European Genomic Data Infrastructure

Deliverable D6.2

Report on European resources and data suitable for inclusion into the GDI

Project Title Grant agreement no	**Genomic Data Infrastructure** Grant agreement 101081813		
Project Acronym (EC Call)	GDI		
WP No & Title	WP6: Data Management		
WP Leaders	Rob Hooft (21. HRI)		
Deliverable Lead Beneficiary	6.1. THL		
Contractual delivery date	31/10/2023	**Actual delivery date**	26/03/2024
Delayed	Yes		
Partner(s) contributing to deliverable	None		
Authors	Markus Perola (THL) Tero Hiekkalinna (THL)		
Contributors	Vilho Heikkinen (THL)		
Reviewers	Janis Klovins (LV), Hedi Peterson (EE)		

Log of changes

Date	Mvm	Who	Description
09/02/2024	0V1	Markus Perola (THL)	Main text editing

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

28/02/2024	0V2	Markus Perola (THL) Tero Hiekkalinna (THL)	Main text editing and formatting
04/03/2024	0V3	Tero Hiekkalinna (THL) Vilho Heikkinen (THL)	Table editing and formatting
08/03/2024	0V4	Tero Hiekkalinna (THL)	Table 1 legend editing
26/03/2024	1VO	Mercedes Rothschild Steiner (ELIXIR Hub)	Final version submitted to the EC Portal

Table of contents

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 report outlines the contributions of European countries to the Genomic Data Infrastructure (GDI), specifying the expected number of samples, the mix of legacy Whole Genome Sequences (WGS) and new DNA samples, sequencing technologies used, and the infrastructure for managing, storing, and sequencing these samples.

Work Accomplished

Below we detail the potential national legacy contributions to the GDI. The work was done with collaboration with the Genome of Europe - initiative (GoE) utilising its questionnaire with comparison to previously collected data. This was done partly for the reason to avoid overlapping sets of questions being sent to the same people during the same time period which probably would have decreased the response frequency for both related efforts.

Data from 27 European countries has been aggregated, while only two countries provided no answers for the GoE questionnaire. These contributions include both legacy WGS, which some of may require reconsenting, and potential newly collected DNA or to be collected samples specifically for 1+MG and GoE. The report highlights the use of both short-read and long-read sequencing technologies across different countries, with work involving a range of biobanks, research institutes, and academic institutions. The capacities for sequencing and data processing vary, with some countries outsourcing these tasks due to limited local capacity.

Conclusion

The Deliverable 6.2 Report provides a detailed account of the efforts by European countries to compile WGS data for the GDI and sample collections for further use. The number of legacy WGS samples available for GDI and GoE was found to be about 27.000. However, the seemingly simple question (how many WGS?) is in real life quite complex and can be duly answered only at the time of data collection.

This is a continuing, collective effort to collect metadata about valuable genomic data, noticing varying national capacities and resources. The combination of legacy WGS alongside new data, and the adoption of both short-read and long-read sequencing technologies, is expected to enrich the GDI as well as 1+MG and also support other genomic research and healthcare advancements across Europe.

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:

[Select 'Yes' (at least one) if the deliverable contributed to the key result, otherwise select 'No'. For more details of project outcomes, see here]

	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.	No
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 (national, European, global) to support the GDI towards its sustainability, including data controllers, patients, citizens, data users, service providers	No

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

(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.	Yes
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

Deliverable scope

By aggregating and analysing data from various European countries for inclusion in the GDI, Deliverable 6.2 indirectly supports the establishment of a sustained coordination mechanism. Understanding the resources, data types, and sequencing capabilities across countries is fundamental for the ongoing coordination of the GDI and the Genome of Europe (GoE) project. This foundational work facilitates collaboration and coordination among member states, which is essential for the success of the 1+MG initiative. Understanding the resources, data types, and sequencing capabilities across countries is fundamental for the ongoing coordination of the Genomic Data Infrastructure (GDI). This foundational work facilitates collaboration and coordination among member states, which is essential for the success of the GDI and eventually for the 1+MG initiative. Although Deliverable 6.2 primarily addresses the identification and description of genomic resources and data, it implicitly contributes to capacity building by highlighting the existing infrastructures, technologies, and gaps in different countries. This information can be used to identify where capacity building is needed to ensure the establishment, sustainable operation, and successful uptake of the infrastructure. Knowledge of the current capabilities and needs can inform targeted training, investment in sequencing technologies, and data management practices. Thus, D6.2. has a direct relationship with WPs 1, 2 and 9 in GDI.

Methodology

The new data was collected through a questionnaire sent to GoE participating countries. Here, we considered the on-going GoE application overlapping the time for delivering D6.2 and made a conscious decision not to duplicate the questionnaire to key individuals in participating countries but to utilise GoE's questionnaire and compare it to previously collected data, aiming to streamline the data collection process. This approach was adopted to prevent the issue of sending overlapping sets of questions to the same individuals within the same time period. Such overlap could not only reduce the response rate for both initiatives but also potentially lead to data fatigue among respondents, compromising the quality and reliability of the information collected. Repeated inquiries about the same issue are not only unnecessary but can be counterproductive, leading to respondent disengagement and potentially hindering the comprehensive gathering of genomic data. By coordinating efforts and harmonising questionnaires, we aimed to maximise participation and data quality, ensuring a more efficient and effective collection process that respects the time and contributions of all participants. Given the description of the Deliverable: "Report on European resources and data suitable for inclusion into the GDI '' this was geographically sufficient for the aims of this Deliverable, D6.4. In future, we will address the question on a global scale and also re-visit the potential for including clinical genetics-based sequencing data. The approach was green-lighted by both GDI and GoE leaders.

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

The new questionnaire was sent as an attachment to one email per country to the 1-3 representatives in the 1+MG WG12 mailing list.

The contents of the new questionnaire in green:

****

To plan our proposal, we ask you to update the specifics on your possible contribution to the first part (100k) of the GoE. We have attached information you provided earlier, any additional WGS datasets mentioned in the B1MG data portal, and the projected composition of the 500k GoE.

Given the available funding, we are aiming for 100k WGS samples as a deliverable of this project, and thus ask you to consider 20% of the projected contribution of your country (of the original 500k) for this proposal. Data will be collected along 3 scenarios. Priority will be given to 1) existing WGS datasets that could be adopted for the GoE, followed by 2) the generation of novel WGS from existing biobank samples, and 3) novel recruitment of participants if specific samples are not available (for example from specific population sub-groups).

Through this survey, we ask you to provide specific information on how the 20% of WGS data for your country could be collected during the 42 months covered by the proposal (1-9-2024 until 1-3-2028). Specifically, we would like to estimate the expected costs of collecting the data, and if/how the co-funding of 50% can be realised. We understand that it is/can be difficult to provide exact numbers, workflows and costs at this stage. This will provide a starting point. Returns will be treated as confidential and preliminary. Please try to provide this information in the next two weeks.

Based on the returned information, we will create the specifics of a plan for the GoE proposal. We will feed this back to the consortium, and then include information on existing WGS data (preferred, as this will leave more budget for WGS) and balance which additional samples can be collected where, and work out other details and alignments needed. Please fill out the questions below by copying the responses from your earlier survey if those answers are still up-to-date.

We outline below 3 distinct routes to assemble the minimal number of 100,000 WGS data. Please read through each to understand the type of samples and data we would like to include. If anything is unclear, please contact the surveyors - or include the thing you are unclear about and add a comment flagging your question, so we can come back to it.

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

1. Existing WGS data

Additional information on sample specifications: Generally, population cohorts/biobanks representative of the general population would be the preferred source of samples. The samples should be unselected for a particular disease. So controls in a case-control design are eligible, but the cases are not. Samples should be unrelated as much as possible. We would like to avoid controls from family studies. Parents of a de-novo trio design would qualify. A large series of rare disease patients could be considered, assuming they are different diseases (to limit enrichment of disease-related variants or haplotypes). Other samples could be flagged but included, so please be inclusive and flag if there is any dataset you are unsure about.

Additional information on technical specifications: WGS data would need to have 30x average coverage and cover >90% of the genome (basically a standard clinical or research WGS, WES or panels are not accepted). Different sequencing technologies can be provided. The BAM/CRAM files or equivalent don't need to be shared for the currently envisioned applications of GoE, although the potential for re-analyses for future applications could be considered. For this phase, we need a VCF/pVCF/gVCF or equivalent file.

Additional information on processed data (e.g., VCF-level) sharing and consent: Currently, we see three options through which the processed WGS data could be accessed: 1) by providing federated access to the GoE VCF files through the GDI* technical infrastructure (note: individual-level data would stay at the national node), 2) by sharing individual level data (a 'frozen' dataset-level VCF file) outside of GDI data infrastructure but within the GoE consortium and covered by the GoE consortium agreement, or: 3) For analyses to take place locally on each GoE dataset, with GoE templated scripts and sharing of aggregated and anonymised results, such as frequencies of variants within a population.

Note that option 1 above is preferred and what we aim for with GDI. Options 2 and/or 3 ensure maximum flexibility, to enable analytical questions that option 1 cannot accommodate.

*For those not familiar, GDI stands for Genomic Data Infrastructure and provides technical solutions to data access, sharing and federated analyses. They typically do not cover hardware, such as data storage or computation servers, or processing of data, such as alignment or variant calling.

For existing/legacy data, informed consent under ethics regime should cover the scope planned activities in the project, or reconsenting should be possible. Please note that if you intend to base the processing of personal data in the context of the project on consent, you will likely have to reobtain a specific consent anyway. At minimum, you need to inform subjects about the project. Where you build on legacy cohorts but no genomic sequencing was done before, genomic sequencing may

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

also require an informed consent in itself. You will likely have to obtain an ethics approval for the participation in the project.

a. Given the details in the table of expected sample numbers per country - country of origin, how many individual datasets could be collected from existing/legacy WGS data?

b. Provide details (short paragraph, responsible institute, cohort name, sample size, sample ancestries, platform type, how to get access) on existing/legacy WGS datasets.

c. What are the costs (e.g. reconsenting, ethics approval, data storage etc) related to the inclusion of this WGS data in the GoE ? (please provide a ballpark estimate)

2. New WGS data from existing biobanks and cohorts

a. If newly generated WGS data are required to reach the 20% suggested in WG12, how would this be organised in your country?

Please provide details on the numbers of samples you would want to recruit from specific biobanks, the institutes involved, and the expected costs on sample handling, sequencing, data processing (alignment and variant calling), and WGS data storage.

b. Could you outline the expected costs (recruitment, consenting, DNA isolation, sequencing costs, data processing and storage) in regards to this collection?

We suggest that individual institutes or biobanks are responsible for the collection and storage of BAM/CRAM data (which could perhaps be used to cover the 50% matching). We understand that providing exact numbers may be difficult, but we need a rough estimate to outline the plan in the proposal. We will get back on the details and there will be room for adjustments if needed. Sequencing could be done in-country, or outsourced, as long as it stays within Europe (i.e., the 26 1+MG signatory member states). Please note that we are negotiating with technology providers and trying to get an umbrella agreement on costs with the main providers. Reach out to us if this is something useful to you.

3. **Novel sample recruitment to generate new WGS data**

a. Lastly, if novel recruitment of samples is required - for example to collect the minority country of origin samples - please outline the expected logistics and costs - similar to how it was asked in the previous questions.

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

Additional questions/comments
A suggested workflow during the WG12 meeting was to recruit/collect biological samples per institute, perform DNA isolation on the institute/country level, and perform sequencing in a subset of available sequencing centers, at least in the beginning of the project. We also foresee the possibility of some countries outsourcing the sequencing to sequencing centers in other countries, so please also indicate if you would need such an option for your samples

b. Can DNA isolation and/or sequencing be done within your country? At one institute or several? Can you name the institutes/partners that are likely going to be involved in this process? (we will ask for administrative details separately - just name the partners here).

c. Do you have a preference for sequencing technology (e.g., short or long read sequencing)? In case of long-read sequencing, have you created WGS using a long-read technology, and would you be interested in exploring this further?

d. Indicated earlier, costs from sequenced WGS data processing and longer-term storage (specifically of the large BAM/CRAM files) need to be included in this proposal. We hope a part of this can be included as matching from the institutes. Can you indicate if there are facilities for the processing and storage of the WGS data collected for GoE? (if not already mentioned in one of the prior questions)

****

End of the questionnaire

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

4. Description of work accomplished

New Questionnaires

The above questionnaire was sent to the individuals on 1+MG WG12 mailing list in June 2023. Remainders were sent to every non-responder, several times if necessary, up to the end of October 2023. Altogether 27 countries responded.

Legacy questionnaire data

We obtained the questionnaire data collected for 1+MG project from the European Commission.

https://ec.europa.eu/eusurvey/runner/1plusMG_Survey2020

This data has been earlier visualised in:

https://dashboard.onemilliongenomes.eu/

Table 1. Numbers from the 2020 legacy questionnaire, of which only WGS data is used (WES or cancer genome WGS are available but not used).

Country	WGS	GWAS	Usability
Austria	0	0	
Belgium	0	0	
Bulgaria	0	0	
Croatia	0	0	
Cyprus	0	0	
Czechia	1837	60	*
Denmark	10100	0	*
Estonia	3000	0	
Finland	4700	393999	*
France	0	0	
Germany	0	16760	*

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

Greece	0	0	
Hungary	87	0	*
Ireland	0	500	*
Italy	1757	27259	*
Latvia	20	450	
Lithuania	0	1000	*
Luxembourg	0	0	
Malta	293	0	*
Netherlands	4550	135420	
Norway	0	0	
Portugal	1	1300	*
Romania	0	0	
Slovenia	0	0	
Spain	5338	56663	*
Sweden	0	0	
United Kingdom	79008	0	

* Usability of some WGS/GWAS data may not be sure (may require re-consenting).

Collaboration

The new questionnaire study was done together with GoE

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

Problems

There was no response from two countries, Romania and the UK, for the new questionnaire. Not having UK data in the current collection is regrettable, especially considering the wealth of genomic data available from leading initiatives such as the UK Biobank, Genomics England, and the NHS's Our Future Health program. The UK Biobank, with its extensive collection of genetic and health information from half a million UK participants, represents a valuable resource for understanding the determinants of a wide range of diseases. Similarly, Genomics England, through the 100,000 Genomes Project, has made significant strides in applying whole genome sequencing technology to improve disease diagnosis and treatment, focusing on rare diseases, certain cancers, and infectious diseases. The NHS's Our Future Health, aiming to be the UK's largest research program, seeks to prevent, detect, and treat diseases by analysing the health data of up to 5 million people. The integration of such comprehensive datasets, two with WGS data together close to 600.000 individuals, could have significantly enhanced the scope and depth of the Genomic Data Infrastructure, providing richer insights into genetic variations and their implications for health and disease across a broader European context. Our plan is that during the global part of this review (D6.4), pathways will open for the inclusion of UK data as a part of the consortium, enriching the collective understanding and enabling more comprehensive genomic research and healthcare advancements.

Results

Here we show summaries of the country-specific responses and at the end tabulated results of the legacy (2020) data and new data. Not all the data received is presented in this report but of course archived for further use.

Country-specific summaries of the questionnaires

Austria: 2,113 expected, no legacy samples, 2,150 samples will be collected through a BBMRI biobank with 75% short-read sequencing (NovaSeq) and 25% long-read sequencing. Project/sample management and storage is located in Innsbruck, Recruitment and sequencing will be carried out by all four participating academic institutions.

Belgium: 2,903 expected, 100 legacy samples, derived from healthy parents of rare disease patients, remainder collected from Belgian Genome Biobank (n=10,500). Sequencing will be done at the KU Leuven and include a small portion lrGS.

Bulgaria: 1,853 expected, no legacy samples, de novo recruitment, participants will sign the informed consent at recruitment, about 2,000 will be collected from BBMRI.bg, approximately 90% will be done with short-read sequencing with Novseq6000 and 10% will be long read sequencing.

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

Croatia: 1,257 expected, 60 legacy, recruit 1,200 from local clinical cohorts, and The Institute for Anthropology (INANTRO) for samples from the Croatian Islands.

Cyprus: expected 310 WGS: Cyprus: 245, Romania: 10, Russia: 10, United Kingdom: 10, Greece: 35; no legacy samples; recruited from national biobank (~8,000 participants)

Czech Republic: expected 2,946, ~300 legacy WGS data (re-consent required, from ACGT and ENIGMA efforts, contain ~2,000 individuals), ~2,700 to be recruited from nationally collected cohorts (CELSPAC, KDIOVIZE and NCMG, consisting of ~3,500 participants, additional cohorts to be added).

Denmark: 1,414 expected samples, of these 200 will also be performed as long read WGS (ONT), no legacy samples/legacy data, 1,414 samples collected from participants in a cohort from the Statens Serum Institute.

Estonia: expected 368, no legacy data, 600 samples collected from Estonian Biobank (n=200,000).

Finland: Expecting about 6,500 legacy WGS samples, depending on eligibility of sample selection and consenting. Additional samples will be recruited via the THL Finnish biobank (n=40,000).

France: Legacy data: 10,000 srGS 30X: 5,000 from the general population in France (POPGEN and FranceGenRef) + 5,000 parents of children with various rare diseases (French Genomic Medicine Initiative). New data to sequence: 4,000 novel short-read 30X WGS that will be sequenced during the project and funded by the French Government (French Genomic Medicine Initiative) + approx. 3,000 new de novo sequencing (including 2,900 short read & 100 long read).

Germany. Samples: expected 19,400 of these 17,000 srGS (Illumina) and 2,400 lrGS (ONT), cohorts: National cohort of Germany (NAKO): 19,000, and 400 (200 pairs) of monozygotic twins (Tübingen). Sequencing: Institute of Medical Genetics and Applied Genomics (IMGAG) Tübingen, 1 x NovaSeqXplus, 1 x Novaseq6000, 2 x PromethION P24/P48 for ONT-based lrGS and for T2T (WP2). Processing: IMGAG Tübingen via GHGA / backup at NAKO Munich. Access via the German Human Genome Archive (GHGA). Uses: IMGAG is a member of GHGA, CanHeal for cancer, SOLVE-RD and ERDERA for Rare Diseases, and the European lrGS diagnostic consortium.

Greece: Samples: 2,500 from the HYDRIA cohort (national representative samples of men and women) and 500 from newly collected samples from parents of children with various rare diseases (Laboratory of Medical Genetics of the University of Athens) with 30x WGS Illumina. Also, 200 legacy samples from the HYDRIA cohort (30x WGS Illumina). Samples: U-PGx cohort: ~1,500 samples from psychiatric patients of various indications.

Hungary: will start out as an observer, data generated in the country will be made available for GoE where possible.

Ireland: sample; 2,000 WGS total, comprising 1,238 from existing biobanks and 762 newly recruited (targeting isolates and underrepresented communities). These will be approximately 90% short-read and 10% long read sequences. We will seek to include an Irish sample in new T2T references for Europe.

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

Italy: Samples: 15,800 WGS expected, of which 15,050 short-read (Illumina) and 750 long-read (ONT). Legacy data: 2,400 short-read WGS 30-40×. De novo sequencing: top-up from WGS 20x to 30x for 5,000 subjects in the MOLI-SANI population-based cohort of the IRCCS Neuromed in Pozzilli;

remaining 8,400 samples randomly collected from ten Italian population-based biobanks (n = 120,000). Sequencing and Processing: The Human Technopole Foundation in Milan and CNR – IBIOM Institute in Bari, 3 NovaSeqXplus and 2 PromethIon systems.

Latvia: 3,000 legacy samples (>30X coverage with short-read sequencing technique) will be provided using the resources of the Genome Database of Latvian Population, additionally around 300 samples for lrGS are expected to be generated during the

Lithuania: no legacy data is available. WGS data will be collected under the Lithuanian Genome project, which launched in 2023, aiming to collect 1,570 WGS samples in total.

Luxembourg: no legacy WGS data. Up to 500 samples are recruited from existing biobanks (ORISCAV-LUX, n=1,374 and ~3,000 controls from control populations in clinical studies), following necessary re-consenting. Sequencing will be done at the LuxGen Genome Center using NovaSeq6000.

Malta: sample collection will be done from scratch, starting from the national DWARNA project to collect samples from the general population. Sequencing will be outsourced as capacity is not currently available.

Netherlands: Samples: 5,000 short read WGS and 1,000 long read WGS, from 5 cohort studies (ERGO, LifeLines, Helius, NLTwins Register, Maastricht Study); no legacy data (but maybe 5,000 from Hartwig Medical Foundation); De novo sequencing: Erasmus MC, NovaSeqX 30x short read, ONT Promethion P24-48 for long read and for T2T. Processing: Erasmus MC via Health RI node (likely together with UMCG).

Norway: 1,300 short-read 30X WGS and 46 30X long-reads WGS. The samples will be retrieved from Biobank Norway. DNA is available. No legacy data is available. Sequencing will be done at NorSeq -The Norwegian Consortium for Sequencing and Personalized Medicine.

Poland: 10,000 samples, legacy data:3,000(PASICHB), 400 (UNILODZ access via Polish FEGA Node),existing collections of samples POPULOUS: 8,000 anonymised samples (UNILODZ), 6,400 samples will be recruited from both collections, Sequencing: Illumina NovaSeq 6000 (UNILODZ), NovaSeq X Plus (PASICHB); IT infrastructure storage: available: 1PB (long storage, Polish FEGA Node) 1PB + 36TB ALL FLASH 250 000 IOPS; Illumina Dragen; virtualization cluster.

Portugal: Samples: expected 3,000 from existing datasets in Biobanks and new collection; some reconsenting needed. No legacy data available. Sequencing to be carried out at the National Institute of Health Doutor Ricardo Jorge (INSA), NovaSeq X Plus System.

Slovenia: Samples: Expected 2,440 WGS (488 WGS in first round of GoE), 67 legacy WGS data (reconsenting required). Already collected 479 blood/DNA samples of the general population with

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

consent to be included into the final GoE dataset. Sequencing: Illumina NovaSeq 6000, Promethion P24 and IT infrastructure for bioinformatics (processing capacity, 500TB of short-term storage and 500TB of long-term storage).

Spain: Samples: Expected 12,000 WGS, no legacy samples. Participants will be recruited at the National IMPaCT cohort coordinated by CIBER. This general population cohort is being established in primary health care centres across the whole country funded by ISCIII grants. The first 12,000 participants will be included in the final GoE dataset. Participants will sign the informed consent at recruitment. Samples will be stored at ISCIII biobank. Sequencing will be carried out by CNAG using short-read technology (Illumina).

Sweden: Samples: Expected 2,600 WGS. 1,400 legacy samples (Illumina) from the SweGen dataset and other data collections. 1,200 long-read WGS to be obtained from biobanked blood samples. Sequencing: The SciLifeLab National Genomics Infrastructure (NGI) will perform lrGS on PacBio Revio and ONT PromethION systems. NGI also has multiple Illumina NovaSeqX instruments that can be used within the GoE project.

Turkey: Samples: expected 1,500 from existing database in IBG Biobank.

Table 2 Summary of the new GDI Deliverable 6.2. Questionnaire

Country	In GDI	GoE Proposal expected	GoE Proposal legacy
Austria		2113	No legacy samples
Belgium	Yes	2903	100
Bulgaria	Yes	1853	No legacy samples
Croatia	Yes	1257	60
Cyprus		310	No legacy samples
Czechia	Yes	2945	300
Denmark	Yes	1414	No legacy samples

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

Estonia	Yes	368	No legacy samples
Finland	Yes	7400	4665
France	Yes	7000	10000
Germany	Yes	19400	No legacy samples
Greece		2500	200
Hungary		No expected	No legacy samples
Ireland	Yes	2000	No legacy samples
Italy	Yes	15800	2400
Latvia	Yes	300	3000
Lithuania	Yes	1570	No legacy samples
Luxembourg	Yes	500	No legacy samples
Malta		No expected	No legacy samples
Netherlands	Yes	6000	5000
Norway	Yes	1346	No legacy samples
Poland		10000	3400
Portugal	Yes	3000	No legacy samples
Romania		No response	No response

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

Slovenia	Yes	2440	67
Spain	Yes	12000	No legacy samples
Sweden	Yes	2600	1400
Turkey*		1500	No legacy samples
United Kingdom		No response	No response
Total (GDI participants)		92096	26992

*Turkey joined in to the consortium at the stage of GoE application

5. Discussion

We have systematically assessed the contributions of European countries towards the Genomic Data Infrastructure (GDI), highlighting the pivotal role of collaborative efforts across nations. It demonstrates the diverse landscape of genomic data collection, processing, and storage across Europe, with significant variances in capabilities, technologies, and approaches. The collective effort to harmonise and integrate legacy whole genome sequences (WGS) with newly collected DNA samples underscores the importance of interoperability and standardisation in genomic data management. Challenges such as re-consenting, ethical considerations, and data privacy have been identified, necessitating ongoing dialogue and consensus-building among stakeholders. The adoption of both short-read and long-read sequencing technologies reflects the evolving nature of genomic research, which demands flexible and forward-thinking data infrastructure.

6. Conclusions & Impact

The Deliverable 6.2 Report represents a significant stride towards establishing a comprehensive, inclusive, and sustainable genomic data infrastructure in Europe. By aggregating and analysing data from 27 European countries, it lays the groundwork for enhanced genomic research and healthcare

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

applications. The identified legacy WGS samples and the integration of new DNA sequencing technologies enrich the GDI, facilitating cross-border research collaborations and healthcare advancements. The findings underscore the necessity for sustained coordination, capacity building, and investment in genomic data management infrastructure across Europe. The collaborative effort set forth in this report not only advances the objectives of the GDI but also exemplifies the collective commitment to advancing genomic science and healthcare at a continental scale.

7. Next steps

1. Enhance Data Integration from Clinical Genome Sequencing: We will extend to the integration of data from clinical genome sequencing into the GDI. This includes establishing protocols for the efficient and secure sharing of clinically relevant genomic data, enhancing the database's utility for medical research and patient care. This approach will be implemented in the following points 2 and 3.
2. Adopt a Global Approach: We will expand collaboration beyond European borders to include potential sources of genomic data worldwide (D6.4). This global approach aims not only to enrich the GDI with diverse genomic data but also to ensure that the initiative actively promotes diversity and inclusion. By engaging with populations from varied genetic backgrounds, especially those underrepresented in current genomic databases, we can foster a more comprehensive understanding of genetic factors in health and disease. This commitment to diversity and inclusion will enhance the relevance and applicability of genomic research, ensuring that advancements in healthcare and precision medicine benefit all global communities equitably.
3. Re-visit European Countries, Especially the UK: We will conduct a focused review of genomic data infrastructure developments in European countries, with particular attention to the UK. This review should assess progress, identify gaps, and highlight opportunities for further integration and collaboration, both for research-originated and clinical genomes.
4. Ethical and Legal Frameworks: In the development and refinement of ethical and legal frameworks, we will actively collaborate with WP9 Ethics Requirements to address the unique challenges that arise from including diverse populations and cultures in genome sequences. This partnership will focus on crafting guidelines for inclusion of new genome data in GDI in a way that respects the wide array of ethical considerations inherent to different groups, ensuring that data privacy, consent, and international data sharing policies are sensitive to cultural differences. This collaborative approach will not only enhance the robustness of our frameworks but also promote trust and inclusivity in genomic research.

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

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