Multi-omics and cross-modal data deposition:** We will enable data deposition for multi-omics on the technical side through expansion of the metadata models (see below, B4.M3), but driven through direct collaborations. Together with other NFDI consortia, we will enable the deposition of muti-modal datasets to be managed across NFDI infrastructures. NFDI4Immuno aims to build a FAIR-based federated repository for all data describing the state of the immune system. In a collaborative effort with the NFDI4Immuno, GHGA will provide the infrastructure for managing the corresponding human omics data. As part of this collaboration, we will also establish solutions to link human omics data with bespoke immunological data (e.g., clinical phenotypes and FACS data; Nahnsen is involved in both consortia). Similarly, with NFDI4BIOIMAGE we intend to enable linked submissions, where sequencing data from multi-omics data is stored and GHGA, whereas the imaging data are held in NFDI4BIOMAGE. Aligned metadata models and linkage of accessions will enable mutual discoverability, and thus expose data from the same samples to both the omics and imaging communities (cf. also B2.M3). With NFDI4Microbiota, we will similarly begin the alignment of metadata to gain consensus on the development of standards for the human and microbiome data types. Furthermore, we are tightly collaborating with the data platform of the German Centre for Infection Research (DZIF), where all infection-related host data (human) will be stored in GHGA. Here, in collaboration with the DZIF consortium, we will establish the linkage between host omics data and microbiological data. We will also offer services to execute workflows developed by NFDI4Immuno and other NFDIs (pipelines for AIRR - Adaptive Immunoreceptor Repertoire data) within the GHGA SPE. Outside the NFDI, we will collaborate with the Network University Medicine (NUM) to align metadata standards and deposit the NAPKON omics data (genomics, transcriptomics, epigenomics, proteomics, metabolomics of 2,000 COVID-19 patients).

Alignment with the Base4NFDI services TS4NFDI and KGI4NFDI: We will consequently incorporate base services to improve the handling and interoperability of metadata. Adoption of the TS4NFDI terminology service will support the automated curation, harmonisation, and mapping of GHGA terminologies. Similarly, we will adopt standardised data vocabularies into GHGA and implement Beacon interoperability. The adoption of Base4NFDI services, such as the Terminology Service (TS4NFDI) and Knowledge Graph Infrastructure services (KGI4NFDI), will also aid the alignment of metadata schemas between GHGA and

NFDI4Health. Building on Base4NFDI services, GHGA and NFDI4Health will establish a knowledge graph registry of metadata schemas, which will accelerate the integration between the two consortia. **Alignment with MV GenomSeq and the Medical Informatics Initiative:** GHGA will continue to proactively engage with the working groups of BfArM to define (and develop) the metadata standards for MV GenomSeq, which are essential for metadata transfer of research-consented data to GHGA (cf. B4.M3 and A2.M4). This activity will be synergistic with our involvement in the interoperability working group of the Medical Informatics Initiative, specifically the task forces for oncological data and rare disease data. The continued participation in these national forums will ensure that GHGA metadata remains interoperable with national standards for structuring clinical data. Visible roles in these working groups of GHGA co-spokespersons, participants and team leads will facilitate the necessary alignment, which will feed into the GHGA metadata implementation (B4.M3).

5.8.2 Measure B4.M2: International Alignment

Consortium Member	Contribution
Korbel (EMBL)	Co-spokesperson of the TA, coordination of B3.M2
Nahnsen (EKUT)	Co-spokesperson of the TA, coordination
Bork (EMBL)	Coordination with EOSC
Kohlbacher (EKUT)	Coordination and alignment with GDI, FEGA, EOSC
Stegle (DKFZ)	Coordination and alignment with GDI, B1MG, FEGA, GA4GH
Lablans (DKFZ)	Alignment with BBMRI-ERIC
Keane (EMBL-EBI)	Alignment with EGA network
Eufinger (DKFZ)	Team Lead Administration, International Networking

Goals: Alignment with international activities and infrastructures to ensure coherence of metadata schema and standards. Alignment with FEGA: The Federated European Genome Phenome Archive (FEGA), a cornerstone resource for discovery and access of sensitive human omics and associated data consented for secondary use, provides a unique opportunity for GHGA to amplify its international impact and reach. Consequently, GHGA has joined this effort as one the founding international FEGA nodes. In order to be able to adhere to German data protection (which is sensitive to the possibility of personal information contained in metadata), and to enable parallel metadata exchange with FEGA and additional European networks, we have decided to not adopt the local EGA software but instead build our own solutions and integrate with FEGA via common interfaces and metadata exchange (c.f. 4.4), which are currently being tested. Interoperability with FEGA metadata standards will be paramount to secure our role in this network. We therefore aim at extremely close engagement and alignment with FEGA, by attending regular consortium meetings, providing frequent updates on the progress of FEGA Germany and exchanging and aligning concepts for standards and data. Should any important conceptual gaps be revealed through this international alignment activity, these will be tackled swiftly with assistance of the workforce of B4.M2. The GHGA team is actively conducting meetings/workshops with national and international stakeholders such as GA4GH, NAKO,

ENA, Beacon, GDI to obtain feedback on the ongoing efforts related to schema and model development, maintenance and interoperability and how the model can be best leveraged across national and international consortia. **Alignment with GDI/1+MG:** GDI and 1+MG have initiated activities gearing up to establish an infrastructure to share human genomic data in Europe. These projects are developing important concepts to realise secure access to genomics and the corresponding clinical data across Europe in the future and involve a large number of European stakeholders. For this reason, B4.M2 will ensure active alignment with GDI/1+MG, learning from ongoing discussions, and informing the European community about our advances and the formats and standards used. We strive to establish full metadata interoperability with GDI, thus enabling GHGA data to be exposed in FEGA and GDI. **Alignment with GA4GH and GSC:** The Global Alliance for Genomics and Health (GA4GH) and the Genomic Standards Consortium (GSC) are two internationally renowned organisations dedicated to harmonising data and metadata standards across the global genomics community. GHGA will actively participate in the meetings and working groups of these consortia to present our models and engage with their communities. Our involvement will serve dual-purpose: positioning GHGA as a valued partner, and leveraging the best practices from these leading organisations. By integrating the insights and standards developed by GA4GH and GSC, GHGA will ensure adherence to internationally accepted standards, thereby enhancing the interoperability and reliability of our data. This alignment will be facilitated by our membership in the national initiatives forum of GA4GH. This strategic alignment will not only bolster GHGA's credibility on the global stage, but also ensure that our metadata schema and standards are at the forefront of international genomic data sharing initiatives. **Alignment with EOSC:** Integration into the European Open Science Cloud (EOSC) ecosystem will significantly enhance the impact of GHGA on the scientific community. Firstly, as a member of the EOSC Association (EOSC-A), EMBL plays a crucial role in this integration. EMBL personnel involved in GHGA will participate in EOSC member assemblies and managerial meetings, task forces such as the Health Data Task Force of EOSC-A, ensuring we stay on top of the latest developments and strategic plans within EOSC. This alignment will facilitate the harmonisation of standards and metadata formats at the European level, benefiting GHGA and its stakeholders. Secondly, in the initial GHGA funding phase, the FAIR-IMPACT initiative has been instrumental, and GHGA has actively engaged with it to promote the adoption of FAIR principles. A GHGA representative has been selected among the twelve onboarded EOSC FAIR Champions, who act as ambassadors to engage their communities, advocate for project results, and facilitate national roadshows in Germany. This representative will play a pivotal role in analysing and shaping FAIR data policies and practices, identifying technological gaps, and developing standards for data management across scientific disciplines. Through these efforts, GHGA

aims to customise FAIR data practices, organise national workshops, and advocate for FAIR principles through webinars and community outreach. By aligning with EOSC's ambition to develop a 'Web of FAIR Data and Services' for science in Europe, GHGA will ensure that its datasets are more FAIR-compliant, thereby enhancing their discoverability and usability.

5.8.3 Measure B4.M3: Metadata Model Maintenance, Development and Alignment

Consortium Member	Contribution
Nahnsen (EKUT)	B4.M3 coordination and metadata development
Korbel (EMBL)	International reachout
Ulas and Schultze (DZNE)	Coordination of metadata development
Robinson (BIH)	Integration of Phenopackets standard
Kohlbacher (EKUT)	Interaction with PRIDE
Gagneur (TUM)	Data processing beyond genomics
Lablans (DKFZ)	Record Linkage Concepts
Kirli (DKFZ)	Team Lead Product Management
Menges (DKFZ)	Team Lead Data Stewardship

Goals: Maintain and further develop the GHGA metadata model while keeping it aligned with the national and international activities. Maintenance of the metadata model: The GHGA metadata model is fully open-source and implemented using the Linked Data Modelling Language (LinkML). The schema is available as a YAML file in our public GHGA Metadata GitHub repository and serves as the single source of truth for services related to metadata handling, auto-generation of submission spreadsheets, or validation. To facilitate future maintenance, the metadata model will be migrated to schemapack, a GHGA-developed linked data modelling framework (see 4.2). The GHGA User Documentation provides detailed information, such as descriptions of the entities and attributes and is compatible with the EGA metadata model. To ensure interoperability, single slots in the model are controlled using community-accepted ontologies, such as HPO or DUO, and customised controlled vocabularies. **Alignment with FEGA and GDI:** We will ensure that the GHGA metadata model is compatible with a minimal FEGA model. The FEGA Metadata Working Group, with GHGA's active participation, aims to establish standards for metadata to ensure FAIR data principles are met. The group will define what constitutes public versus non-public metadata, and customise the model for specific FEGA node requirements, which will be incorporated into the GHGA metadata model. Workshops with FEGA colleagues will be organised to create necessary user documentation and identify gaps in the metadata model. Similar modes of interactions are planned with GDI, and we will play an active role in the ongoing definition of the first GDI metadata model. Active collaboration with key stakeholders in FEGA and GDI, and in particular nodes who are active in both networks, will ensure alignment and compatibility between the GHGA metadata model and these networks. We will interact with the B4.M1 and B4.M2 to harmonise the national and international harmonisation efforts by defining shared goals that are laid out in shared roadmaps. **Extension to multi-omics and other omics modalities:** The metadata

model will be continually improved, aligned with the release cycle of the GHGA software components (e.g., portal, backend data services, data stewardship toolkit) and the activities described in B4.M1 and B4.M2, and integrated in the update, test, and release cycle of these components (developed in A3). As part of the metadata extension, we will extend the model to describe multi-omics datasets, e.g. combining genomics with other omics modalities (e.g., transcriptome, proteome, etc), but also imaging data (see B4.M1). This enhancement will open the door for broader collaboration with other consortia and foster the inclusion of diverse data types (e.g. proteome) and also facilitate multi-omics data processing (collaboration with B1 and B2). A collaboration with the PRIDE Archive (EMBL-EBI) and FEGA Spain (CRG) has already been initiated to align different models for controlled access to human proteomics data within FEGA and PRIDE. **Extension for structured clinical data:** We will expand the GHGA metadata model to align with the NFDI4Health schema, as well as with other critical standards such as MV GenomSeq, HL7/FHIR, MII KDS, and the Phenopackets standards. By identifying commonalities and integrating these standards, we will develop a unified metadata framework that supports comprehensive data interoperability. The initiative includes strengthening collaboration with the NFDI4Health Metadata Working Group through workshops and co-authoring a white paper on the integration process. Additionally, we will leverage the expertise of GHGA participant Peter Robinson to align with the Phenopackets standards, which is widely used in rare disease, ensuring robust representation of phenotypic data. We will also include the necessary extensions to support record linkage across data infrastructures (privacy- preserving record linkage or direct linking via identifiers) that will be essential to reference data across infrastructures (cf. TAs B). This expanded metadata model will enhance the usability of GHGA data for research and healthcare applications.

Tasks and Deliverables

Task	Deliverables	Due Date
B4.M1.T1 Mapping and alignment of GHGA metadata schema with NFDI4Health schema	Fully aligned metadata schemata announced	M6
B4.M1.T2 Automatic integration of GHGA metadata into NFDI4Health study portal	First pilot data set tested and integrated	M18
B4.M1.T3 Integrating data sets from other platforms in GHGA Archive	First NFDI4Immuno, BIOIMAGE, DZIF data sets listed in GHGA Archive	M36
B4.M1.T4 Aligning with the Base4NFDI services TS4NFDI and KG4NFDI	Entry in knowledge graph registry listed	M12
B4.M1.T5 Aligning metadata schemata with MV GenomSeq and MII	Joint metadata concept published	M12
B4.M2.T1 Identifying metadata related gaps and extending support to other FEGA nodes	Regular EGA working group meetings attended	M6, M18
B4.M2.T2 Foster compatibility between EGA and GHGA metadata models	Workshop on metadata alignment hold	M18
B4.M2.T3 Engaging with international initiatives for metadata standard alignment	Attendance at GDI/1+MG, GA4GH and GSC working group meetings	M6, M12, …, M60

Task	Deliverables	Due Date
B4.M2.T4 Promoting FAIR principles via the EOSC network	First participation in EOSC governance meeting	M6
B4.M2.T5 Establishing metadata link to GDI	Metadata exchange w. GDI established	M24
B4.M3.T1 Migration from LinkML-based metadata model to schemapack	Metadata model fully migrated	M6
B4.M3.T2 Curating and maintaining of the metadata model	Update on metadata model released and communicated to GHGA users	M48
B4.M3.T3 Integrating of other omics and imaging datasets through the introduction of abstract classes	First multi-omics metadata model operational	M12
B4.M3.T4 Publishing a unified metadata framework with other data initiatives	White Paper published	M48

Dependencies, Interactions, Risks, and Mitigation Strategies

Interactions: We intend to interact with TA A1 to A4 on metadata model development, architecture and data stewardship. Further with B2 B3 on ELSI and workflows and with C2 on training. The task area will be the area to interact with other NFDIs and international initiatives. **Dependencies:** Our goals depend on the progress of A1, A3, A4, B2, and C2. **Risks and mitigation:** A significant risk for this task area is the lack of interoperability with other NFDIs (e.g. NFDI4Health). This may arise if metadata models are developed in full independence and without alignment. Furthermore, we see additional risks if the interactions with EOSC, EGA, and the 1+MG initiative are reduced. For the mitigation of these risks, our measures include strategies to prioritise close interaction with all initiatives and foster collaboration from both the operational and the TA Lead level.

Justification of Requested Funds

As outlined in 7.8, we will be applying for two positions for this TA. One position will be hired as a metadata officer (EKUT, Nahnsen) and one position will oversee national and international connectivity of GHGA (EMBL, Korbel). Both positions will be cross-connected to the overall consortium, especially with A3, all B Areas, and C2.

5.9 TA B5: Legal and Ethical Issues

Overview of the Task Area

B5 develops the governance framework for GHGA. Building on the results from the first funding period, B5 will continue to provide ethical and legal guidance, develop ethics training for GHGA staff and other professionals working with genome data, explore ways to sustainably link GHGA to patient groups, and address key issues to prepare GHGA for legal alignment with the European Health Data Space (EHDS). These include compliance with EHDS rules and standards, preparation for integration with EHDS infrastructure, and adherence to national requirements.

5.9.1 Measure B5.M1: Legal Advice / EHDS and GHGA

Consortium Member	Contribution
Molnár-Gábor (UHD)	Lead PI
Parker (DKFZ)	Team Lead Data Protection and Legal

Goal: Continuously align the legal basis for GHGA's operation to changing legal requirements and ensure robust legal compliance. It is important to regularly review GHGA's current research infrastructure to assess whether the planned aspects of storage and data processing have been realised as intended. This will enable alignment with the European Health Data Space (EHDS), including international data transfer. Alignment requires certain characteristics of an entity. For this reason, it is necessary to check whether GHGA fulfils these requirements and which conditions must be met. The review will compare the current state of GHGA's infrastructure with the plans outlined in official documents such as the Project Proposal and Governance document. This includes examining data management practices, security measures, legal and ethical frameworks, and interoperability with European initiatives. The outcomes of this review will be vital for identifying areas that may need adaptation to align with further emerging data infrastructures at European and international level. The first important consideration is whether GHGA requires its own legal personality to align with the EHDS. If it does not, is obtaining legal personality still necessary to facilitate effective alignment with the EHDS? Second, it is essential to evaluate whether GHGA's role as a designated data processor for research data and personal metadata aligns with the provisions outlined in the EHDS Regulation. A further crucial point to examine is whether GHGA meets the criteria of a 'data holder' as defined in the EHDS Regulation (Art. 2(2)(y) EHDS-P). If so, this classification would entail various obligations that GHGA must fulfil. Additional duties as outlined in Art. 41 EHDS-P will be assessed. By proactively assessing its position and potential responsibilities, GHGA can implement necessary changes to ensure compliance, avoid penalties, and position itself as a responsible participant in the European health data ecosystem. This approach not only mitigates risks but also enhances GHGA's reputation and preparedness to contribute effectively within the EHDS framework.

5.9.2. Measure B5.M2: Legal Embedding in the National Infrastructure (GDNG)

Consortium Member	Contribution
Molnár-Gábor (UHD)	Lead PI
Parker (DKFZ)	Team Lead Data Protection and Legal

Goal: Continuous alignment of the legal basis of GHGA with evolving national and European legislation. The recently adopted Health Data Utilisation Act (Gesundheits-datennutzungsgesetz, (GDNG)) is a first step towards the national realisation of the EHDS. The already applicable GDNG will be examined and the implications for GHGA assessed. Additionally, the necessary changes in data protection within the infrastructure of GHGA will have to be initiated. Emerging legislation such as the planned Research Data Act (FDG, (Forschungsdatengesetz)) will be observed. The emerging legislation on the national health data infrastructure will make use of possible regulatory discretion of the national legislator. Additionally, it will be up to the national legislator to realise the directly binding provisions of

the EHDS Regulation and fit it into its national administrative infrastructure. It is of decisive importance to investigate the related decisions of the national legislator to assess the compliance options and roles for GHGA in the national infrastructure.

As the legislation is not yet finalised and is expected to emerge throughout the second funding phase of the NFDI, this action includes a dynamic exchange with policy makers and legislators, a comparative legal analysis following the evolution of the implementation of EU law in other Member States, and the development of a forward-looking governance and legal framework for the GHGA that can be adapted to emerging compliance requirements. Due to this situation, a forward-looking legal analysis is required.

This proactive approach will enhance GHGA's ability to contribute to, and benefit from, broader European genomic data initiatives, advancing its mission of facilitating secure and legally compliant genomic data sharing on the national level. Regarding the FDG and GDNG, we will also ensure close exchange within the NFDI as these legislations have strong relevance for many other consortia.

5.9.3 Measure B5.M3: Integrated Ethics

Consortium Member	Contribution
Eva C Winkler (UHH)	PI
Schickhardt (NCT-HD / DKFZ)	Advice on ethical governance
Bruns (UHH)	Coordination of ethics activities

Goal: Develop an ethics framework and training course to help GHGA staff and other professionals working with genome data to better identify and address ethical issues related to their work and to make ethics part of the everyday language that technical and scientific staff are comfortable using. There is a growing consensus that the complexity and subtlety of ethical issues related to sensitive research data require a unique set of skills in ethical analysis and reflection. B5.M3 aims to support GHGA staff and other professionals working with genome data in acquiring those skills. By the end of the current funding period, we will have developed ethical guidelines on informed consent, standards for Data Access Committees, and a code of ethics for GHGA. Building on this work, B5.M3 has three main objectives: (i) Develop and maintain an **ethics framework** that includes an introduction to the ethical issues around genome data, a mission statement and code of ethics for GHGA, and guidelines on specific topics. Apart from integrating the existing guidelines, B5.M3 will also look into further ethical issues, such as alternative ethical and legal bases beyond opt-in consent, social justice, and sustainability and environmental impact. (ii) B5.M3 will also develop a **complementary training course** targeted at GHGA staff and other professionals working with genome data that introduces participants to the ethics of data generally and the GHGA guidelines and code of ethics specifically (alongside B3.M4 and B3.M5). Discussing fictional case studies will allow participants to acquire genuine skills in ethical analysis and reflection. The training course will be provided on a

regular basis (e.g., annually) and evaluated to better track changes in participants' skill to address ethical questions. (iii) Finally, B5.M3 will aim to strengthen GHGA's position as a **global contributor to debates in the ethics of genome and research data**. B5.M3 will allow GHGA staff to continue leading the Task Force Ethics within the ELSA Section of the NFDI and to keep contributing to international policy making as well as academic debates through peer-reviewed publications. In particular, B5.M3 will intensify collaboration with the Regulatory and Ethics Workstream of GA4GH to achieve better international alignment and greater visibility of GHGA's work in this area.

5.9.4 Measure B5.M4: Patient Engagement

Consortium Member	Contribution
Winkler (UHH)	Lead PI
Schickhardt (NCT HD / DKFZ)	Ethical considerations
Bruns (UHH)	Coordination of ethics activities
Eufinger (DKFZ)	Team Lead, Governance coordination
Träger (DKFZ)	Team Lead Communications, Patient Communication

Goal: Identify sustainable ways to involve diverse patient groups in the governance of GHGA and link GHGA to existing and emerging patient engagement structures. In the current funding period, GHGA has conducted the PaGODA study [15] to explore the needs and interests of patients with regard to the sharing and use of their genome data for research. A key outcome of the study was that patient advisors and Patient Advisory Boards were deemed the best way to involve patients in the governance of data infrastructures such as GHGA. Following up on this outcome, B5.M4 has two main objectives: (i) To define the best ways for GHGA to sustainably involve patients, either through working together with existing Patient Advisory Boards (e.g., at DKFZ and NCT) or recruiting new patient advisors to link GHGA to its target patient communities (cancer and rare diseases) and possibly other patient communities involved in genomics. In this regard, B5.M4 will also focus on linking GHGA's efforts to emerging patient engagement structures, such as in MV GenomSeq. (ii) To work together with patient advisors to develop key performance indicators (KPIs) to continuously evaluate GHGA's patient engagement strategy (B3.M3) and identify conditions for what should count as successful patient engagement. On this basis, B5.M4 will develop an advanced patient engagement concept for GHGA. B5.M4 will allow GHGA to contribute to the setting of high standards and establishment of best practices regarding the involvement of patients in research data infrastructures in Germany and beyond. Meaningful and sustainable patient engagement will also contribute to the trustworthiness and social acceptability of GHGA and strengthen GHGA's position as a project in the public interest.

Tasks and Deliverables

Task	Deliverables	Due Date
B5.M1.T1 Reviewing the GHGA Data Infrastructure	Legal evaluation of the current infrastructure completed	M6

Task	Deliverables	Due Date
B5.M1.T2 Adapting GHGA's legal framework to EHDS requirements	Compliance measures developed to align with EHDS (including the adaptation of the contractual and governance framework)	M24
B5.M1.T3 Extending legal framework to international data sharing	Governance framework for international data sharing developed	M30
B5.M2.T1 Aligning with national legislation	Compliance measures developed to align with emerging national legislation	M36
B5.M2.T2 Developing forward looking governance model	Risk governance framework for the alignment with existing infrastructures finalised (1+MG, FEGA)	M48
B5.M2.T3 Interacting with policy makers for promoting GHGA's mission	Ongoing communication with policymakers, legal comparison across EU Member States, adaptive contractual framing	M60
B5.M3.T1 Developing ethics framework for GHGA staff and professionals	First version of ethics framework completed	M12
B5.M3.T2 Developing training on applied ethics for the use of personal data	Training materials and course design completed	M24
B5.M3.T3 Contributing to global ethical standards	First peer-reviewed article published	M24
B5.M4.T1 Developing KPIs for effective patient engagement	KPIs and evaluation of patient engagement strategy completed	M24
B5.M4.T2 Creating an advanced patient engagement concept	Advanced patient engagement concept completed and published	M36

Dependencies, Interactions, Risks, and Mitigation Strategies

In B5.M1 interaction is planned with A1, A2, B1, and B2 in order to ascertain an appropriate representation of the actual state and development of GHGA on all levels. The legal measure B5.M2 will also be based on close cooperation with the operations TAs (A1, A2). Their work for the further development and establishment of GHGA will be aligned with all actions in B5.M1 and B5.M2. B1 and B2 will inform the integration (B5.M1), B4 will guide international management and governance (B5.M1 and B5.M2). Measures B5.M3 and B5.M4 will closely interact with B3 to align GHGA's patient engagement concept and strategy with GHGA's patient communication activities and to provide training to GHGA staff. B5.M3 will also interact with A4 to identify the training needs of data stewards and technical staff and align with the training provided through B3.M5. Moreover, B5.M3 will contribute to national alignment with the NFDI (B4.M1) and international alignment (B4.M2), in particular with GA4GH, regarding the ethical governance of research data.

The risk of unknown and unexpected factors affecting the link to the EHDS and the national infrastructure will be mitigated through forward-looking legal analyses and comparative law between Member States, as well as stakeholder consultation with EU and national policy makers.

Justification of Requested Funds

As outlined in 7.9, we will be applying for one position for the legal measures (B5.M1 and B5.M2, UHD, Molnar-Gabor), one position to continue the ethical work in GHGA and NFDI (B5.M3, UHH, Winkler) and half a position to conceptualise and coordinate the patient engagement efforts (B5.M4).

5.10 TA C1: Flex Funds

Overview of the Task Area

Research data infrastructures have to react to changing needs and thus not all necessary funding measures can be foreseen. We have set aside approximately 10% of the overall DFG funding (equivalent to five FTE) to allow agile and timely reactions to new developments and external impulses. These funds will be complemented by unspent funds, e.g. due to delays in the recruitment at institutions, in order to maximise the number of additional developments that can be implemented.

In the previous funding phase, we have successfully established this model through annual calls for Innovation & Implementation Projects (IIPs), for which both co-spokespersons and participants of GHGA could apply equally. The focus of these projects was on closing gaps (partially due to the incurred budget cuts, cf. 3.4.2) and connecting GHGA to other infrastructures and projects. In the next funding round, with the GHGA Data Infrastructure having reached a maturation level, we will split this measure up into three separate types of measures, aiming to ensure (i) the capacity to adapt to changing needs in an agile way and (ii) the continuous embedding of GHGA into the research community.

Title	Scope / Budget estimates	Eligibility	Selection process
Innovation and Implementation Projects (IIP)	Extension of functionalities and services of the GHGA project / Funding for positions between 0.5 and 2 years	- GHGA Participants - External researchers aiming to become engaged in GHGA	Public Call Internal review and decision by GHGA SC
Data Mobilisation grants (DMG)	Supporting submission of high value datasets, e.g. for metadata preparation, adaptation of data management / LIMS system to extract GHGA metadata. 5 - 15 k€	- GHGA Participants - External researchers aiming to engage in GHGA	Public Call Internal review and decision by GHGA SC
Internal Flex Funds (IFF)	Agile support for existing GHGA projects in cases of funding shortages, e.g. event organisation, temporary shortages in personnel costs at a member institution / 10k€ max. for a given year	- Only for current GHGA Participants	Internal Call Decision by the GHGA BoD, regular reporting to GHGA SC

A transparent selection process ensures a fair allocation of the funds based on their relevance to community and consortium. The TA is managed by the BoD supported by the GHGA Office (C2), therefore now individual contributions are listed with the measures.

5.10.1 Measure C1.M1: Innovation & Implementation Projects (IIP)

Goals: Selection of innovative projects to support new functionalities and services.

Annual calls for 'Innovation & Implementation Projects' (*IIPs, Flex Funds*) will be sent out via the GHGA website and newsletters (including NFDI and other external newsletters) and brief project descriptions will be collected in a standardised format. Projects should be planned for a short to intermediate period of time (3 - 24 months), with a budget of up to 150,000 €. We

expect typical projects to request less. As of July 2023, 21 Flex Funds projects have been funded with a total volume of 1.2 M€. Topics of the projects range from support for the establishment of ties to other NFDIs, community outreach measures, the development of data protection, legal and training materials, as well as reinforcements of existing measures to compensate for increased needs of resources. Details about these supported proposals have been provided in our recent progress report.

Ranking of these projects will be performed by the BoD and SC (with consideration of conflicts of interest) with the aid (where necessary) of external reviewers or our SAB. The call and review process will be organised by the GHGA Office (C2). Details of this process have been established by the GHGA Office and include (i) the opening of the call to all GHGA members, (ii) collection of short descriptions of the envisioned projects, (iii) internal discussions of the proposal by the BoD, SC and other involved GHGA members, (iv) external consultation where required, (v) if necessary adaptations of the projects in consultation with the applicants and (vi) final decision by the GHGA SC. The GHGA Office will then transfer the funds (including overhead) to the corresponding co-applicant or participant institution(s). In case of assignment of funds to institutions not yet included in the GHGA contractual framework, those institutions will be onboarded appropriately to allow transfer of funds. Recipients of IIPs will report on project progress as part of the regular internal reporting. At the end of the project, a written report on the project results will be made available to all GHGA members.

5.10.2 Measure C1.M2: Data Mobilisation grants (DMG)

Goals: Supporting submission of high value datasets to GHGA. In order to increase the availability of high-value datasets included in the GHGA Data Infrastructure, annual calls for Data Mobilisation grants (DMG) will be distributed similarly to the IIPs described in C1.M1. In contrast to the IIPs, support can only be used for activities directly leading to data submissions into GHGA and may include for example, support for metadata preparation or for the adaptation of local data management and LIMS systems to collect GHGA metadata, thereby enabling streamlined submissions into GHGA. Budgets for DMG are limited to 5-15k€ and an own contribution by the recipient institutions of minimum the same budget is expected. Distribution, selection, and monitoring of the projects will be carried out according to the progress described in C1.M1 above with the GHGA Office being responsible for supporting the decision making by the SC and OC.

5.10.3 Measure C1.M3: Internal Flex Funds (IFF)

Goals: Agile support for existing GHGA projects in cases of funding shortages. While we have had good experiences with the processes for the selection of Flex Funds projects in the first funding phase, however for certain measures, such as support for events organised by GHGA members or mitigation of short-term and limited budget incapacities (e.g., due to

insufficient project funds), the requirement for a leaner decision making process without the need for immediate involvement of project governance other than the BoD has become apparent. We therefore plan to set aside around 20% of the budget in C1 for Internal Flex Funds (IFFs). IFFs will be administered by the GHGA Office and open for all GHGA members to apply permanently for support by the IFF, with support being limited to a maximum of 10 k€ per activity. GHGA Office will collect applications in a standardised format and together with the spokesperson propose a resolution to the BoD, which decides. Usage of the funds is reported to the SC as part of the regular SC meetings.

Tasks and Deliverables

Tasks	Due Date	Deliverables
C1.M1.T1 Yearly calling for projects (IIP)	M12, M24, M36, M48	Call concluded and funds allocated
C1.M2.T1 Yearly calling for projects (DMG)	M12, M24, M36, M48	Call concluded and funds allocated
C1.M3.T1 Establishing revised process for IFF	M6	Process description published and open for application

Dependencies, Interactions, Risks, and Mitigation Strategies

TA C1 will potentially interact with all other TAs, as every GHGA member is eligible for funding. Key interactions will however be with C2, as the GHGA Office will coordinate the review process, monitoring, and reporting. Risks and mitigation strategies will be project-specific and will be evaluated by the BoD and reviewers as part of the review process of each proposal.

Justification of Requested Funds

As outlined in 7.10, the Flex Funds budget, we have divided the budget for this TA into funding for three positions (1,292 k€) and an additional equivalent of two positions (861 k€) for direct costs such as consumables for events, outsourcing costs or other direct costs. Please note that we do not think this distribution final and we will adapt it according to the flexible funding scheme of the DFG. In addition to the budget listed currently, we will also include unspent funds from institutions into this budget as described under C2.M3.

5.11 TA C2: Project Management, Legal, Sustainability

Overview of the Task Area

C2 encompasses all measures directly related to the management of the project, in particular, the internal coordination and communication, the management of finances and contractual arrangements within the consortium, the monitoring of project progress, reporting, establishing, and supporting the governance structures and strategic development of GHGA. As described in 3.4, this TA is operated via the **GHGA Office**, which will support the Board of Directors with day-to-day management and administration of the project. This includes supporting all the bodies of GHGA in conducting their business, acting as a liaison of the consortium to DFG, NFDI and other connected projects, coordinating the periodic project reporting, and organising the legal and data protection affairs of the consortium. The

GHGA Office consists of staff supported by DFG and is supported by additional core services established at DKFZ, including finance, legal, travel, and technical.

5.11.1 Measure C2.M1: Project Management and Governance

Consortium Member	Contribution
Stegle (DKFZ	Speaker of GHGA, overseeing GHGA Governance
Kohlbacher (EKUT)	Co-Speaker of GHGA
Korbel (EMBL)	Member of the current BoD
Winkler (UHH, NCT HD)	Member of the current BoD
Eufinger (DKFZ)	Team Lead Administration

Goals: Continue to execute and to continuously improve the GHGA Governance and Project Management to enable the achievement of the project goals. Since the beginning of the project, the GHGA Office has worked together with the GHGA BoD to establish the current GHGA governance (3.4.1)**.** In the upcoming funding phase, this governance structure will be continued, adapted where necessary, and further professionalised. To enable efficient interaction between the different parts and contributors of the project, we will optimise our internal communication and collaboration tools. Governance boards, TAs, and cross-cutting working groups will be supported in organisation of daily work and suitable meeting schemes. In detail, the PM team will be responsible, in close interaction with B3, for the organisation of events such as Annual Meetings, Team Retreats, and other relevant TA meetings. It also engages, together with B4 in multiple NFDI platforms e.g. via the NFDI Management Circle, the Task Force Evaluation and Reporting, and others. The TA also handles the organisational parts of institutional partnerships of the project, including the management of partner projects such as FEGA, GDI, or NAKO, and also coordinates the formalisation of new collaborations (e.g., via preparation of letters of commitment by the GHGA BoD). C2 is also responsible for the operations of the GHGA Helpdesk and ensures distribution of all incoming inquiries to the responsible TAs. For data access requests, this is carried out by the Data Stewardship team (A4), but based on previous experiences, the helpdesk is also a valuable tool for the management of communication with other stakeholders and for ensuring quality-managed feedback to external inquiries e.g. on our communication and training measures and collection of service user feedback (B3). The GHGA Office is led by the Team Lead Administration (Jan Eufinger cf. Team Structure).

5.11.2 Measure C2.M2: Legal and Data Protection Affairs

Consortium Member	Contribution
Stegle (DKFZ	Speaker of GHGA, responsible for legal processes and DP around GHGA
Kohlbacher (EKUT)	Coordination of legal and DP processes especially with GHGA data hubs
Parker (DKFZ)	Team Lead Data Protection and Legal, overall coordination of all legal and DP processes
Eufinger (DKFZ)	Team Lead Administration

Goals: To provide support to the GHGA Project with regards to legal and data protection matters including developing required contracts and documentation,

obtaining approval from key stakeholders, and monitoring compliance. Legal and data protection concerns are at the heart of GHGA's work and, due to the complexity of the project, coordination and having expertise available in these areas is essential. This coordination also incorporates a significant amount of stakeholder engagement, as GHGA is dependent on approval from institutional stakeholders including Data Protection Officers, Legal Departments, and Management Boards.

In the current funding period, GHGA has created a number of legal documents, both to ensure the efficient collaboration of consortium partners and those needed for customers of the GHGA Data Infrastructure. Legal documents created include Consortium Contracts, Bilateral Contracts with data hubs, Joint Controller Agreements, and Data Processing Contracts. Similarly, with regards to the data protection, an extensive series of documents have been created, including: a Data Protection Framework, Technical and Organisational Measures, Risk Assessments, Standard Operating Procedures, and Terms of Use.

In the second funding period, this work is expected to continue. It will be necessary to monitor GHGA's compliance with legal obligations and policies, as well ensuring that existing documentation is kept up-to-date and adequately addresses state-of-the-art risks. Furthermore, as GHGA evolves to incorporate new functionalities, there will be a need to provide oversight of changes from a legal and data protection perspective, to work with stakeholders to ensure that such changes are approved, and develop new documents that may be required. For example, the planned interaction with the MV GenomSeq and the planned development of GHGA SPE will require a significant change to existing legal structures. This work will be conducted alongside A3.M2, A3.M3, and B5.M2 to ensure that new services and developments adhere to applicable data protection law, and that our legal and data protection documentation adequately reflects the implementation of those services and developments. We will also work towards standard contracts transitioning to a non-negotiation model as the load of requests and depositions grows. Work in C2.M2 will be led by the Team Lead Data Protection and Legal (Simon Parker, cf. Team Structure) who also oversees the Assured project in B3.M5.

5.11.3 Measure C2.M3: Finances, Human Resources and Reporting

Consortium Member	Contribution
Stegle (DKFZ	Spokesperson, overseeing financial management
Kohlbacher (EKUT)	Co-Spokesperson, overseeing financial management
Eufinger (DKFZ)	Team Lead Administration, Implementation

Goals: To enable the GHGA project with an efficient management of resources to achieve its aims. The GHGA Office is responsible for the overall handling of funds and the execution of the budget strategy of the project. The DKFZ, as the coordinating centre, distributes the funds to the partner institutions and is responsible for controlling and financial reporting to the funding organisations. It advises the partners on regulatory affairs and,

following the funding conditions, it annually reports potential needs of budget transfer to later years to the DFG. Regular financial monitoring will be carried out and if tasks cannot be fulfilled at one institution, the TA will coordinate internal re-routing of tasks and resources, potentially via allocating additional funds to the Flex Funds Budget (C1).

Having grown a highly trained GHGA Team in the first funding phase, in the second funding phase key personnel needs to be retained in the project, staff for new tasks needs to be recruited, and leaving staff need to be replaced. The TA will support institutions with HR-related measures such as advice and content for job adverts and other materials. Especially for the recruitment of specialised IT-staff, experience has shown that this often needs several recruitment rounds and usage of multiple channels to attract talents. To increase diversity in the project, we will implement targeted recruitment strategies and regularly report on demographic data and inclusion metrics to the BoD. In cases where regular recruitments are not possible, we will also need to engage with service providers to either provide IT personnel on a contractual basis or to alternatively provide needed IT services directly to the consortium. The PM team will support the respective TAs A1 to A4 in these tasks and organise the necessary procurement and contractual processes.

A further focus will be on the continuous development of internal and external reporting on the GHGA project. As before, annual reports will be collected to ensure visibility of contributions by the TAs and the contributing institutions. We will build up a database structure to aid the collection of KPIs and to fulfil diverse reporting needs for internal monitoring and especially for reporting needs within the NFDI.

5.11.4 Measure C2.M4: Sustainability and Strategic Development

Consortium Member	Contribution
Stegle (DKFZ	Strategic planning and interaction with funders and stakeholders
Kohlbacher (EKUT)	Strategic planning and interaction with funders and stakeholders
Hänisch (BfArM)	Strategic input on the MV GenomSeq and its successor initiatives
Eufinger (DKFZ)	Team Lead Administration, Strategy development

Goals: Develop a sustainable operating model for the GHGA Project and the GHGA Data Infrastructure. Structural activities: In the second funding phase of GHGA, the development of stable operations and continuous development of the provided services, including the contributions as genome data centres in MV GenomSeq (see A2.M4), will be at the centre of activities. To ensure these efforts can be sustained and transferred into a stable infrastructure, GHGA will form a strategy task force to support the BoD and SC in implementing the sustainability plan described in 3.4 and detail this strategy in an internal ‘sustainability strategy whitepaper’. The operational model laid out therein will need to be implemented in multiple steps. Together with other consortia, we will support the development of a long-term business model of NFDI and explore opportunities for long-term institutional base funding for GHGA. On the political level we have started discussing the

creation of a National Institute for Personalized Medicine, which could in the long-term, be the legal entity anchoring GHGA in the German science system. In parallel, we will also explore other sustainable funding models via NFDI e.V.

Research funding: Based on our experience in the first funding round, where we were able to obtain support from several projects (including NAKO, GDI, or MV GenomSeq), GHGA will support its members to realise new research funding opportunities within the communities by contributing (i) expertise and infrastructure on large-scale analytics of human omics data (Artificial Intelligence methods, data mining), (ii) support with ethico-legal issues, and (iii) domain knowledge in genomic medicine to joint grant proposals. The focus of these activities will be on strengthening ties with the communities and establishing GHGA as a partner in implementing data-driven research in genomics.

Infrastructure funding: To maintain the competitiveness of the infrastructure at the data hubs, GHGA will support the data hubs in joint funding initiatives (e.g., federal 91b proposals, renewal proposals within ELIXIR-DE, integration with NFDI-wide infrastructure initiatives) to maintain and renew the physical infrastructure of GHGA. In the first funding period, successful engagement with NAKO, and our role in the GDI project are examples of such activities that create additional income. We will also explore the sustainability of the funding through BMG/BfArM for MV GenomSeq and develop a joint sustainability programme for the data platform beyond MV GenomSeq (as part of the MV GenomSeq initiative, cf. A2.M4).

Scalable and commercial access: In order to enable very large-scale projects and to enable commercial access to the data (on behalf of the data controllers), for example via community SPEs, we will need to provide a fee-based usage model. We will ask for legal advice concerning tax issues and collaboration models with industry and evolve this into a sustainable business model in close collaboration with our European partners from FEGA and GDI (cf. B4).

Tasks and Deliverables

Task	Deliverables	Due Date
C2.M1.T1 Participating at NFDI governance meetings	First meeting attended	M3
C2.M1.T2 Serving GHGA Helpdesk	First tickets answered and closed	M3
C2.M2.T1 Establishing collaboration contract for GHGA 2nd funding period	Contract in place	M3
C2.M2.T2 Establishing legal framework for SPE and Atlas phase	Legal Framework document published	M9
C2.M2.T3 Establishing legal framework for MV GenomSeq	Legal Framework negotiated and executed	M6
C2.M3.T1 Conducting annual reporting (financial, HR and diversity, scientific)	Annual Report submitted	M12, M24, …, M60
C2.M3.T2 Submission of GHGA Progress Report to DFG	Report submitted, depending on timeline by DFG	M30
C2.M3.T3 Establishing and maintaining a KPI	Database set up with KPIs from first	M3

Task	Deliverables	Due Date
database	funding period	
C2.M4.T1 Developing sustainability concept	Sustainability whitepaper published	M15
C2.M4.T2 Enabling fee-based commercial access	Legal opinion obtained and initial usage fees calculated	M27

Dependencies, Interactions, Risks, and Mitigation Strategies

The TA will work with all other TAs and other entities of the consortium and support them in all administrative and coordinating matters. Risks in this TA are in particular, insufficient engagement of institutions or a lack of awareness of responsibilities due to the highly distributed nature of the project. As before, the GHGA Office will counteract these by transparent internal communication methods and regular exchange via the established governance structures. Furthermore, the management of the multiple partner projects on the national- (NFDI, MV GenomSeq, NAKO and others) and international-level (FEGA, GDI, GA4GH) requires good oversight and efficient management of resources. We will ensure this by regular exchange between the involved GHGA members and prioritisation of high relevance topics. With respect to financial management, the often suboptimal funding conditions with limited flexibility to shift funds to later years combined with the often difficult recruiting situation, especially in IT-related areas will continue to pose a medium risk to the project, likely to lead to delays in the filling of positions and consequently to delays in the overall project timeline. There is also a certain risk that due to the late funding decisions on the second phase of the project, a certain fraction of the team might seek other opportunities outside of the project. Working together with the involved institutions we have already started measures to avoid this and to retain key staff. For any necessary new or re-recruitments, we will aim to mitigate this risk by a coordinated (re-)recruitment strategy at the beginning of the second funding phase, ensuring that the optimal workforce can be maintained.

For the financial management, constraints given by the funding model in NFDI might again lead to the risk of the accumulation of a certain fraction of unspent funds in the beginning of the project and potentially cuts in the transfer of unspent funds to the following years. As described above, we will work with all institutions to avoid larger accumulation of unspent funds and, if necessary, work together with the funding organisation to develop strategies for a transfer of funds to later years to enable the project to increase efforts to catch up with delays caused.

Justification of Requested Funds

As outlined in 7.11, we will be applying for 2.5 positions for this TA to be funded via DFG. Those include funding for the Team Leads for Administration and for Legal and Data Protection as well as half a position for administrative support (all at DKFZ). Work in this area is complemented by own contributions from DKFZ and EKUT, each adding one position to this TA. Besides the personnel costs, for efficiency reasons, we have also budgeted 420 k€ for direct costs in this TA. These budgets will be administered by the GHGA office to

support central activities and events (Annual meetings, Outreach, etc.) and necessary outsourcing activities (e.g., website and help desk hosting and maintenance) for all Task Areas.

6 Additional Aspects

6.1 Equal opportunity and diversity

Recruitment and retention strategy

All institutions in GHGA follow established processes to ensure transparent and fair recruitment processes. We have taken great care to ensure that we are open to part time employment and implement family-friendly strategies for meetings. The GHGA Team is also very international with a large number of team members having a non-German background. GHGA has supported this inclusive approach by seeing itself as an international organisation with English as the primary working language and all documentation including most legal contracts being primarily developed in English. This has helped us also in recruitment, as international colleagues in other settings often feel they are not fully integrated due to language barriers. To develop and retain our staff, we have also implemented mechanisms to empower team members by giving coordinating roles relatively early in the project (e.g., through the establishment of the Operations Committee/Team Leads Committee (cf. 3.4.1 - Team Structure).

Status, progress, and monitoring

We have been aiming at increasing the percentage of female persons on all levels, in particular by including a larger fraction of PIs that can also act as role models for more junior staff. In the new funding period, we have managed to increase the ratio of female co-spokespersons from 19% to 27% (compared to the original proposal) and among the participants from 10% to 18%. On the level of GHGA staff, the team currently has an almost equal gender balance (43% female), which is a relatively high percentage of female staff in an IT-focused, technology-driven project. The project management will report on gender equality and diversity issues to the BoD on request and to the steering committee at each annual meeting (C2.M3) to both create an increased awareness and to collect ideas on improving diversity and equality issues in the project.

6.2 Further comments

None.

B-2 Part 2 Funding

7 Funding Request for Individual Task Areas

Please note these remarks concerning all following chapters:

- (*) these years are not covered by the current agreement between Germany's federal and state governments and will be subject to further evaluation, as soon as it has been established how much funding is actually available
- (**) this corresponds to the DFG staff category "Postdoctoral researchers and comparable"
- (***) this corresponds to the DFG staff category "Doctoral researchers and comparable"
- Staff listed under "Individuals with a doctoral degree (**)" also includes experienced (IT) professionals that sometimes do not hold a doctoral degree but are paid similarly.

7.1 TA A1: Operations - Central

Table 7.1.1: Funding Request for TA A1 per Institution

Institution	2025 (Oct-Dec)	2026	2027	2028	2029(*)	2030(*) (Jan-Sep)	Total
DKFZ	118,150 €	472,600 €	447,600 €	447,600 €	447,600 €	329,450 €	**2,263,000 €**
EKUT	21,525 €	86,100 €	86,100 €	86,100 €	86,100 €	64,575 €	**430,500 €**
Total	**139,675 €**	**558,700 €**	**533,700 €**	**533,700 €**	**533,700 €**	**394,025 €**	**2,693,500 €**

Table 7.1.2: Funding Request for TA A1 by Funding Category

Staff by category	2025 Oct-Dec	2026	2027	2028	2029(*)	2030(*) Jan-Sep	Total
	Number of persons (full-time equivalents)						
Individuals with a doctoral degree (**)	9 PM	36 PM	36 PM	36 PM	36 PM	27 PM	**180 PM**
Other staff	6 PM	24 PM	24 PM	24 PM	24 PM	18 PM	**120 PM**
	Totals in € (Project Funds)						
Personnel	114,675 €	458,700 €	458,700 €	458,700 €	458,700 €	344,025 €	**2,293,500 €**
Direct project costs	25,000 €	100,000 €	75,000 €	75,000 €	75,000 €	50,000 €	**400,000 €**
Total	**139,675 €**	**558,700 €**	**533,700 €**	**533,700 €**	**533,700 €**	**394,025 €**	**2,693,500 €**

7.2 TA A2: Operations - Data Hubs

Table 7.2.1: Funding Request for TA A2 per Institution

Institution	2025 (Oct-Dec)	2026	2027	2028	2029(*)	2030(*) (Jan-Sep)	Total
DKFZ	26,526 €	96,100 €	96,100 €	96,100 €	96,100 €	69,576 €	**480,502 €**
EKUT	37,289 €	139,150 €	139,150 €	139,150 €	139,150 €	101,864 €	**695,753 €**
MDC	26,526 €	96,100 €	96,100 €	96,100 €	96,100 €	69,576 €	**480,502 €**
MRI	26,526 €	96,100 €	96,100 €	96,100 €	96,100 €	69,576 €	**480,502 €**
TUD	26,526 €	96,100 €	96,100 €	96,100 €	96,100 €	69,576 €	**480,502 €**
UzK	37,289 €	139,150 €	139,150 €	139,150 €	139,150 €	101,864 €	**695,753 €**
Total	**180,682 €**	**662,700 €**	**662,700 €**	**662,700 €**	**662,700 €**	**482,032 €**	**3,313,514 €**

Table 7.2.2: Funding Request for TA A2 by Funding Category

Staff by category	2025 Oct-Dec	2026	2027	2028	2029(*)	2030(*) Jan-Sep	Total
	Number of persons (full-time equivalents)						
Individuals with a doctoral degree (**)	21 PM	84 PM	84 PM	84 PM	84 PM	63 PM	**420 PM**
	Totals in €						
Personnel	150,682 €	602,700 €	602,700 €	602,700 €	602,700 €	452,032 €	**3,013,514 €**
Direct project costs	30,000 €	60,000 €	60,000 €	60,000 €	60,000 €	30,000 €	**300,000 €**
Total	**180,682 €**	**662,700 €**	**662,700 €**	**662,700 €**	**662,700 €**	**482,032 €**	**3,313,514 €**

7.3 TA A3: Architecture & Development

Table 7.3.1: Funding Request for TA A3 per Institution

Institution	2025 (Oct-Dec)	2026	2027	2028	2029(*)	2030(*) (Jan-Sep)	Total
DKFZ	46,575 €	186,300 €	186,300 €	186,300 €	186,300 €	139,725 €	**931,500 €**
EKUT	21,525 €	86,100 €	86,100 €	86,100 €	86,100 €	64,575 €	**430,500 €**
EMBL	21,525 €	86,100 €	86,100 €	86,100 €	86,100 €	64,575 €	**430,500 €**
Total	**89,625 €**	**358,500 €**	**358,500 €**	**358,500 €**	**358,500 €**	**268,875 €**	**1,792,500 €**

Table 7.3.2: Funding Request for TA A3 by Funding Category

Staff by category	2025 Oct-Dec	2026	2027	2028	2029(*)	2030(*) Jan-Sep	Total
	Number of persons (full-time equivalents)						
Individuals with a doctoral degree (**)	9 PM	36 PM	36 PM	36 PM	36 PM	27 PM	**180 PM**
Other staff	3 PM	12 PM	12 PM	12 PM	12 PM	9 PM	**60 PM**
	Totals in € (Project Funds)						
Personnel	89,625 €	358,500 €	358,500 €	358,500 €	358,500 €	268,875 €	**1,792,500 €**
Total	**89,625 €**	**358,500 €**	**358,500 €**	**358,500 €**	**358,500 €**	**268,875 €**	**1,792,500 €**

7.4 TA A4: Data Stewardship - Central and Data Hubs

Table 7.4.1: Funding Request for TA A4 per Institution

Institution	2025 (Oct-Dec)	2026	2027	2028	2029(*)	2030(*) (Jan-Sep)	Total
DKFZ	78,863 €	315,450 €	315,450 €	315,450 €	315,450 €	236,588 €	**1,577,251 €**
EKUT	10,763 €	43,050 €	43,050 €	43,050 €	43,050 €	32,288 €	**215,251 €**
MDC	10,763 €	43,050 €	43,050 €	43,050 €	43,050 €	32,288 €	**215,251 €**
TUD	10,763 €	43,050 €	43,050 €	43,050 €	43,050 €	32,288 €	**215,251 €**
TUM	10,763 €	43,050 €	43,050 €	43,050 €	43,050 €	32,288 €	**215,251 €**
UzK	10,763 €	43,050 €	43,050 €	43,050 €	43,050 €	32,288 €	**215,251 €**
Total	**132,678 €**	**530,700 €**	**530,700 €**	**530,700 €**	**530,700 €**	**398,028 €**	**2,653,506 €**

Table 7.4.2: Funding Request for TA A4 by Funding Category

Staff by category	2025 Oct-Dec	2026	2027	2028	2029(*)	2030(*) Jan-Sep	Total
	Number of persons (full-time equivalents)						
Individuals with a doctoral degree (**)	15 PM	60 PM	60 PM	60 PM	60 PM	45 PM	**300 PM**
Other staff	3.0 PM	12.0 PM	12.0 PM	12.0 PM	12.0 PM	9.0 PM	**60 PM**
	Totals in €						
Personnel	132,678 €	530,700 €	530,700 €	530,700 €	530,700 €	398,028 €	**2,653,506 €**
Total	**132,678 €**	**530,700 €**	**530,700 €**	**530,700 €**	**530,700 €**	**398,028 €**	**2,653,506 €**

7.5 TA B1: Community Driver Projects

Table 7.5.1: Funding Request for TA B1 per Institution

Institution	2025 (Oct-Dec)	2026	2027	2028	2029(*)	2030(*) (Jan-Sep)	Total
NAKO	10,763 €	43,050 €	43,050 €	43,050 €	43,050 €	32,288 €	**215,251 €**
TUM	12,525 €	50,100 €	50,100 €	50,100 €	50,100 €	37,575 €	**250,500 €**
UKT	21,525 €	86,100 €	86,100 €	86,100 €	86,100 €	64,575 €	**430,500 €**
Total	**44,813 €**	**179,250 €**	**179,250 €**	**179,250 €**	**179,250 €**	**134,438 €**	**896,251 €**

Table 7.5.2: Funding Request for TA B1 by Funding Category

Staff by category	2025 Oct-Dec	2026	2027	2028	2029(*)	2030(*) Jan-Sep	Total
	Number of persons (full-time equivalents)						
Individuals with a doctoral degree (**)	5 PM	18 PM	18 PM	18 PM	18 PM	14 PM	**90 PM**
Other staff	1,5 PM	6 PM	6 PM	6 PM	6 PM	4,5 PM	**30 PM**
	Totals in € (Project Funds)						
Personnel	44,813 €	179,250 €	179,250 €	179,250 €	179,250 €	134,438 €	**896,251 €**
Total	**44,813 €**	**179,250 €**	**179,250 €**	**179,250 €**	**179,250 €**	**134,438 €**	**896,251 €**

7.6 TA B2: Community Data Services

Table 7.6.1: Funding Request for TA B2 per Institution

Institution	2025 (Oct-Dec)	2026	2027	2028	2029(*)	2030(*) (Jan-Sep)	Total
BIH	10,763 €	43,050 €	43,050 €	43,050 €	43,050 €	32,288 €	**215,251 €**
DKFZ	10,763 €	43,050 €	43,050 €	43,050 €	43,050 €	32,288 €	**215,251 €**
DZNE	21,525 €	86,100 €	86,100 €	64,575 €	- €	- €	**258,300 €**
TUM	23,288 €	93,150 €	93,150 €	93,150 €	93,150 €	69,863 €	**465,751 €**
Grand Total	**66,339 €**	**265,350 €**	**265,350 €**	**243,825 €**	**179,250 €**	**134,439 €**	**1,154,553 €**

Table 7.6.2: Funding Request for TA B2 by Funding Category

Staff by category	2025 Oct-Dec	2026	2027	2028	2029(*)	2030(*) Jan-Sep	Total
	Number of persons (full-time equivalents)						
Individuals with a doctoral degree (**)	8 PM	30 PM	30 PM	27 PM	18 PM	14 PM	**126 PM**
Other staff	1.5 PM	6.0 PM	6.0 PM	6.0 PM	6.0 PM	4.5 PM	**30 PM**
	Totals in €						
Personnel	66,339 €	265,350 €	265,350 €	243,825 €	179,250 €	134,439 €	**1,154,553 €**
Total	**66,339 €**	**265,350 €**	**265,350 €**	**243,825 €**	**179,250 €**	**134,439 €**	**1,154,553 €**

7.7 TA B3: Outreach and Training

Table 7.7.1: Funding Request for TA B3 per Institution

Institution	2025 (Oct-Dec)	2026	2027	2028	2029(*)	2030(*) (Jan-Sep)	Total
DKFZ	53,033 €	212,130 €	212,130 €	212,130 €	212,130 €	159,098 €	**1,060,651 €**
EKUT	21,526 €	86,100 €	86,100 €	86,100 €	86,100 €	64,576 €	**430,502 €**
HMGU	21,525 €	86,100 €	86,100 €	86,100 €	86,100 €	64,575 €	**430,500 €**
UDS	21,525 €	86,100 €	86,100 €	86,100 €	86,100 €	64,575 €	**430,500 €**
UKT	10,763 €	43,050 €	43,050 €	43,050 €	43,050 €	32,288 €	**215,251 €**
Total	**128,370 €**	**513,480 €**	**513,480 €**	**513,480 €**	**513,480 €**	**385,110 €**	**2,567,404 €**

Table 7.7.2: Funding Request for TA B3 by Funding Category

Staff by category	2025 Oct-Dec	2026	2027	2028	2029(*)	2030(*) Jan-Sep	Total
	Number of persons (full-time equivalents)						
Individuals with a doctoral degree (**)	14 PM	58 PM	58 PM	58 PM	58 PM	43 PM	**288 PM**
Other staff	3 PM	12 PM	12 PM	12 PM	12 PM	9 PM	**60 PM**
	Totals in € (Project Funds)						
Personnel	128,372 €	513,480 €	513,480 €	513,480 €	513,480 €	385,112 €	**2,567,404 €**
Total	**128,372 €**	**513,480 €**	**513,480 €**	**513,480 €**	**513,480 €**	**385,112 €**	**2,567,404 €**

7.8 TA B4: National and International Connectivity and Metadata Alignment

Table 7.8.1: Funding Request for TA B4 per Institution

Institution	2025 (Oct-Dec)	2026	2027	2028	2029(*)	2030(*) (Jan-Sep)	Total
EKUT	21,525 €	86,100 €	86,100 €	86,100 €	86,100 €	64,575 €	**430,500 €**
EMBL	21,526 €	86,100 €	86,100 €	86,100 €	86,100 €	64,576 €	**430,502 €**
Total	**43,050 €**	**172,200 €**	**172,200 €**	**172,200 €**	**172,200 €**	**129,150 €**	**861,002 €**

Table 7.8.2: Funding Request for TA B4 by Funding Category

Staff by category	2025 Oct-Dec	2026	2027	2028	2029(*)	2030(*) Jan-Sep	Total
	Number of persons (full-time equivalents)						
Individuals with a doctoral degree (**)	6 PM	24 PM	24 PM	24 PM	24 PM	18 PM	**120 PM**
	Totals in € (Project Funds)						
Personnel	43,051 €	172,200 €	172,200 €	172,200 €	172,200 €	129,151 €	**861,002 €**
Total	**43,051 €**	**172,200 €**	**172,200 €**	**172,200 €**	**172,200 €**	**129,151 €**	**861,002 €**

7.9 TA B5: Legal and Ethical Issues

Table 7.9.1: Funding Request for TA B5 per Institution

Institution	2025 (Oct-Dec)	2026	2027	2028	2029(*)	2030(*) (Jan-Sep)	Total
UHD	21,526 €	111,100 €	111,100 €	111,100 €	111,100 €	64,576 €	**530,502 €**
UKHD	32,288 €	129,150 €	129,150 €	129,150 €	129,150 €	96,863 €	**645,751 €**
Total	**53,815 €**	**240,250 €**	**240,250 €**	**240,250 €**	**240,250 €**	**161,439 €**	**1,176,253 €**

Table 7.9.2: Funding Request for TA B5 by Funding Category

Staff by category	2025 Oct-Dec	2026	2027	2028	2029(*)	2030(*) Jan-Sep	Total
	Number of persons (full-time equivalents)						
Individuals with a doctoral degree (**)	8 PM	30 PM	30 PM	30 PM	30 PM	23 PM	**150 PM**
	Totals in € (Project Funds)						
Personnel	53,814 €	215,250 €	215,250 €	215,250 €	215,250 €	161,439 €	**1,076,253 €**
Direct project costs	- €	25,000 €	25,000 €	25,000 €	25,000 €	- €	**100,000 €**
Total	**53,814 €**	**240,250 €**	**240,250 €**	**240,250 €**	**240,250 €**	**161,439 €**	**1,176,253 €**

7.10 TA C1: Flex Funds

Table 7.10.1: Funding Request for TA C1 per Institution

Institution	2025 (Oct-Dec)	2026	2027	2028	2029(*)	2030(*) (Jan-Sep)	Total
Flex Funds	107,625 €	430,500 €	430,500 €	430,500 €	430,500 €	322,875 €	**2,152,500 €**
Total	**107,625 €**	**430,500 €**	**430,500 €**	**430,500 €**	**430,500 €**	**322,875 €**	**2,152,500 €**

Table 7.10.2: Funding Request for TA C1 by Funding Category

Staff by category	2025 Oct-Dec	2026	2027	2028	2029(*)	2030(*) Jan-Sep	Total
	Number of persons (full-time equivalents)						
Individuals with a doctoral degree (**)	9 PM	36 PM	36 PM	36 PM	36 PM	27 PM	**180 PM**
	Totals in € (Project Funds)						
Personnel	64,575 €	258,300 €	258,300 €	258,300 €	258,300 €	193,725 €	**1,291,500 €**
Direct project costs	50,000 €	170,000 €	170,000 €	170,000 €	170,000 €	131,000 €	**861,000 €**
Total	**114,575 €**	**428,300 €**	**428,300 €**	**428,300 €**	**428,300 €**	**324,725 €**	**2,152,500 €**

7.11 TA C2: Project Management, Legal, Sustainability

Table 7.11.1: Funding Request for TA C2 per Institution

Institution	2025 (Oct-Dec)	2026	2027	2028	2029(*)	2030(*) (Jan-Sep)	Total
DKFZ	78,275 €	313,100 €	313,100 €	313,100 €	313,100 €	254,825 €	**1,585,500 €**
Total	**78,275 €**	**313,100 €**	**313,100 €**	**313,100 €**	**313,100 €**	**254,825 €**	**1,585,500 €**

Table 7.11.2: Funding Request for TA C2 by Funding Category

Staff by category	2025 Oct-Dec	2026	2027	2028	2029(*)	2030(*) Jan-Sep	Total
	Number of persons (full-time equivalents)						
Other staff	7.5 PM	30 PM	30 PM	30 PM	30 PM	22.5 PM	**150 PM**
	Totals in € (Project Funds)						
Personnel	58,275 €	233,100 €	233,100 €	233,100 €	233,100 €	174,825 €	**1,165,500 €**
Direct project costs	20,000 €	80,000 €	80,000 €	80,000 €	80,000 €	80,000 €	**420,000 €**
Total	**78,275 €**	**313,100 €**	**313,100 €**	**313,100 €**	**313,100 €**	**254,825 €**	**1,585,500 €**

8 Overall Funding Request

Table 8.1: Overall Funding Request by Task Area (Project Funds in €)

Task Area	2025 Oct-Dec	2026	2027	2028	2029 (*)	2030 Jan-Sep (*)	Total
A1. Operations - Central	139,675	558,700	533,700	533,700	533,700	394,025	**2,693,500**
A2. Operations - Data Hubs	180,682	662,700	662,700	662,700	662,700	482,032	**3,313,514**
A3. Architecture & Development	89,625	358,500	358,500	358,500	358,500	268,875	**1,792,500**
A4. Data Stewardship	132,678	530,700	530,700	530,700	530,700	398,028	**2,653,506**
B1. Community Driver Projects	44,813	179,250	179,250	179,250	179,250	134,438	**896,251**
B2. Community Data Services	66,339	265,350	265,350	243,825	179,250	134,439	**1,154,553**
B3. Outreach and	128,372	513,480	513,480	513,480	513,480	385,112	**2,567,404**

Task Area	2025 Oct-Dec	2026	2027	2028	2029 (*)	2030 Jan-Sep (*)	Total
Training							
B4. Internat. Connect. & Metadata	43,051	172,200	172,200	172,200	172,200	129,151	**861,002**
B5. ELSI	53,814	240,250	240,250	240,250	240,250	161,439	**1,176,253**
C1. Flex Funds	114,575	428,300	428,300	428,300	428,300	324,725	**2,152,500**
C2. Management	78,275	313,100	313,100	313,100	313,100	254,825	**1,585,500**
Grand Total	**1,071,899**	**4,222,530**	**4,197,530**	**4,176,005**	**4,111,430**	**3,067,089**	**20,846,483**

Table 8.2: Overall Funding Request by Institution (Project Funds in €)

Institution	2025 Oct-Dec	2026	2027	2028	2029 (*)	2030(*) Jan-Sep	Total
BIH	10,763	43,050	43,050	43,050	43,050	32,288	**215,251**
DKFZ	412,185	1,638,730	1,613,730	1,613,730	1,613,730	1,221,550	**8,113,655**
DZNE	21,525	86,100	86,100	64,575	0	0	**258,300**
EKUT	134,153	526,600	526,600	526,600	526,600	392,453	**2,633,006**
EMBL	43,051	172,200	172,200	172,200	172,200	129,151	**861,002**
Flex Funds	114,575	428,300	428,300	428,300	428,300	324,725	**2,152,500**
HMGU	21,525	86,100	86,100	86,100	86,100	64,575	**430,500**
MDC	37,289	139,150	139,150	139,150	139,150	101,864	**695,753**
MRI	26,526	96,100	96,100	96,100	96,100	69,576	**480,502**
NAKO	10,763	43,050	43,050	43,050	43,050	32,288	**215,251**
TUD	37,289	139,150	139,150	139,150	139,150	101,864	**695,753**
TUM	46,576	186,300	186,300	186,300	186,300	139,726	**931,502**
UDS	21,525	86,100	86,100	86,100	86,100	64,575	**430,500**
UHD	21,526	111,100	111,100	111,100	111,100	64,576	**530,502**
UKHD	32,288	129,150	129,150	129,150	129,150	96,863	**645,751**
UKT	32,288	129,150	129,150	129,150	129,150	96,863	**645,751**
UzK	48,052	182,200	182,200	182,200	182,200	134,152	**911,004**
Grand Total	**1,071,899**	**4,222,530**	**4,197,530**	**4,176,005**	**4,111,430**	**3,067,089**	**20,846,483**

Table 8.3: Overall Funding Request by Funding Category

Staff by category	2025 Oct-Dec	2026	2027	2028	2029(*)	2030(*) Jan-Sep	Total
	Number of persons (full-time equivalents)						
Individuals with a doctoral degree (**)	102.9 PM	411.6 PM	411.6 PM	408.6 PM	399.6 PM	299.7 PM	**2,034.0 PM**
Other staff	25.5 PM	102.0 PM	102.0 PM	102.0 PM	102.0 PM	76.5 PM	**510.0 PM**
	Totals in €						
Personnel	946,899	3,787,530	3,787,530	3,766,005	3,701,430	2,776,089	**18,765,483**
Direct project costs	125,000	435,000	410,000	410,000	410,000	291,000	**2,081,000**
Total	**1,071,899**	**4,222,530**	**4,197,530**	**4,176,005**	**4,111,430**	**3,067,089**	**20,846,483**

Description and Summary of Contributions by (Co-) Applicants

In the first funding phase, GHGA has received generous support from the participating institutions. Firstly, both in the form of staff exclusively working for GHGA but also via existing staff dedicating a significant amount of their working time to GHGA. Secondly, operations of the GHGA Data Infrastructure relies entirely on infrastructures financed via other means, namely institutional own contributions and related third-party funded measures. For the second funding phase this support will be continued with the majority of the contributions being provided by the institutions operating GHGA data hubs. Overall, the support for the next funding phase includes 5 M€ for GHGA exclusive staff (11.5 FTE), at least 2.3 M€ for additional support staff and over 20 M€ for investment and operations costs for the data hub infrastructures. Further details are listed below.

Own Contributions - Data Hub Infrastructure

The GHGA data hubs contribute significant storage and compute infrastructure to the project, as physical infrastructure is not fundable within the NFDI. Each data hub is operating high-performance computing infrastructures that provide storage and backup for the research data (S3-compatible) and compute infrastructure (e.g., for data management, quality control, and SPEs). Investments for these infrastructures come from different sources (primarily core funding of the institutions, state funding, DFG, BMBF). Operating costs of the infrastructure are borne by the co-applicant institutions (e.g., power, cooling, maintenance). In order to determine the contribution of the institution, we assumed a harmonised full-cost model across all data hubs taking into account investments and running cost. Overall, the data hubs will contribute infrastructure to store 55 PB (net) of research data until 2030 with an estimated overall cost **equivalent of 22.5 M€ over the requested funding period.**

Own Contributions - Staff

Overall, participant institutions have committed to dedicate up to 11.5 FTE for the next funding phase, equivalent to almost 5 M€. In addition a minimum of another 10 positions will be supporting GHGA.

Institution	FTE p.a. - GHGA Exclusive Staff	Personnel Costs GHGA Exclusive Staff (for 5y)	In-Kind Contributions existing Staff	Personnel Costs GHGA In-kind Staff
DKFZ	5.2 FTE	2,221 k€	4.0 FTE	1,722 k€
EKUT	2.0 FTE	861 k€	4.0 FTE	344 k€
TUM	1.3 FTE	573 k€	.6 FTE	47 k€
MDC / BIH	1.0 FTE	431 k€	.3 FTE	22 k€
UKT	1.0 FTE	431 k€	.3 FTE	k€
TUD	.5 FTE	215 k€	.4 FTE	31 k€
UzK	.5 FTE	215 k€	.6 FTE	54 k€
HMGU	n.a.	k€	.3 FTE	k€
UHD	n.a.	k€	.1 FTE	9 k€
UKHD	n.a.	k€	.1 FTE	10 k€
Total	**11.5 FTE**	**4,946 k€**	**10.5 FTE**	**2,239 k€**

Appendix

A1 - Bibliography and list of references

*References in **bold** are based on work of the GHGA consortium.*

1. Freeberg MA, Fromont LA, D'Altri T, et al. The European Genome-phenome Archive in 2021. *Nucleic Acids Res*. 2022;50(D1):D980-D987. doi:10.1093/nar/gkab1059
2. Lappalainen I, Almeida-King J, Kumanduri V, et al. The European Genome-phenome Archive of human data consented for biomedical research. *Nat Genet*. 2015;47(7):692-695. doi:10.1038/ng.3312
3. **Apondo E, Bruns A, Züger A, et al. Patient Involvement in the Governance of the German Human Genome-Phenome Archive (GHGA): Patients' perspectives and recommendations for implementation. Published online June 30, 2023. Accessed July 27, 2023. https://zenodo.org/record/8099346**
4. Birney E, Vamathevan J, Goodhand P. Genomics in healthcare: GA4GH looks to 2022. Published online October 15, 2017:203554. doi:10.1101/203554
5. Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM). Zulassungskriterien GRZ. Published online April 24, 2024. Accessed July 31, 2024. https://www.bfarm.de/SharedDocs/Downloads/DE/Forschung/modellvorhaben-genomsequenzierung/Zulassungkriterien-GRZ-final.html
6. **Bruns A, Benet-Pages A, Eufinger J, et al. Consent Modules for Data Sharing via the German Human Genome-Phenome Archive (GHGA). *Zenodo*. Published online July 13, 2022. doi:10.5281/zenodo.6828131**
7. Der Bundesbeauftragte für den Datenschutz und die Informationsfreiheit. Tätigkeitsbericht 2021 - 30. Tätigkeitsbericht für den Datenschutz und die Informationsfreiheit. Published online 2022. Accessed September 15, 2023. https://www.bfdi.bund.de/SharedDocs/Downloads/DE/Taetigkeitsberichte/30TB_21.html
8. **Parker S, Deschler K, Eufinger J, et al. GHGA Legal Concept White Paper. *Zenodo*. Published online September 28, 2023. doi:10.5281/zenodo.8387734**
9. **GHGA Consortium. Letter of Intent within the Call for National Research Data Infrastructures (NFDI) Renewal Proposal 2024 - German Human Genome-Phenome Archive. Published online June 18, 2024. https://www.dfg.de/resource/blob/337484/c368b0478faaefe1ff78e2c0a0f9f491/2024-ghga-nfdi-data.pdf**
10. **Deschler K, Eufinger J, Fluck J, et al. Memorandum of Understanding between NFDI4Health and GHGA. Published online July 23, 2024. doi:10.5281/zenodo.12799244**
11. Jacobsen JOB, Baudis M, Baynam GS, et al. The GA4GH Phenopacket schema defines a computable representation of clinical data. *Nat Biotechnol*. 2022;40(6):817-820. doi:10.1038/s41587-022-01357-4
12. **Iyappan A, Mauer K, Menges P, et al. Metadata Schema for the German Human Genome-Phenome Archive. Published online September 13, 2023. Accessed July 28, 2024. https://zenodo.org/records/8341224**
13. **GHGA Consortium. ghga-de/ghga-metadata-schema: Metadata schema for the German Human Genome-Phenome Archive (GHGA). Accessed August 1, 2024. https://github.com/ghga-de/ghga-metadata-schema**
14. **GHGA Consortium. The German Human Genome-Phenome Archive - GHGA Brochure. Published online September 19, 2023. doi:10.5281/zenodo.8359328**
15. **Apondo E, Bruns A, Züger A, et al. Patient Involvement in the Governance of the German Human Genome-Phenome Archive (GHGA): Patients' perspectives and**

recommendations for implementation. *Zenodo*. Published online June 30, 2023. doi:10.5281/ZENODO.8099345
16. Gemeinsame Wissenschaftskonferenz von Bund und Ländern. Bund-Länder-Vereinbarung zu Aufbau und Förderung einer Nationalen Forschungsdateninfrastruktur vom 26. November 2018. Published online 2018. Accessed July 22, 2024. https://www.gwk-bonn.de/fileadmin/Redaktion/Dokumente/Papers/NFDI.pdf
17. Ebert B, Domisch S, Henzen C, et al. When Data Crosses Borders – Join Forces! Multidisciplinary Use Cases Within NFDI. *Proceedings of the Conference on Research Data Infrastructure*. 2023;1. doi:10.52825/cordi.v1i.341
18. Bundesministerium für Bildung und Forschung - BMBF. Deutschland tritt Genomprojekt der EU bei - BMBF. January 16, 2020. Accessed January 18, 2021. https://www.bmbf.de/de/deutschland-tritt-genomprojekt-der-eu-bei-10676.html
19. Rehm HL, Page AJH, Smith L, et al. GA4GH: International policies and standards for data sharing across genomic research and healthcare. *Cell Genomics*. 2021;1(2). doi:10.1016/j.xgen.2021.100029
20. Phillips M, Molnár-Gábor F, Korbel JO, et al. Genomics: data sharing needs an international code of conduct. *Nature*. 2020;578(7793):31-33. doi:10.1038/d41586-020-00082-9
21. Dietmar Hopp Foundation funds development of new cancer therapies for children at the KiTZ with around 21 million euros. January 30, 2023. Accessed July 3, 2023. https://www.kitz-heidelberg.de/en/the-kitz/kitz-newsroom/kitz-news/detail/dietmar-hopp-stiftung-foerdert-entwicklung-neuer-krebstherapien-fuer-kinder-am-hopp-kindertumorzentrum-heidelberg-mit-rund-21-millionen-euro
22. Zurek B, Ellwanger K, Vissers LELM, et al. Solve-RD: systematic pan-European data sharing and collaborative analysis to solve rare diseases. *Eur J Hum Genet*. 2021;29(9):1325-1331. doi:10.1038/s41431-021-00859-0
23. Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM). Zulassung klinischer Datenknoten (KDK) und Genomrechenzentren (GRZ) - Modellvorhaben Genomsequenzierung. June 27, 2024. Accessed August 1, 2024. https://www.bfarm.de/DE/Das-BfArM/Aufgaben/Modellvorhaben-Genomsequenzierung/Aktuelles/_node.html
24. Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. 2016;3(1):160018. doi:10.1038/sdata.2016.18
25. Senf A, Davies R, Haziza F, et al. Crypt4GH: a file format standard enabling native access to encrypted data. *Bioinformatics*. 2021;37(17):2753-2754. doi:10.1093/bioinformatics/btab087
26. GHGA Consortium. ghga-de/hexkit. Published online July 31, 2024. Accessed August 2, 2024. https://github.com/ghga-de/hexkit
27. ghga-de/ghga-datasteward-kit. Published online July 25, 2024. Accessed July 28, 2024. https://github.com/ghga-de/ghga-datasteward-kit
28. Chue Hong NP, Katz DS, Barker M, et al. FAIR Principles for Research Software (FAIR4RS Principles). Published online May 24, 2022. doi:10.15497/RDA00068
29. Yépez VA, Mertes C, Müller MF, et al. Detection of aberrant gene expression events in RNA sequencing data. *Nat Protoc*. 2021;16(2):1276-1296. doi:10.1038/s41596-020-00462-5
30. Jin Y, Schäffer AA, Sherry ST, Feolo M. Quickly identifying identical and closely related subjects in large databases using genotype data. *PLoS One*. 2017;12(6):e0179106. doi:10.1371/journal.pone.0179106
31. Lamnidis TC, Majander K, Jeong C, et al. Ancient Fennoscandian genomes reveal origin and spread of Siberian ancestry in Europe. *Nat Commun*. 2018;9(1):5018. doi:10.1038/s41467-018-07483-5
32. Avsec Ž, Kreuzhuber R, Israeli J, et al. The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nat*

Biotechnol. 2019;37(6):592-600. doi:10.1038/s41587-019-0140-0
33. Wagner N, Çelik MH, Hölzlwimmer FR, et al. Aberrant splicing prediction across human tissues. *Nat Genet*. 2023;55(5):861-870. doi:10.1038/s41588-023-01373-3
34. Clarke B, Holtkamp E, Öztürk H, et al. Integration of variant annotations using deep set networks boosts rare variant association genetics. Published online October 26, 2023:2023.07.12.548506. doi:10.1101/2023.07.12.548506

A 2 - 5 : see separate Appendix document