European Genomic Data Infrastructure

Deliverable D7.1

Use case demonstrator package

Project Title Grant agreement no	**Genomic Data Infrastructure** Grant agreement 101081813		
Project Acronym (EC Call)	GDI		
WP No & Title	WP7: GDI use cases		
WP Leaders	Alfonso Valencia (37. BSC) Salvador Capella-Gutierrez (37. BSC) Marc Van Den Bulcke (9. SC) Oliver Stegle (20. DKFZ)		
Deliverable Lead Beneficiary	37. BSC		
Contractual delivery date	29/02/2024	**Actual delivery date**	31/05/2024
Delayed	Yes		
Partner(s) contributing to deliverable	BSC, IRCCS, UB, HU, BioData.pt, HSR, CNAG, UT, UL, Health RI, Erasmus MC, DKFZ		
Authors	Laura Portell-Silva (BSC)		
Contributors	Carles Hernandez-Ferrer (BSC) Salvador Capella-Gutierrez (BSC) Sergi Aguiló (BSC) Domenico Coviello (IRCCS)		

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

	Macha Nikolski (UB) Andreas Scherer (HU) Jorge Oliveira (BioData.pt) Gabriele Bucci (HSR) Sergi Beltrán (CNAG) Priit Kleeman (UT) Marko Arrak (UT) Marco Morelli (HSR) Aedin Culhane (UL) Jeroen van Rooij (Erasmus MC)
Reviewers	Rob Hooft (Health RI) Luiz Gadelha (DKFZ)

Log of changes

Date	Mvm	Who	Description
22/04/2024	v0.1	Laura Portell-Silva	Initial version
13/05/2024	v0.2	Laura Portell-Silva	Deliverable sent to WP7
17/05/2024	v0.3	Laura Portell-Silva	Deliverable sent to reviewers
20/05/2024	v0.3	Mercedes Rothschild Steiner (ELIXIR Hub)	Deliverable sent to GDI MB for review
30/05/2024	v0.4	Laura Portell-Silva	Addressed comments
30/05/2024	V05	Laura Portell-Silva	Final version sent to the coordination

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

05/06/2024	1V0	Mercedes Rothschild Steiner (ELIXIR Hub)	Final version submitted to EC Portal

Contents

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

European Genomic Data Infrastructure

1. Executive Summary

This deliverable focuses on creating a use case demonstrator package, which includes an initial set of data collections tailored for specific use case scenarios within the European Genomic Data Infrastructure (GDI). Given the current lack of real data in the GDI nodes, efforts were concentrated on gathering synthetic and other available real data that closely align with intended use cases to facilitate comprehensive testing of the GDI infrastructure.

Seven datasets have been identified and made accessible for GDI use cases, with five specifically tailored to GDI requirements and two offering more generalised data. An additional six datasets are being generated and are expected to be available soon. Key datasets include the Genome of Europe, featuring real data from the Genome of the Netherlands project, and 1+MG/B1MG use cases, offering synthetic datasets for rare diseases.

The identified datasets are ready for request and usage within the GDI nodes. Most datasets contain Variant Calling Format (VCF) files essential for discoverability via Beacon v2. Additionally, these datasets have been used in the MS7 demonstrator to assist the nodes in testing the infrastructure.

Mapping the datasets to key questions ensures they address essential inquiries for different use cases, such as genetic variant lookup, recalibrating polygenic risk scores, and medication side effects. In addition, different federated processing scenarios could be tested using this data, like screening for common variants across populations and processing cancer datasets with variant calling workflows.

This collection of datasets will allow Pillar II to test the infrastructure, making sure that it is robust and usable by all GDI nodes.

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

2. Contribution towards project outcomes

With this deliverable, the project has reached or the deliverable has contributed to the following project outcomes:

	Contributed
Outcome 1 Secure federated infrastructure and data governance needed to enable sustainable and secure cross border linkage of genomic data sets in compliance with the relevant and agreed legal, ethical, quality and interoperability requirements and standards based on the progress achieved by the 1+MG initiative.	**Yes**
Outcome 2 Platform performing distributed analysis of genetic/genomic data and any linked clinical/phenotypic information; it should be based on the principle of federated access to data sources, include a federated/multi party authorisation and authentication system, and enable application of appropriate secure multi-party and/or high-end computing, AI and simulation techniques and resources.	**Yes**
Outcome 3 Clear description of the roles and responsibilities related to personal data and privacy protection, for humans and computers, applicable during project lifetime and after its finalisation.	**No**
Outcome 4 Business model including an uptake strategy explaining the motivation, patient incentives and conditions for all stakeholders at the different levels	**No**

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

(national, European, global) to support the GDI towards its sustainability, including data controllers, patients, citizens, data users, service providers (e.g., IT and biotech companies), healthcare systems and public authorities at large.	
Outcome 5 Sustained coordination mechanism for the GDI and for the GoE multi-country project launched in the context of the 1+MG initiative.	**Yes**
Outcome 6 Communication strategy – to be designed and implemented at the European and national levels.	**No**
Outcome 7 Capacity building measures necessary to ensure the establishment, sustainable operation, and successful uptake of the infrastructure.	**No**
Outcome 8 Financial support to the relevant stakeholders to enable extension, upgrade, creation and/or physical connection of further data sources beyond the project consortium or to implement the communication strategy and for capacity-building.	**No**

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

3. Methods

This deliverable pertains to the creation of a use case demonstrator package comprising the initial set of data collections tailored for specific use case scenarios. Given the current absence of real data in the GDI nodes to be used by the federation, our efforts have been directed towards collecting synthetic and real data openly available and closely aligned with the intended use cases that can be used to facilitate thorough testing of the GDI infrastructure.

A collaborative approach was adopted within Work Package 7 to compile this data, which encompasses the GDI use cases. First, a spreadsheet[1] was circulated among the participants involved in these use cases. The information requested to be included in this document included general information about the datasets, which relates to:

- Dataset name
- URL
- Metadata
- Genomic data types
- Number of samples
- If applicable, phenotypic data
- Size (GB)

After that, supplementary data access and usage details were incorporated to ensure that all pertinent information essential for Pillar II to use these datasets was included. Such information encompasses:

- Possibility to use Beacon as a discoverability mechanism
- Special conditions for access and usage of the data
- Data access technologies (e.g., flat file access, htsget)
- Potential use for federated processing use cases

This effort is situated within the broader framework of GDI, which is structured in three pillars. Pillar I focuses on data governance and policy frameworks, Pillar II on technological infrastructure, and Pillar III on use cases and applications. Our work is specifically carried out under Pillar III, particularly within WP7, dedicated to use cases. A key objective of WP7 is to contribute towards the testing (WP8 in collaboration with Pillar II) and the deployment of technologies under Pillar II. As such, it is permissible to use synthetic data for these purposes to ensure robust and effective testing and implementation of the technological solutions being developed.

[1] D7.1 use-cases synthethic data

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

4. Description of work accomplished

A total of 7 datasets have been identified and made accessible by the use cases, ready for request and usage within the GDI federation. Within this collection, 5 datasets are specifically tailored to address the requirements of GDI use cases, while the remaining 2 offer more generalised datasets. Additionally, there are 6 additional datasets currently being generated, each aligned with specific use case scenarios, and are anticipated to be available shortly.

4.1 Genome of Europe

4.1.1. Datasets description

Table 1. Genome of Europe datasets

Dataset	Metadata	Genomic Data	Samples	Phenotypic data	Size (GB)
Genome of the Netherlands[2]	EGA	FASTQ / BAM / cVCF	769	Very minimal	61,000

The Genome of the Netherlands (GoNL) project, led by a consortium of top Dutch institutions, is a pioneering effort to explore the genetic diversity within the Dutch population. With contributions from key studies and sequencing conducted by BGI Hong Kong, GoNL has assembled a dataset containing real data from more than 750 individuals, including 250 trios, offering unprecedented insights into Dutch genetic variation. Supported by funding from the Netherlands Organization for Scientific Research and BBMRI-NL, this initiative promises significant advancements in biomedical research and diagnostic technologies, potentially revolutionising healthcare.

Additionally, it is important to note that the Genome of Europe use case, which will involve the sequencing of real data once the project starts, will have the capability to use this data in an aggregated form. This approach allows the project to bypass legal issues associated with using real data since it will not involve identifiable information from specific individuals. Consequently, the aggregated data can be used to test the infrastructure effectively and compliantly.

4.1.2. Data access and usage conditions

The Genome of the Netherlands data[3] related to SNVs, indels, structural variants, *de novo* mutations, and mutation rate map is publicly available and cfan be used under the condition of citation (main

[2] https://nlgenome.nl/
[3] https://nlgenome.nl/menu/main/app-go-nl/download-data

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

paper, additional papers when appropriate). Individual-level sequence data and/or variant calls can be requested. Data can be requested through EGA and nlgenome.nl[4].

4.2 1+MG/B1MG use cases

4.2.1. Datasets description

Table 2. 1+MG/B1MG use cases datasets

Dataset	Metadata	Genomic Data	Samples	Phenotypic data	Size (GB)
B1MG Rare Diseases[5]	EGA	FASTQ / BAM / cVCF	18	Phenopackets, PED	6,000
1+ Million Genomes	-	VCF	1,017,219	-	20,000 - 30,000

From the 1+MG/B1MG use cases, 2 datasets focusing on rare diseases have been incorporated. The first one, the B1MG Rare Diseases dataset, aims to facilitate the development of technical implementations for rare disease data integration, analysis, discovery, and federated access. This dataset was created by the B1MG Rare Diseases WG, which is dedicated to furnishing publicly accessible human datasets tailored for the study of rare diseases. By leveraging public human genomic background and incorporating real disease-causing variants through *in silico* insertion, this dataset can be used for testing purposes, circumventing ethical and legal concerns associated with sensitive human data usage.

This synthetic dataset includes clinical and genomic data from 6 rare disease cases. It consists of 18 whole genomes (6 index cases with their parents) with genetic backgrounds based on publicly available human data sequenced in the Illumina Platinum initiative[6] and made available by the HapMap project[7]. In each case, real causative variants correlating with the phenotypic data provided were spiked in.

The cases included in this synthetic dataset correspond to different types of disorders[8], and for each case, one will be able to download the following data: clinical information (GA4GH Phenopackets per

[4] https://nlgenome.nl/menu/main/app-go-nl/request-access
[5] https://ega-archive.org/datasets/EGAD00001008392
[6] https://genome.cshlp.org/content/27/1/157
[7] https://www.genome.gov/10001688/international-hapmap-project
[8] https://ega-archive.org/studies/EGAS00001005702

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

individual and pedigree per family), raw genomic data (FASTQ and BAMs) and processed genomic data (VCFs).

The second dataset[9] was generated by the Finnish Institute for Health and Welfare (THL) and Finnish CSC - IT Center for Science Ltd (CSC). The data used for the simulation were publicly available whole genome sequences, but in the simulations, they were formed into synthetic genomes, no longer representing real people. The simulation was done with CSC's LUMI supercomputer and it was done for the EU's 1+MG Initiative. Data is provided in the VCF format and it includes five use cases, from cancer to rare diseases.

4.2.2. Data access and usage conditions

When using the B1MG Rare Diseases dataset, the following should be acknowledged: the RD-Connect GPAP (https://platform.rd-connect.eu/), EC H2020 project EJP-RD (grant # 825575), EC H2020 project B1MG (grant # 951724) and Generalitat de Catalunya VEIS project (grant # 001-P-001647). The dataset is publicly available for non-commercial research and educational activities. To access the data, one must submit a request to the Data Access Committee. Requests are accepted from everyone without needing any specific agreement with the EGA or signing any document with data generators. Therefore, this data can be considered open.

The 1+ Million Genomes data will be available in the Finnish Federated European Genome-phenome Archive (FI-FEGA), maintained by CSC. The data generated is now being transferred to CSC, and approval has been received from the Ministry of Social Affairs and Health of Finland for this data to be used within GDI.

4.3 Infectious Diseases

4.3.1. Datasets description

Table 3. Infectious diseases datasets

Dataset	Metadata	Genomic Data	Samples	Phenotypic data	Size (GB)
WG11 Synthetic Data (Case 1)	-	VCF	-	-	50.000

[9] https://www.elixir-finland.org/en/a-million-european-genomes/

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

The GDI Infectious Diseases use case includes only one dataset, which the 1+MG WG11 is currently generating. After completion, this dataset will be deposited in a pertinent repository and made available to the wider research community. This dataset will serve as an important tool for the research of infectious diseases in two different scenarios.

In the first scenario (infectious diseases use case 1), the dataset can be used to perform a GWAS analysis, identifying variants determining the severity of COVID-19 disease progression. Specifically tailored for this purpose, the dataset is structured to ensure that six published variants are discoverable when employing the workflow developed by the 1+MG WG11.

In a healthcare setting called infectious diseases use case 2, the same dataset will serve a different yet equally significant role. Here, it facilitates lookup scenarios without necessitating access to the entire genome by healthcare professionals. This enables them to swiftly identify patients with specific variants, tailor treatments accordingly, and potentially predict disease progression, thereby streamlining healthcare delivery.

4.4 Cancer Research

4.4.1. Datasets description

Table 4. Cancer Research datasets

Dataset name	Metadata	Genomic Data	Samples	Phenotypic data	Size (GB)
EOSC4Cancer Synthetic Colorectal Cancer Genomic data[10]	EGA	FASTQ/ BAM / MAF / VCF	8	-	430
Stage II/III colorectal cancer dataset[11]	NCBI	FASTQ	114 patients	Custom metadata	6,430

[10] https://ega-archive.org/studies/EGAS50000000190
[11] https://www.ncbi.nlm.nih.gov/bioproject/PRJNA689313

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

B1MG Cancer Dataset (Melanoma single patient) [12] [13]	Zenodo[14]	VCF/BED	2	MDC/Phenopackets	0.12
B1MG Cancer Dataset (NSCLC single patient)[15]	-	VCF/CRAM	2	MDC/Phenopackets	70
EOSC4Cancer UiO Synthetic Dataset	-	-	-	-	-
EOSC4Cancer Synthetic Clinical + variants dataset	-	Part of the patient's data in JSON format	100 patients	-	-

The GDI Cancer Research use cases encompass 6 datasets, with 3 already accessible and 3 currently being generated.

The EOSC4Cancer[16] Synthetic Colorectal Cancer Genomic data have been meticulously created to emulate real cancer data, including mutations observed in actual Colorectal Cancer (CRC) patients from the PCAWG dataset. This development process involved simulating reads from each clone individually and, subsequently, mixing them based on their clonal proportions within each sequence. Comprising 8 samples, this dataset offers access to chromosome sequencing data based on paired-end FASTQ format files, VCF files, genomic variants data in MAF format, and binary alignment map (BAM) data along with their respective indices (BAI) for four simulated patients.

The Stage II/III colorectal cancer dataset includes real data of gene expression and whole exome sequencing of 114 stage II/III colorectal cancer patients with poor outcomes.

The B1MG Cancer Dataset (Melanoma single patient) is a synthetically generated paired tumour/normal dataset from a single patient. It includes the VCF and BED formats files of a common

[12] https://bioinformatics.nygenome.org/wp-content/uploads/CancerCellLines/COLO-829-NovaSeq--COLO-829BL-NovaSeq.snv.indel.final.v6.annotated.vcf

[13] https://bioinformatics.nygenome.org/wp-content/uploads/CancerCellLines/COLO-829-NovaSeq--COLO-829BL-NovaSeq.cnv.annotated.v6.final.bed

[14] https://doi.org/10.5281/zenodo.11048418

[15] https://doi.org/10.5281/zenodo.11203957

[16] https://eosc4cancer.eu/

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

cancer cell line (COLO-829). The metadata attached to the dataset can be found in Zenodo and adheres to the MDC/Phenopackets standard.

The remaining three datasets are currently in the process of being generated and will soon be made available through a pertinent publicly available repository.

4.4.2. Data access and usage conditions

The EOSC4Cancer Synthetic Colorectal Cancer Genomic data follows the same access policy as the open-access datasets archived at the EGA. As mentioned earlier, one must submit a request to the Data Access Committee to access this open data at EGA. Requests are accepted from everyone without needing any agreement with the EGA or signing any document with the data generators. Therefore, this data can be considered open.

The Stage II/III colorectal cancer dataset is publicly accessible. The B1MG Cancer Dataset (Melanoma single patient) is publically available, and the B1MG Cancer Dataset (NSCLC single patient) is publicly available as well, but files are restricted to users with access.

The EOSC4Cancer Synthetic Clinical + variants dataset has been generated from the MIMIC clinical data and colorectal cancer data from AACR Genie. Access to MIMIC data requires authorisation due to its controlled access policy.

4.5 Other Datasets

4.5.1. Datasets description

Table 5. Other datasets

Dataset name	Metadata	Genomic Data	Samples	Phenotypic data	Size (GB)
CINECA UK1[17]	EGA	FASTQ / BAM / VCF	2,521	Derived from UKBiobank	2,600
EGA test dataset[18]	EGA	CRAM / BAM / VCF / BCF	2,508	Minimal metadata model	1,300

These 2 datasets serve as foundational resources crucial for the development of technical frameworks supporting cohort data discovery, harmonisation, access, and federated analysis, all of which are integral components of the GDI framework.

[17] https://ega-archive.org/datasets/EGAD00001006673
[18] https://ega-archive.org/datasets/EGAD00001003338

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

The CINECA dataset consists of 2,521 samples with genetic data based on the 1,000 Human Genomes Project data[19] and synthetic subject attributes and phenotypic data derived from UK Biobank[20]. It includes FASTQ, BAM and VCF formats files.

The EGA test dataset consists of 2,508 samples also from the 1,000 Human Genomes Project. This dataset includes diverse data types, including Variant Calling Format (VCF, or its binary counterparts BCF) files, both joint and split, exome sequencing CRAM files and whole genome sequencing CRAM/BAM files. Additionally, multiple files were sliced to create shorter files, facilitating a quicker download.

4.5.2. Data access and usage conditions

Both datasets are openly available in the EGA. Therefore, they are not subject to controlled access and, as a result, may be distributed without the requirement of a data access application. As mentioned before, one must submit a request to the Data Access Committee to access the open data at the EGA. Requests are accepted from everyone without needing any agreement with the EGA or signing any document. Therefore, this data can be considered open. However, you must complete a simple step of requesting access (by clicking the button) and waiting for the Helpdesk to approve the request, even though approval is always granted.

In support of FAIRness in data sharing, the CINECA dataset is made freely available under the Creative Commons Licence (CC-BY), so the preamble has to be included with this dataset and that the CINECA project (funding: EC H2020 grant 825775) acknowledged when used. In addition, the CINECA dataset includes different DUO codes that restrict the use of this data to several conditions:

- DUO0000019 (publication required): This data use modifier indicates that the requester agrees to make results of studies using the data available to the larger scientific community.
- DUO0000026 (user-specific restriction): This data use modifier indicates that use is limited to use by approved users.
- DUO0000028 (institution-specific restriction): This data use modifier indicates that use is limited to use within an approved institution
- DUO0000042 (general research use): This data use permission indicates that use is allowed for general research use for any research purpose.

[19] https://www.nature.com/articles/nature15393
[20] https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001779

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

5. Results

As mentioned before, 7 datasets have been identified and made accessible by the use cases, ready for request and use within GDI. Since all the data included here is publicly available, encryption is not mandatory, but if required for development purposes, it can be encrypted using GA4GH crypt4gh[21].

Tables 1-5 illustrate that many of the datasets contain Variant Calling (VCF) format files, crucial for discoverability through Beacon v2. However, to facilitate discovery and querying via Beacon v2 and the GDI Beacon Network, these VCF files need conversion to a Beacon-Friendly format (BFF). Similarly, phenotypic data associated with each dataset must be transformed into the Beacon Schema for compatibility with Beacon v2 queries.

The data outlined in this document is presently available as a data package, accessible for local download or through the EGA infrastructure. For datasets already hosted on EGA, the advancing capabilities within the GDI Starter Kit will facilitate easy access and mobilisation using its components, such as GA4GH htsget[22].

Consequently, these datasets have been used in the milestone MS7 demonstrator to assist the nodes in testing the infrastructure. Table 6 presents the dataset selected by each node for inclusion in the first demonstrator of the GDI federation.

Table 6. Datasets used by the nodes that are part of the MS7 demonstrator.

Finland	Luxemburg	Netherlands	Norway	Portugal	Spain	Sweden
CINECA	B1MG Rare Diseases	B1MG Rare Diseases	EOSC4Cancer UiO Synthetic Dataset	Stage II/III colorectal cancer dataset and CINECA subset	EOSC4Cancer Synthetic Colorectal Cancer Genomic data	B1MG Rare Diseases

[21] https://crypt4gh.readthedocs.io/en/latest/
[22] https://samtools.github.io/hts-specs/htsget.html

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

5. Discussion

5.1. Mapping to prototypical questions

Once the datasets from the use cases have been identified, aligning them with the key questions outlined in MS26 at the project's outset is crucial. This ensures we can assess whether the data at hand can effectively address these inquiries.

Three key questions have been formulated for the Genome of Europe use case. Using the Genome of the Netherlands and the 1+ Million Genomes dataset, Genome of Europe aims to address two of these questions: individual genetic variant lookup and recalibrating polygenic risk scores (PRS). In the context of this use case, researchers aim to divide the 1+ Million Genomes dataset into different nodes to test the PRS calculation in a federated manner, thereby assessing the robustness of the infrastructure for such analyses. This question is related to the work done in GDI WP8, which can be found in deliverable 8.8 that includes a technical demonstrator on PRS calculation in a federated manner[23]. The third question, which pertains to ancestry-specific imputation, has been excluded from the scope of GDI at the moment.

Within the 1+MG/B1MG framework, two essential questions emerge, both of which find answers within the datasets contributed by the respective members. The first query, concerning why individuals with disease-specific genes may not manifest the associated conditions, finds its solution in the 1+ Million Genomes dataset. Meanwhile, the B1MG Rare Diseases dataset can be used to answer the question regarding the side effects of medications caused by some gene variants.

Moving to the domain of infectious diseases, the formulated questions can be addressed once the relevant dataset becomes available. These two questions approach the use case from different perspectives. The first question is research-focused, specifically on Genome-Wide Association Studies (GWAS), aiming to validate risk variants for severe COVID-19. The second question is more healthcare-oriented, seeking to identify variants that may guide prognosis and/or treatment.

However, the cancer use case presents a challenge. While two questions were formulated, their resolution may not solely rely on existing datasets. The questions need a database containing known variants responsible for tumour regrowth or therapy resistance, whose existence remains uncertain. In light of this, potential reformulation of the inquiry may be necessary to align it with the available data resources.

[23] https://doi.org/10.5281/zenodo.10887366

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

5.2. Federated processing

The richness and diversity of these datasets offer an exciting opportunity to explore the capabilities of federated processing and federated learning, as outlined in WP8. Here are potential scenarios that exemplify the breadth of possibilities:

- Scenario 1. Screening for common variants across European populations using the general population datasets of "1+ Million Genomes" and "genomes of the Netherlands," using two or more of the national nodes.
- Scenario 2. Processing the 114 participants of the "Stage II/III colorectal cancer dataset" with a variant calling pipeline, distributing the dataset across multiple national nodes.
- Scenario 3. Within the domain of rare diseases, combining the datasets "B1MG Rare Diseases" with "1+ Million Genomes" to compare each participant of the former dataset to the entire population of the latter, identifying personalised variants.

These simplified examples serve not only to showcase the federated analytical capabilities but also to stress-test the robustness of the infrastructure under development within the project, which will be essential for Pillar II in the future. By exploring such scenarios, we can identify technological gaps and address pending needs, thus contributing to the continual enhancement of our solutions.

In addition, we can envision using some datasets to test the Minimal Viable Product (MVP) from Pillar II with simple analytical questions, such as "Do you have similar variants and/or phenotypes to the ones that I have (rare diseases)?" and more complex scenarios that could involve running workflows or using distributed PRS analysis.

5.3. European Genome Dashboard

In addition to the synthetic and open data previously mentioned, data gathered by the 1+MG initiative in various European countries is also available. The dashboard in the 1+MG webpage[24] displays the current number of genomic datasets across the 1+MG Member States. This real data will be subject to access conditions depending on the repository where it is stored. Therefore, its use for GDI will have to wait until a governance model for the EDIC is established, as using real data will involve significant legal considerations.

[24] https://dashboard.onemilliongenomes.eu/

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

6. Conclusions & Impact

This collaborative effort within GDI has led to the successful identification of seven datasets. These datasets are instrumental in advancing genomic research as they cover a broad range of applications, including genetic diversity in European populations, rare diseases, infectious diseases, and cancer research.

Each dataset meets the specific requirements of the GDI use cases, ensuring they effectively address essential questions in genomic science. Several datasets are versatile, serving multiple use cases by answering a range of prototypical questions, which highlights their robustness and wide applicability within the genomic research community.

The exercise of mapping prototypical questions to datasets has revealed the alignment of formulated questions with available data and identified areas for improvement, such as the need to reformulate questions related to cancer data to better match the data available.

All identified datasets are openly accessible, adhering to the principles of FAIR (Findable, Accessible, Interoperable, Reusable), ensuring researchers can readily access and use the data for various genomic studies. The use of standard formats like VCF and Phenopackets also allows efficient and effective data usage across the genomic research community.

7. Next steps

The work accomplished for this deliverable lays a strong foundation for future research and development within GDI. However, several areas warrant further exploration and enhancement:

- Expansion of Datasets: The generation and inclusion of additional datasets, for instance those addressing infectious diseases and specific cancer types, will be crucial in broadening the scope of research.
- Federated Learning: Continued development and testing of federated learning frameworks will ensure robust, scalable solutions for genomic data analysis. Include an estimation of the compute resources that are needed to do these tests, to be sure that the nodes included in the federation can support it.
- Legal and Ethical Considerations: Establishing a clear governance model for the use of real genomic data across European countries will be essential in addressing legal and ethical challenges, which is being addressed by Pillar I.
- Technological Advancements: Ongoing technological advancements to make the GDI products portfolio robust and maintainable by all GDI nodes, will further facilitate secure and efficient data sharing and analysis.

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

By addressing these areas, we can enhance the impact and reach of GDI, fostering a collaborative environment that drives innovation in genomic data integration and analysis across Europe.

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.