European Genomic Data Infrastructure

Deliverable D1.4

Genome of Europe plan

Project Title Grant agreement no	**Genomic Data Infrastructure** Grant agreement 101081813		
Project Acronym (EC Call)	GDI		
WP No & Title	WP1: Coordination & Support		
WP Leaders	Hannah Hurst (1. ELIXIR Hub)		
Deliverable Lead Beneficiary	2. ERASMUS MC		
Contractual delivery date	30/04/2023	**Actual delivery date**	08/06/2023
Delayed	Yes		
Partner(s) contributing to deliverable	UTARTU, HealthRI		
Authors	EMC		
Contributors	Jeroen van Rooij (EMC), André Uitterlinden (EMC)		
Acknowledgements	Andres Metspalu (UTARTU), Mariliis Vaht (UTARTU), Rob Hooft (HealthRI)		
Reviewers	-Tommy Nyronen -Alfonso Valencia		

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

European Genomic Data Infrastructure

Log of changes

Date	Mvm	Who	Description
14/04/2023	First draft	Jeroen van Rooij (EMC)	First draft of complete document, left comments on things to be checked.
16/04/2023	revision	André Uitterlinden (EMC)	Revision of first draft
08/06/2023	Final	André Uitterlinden (EMC)	Version ready for submission following Management Board review

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

The Genome of Europe (GoE) is WG12 of the 1+MG initiative, started in 2020 and is coordinated by EMC/Netherlands (Uitterlinden, van Rooij) and UTARTU/Estonia (Metspalu, Vaht). It is aiming to deliver >500,000 whole genome sequences (WGS) as a reference database to represent genetic diversity across Europe, and was chosen as a use case in the GDI project. This deliverable for GDI is a short report on the alignment of GoE working group planning as use case with GDI. In particular we provide an update on recommendations derived from the GoE WG12 discussions regarding: a) composition of the samples contributing to GoE; b) contributions expected/requested per country regarding number of samples; c) WGS technology and data requirements; and d) sample recruitment into GoE. The recommendations will facilitate and structure further discussion and decision-making in GoE and impact the GDI project. The GoE-specific recommendations align with other existing 1+MG working groups regarding ELSI, data standards, data quality and technical infrastructure. We suggest pilot studies and use-cases derived from GoE discussions based on the data collection within GoE. The GoE pilot studies and use cases will ensure the GDI infrastructure is developed in accordance with GoE user needs. The storage of the data generated within GoE (estimated to be ~150-300 petabyte raw WGS data) but also data generated in other WG's, is an issue that deserves close attention within GDI. We also point to –once GoE will be finished- what other datatypes will be generated as a sequela (e.g., multi-million array genotypes across various biobanks and cohort studies in WG10), and what analyses should be accommodated within the Genomic Data Infrastructure.

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.	No
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 finalization.	Yes
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

The GoE use case was discussed with GDI members by the GDI WP6 leader, to explain what the status of GoE is, what GoE is doing and designing, and discuss possible (more technical) pilot studies to work with WGS data. Such pilot studies are needed to explore the needs and (im)possibilities for storage and access to WGS data generated by GoE.

Yet, GoE is also discussing the uses of WGS data that GoE will generate to allow the 1+MG having (clinical) impact, under the name of Use Cases. See the "work accomplished" paragraph for a more detailed overview of the pilot studies and use cases.

The below introduction and use cases (I, III, and IV) were presented and discussed by the WP leader, as well as ideas on data-flow and federated execution of the GoE use cases. We plan further discussing the details of how the GDI infrastructure develops, and how WP6 may provide feedback to GDI members to ensure that the GoE pilot studies and use-cases can be executed on GDI solutions.

4. Description of work accomplished/Results

4.1 Short recap of the Genome of Europe design. One of the objectives of GoE is to contribute to the 1+MG effort initialized in 2018 when 24 signatory countries declared to make the data accessible for the genomes of at least 1,000,000 European citizens. This will require existing genomes to be made available for access and new genomes to be created by whole genome sequencing. The rationale for creating a GoE is to have a large and European-wide collection of subgroup-specific reference datasets as "normal or Reference" genomes. These can be used a) to document and quantify genetic diversity and heterogeneity across European populations, such as b) to interpret potentially clinical or pathogenic genetic variants in comparison to disease-specific genomes, but also c) such as to recalibrate genetic risk profiles to ancestral backgrounds, which is important for developing population specific polygenic risk scores, and d) as a reference panel for imputations in lower resolution but larger scale array genotyping efforts (1+MG WG10), these use cases are further described below. Overall, the GoE cohort will create the reference dataset for genomic health programs of the European countries. Initially, minimal/no phenotype information is required for GoE samples, being: age at blood draw, sex and country of origin, while the inclusion of more phenotypic characteristics is clearly encouraged, as this enhances the added value of the datasets from user perspective. Data will be collected following the 1+MG Trust Framework, including ethical-legal issues, data quality and format, metadata and technical inter-operability aspects.

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

4.2 Composition of the GoE members. Signatory and observer states participating in the 1+MG were asked before the start of GoE/WG12 to nominate two expert delegates each to participate in the GoE discussions. Currently, 23 countries are represented in the GoE, with only Austria, Cyprus, Greece and Poland not included yet. France and Ireland recently changed their status from observer countries to signatory countries, while Switzerland is the only participating observer country in GoE now.

4.3 GoE recommendations: We have established recommendations for execution of generating 500,000 WGS datasets within GoE based on the discussions held in and around the WG12/GoE, which included making several inventories by questionnaire across the members of WG12. Such recommendations are important for the GDI project in that it will direct the development of solutions for GoE. From the GoE, a document detailing these recommendations was released in November 2022. Briefly, these recommendations regarded:

1) Number of samples per country to contribute to GoE, The WG12 reached consensus on making this proportional to the number of inhabitants per country with larger countries delivering more samples than smaller countries, to be as representative as possible for the European population as a whole.
2) Minimum number of samples to contribute. The WG12 reached consensus on a minimum of 1,000 samples per group of samples to be defined as a subgroup of the European population. This ensures detection of rare frequency variants and a robust sample size to ensure comprehensive detection of more common variants.
3) Ancestral composition of GoE, The GoE coordination group asked the members to provide information on composition of their population in terms of people by "country of origin" as a pragmatic proxy for ancestry (which is in essence a genetically defined concept). Keeping the total number of 500,000 in GoE and the minimum number per subgroup in mind, this resulted in identification of 40 subgroups in GoE to be contributed by the 23 members. These 40 subgroups include ancestries outside of Europe (e.g., Turkey, Morocco, Algeria, Russian), and subgroups not having a separate country (e.g., Roma). We excluded ancestries which are knows to be represented by genome programs in their own home country (e.g., UK, China).
4) Suggested contribution per ancestry per country. From the above recommendations and information we provided a suggested contribution per ancestry per participating country. This took into account the 40 subgroups to be contributed by various member states to distribute the workload and to account -to some extent- for variability in ancestral background of the various subgroups in different European countries.
5) Inventory of sequencing technology, capacity, existing data and funding. The GoE coordination group asked members by questionnaire to provide information of capacity for sequencing technology, existing WGS data that could be used for GoE, and the local funding situation to contribute to GoE. We will leave the funding situation outside this deliverable for

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

the moment, and focus on the WGS technology which aspect is important for the GDI project. The overwhelming majority of members have capacity for sequencing based on Illumina short read technology, and some members had already generated some datasets with that technology, which might be eligible for inclusion in GoE (so-called "legacy data"). In addition, some members were using MGI short read technology and several others had long read technology available as capacity (i.e., PacBio, Oxford Nanopore Technology/ONT). In addition some members were active in exploring use of the latest sequencing technology (e.g., Element, Ultima, Illumina long read chemistry). The GoE members reached consensus on using the Illumina short read technology as a starting point for GoE, but also to allow WGS data generated with other technology to be included in GoE, and explore the use of long read technology at a later stage. The inclusion of several different WGS technologies will of course be a challenge for creating a GDI database of GoE but by being inclusive will allow more data to be included and will be a better preparation to reflect reality at a later stage, when technology progresses and members will make own choices in that respect.

4.4 Suggested Pilot Studies and Use Cases, for envisioned GoE data applications in GDI.

The WP6 GoE group is planning some pilot studies to be undertaken within the GDI context, which are of a more technical nature and need further discussion within GDI and therefore may be subject to change depending on feedback.

Focusing on future applications of GoE data we also here present use cases to highlight possible uses of GoE data and for which GDI will develop solutions. Use cases I, II, and III require just accessing the genetic information per se, and transferring extractions of such genetic data across institutes and borders. Use case IV will also involve use of phenotypic/disease information in local biobanks/cohort studies and genetic data based on genotyping arrays (and not necessarily WGS) and represents a next level of complexity. We briefly describe the concept behind each pilot study and use case below.

4.4.1 Pilot studies:

A. *Exchange of synthetic WGS data.* Several GoE members have been working on creating synthetic datasets of WGS data based on existing databases containing European WGS data. In particular, there are Finnish data in the 1000genomes reference dataset which have been used by the group of Markus Perola/Terro Hiekkalinna to create a large number of so-called synthetic genomes by scrambling and reconstituting such genomes. Such synthetic WGS data will be used to explore the possibilities of storing and exchanging (extracts of) such WGS data across institutes and borders with a limited set of GoE members (e.g., Finland, Estonia, Netherlands). This will allow GDI to understand possibilities and hindrances regarding exchange of WGS data and/or extractions thereof.

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

B. *Exchange of real WGS data.* Similarly, several GoE members have already limited WGS datasets with appropriate consent and/or samples that are not subject to GDPR restrictions (e.g., DNA samples from Coriell collection) allowing access to and exchange with scientific collaborators, which could be used for exploring the *real-world* possibilities and hindrances regarding exchange of (extractions of) WGS data, as a follow up to the pilot with synthetic data. This will involve a few more GoE members (than in pilot A) who indicated to have such WGS data available within the near future.

4.4.2 GoE Use Cases:

Use-case I – Look-ups of individual variants across the GoE. As the simplest form of a genetic query, this use case is assessing the frequency of a specific genetic variant of interest across the populations of Europe. This query is interesting to researchers and clinicians, when having observed a (rare) genetic variant in a patient and wanting to evaluate the possibility of it being pathogenic. The use case should allow to develop the workflow of a straightforward query, and troubleshoot on connecting GoE data through existing or novel infrastructure. As well as address the ELSI challenges in approaching GoE data internationally, making it findable, FAIR, etc.

Use-case II – Creating a Principal Component Analysis (PCA) of genetic variation across GoE WGS data. Once sufficient GoE WGS data has been made accessible we can perform a PCA analysis to discover how many and which sub-groups can be distinguished based on purely genetic information, how it relates to the self-reported information on country of origin, and how the inter-relationship of the sub-groups looks like. Based on the current, relatively modest amounts of WGS data in reference populations, this is now possible for the large continental groups (i.e., distinguishing White Europeans from Africans from Asians) and based on relatively small number of genetic variants from SNP microarray genotyped populations also in some cases at a somewhat higher geographic resolution. With GoE WGS data becoming available for many if not all of the 40 subgroups now defined in GoE, this PCA can be done at an unprecedented level of precision and resolution leading to new insights in European genetic diversity and relationships, and will lead to redefining ancestry for these populations which is important for the following use-cases. Knowing the correct ancestry is important in clinical settings given the genetic nature of most if not all diseases and the wide variety in incidence and expression of diseases for different ancestry groups. Yet, one can also foresee this PCA data will be important beyond clinical applications, such as for identification purposes in forensic settings, and thus such data access needs to be discussed extensively also in the GDI.

Use-case III – Generating reference panels for ancestry-specific imputation. Imputation is a process whereby genotyped datasets (e.g., biobanks genotyped with SNP micro arrays with ~1 million genetic variants) can be enriched to > 100 million variants by aligning them with so-called reference datasets (consisting of whole genome sequenced genomes with ~3 billion genotypes). This pilot study will

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

use GoE reference genomes (GoE as a whole or just from certain subgroups) to impute other genetic datasets against (e.g., SNP microarray genotyped biobanks/cohort studies generated by members). The currently existing imputation reference genome datasets (1000G, haplotype reference consortium (HRC) panel, TopMed) mostly consist of WGS data on samples from outside Europe, mostly USA, and are not including the rich genetic variation that exists within Europe. Moreover, GDPR regulations prohibit the direct access to such imputation reference datasets and these data are not released from USA to Europe (e.g., TopMed data).
The GoE WGS reference dataset will be used to ascertain correlations between individual variants observed within and across ancestries/subgroups in GoE, and constructing these into so-called haplotypes (combinations of correlated genotypes) that inform scientists on which genetic variants to measure (by array genotyping), and how to infer the unmeasured genotypes (by imputation to the appropriate reference panel). The resolution (i.e., number of genotyped variants imputed) of these imputations increases with sample size, and population diversity, and thus requires the largest amount of data to be accessed simultaneously. Similar to the previous use-cases, this could be done at a national level, but doing this on the GoE as a whole has the obvious advantage of capturing much more of the genetic variation including all subgroups in GoE. It will require access to the full genotype data of all GoE participants and might involve the creation of a central European imputation service (and server handling the data, located within the EU).

Use-case IV – Providing population distributions for polygenic risk scores (PRS). In genetic medicine, the so-called polygenic risk scores (PRS) are considered valuable additions to early detection and/or prevention of diseases. These scores are the risk-weighted sums of hundreds or thousands of genetic variants associated with a specific disease, the products of genome-wide association studies (GWAS). To evaluate the risk of disease in a single person, its weighted sum (polygenic score) should be compared to a reference population of the same ancestry. This use case is more complex than the first, because it requires a query of multiple variants and across multiple samples (all of a single ancestry). However, it also provides the option to share data at different levels; open access to all variants for all samples would be easiest, but we could also share the polygenic score value of GoE participants (which can be anonymized, as it does not contain individual variant genotypes) or even just the distribution curve of this score in that ancestry. Calculating the scores or the reference distributions could be done locally, nationally or internationally, depending on the ELSI and IT guidelines.

5. Conclusions & Impact

We have described pilot studies and use cases but it is of course still an early phase for GDI to come up with concrete solutions for GoE initiatives. Importantly, since there is no funding for GoE as a project as of yet (April 2023) the discussions have remained somewhat theoretical of course with

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

little impact for GoE. We have formulated draft requirements for infrastructure from the GoE perspective and we think the pilot studies A and B we propose can be executed at a relatively short term. The impact of GDI for GoE is limited therefore at this point in time, but this is likely to change soon (see below under Next Steps).

6. Next steps

As indicated before, the Genome of Europe as a separate working group WG12 within the 1+MG initiative is proceeding with the design and logistics of getting the GoE project up and running. There are various milestones reached there already, see e.g., the recommendations listed in the 1+MG WG12 (GoE) progress report, and we expect data flow to start within the next two years based on initiatives in several individual countries, and some existing WGS legacy data.

In addition, GoE has been approached by several industrial parties working on WGS data infrastructure (e.g., Illumina, BC Platforms, Nexus) who are offering data infrastructure solutions and help (at fee for service of course). They however also complained about the lack of involvement with the GDI project as a whole, and the lack of clear communication channels to initiate discussions if and how they can be involved in GDI and/or GoE. We therefore suggest organizing discussions with GDI, GoE, and such industrial parties to see if and how they can contribute.

We noted, somewhat to our surprise, in our initial discussion with GDI representatives that the simple storage of the WGS data generated within GoE (estimated to be ~150-300 petabyte raw WGS data across all members) but also data generated in other WG's, is not planned or funded under GDI. This is an issue that deserves close attention within GDI, but also within the member states of the GoE WG12, and in discussions of GDI with the use cases, such as GoE.

Importantly, a new situation has developed since there is now (April 2023) a call for proposals in the Digital Europe program in which the Genome of Europe is specifically described. So, the GoE WG12 project group is preparing to submit a proposal and we therefore, upon successful review and approval, expect somewhere in 2024 to start generating substantial amounts of WGS data through the then funded GoE project. This will also create for GDI a new situation for which further discussions will be needed.

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