Beyond One Million Genomes

D4.3 Secure cross-border data access roadmap - 2v0

Project Title (grant agreement No)	Beyond One Million Genomes (B1MG) Grant Agreement 951724
Project Acronym	B1MG
WP No & Title	WP4 - Federated Secure Cross-border Technical Infrastructure
WP Leaders	Tommi Nyrönen (CSC), Ilkka Lappalainen (CSC), Bengt Persson (UU), Sergi Beltran (CNAG-CRG)
Deliverable Lead Beneficiary	4 - CSC
Deliverable	D4.3 - Secure cross-border data access roadmap - 2v0
Contractual delivery date	31/05/2023
Actual delivery date	27/10/2023
Delayed	Yes
Authors	Dylan Spalding (CSC), Tommi Nyrönen (CSC), Riku Riski (CSC)
Contributors	
Acknowledgements	Christophe Trefois (LNDS)

Beyond One Million Genomes
B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

(not grant participants)	
Deliverable type	Report
Dissemination level	Public

Document History

Date	Mvm	Who	Description
27/09/2023	0v1	Dylan Spalding (CSC)	Initial draft written and circulated to WP participants for feedback
19/10/2023	0v2	Nikki Coutts (ELIXIR Hub)	Circulated to OG, Stakeholders and GB for review
27/10/2023	0v3	Dylan Spalding (CSC)	Final comments closed
27/10/2023	1v0	Nikki Coutts (ELIXIR Hub)	Version uploaded to the EC Portal

Table of Contents

1. Executive Summary 4
2. Contribution towards project objectives 5
3. Methods 6
4. Description of work accomplished 7
4.1 Standards 7
4.1.1 Beacon 7
4.1.2 Phenopackets 7
4.1.3 Passport 8
4.1.4 Authentication and Authorisation Infrastructure 8
4.1.5 Data Use Ontology 8
4.1.6 Htsget 8
4.1.7 Crypt4GH 9
4.1.8 File Formats 9
4.1.9 Task Execution Service, Workflow Execution Service, and Tool Registry Service 9
4.1.10 refget 10
4.1.11 Data Repository Service 10
4.2.1 Proof-of-Concept 10
4.2 Interoperability 12
4.2.1 Interoperability between use-cases 13
4.2.2 Interoperability between data spaces 13
4.2.2.1 SIMPL 13
4.2.2.2 eID 14

B1MG has received funding from the European Union’s Horizon 2020 Research and Innovation programme under grant agreement No 951724

4.3 Alignment with GDI 15
4.3.1 GDI 15
4.3.2 Genome of Europe 16
4.3.3 Testing and Deployment 17
5. Results 17
6. Discussion 23
7. Conclusions 23
8. Next steps 23
9. Impact 23
10. References 23

B1MG has received funding from the European Union’s Horizon 2020 Research and Innovation programme under grant agreement No 951724

1. Executive Summary

This document outlines the updated roadmap of the EU 1+ million genomes initiative for the deployment of a network of nodes allowing secure cross border genomic and phenotypic data access, primarily as part of the Genomic Data Infrastructure project. The document outlines the proposed initial infrastructure, including the standards, application, and technologies within the infrastructure, and describes a proof-of-concept that was used to demonstrate the proposed infrastructure, and elicit feedback from a range of stakeholders. The document then goes on to outline the key technology (such as eID and SIMPL), standards (e.g. ISO, GA4GH) and policy (Data Protection by Default and Design principles, 1+MG trust framework) developments at a European level to help enable interoperability between different data spaces (e.g European Health Data Space). Work by 1+MG experts ensures relevance to the Genomic Data Infrastructure and improved alignment between prospective European data spaces. A roadmap is presented which is aligned with the milestones and deliverables within the Genomic Data Infrastructure project. The document references the existing Genomic Data Infrastructure helpdesk roadmap, which defines how operations will be set up following the FitSIM standard.

2. Contribution towards project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives/key results:

[Select 'Yes' (at least one) if the deliverable contributed to the key result, otherwise select 'No'.]

	Key Result No and description	Contributed
Objective 1 Engage local, regional, national and European stakeholders to define the requirements for cross-border access to genomics and personalised medicine data	**1.** B1MG assembles key local, national, European and global actors in the field of Personalised Medicine within a B1MG Stakeholder Coordination Group (WP1) by M6.	No
	2. B1MG drives broad engagement around European access to personalised medicine data via the B1MG Stakeholder Coordination Portal (WP1) following the B1MG Communication Strategy (WP6) by M12.	No
	3. B1MG establishes awareness and dialogue with a broad set of societal actors via a continuously monitored and refined communications strategy (WP1, WP6) by M12, M18, M24 & M30.	No
	4. The open B1MG Summit (M18) engages and ensures that the views of all relevant stakeholders are captured in B1MG requirements and guidelines (WP1, WP6).	No
Objective 2 Translate requirements for data quality, standards, technical infrastructure, and ELSI into technical specifications and implementation guidelines that captures European best practice	**Legal & Ethical Key Results**	
	1. Establish relevant best practice in ethics of cross-border access to genome and phenotypic data (WP2) by M36	No
	2. Analysis of legal framework and development of common minimum standard (WP2) by M36.	No
	3. Cross-border Data Access and Use Governance Toolkit Framework (WP2) by M36.	No
	Technical Key Results	
	4. Quality metrics for sequencing (WP3) by M12.	No
	5. Best practices for Next Generation Sequencing (WP3) by M24.	No
	6. Phenotypic and clinical metadata framework (WP3) by M12, M24 & M36.	No
	7. Best practices in sharing and linking phenotypic and genetic data (WP3) by M12 & M24.	No
	8. Data analysis challenge (WP3) by M36.	No
	Infrastructure Key Results	
	9. Secure cross-border data access roadmap (WP4) by M12 & M36.	Yes
	10. Secure cross-border data access demonstrator (WP4) by M24.	No

Objective 3 Drive adoption and support long-term operation by organisations at local, regional, national and European level by providing guidance on phased development (via the B1MG maturity level model), and a methodology for economic evaluation	**1.** The B1MG maturity level model (WP5) by M24.	No
	2. Roadmap and guidance tools for countries for effective implementation of Personalised Medicine (WP5) by M36.	No
	3. Economic evaluation models for Personalised Medicine and case studies (WP5) by M30.	No
	4. Guidance principles for national mirror groups and cross-border Personalised Medicine governance (WP6) by M30.	No
	5. Long-term sustainability design and funding routes for cross-border Personalised Medicine delivery (WP6) by M34.	No

3. Methods

This document describes the final evolution of the 1+MG roadmap to provide the technical infrastructure required to support secure cross border access to genetic and associated phenotypic, clinical, and pedigree data with the European Union.

Monthly coordination meetings were held for WP4 and WG5 throughout the project, to which other WPs and WGs were invited. This ensured the infrastructure proposed by WP4 supports the requirements of the other work packages within Beyond 1 Million Genomes (B1MG), as well as those of the 1+ Million Genomes (1+MG) working groups.

A proof-of-concept (PoC) demonstrator (https://youtu.be/6MtIJA4xXdU) was created, in collaboration with 1+MG WG8 (rare disease). The PoC allowed feedback and reactions from different WPs and WGs, and also demonstrated the capabilities of the emerging infrastructure services and allowed the determination of gaps where these exist. The PoC was demonstrated to WGs 9, 10 and 11. The PoC was then extended to support WG9 (cancer) use case, with different data discovery and processing functionalities needed in data management and analysis, specific to a highlighted cancer data case in melanoma.

The five core functionalities for 1+MG technical infrastructure required by the initiative were outlined in this work and described in a scoping paper, and include data discovery, data reception, access management, storage and interfaces, and processing or data analysis. Standards were proposed in the original PoC to enable these functionalities, The PoC and the five functionalities allowed a step towards a service architecture for technical infrastructure required to support secure cross border access. The next step is to deploy European services outlined in this architecture, made in the Starter Kit for the European Genomic Data Infrastructure (GDI), first released in June 2023. GDI is tasked with the deploying of the proposed infrastructure recommended by B1MG. The overall process is endorsed by the 1+MG member states EU special group.

4. Description of work accomplished

4.1 Standards

As described in the earlier version of the roadmap (D4.1 Secure cross-border data access roadmap)[1], the infrastructure will make use of standards from the Global Alliance for Genomics and Health (GA4GH), outlined below, which were used as the basis for a proof of concept (PoC) as part of D4.2[2] Secure Access Demonstrator. European experts are actively contributing to creating and maintaining GA4GH standards, helping to ensure the European use cases are supported.

4.1.1 Beacon

The Beacon[3] standard is a GA4GH standard that was originally designed to support queries on alleles (gene variants). Beacon aims to protect the identity of the participants by returning a response based on the existence of an allele within a cohort as a boolean or count dependent on the user's access level. This allows users to perform discovery queries without having access to the data a priori. The standard was updated to version 2 in April 2022, which allows record level queries, queries on phenotype and assay as well as extended variant queries, while maintaining the restrictions of the original standard when data protection principles require it. . For example, registered users who have an authenticated electronic identity and a 'bona-fide'[4] status within the LifeScience AAI[5] (see Passports section below) could access more information than anonymous users, depending on the settings of the individual beacon. The beacon standard allows a network of beacons to be queried using a single query, supporting federated discovery. Further updates to the standard with European contributors are expected in the near future.

4.1.2 Phenopackets

Transfering phenotypic data between different resources in a standardised way is a requirement to ensure that both phenotypic queries and data transfer are consistent across a federated network. Phenopackets[6] provide the information model that different levels of clinical and phenotypic data require in order to be exchanged. Phenopackets are a GA4GH standard, and also an ISO standard[7] for the exchange of phenotypic data. One of the benefits of the use of standards is evidenced by the loading of phenotypic data into a deployment of the Beacon reference implementation, the use of a standardised representation means the same loading and annotation pipeline can be used.

4.1.3 Passport

The Passport[8],[9] is a GA4GH standard that defines a standard way to represent the role and access rights of a particular user. It can be used to indicate if a particular user can access

[1] https://doi.org/10.5281/zenodo.6139231
[2] https://doi.org/10.5281/zenodo.7590822
[3] https://github.com/ga4gh-beacon/specification
[4] https://www.nature.com/articles/s41431-018-0219-y
[5] https://lifescience-ri.eu/ls-login/
[6] https://github.com/phenopackets/phenopacket-schema
[7] https://www.iso.org/standard/79991.html
[8] https://doi.org/10.1016/j.xgen.2021.100030
[9] https://github.com/ga4gh-duri/ga4gh-duri.github.io/blob/master/researcher_ids/ga4gh_passport_v1.md

Beyond One Million Genomes
B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

registered access resources, such as the Beacon, using the 'bona-fide' researcher attribute as implemented by ELIXIR[10]. It also details which datasets the user has full controlled access to (via Controlled Access Visas), their affiliation(s), and their roles. Within a federated infrastructure, the LinkedIdentities visa facilitates the linking of identities of an individual from multiple identity providers, where in the 1+MG setting at least one must be of a high level of assurance (e.g. eID). Hence taken together, user identity (e.g. Life Science Login service) and access permission (Passport integration) allows a distributed ecosystem of data nodes and trusted computing environments to together deliver services for a user, who in turn may login to these services with a single, e.g. institutional, identity.

4.1.4 Authentication and Authorisation Infrastructure

For GA4GH Passports to work, there must be a standard for the infrastructure that details how to define the user's identity, and how to pass this information between federated service providers (e.g. data holder for 1+MG virtual cohort) in a consistent way. The GA4GH approved an Authentication and Authorisation Infrastructure (AAI) OpenID Connect (OIDC) profile[11] for this purpose, which extends the existing OpenID Connect[12] (OIDC) standard to ensure maximum compatibility with existing identity providers and relying parties.

4.1.5 Data Use Ontology

There are a diverse range of conditions which different cohorts apply to restrict use of their data. These can range from geographical restrictions, to restricting research for particular purposes, to requiring results to be published. To standardise the way these different requirements were communicated in a machine-readable way, the Data Use Ontology[13] (DUO) was developed. This ensures that equivalent data use conditions may be expressed in the same way across different cohorts in a cross-border setting, such as the 1+MG initiative. This is especially important when the cohorts are based in different legal jurisdictions. The DUO also facilitates data discovery, for example via Beacon so when a researcher is investigating heart disease, the DUO can be used to ensure that only datasets that can be used for research into heart disease are returned as possible cohorts of interest.

4.1.6 Htsget

Htsget[14] is a data streaming specification to allow the secure and encrypted transfer of genomic data between locations over https, delivering the data in plaintext ready for use. The standard supports range queries, for example allowing the user to just stream a region of a genome with a gene of interest from a whole genome CRAM or BAM file, instead of downloading the whole file. Htsget technology is important also for data protection and minimisation purposes.

4.1.7 Crypt4GH

Genomic data contains sensitive data on individuals, and as such preventing accidental disclosure of these data is extremely important. Crypt4GH[15] is a file format that stores such data

[10] https://elixir-europe.org/services/compute/aai/bonafide
[11] https://ga4gh.github.io/data-security/aai-openid-connect-profile
[12] https://openid.net/specs/openid-connect-core-1_0.html
[13] https://github.com/EBISPOT/DUO
[14] http://samtools.github.io/hts-specs/htsget.html
[15] http://samtools.github.io/hts-specs/crypt4gh.pdf

in an encrypted state. It supports fast decryption of the data, and also indexing allows random access of the files without having to decrypt the whole file. Such advantages are leveraged by other standards, such as htsget which ensures data are encrypted in transit but delivered to the user in plaintext. State-of-the-art does not yet allow encrypted genomic data to be delivered to the user via htsget, although other traditional methods such as Aspera[16] or ftp[17] can be used for this purpose, but a requirement to transfer data to the user in an encrypted format via htsget has been identified and thus may be expected to become available for the 1+MG initiative in the near future.

4.1.8 File Formats

GA4GH standards include the VCF, BAM, and CRAM[18] file formats as well as phenotype schemas such as phenopackets. These ensure that genomic and phenotype information is consistently represented, and help underpin the bioinformatic analysis tools and APIs. For example, using such consistent file formats allows the implementation of htsget on Crypt4GH encrypted CRAM files, allowing secure federated access to remote genetic data. Additionally the Beacon reference implementation utilises the VCF file format to load data into the back-end database.

4.1.9 Task Execution Service, Workflow Execution Service, and Tool Registry Service

When data may not leave a certain jurisdiction or compute resource, the analyses need to travel to the data (data visitation). However, the compute resources (such as secure processing environments (SPEs) or trusted research environments (TREs) that these data reside on may be heterogeneous and require different ways of describing the required analysis, for example on different cloud platforms or high-performance computing systems. The Task and Workflow Execution Services facilitate federated computation across these diverse resources. The Workflow Execution Service[19] (WES) provides a standardised way of representing a workflow; a set of distinct tasks that must be performed to complete an analysis, which can be fed into a workflow engine which manages the workflow. The Task Execution Service[20] (TES) provides a standardised way of defining the tasks the workflow engine manages, allowing a particular workflow engine to manage the workflows across different compute resources or SPEs. The Tool Registry Service[21] (TRS) allows the same operation to be performed on these different compute resources or SPEs, by standardising methods to list, search, and retrieve tools from different registries for different environments. Compute services that are eligible for 1+MG initiative job execution must connect to the same user identity and access process as the data access services.

4.1.10 refget

Genomic reference sequences are required to allow the generation of whole genome or exome data. There are a range of different reference sequences in use, and many of these sequences are known by multiple names. Refget utilises a unique identifier derived from the sequence itself to ensure that the correct sequence is used. The refget[22] protocol allows the reference sequence

[16] https://www.ibm.com/products/aspera
[17] https://datatracker.ietf.org/doc/html/rfc959
[18] https://github.com/samtools/hts-specs
[19] https://ga4gh.github.io/workflow-execution-service-schemas/docs/
[20] https://github.com/ga4gh/task-execution-schemas
[21] https://ga4gh.github.io/tool-registry-service-schemas/
[22] https://doi.org/10.1093/bioinformatics/btab524

to be obtained unambiguously, and subsequently allows file formats such as CRAM to compress genomic data reducing storage costs.

4.1.11 Data Repository Service

There are many data repositories that store genomic data, but these often have different methods of access. The Data Repository Service[23] (DRS) is a standardised way to retrieve a dataset irrespective of the underlying architecture of the repository.

4.2.1 Proof-of-Concept

As part of D4.2, a Proof-of-Concept (PoC) was developed, primarily with the rare disease use case (WG8). This was leveraged to elicit feedback from other WGs, especially WG9 (Cancer) to determine areas where the proposed infrastructure for the rare disease use case did not work for WG9. Two issues were identified:

1. the fact that approval of the Beacon Version 2 standard was delayed so the PoC relied on an implementation of the Beacon version 1, and hence could not support the WG9 use case,
2. and the specific nature of the 'processing' functionality provided by the Genome-phenome Analysis Platform which was not applicable for the WG9 use case.

Feedback from WG2 (ELSI) also resulted in changes in the way the discovery functionality was performed. These included removing the use of MatchMaker Exchange[24] (MME), as data discovery was seen as a single step to define a virtual cohort, supporting the data minimisation principle, and the requirement to perform Beacon searches at the registered level, not the anonymous level. Additionally feedback on how the term 'processing' as described in the PoC and within the GDPR was clarified for better compliance with data protection principles.

To address these issues, the initial discovery query could be done on the 1+MG User Portal, which uses aggregate or anonymous data to describe the data within the infrastructure. As Beacon version 2 was approved before cancer PoC work, it was leveraged to link genotype, phenotype, and treatment queries across a Beacon Network. For the cancer use case a discovery query linking the BRAF V600E biomarker for melanoma, the standard treatment path using vemurafenib, and variants in the PTEN gene was suggested as variants in the PTEN gene can confer resistance to BRAF inhibitors, such as vemurafenib. This means that a standard treatment path for melanoma using vemurafenib for BRAF positive cancer would not be suitable for the individual, since the second mutation (PTEN) would render the treatment with the drug less efficacious. Therefore a Beacon search for an individual with a BRAF V600E biomarker who has been treated with vemurafenib and has a variant in the PTEN gene was demonstrated in the cancer PoC. If suitable data were available in 1+MG virtual cohort for a use case like this, a data access request could be made to the relevant Data Access Committee (DAC) who would then grant or deny access. Subsequently data will help with correct cancer analysis by medical research, and eventually influence the choice of the right treatment paths by medical

[23] https://ga4gh.github.io/data-repository-service-schemas/preview/release/drs-1.2.0/docs/
[24] https://www.matchmakerexchange.org/

professionals. If access was granted, the user could log into a suitable SPE and perform some form of data authorised data analysis, for example using cBioPortal[25] for the cancer use case.

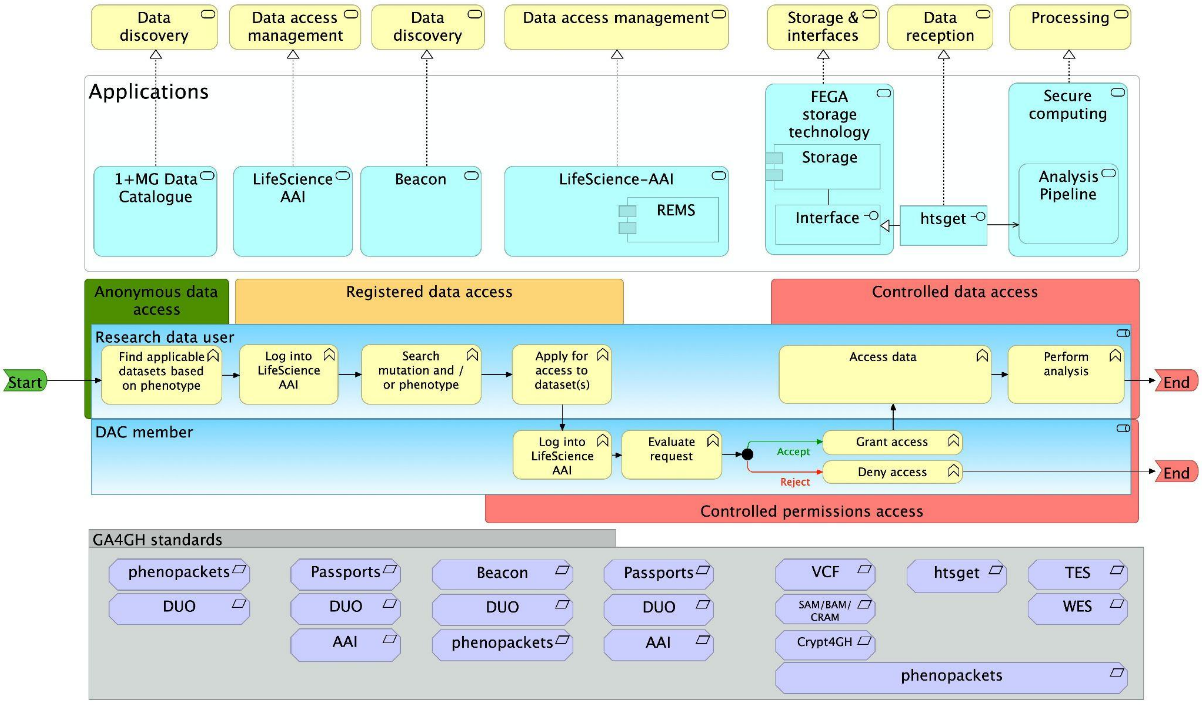


Figure 1: Diagram illustrating the functionalities at the top, the generic infrastructure, research use case, and standards used within the updated PoC at the bottom. The use case in the middle describes how a researcher starts their journey as an anonymous user, logs in to allow discovery queries over individual level data, and finally is granted or denied access. The DAC member utilises the same infrastructure to grant or deny access to these data. The applications that support the research use case, and the defined functionalities, are listed above the use case.

As can be seen from the different tools used for both rare disease (WG8) and cancer (WG9) use cases, the types of data analysis software environments used are distinct. We propose the use of standards, such as WES / TES or Open Containers Initiative[26] (OCI) compatible containers to enable the same technical infrastructure services to support a diverse range of data analysis processes needed to understand genetic and phenotypic data. An example from WG9 was a mutation detection pipeline (Figure 2) which was containerised and run on the secure SD Desktop[27] service at CSC in Finland. During this work it became apparent that such technologies need not be restricted to the final ‘processing’ part of the user story in Figure 1, but can also be used for the data reception and data generation parts, allowing mutation and variation to be called in federated locations using the same pipeline. This ensures the resulting data is harmonised across different nodes.

[25] https://www.cbioportal.org/
[26] https://opencontainers.org/
[27] https://research.csc.fi/-/sd-desktop

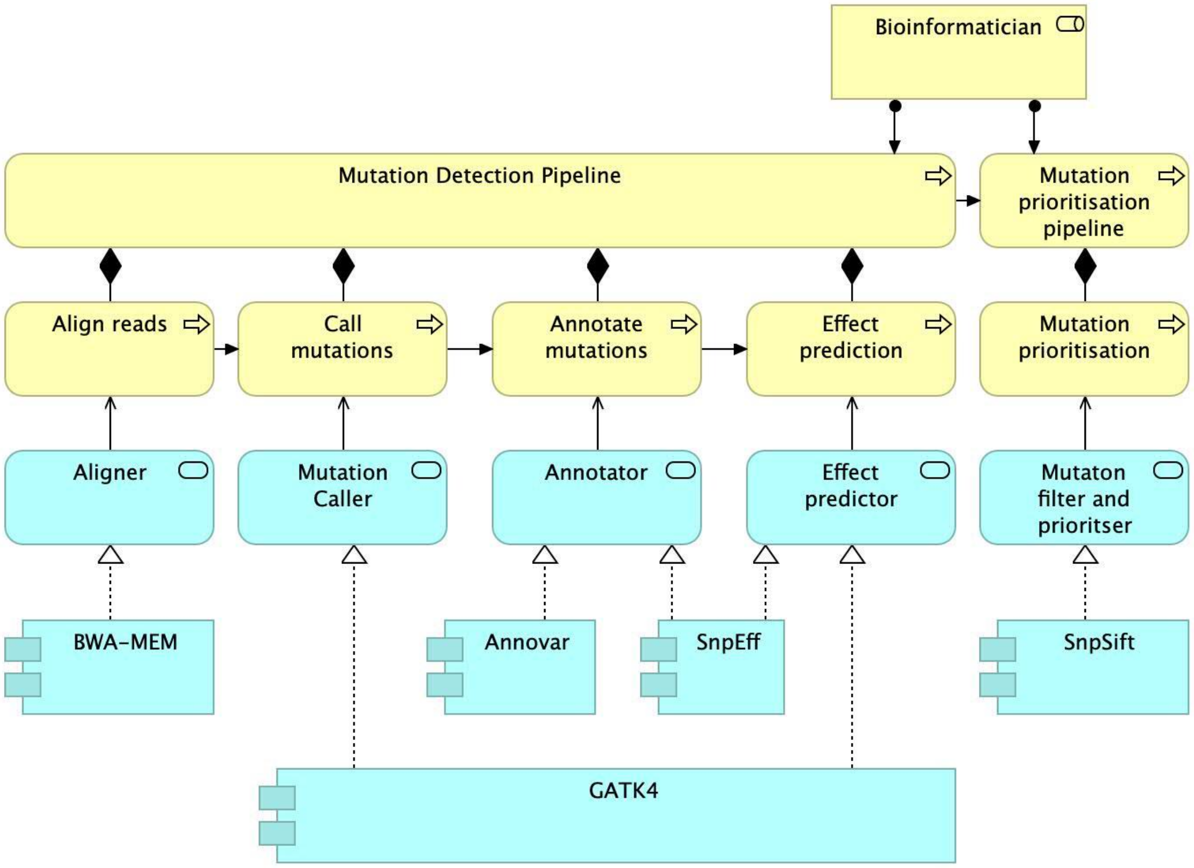


Figure 2: An example of a containerised mutation calling pipeline as defined by WG9. This pipeline was built into a Singularity[28] container and run within an isolated, secure computing environment.

4.2 Interoperability

The use of standards (e.g. data formats, technical/Internet protocols, standard operating procedures) helps to ensure interoperability while allowing for heterogeneity of service implementations within the distributed infrastructure of 1+MG. Additionally these standards also help ensure interoperability with other projects and data spaces, such as the joint action Towards the European Health Data Space[29] (TEHDAS), the Common Infrastructure for National Cohorts in Europe, Canada, and Africa[30] (CINECA) project, the European Joint Programme on Rare Disease[31] (EJP-RD), EOSC4Cancer[32], and the European Health Data Space[33] (EHDS) via the Healthdata@EU[34] Pilot and the Data Spaces Support Centre[35] (DSSC).

[28] https://sylabs.io/
[29] https://tehdas.eu/
[30] https://www.cineca-project.eu/
[31] https://www.ejprarediseases.org/
[32] https://eosc4cancer.eu/
[33] https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en
[34] https://ehds2pilot.eu/
[35] https://dssc.eu/

4.2.1 Interoperability between use-cases

Within 1+MG, an example of interoperability is the use of phenopackets. The deployed infrastructure must support the requirements of the different 1+MG use cases, and therefore the standards for the phenotypic and clinical data cannot be too specialised towards any particular use case at the expense of other use cases. Version 2 of phenopackets increases the interoperability with the Observational Medical Outcomes Partnership Common Data Model[36] (OMOP CDM) compared to version 1, by allowing representation of measurements, medical actions, and time element data to better record longitudinal observations. For rare disease the EJP-RD project has created[37] a semantic model of phenopackets to make phenotypic data more FAIR[38] (Findable, Accessible, Interoperable, Reuseable), and phenopackets have been used within the Solve-RD programme. For the cancer use case collaboration with ICGC-ARGO[39] and mCODE[40] has increased the utility of phenopackets to the cancer community, and work has been ongoing with 1+MG WG9 to map the minimal cancer metadata model to phenopackets to facilitate the sharing (FAIRification) of cancer data within 1+MG. Work is ongoing to make a FHIR[41] (Fast Healthcare Interoperability Resources) implementation guide for phenopackets[42].

4.2.2 Interoperability between data spaces

A “cloud” can be defined as an IT environment that shares and pools IT resources across a network, while the “edge” can refer to devices near the location of the user or the source of the data. Hence edge computing is the process of running workloads on edge devices. However, edge devices can contribute to a cloud if their compute and storage capabilities can be abstracted and shared across a network, but edge computing is distinct in that the workloads run in remote locations outside of the cloud.

4.2.2.1 SIMPL

With the development of a set of common European data spaces, such as European Health Data Space, and GDI, there will be an increased demand for data processing capability, and it is expected that part of this increased demand will come from local processing capacity where there is more control over the data, or the processing is close to the data source. These edge nodes need to be interoperable, both within an infrastructure but also with other data spaces to extract maximum value from the data.

To achieve this, SIMPL[43] smart middleware has been envisaged to allow cloud-to-edge work in Europe. The SIMPL programme is a European level public procurement, using Digital Europe for deployment. SIMPL will be publicly available to be deployed within Europe and outside, is open-source, and intended to be used by both public and private sectors. GAIA-X[44] is a significant

[36] https://www.ohdsi.org/data-standardization/
[37] https://osf.io/ep3xh/download
[38] http://dx.doi.org/10.1038/sdata.2016.18
[39] https://www.icgc-argo.org/
[40] http://dx.doi.org/10.1200/CCI.20.00059
[41] https://www.hl7.org/fhir/
[42] http://phenopackets.org/core-ig/ig/branch/master/index.html
[43] https://digital-strategy.ec.europa.eu/en/news/simpl-cloud-edge-federations-and-data-spaces-made-simple-updated-august-2023
[44] https://gaia-x.eu/

Beyond One Million Genomes
B1MG has received funding from the European Union’s Horizon 2020 Research and Innovation programme under grant agreement No 951724

stakeholder, working towards consensus building of standards for SIMPL. As GDI is funded under Digital Europe programme, it is a priority use case for SIMPL, and can help integrate with SIMPL, with adjustments in SIMPL possible to help integrate the GDI use case.

4.2.2.2 eID

The eIDAS regulation[45] aims to make a cross border European digital identity, or eID, available by 2024. eID is a set of EC provided services and trust framework allowing users to use their national eIDs to access services in other European countries. The Life Science Login (https://lifescience-ri.eu/ls-login/) service will support eID as one form of authentication alongside home organisation identity services. The LS Login service special feature is identity linking. The LS AAI allows linked identities and single sign-on to allow a user to access resources across multiple locations. eID will be interoperable throughout Europe, but the user identity and access metadata supported by eID will need to be augmented with federated authorisation metadata needed in GDI and EHDS data access authorisation procedures. As GDI deploys and becomes interoperable with other data spaces and projects, for example via SIMPL, such as EHDS, GAIA-X etc it is important the GDI investigates and communicates with the eID. In the future there may be full coverage across GDI participating states and the eID network metadata description, which would allow eID to be utilised within the GDI. This is, however, not yet the case according to our knowledge. The ideal situation for GDI would be that a single identity could be used across all data spaces, including GDI, allowing secure cross border access not only to genomic data, and closely related data such as phenotypic, clinical, and lifestyle data, supporting seamless data integration and adding value to these heterogeneous data.

As the eID would be available to any EU citizen, resident , or business, it would facilitate citizen engagement within GDI, from monitoring relevant research outputs which would help increase engagement, towards using eID for dynamic consent. It is proposed that the identity will be widely usable, and also enable users to control their own data, but allowing users to only share that data which they wish to share, supporting the transparency and data minimisation principles of GDPR. This would be done via the eWallet[46], which would also encode and host access permissions of the GA4GH Passport, containing a list of attributes associated with, and resources available to, the user. However, the eWallet would be available with an eID, and the eWallet would able to be used to confirm a wide range of of personal attributes for access to public and private digital services and resources, and as such would be the basis of the interoperability of identities between different EU services and data spaces. With the EC Digital Compass[47] aiming to allow citizens to have access to electronic medical records by 2030, and an uptake of eID by 80% of citizens, ensuring GDI supports eID would facilitate citizen access to their data. This, however, requires that GDI and EHDS data access governance metadata will become supported in eID and eWallet.

[45] https://digital-strategy.ec.europa.eu/en/policies/eidas-regulation
[46] https://digital-strategy.ec.europa.eu/en/policies/eudi-wallet-implementation
[47] https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/europes-digital-decade-digital-targets-2030_en

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

4.3 Alignment with GDI

4.3.1 GDI

The GDI project has been set up to deploy the data infrastructure service proposed by B1MG. This has been demonstrated by the GDI Starter Kit in June 2023, which is an evolved and disease agnostic version of the original B1MG rare disease and Cancer PoC, which was deployed by a set of vanguard nodes within GDI to achieve the GDI milestone MS6[48]. The Starter Kit was split into separate products, each of which helps provide the functionalities required for secure cross-border data access (Table 1). Where possible, implementations of products in the Starter Kit were production-ready versions of software - for example REMS is used in production services by both FINDATA[49] and THL Biobank[50] and administers access rights to data controlled by them. The Starter Kit has been demonstrated performing different types of data analysis, using both htsget to securely transmit data from the data store to an SPE, and TES to execute a task on a remote SPE - both using OCI compatible containers.

GDI is developing larger synthetic datasets, which will enable additional testing of the deployed infrastructure. An example of the synthetic dataset being generated are the data currently being produced by THL which has over 1 million mosaic genomes in VCF format, with 15 phenotypes and pedigree information for up to four generations. A subset of these data are available from CSC and can be requested via the B1MG REMS instance[51] hosted at CSC.

The design and implementation of the 1+MG User Portal referred to in Figure 1 has been assigned to WP4 (European Operations) of the GDI, more specifically as Task 4.1. While not strictly part of the GDI Starter Kit as it will be a single deployed instance, it has two products associated with it - User Portal - Data Catalogue and User Portal - Access Management.

Product	Outline	Production	Function
Sensitive Data Archive	Securely stores data	✔	🗄
LifeScience AAI	Provides a federated Identity	✔	🛢
REMS	Allows data access applications and decisions	✔	🛢
Beacon	Genetic and phenotypic data discovery	✔	🔍
Beacon Network	Federated network of Beacons	✔	🔍

[48] https://zenodo.org/doi/10.5281/zenodo.8074285
[49] https://findata.fi/
[50] https://thl.fi/en/web/thl-biobank
[51] https://1mg-rems.sd.csc.fi/catalogue

Synthetic Data	Artificial anonymous data		🟥
htsget	Secure genetic data distribution		🟥
Containerised Computation	Computation via virtualised portable software packages	✔	🟩
Federated Computation	Interoperable distributed workflows		🟩

While this document is a product of B1MG WP4, which defines the technical infrastructure, the roadmap also has to take account of how the infrastructure will be operated during and after deployment. However, GDI has also a roadmap[52] on how the infrastructure will operate at a European level, based on the FitSM[53], which is a family of standards for IT service management. This includes the planning, deployment, and operation of the service, applicable to federated scenarios. As the GDI roadmap has already been in the project plan it is not detailed here, but the timings and dependencies of the roadmap have been included in the considerations of timings for the proposed roadmap described here.

GDI interacts as a use case for EHDS in HealthData@EU pilot[54] project (EHDS2) via ELIXIR[55] coordination. This helps to ensure that the requirements of the GDI are taken into account as the EHDS takes shape.

4.3.2 Genome of Europe

The Genome of Europe aims to sequence at least 500,000 individuals across Europe to create a genomic reference cohort representative of the different populations within Europe. 1+MG WG12 has made an inventory of expected genomes from each participating country and population subgroups totalling 560,990 WGS at 30x coverage[56], with minimal associated metadata. Such a cohort can be used to analyse genetic diversity across European populations, help interpret clinically relevant genetic variants, adjust genetic profiles to specific populations, and as a reference panel for imputation. It is estimated[57] that there will be potentially between 150 and 300 petabytes of raw WGS data generated by GoE. The GoE is a specific use case for the GDI, and as such it would be good if the data generated would be stored within the GDI. As such, preparations must be made to ensure that the GDI has the capability to receive these data, check the accuracy of the data, and store these data in a timely manner. Therefore confirmation

[52] https://doi.org/10.5281/zenodo.8017873
[53] https://www.fitsm.eu/
[54] https://ehds2pilot.eu/
[55] https://elixir-europe.org/
[56] https://doi.org/10.5281/zenodo.8055610
[57] https://doi.org/10.5281/zenodo.8017856

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

of the volume of expected data at each node or country, and hence the required physical infrastructure must be determined.

4.3.3 Testing and Deployment

There are eight main types of software testing: unit, integration, functional, system, acceptance, performance, end-to-end, and smoke testing. Unit testing tests the individual components of the software to ensure that each component performs its specific function. Integration testing combines the software units and tests them as a group to ensure the interaction of the different units of software performs as expected. Functional testing is related to integration testing, but ensures that the function of the interconnected components meets the requirements, while integration testing may just ensure components communicate without errors. System testing checks if the complete and fully integrated system performs as expected, typically including functional, performance, and security testing, checking that the system as a whole meets the specifications and KPIs. Acceptance testing is performed to ensure the software meets the customer's requirements, and can be split into alpha and beta testing, with beta testing performed by a group of customers or stakeholders. Performance testing tests how the system performs under a specific set of workloads to help ensure a reliable and robust system. Smoke testing are quick and simple checks to ensure the system is running as expected, often after deployment.

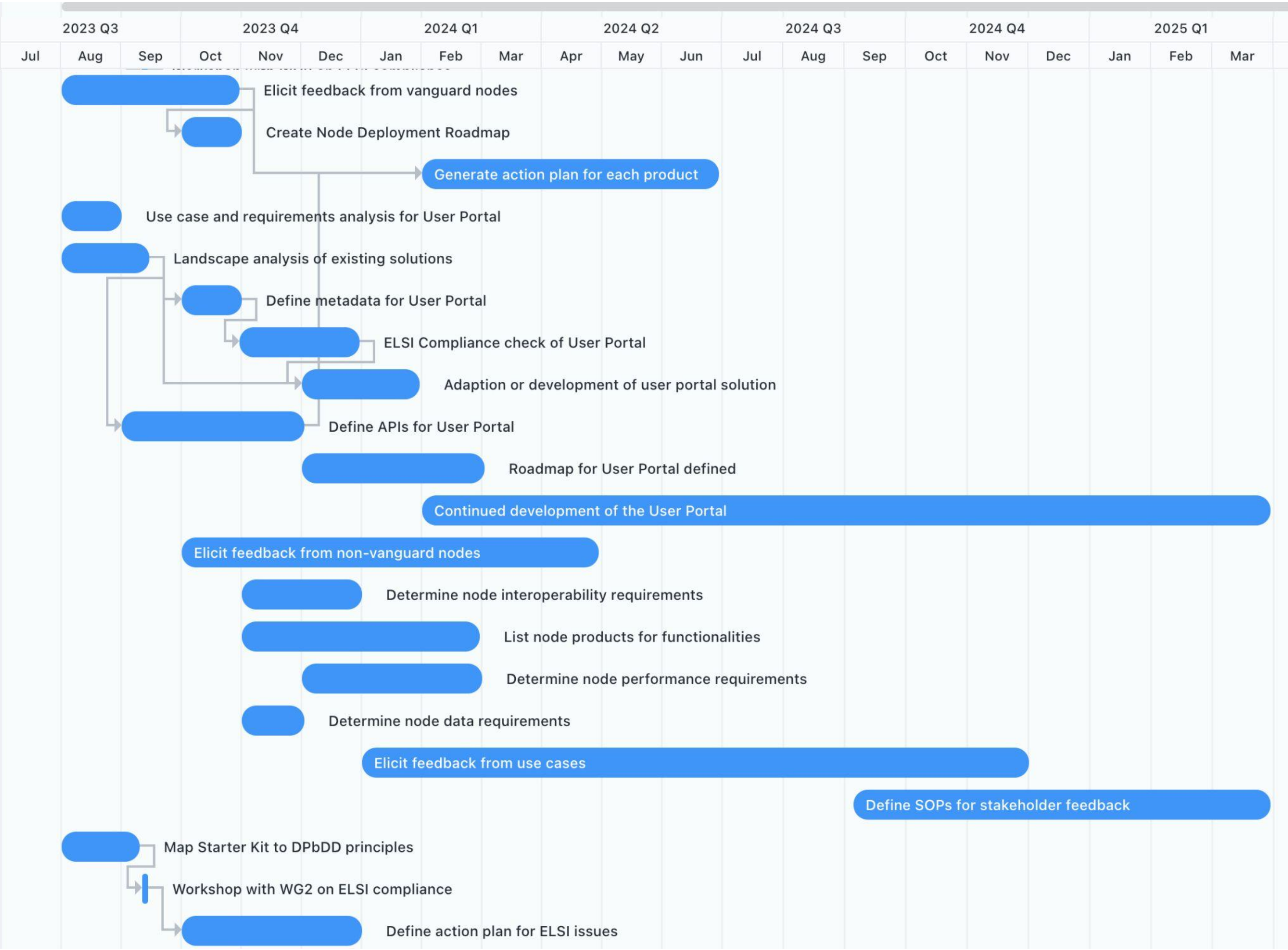
These types of testing should be run at node and network level - a node can be defined as a complete system, or as a unit within the federated network. Therefore when deploying a federated infrastructure these tests must be carried out at node level initially, and subsequently at network level. Additionally each test must be planned, from identifying the requirements of the test, planning and designing the test, to identifying the environment for and executing the test, to measuring the output of the test and determining the pass criteria. These tests have been entered into the roadmap.

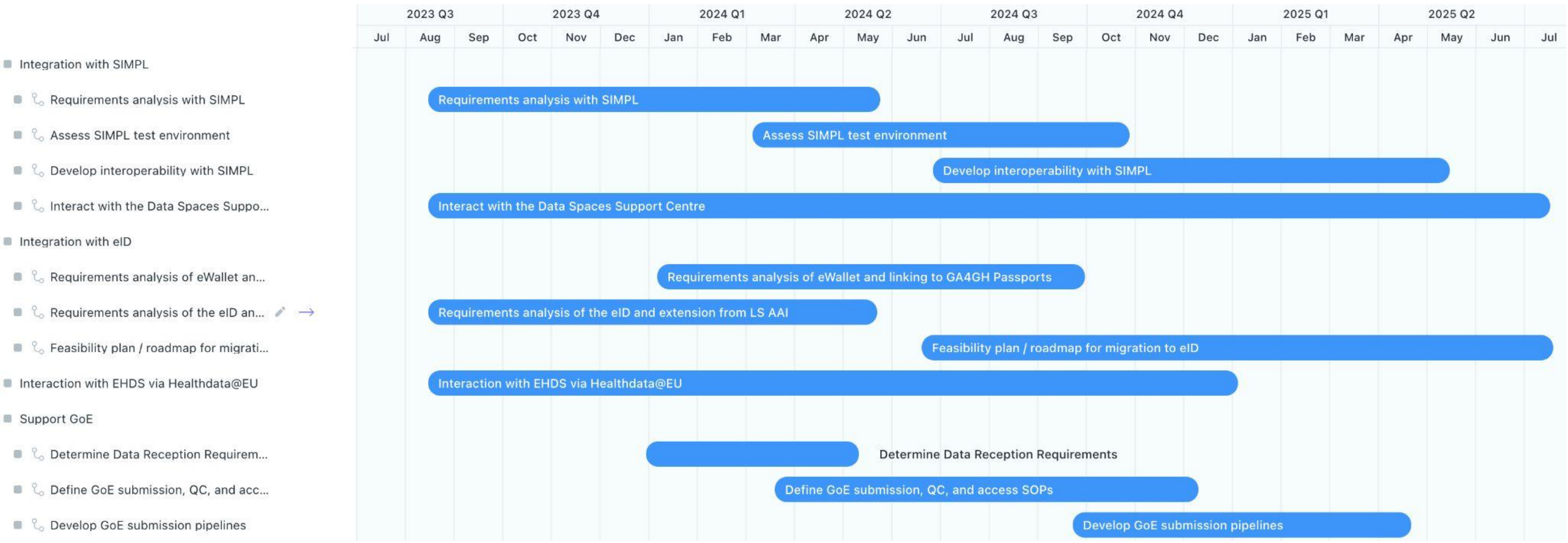
5. Results

The roadmap detailed in the following Gantt chart has been based on the work of section 4, and will leverage the funding within the GDI project to start the deployment, and production level development, of the 1+MG working group recommendations and PoCs described in B1MG. The timings are defined primarily by the GDI requirements, especially the goal that the User Portal and nodes are expected to be operational by Month 30 of the GDI project (Q1 2025).

At a high level, the Gantt details how the PoCs will evolve from a demonstration infrastructure to a production infrastructure ready for real genomic data. This includes the ELSI analysis of the infrastructure, the services, and procedures used to run and maintain the infrastructure, and the support services which enable user and stakeholder interaction, such as the helpdesk. The helpdesk, including SOPs and European level interaction between the nodes, have already been described with the GDI deliverable D4.1, so are referenced from the B1MG roadmap. However the actual methodology of deployment, and the associated timelines based on the GDI timelines, are included in this roadmap.

Also included are the high-level interactions and requirement analysis from other external stakeholders, such as EHDS2, GA4GH, eID and SIMPL; all of which are necessary to try and ensure not only European interoperability, but also where possible, global interoperability and to help present the European requirements, and solutions to a global audience.

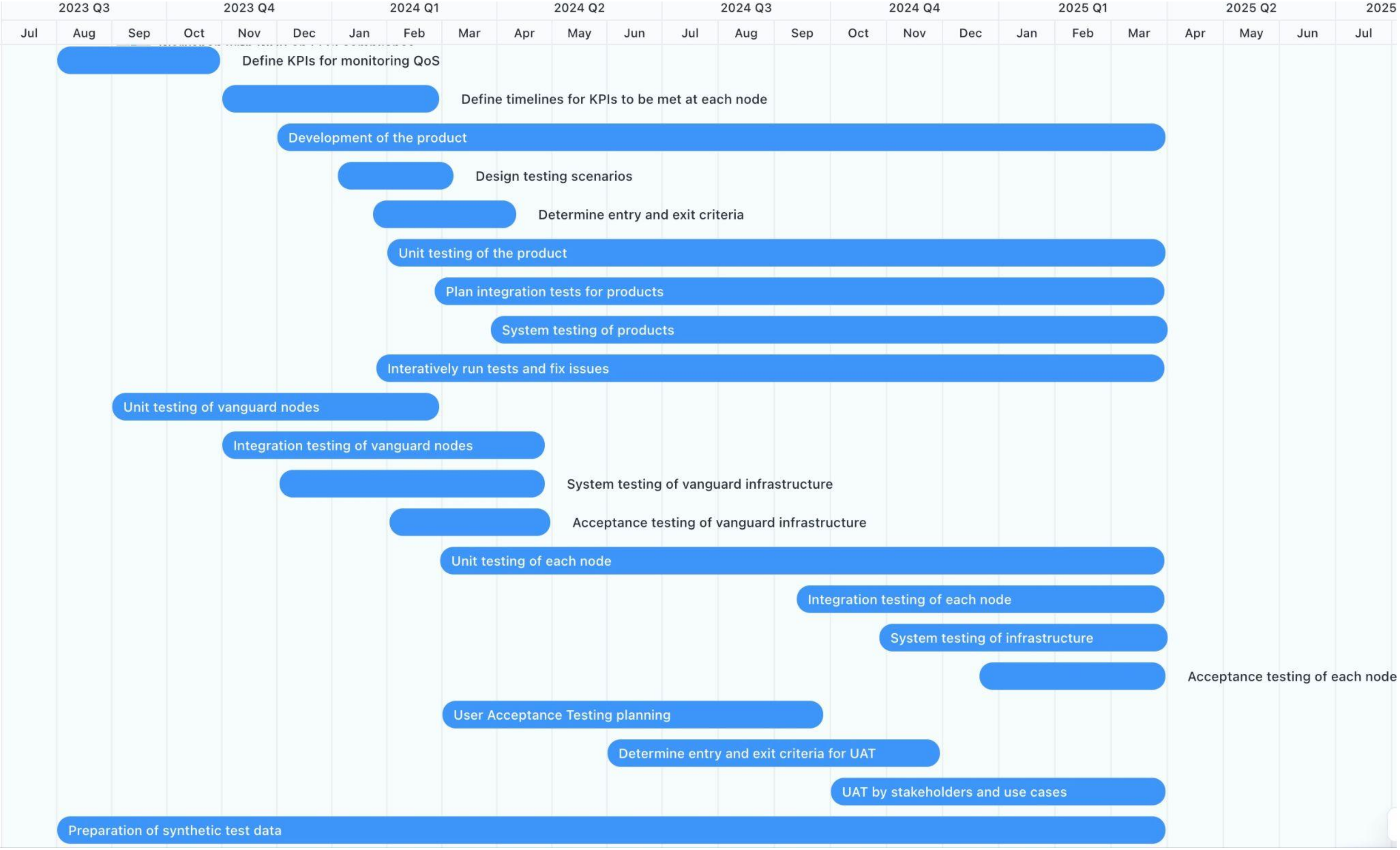

2023 Q3
2023 Q4
2024 Q1
2024 Q2
2024 Q3
2024 Q4
2025 Q1
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Elicit feedback from vanguard nodes
Create Node Deployment Roadmap
Generate action plan for each product
Use case and requirements analysis for User Portal
Landscape analysis of existing solutions
Define metadata for User Portal
ELSI Compliance check of User Portal
Adaption or development of user portal solution
Define APIs for User Portal
Roadmap for User Portal defined
Continued development of the User Portal
Elicit feedback from non-vanguard nodes
Determine node interoperability requirements
List node products for functionalities
Determine node performance requirements
Determine node data requirements
Elicit feedback from use cases
Define SOPs for stakeholder feedback
Map Starter Kit to DPbDD principles
Workshop with WG2 on ELSI compliance
Define action plan for ELSI issues



- Integration with SIMPL
  - Requirements analysis with SIMPL
  - Assess SIMPL test environment
  - Develop interoperability with SIMPL
  - Interact with the Data Spaces Suppo...
- Integration with eID
  - Requirements analysis of eWallet an...
  - Requirements analysis of the eID an...
  - Feasibility plan / roadmap for migrati...
- Interaction with EHDS via Healthdata@EU
- Support GoE
  - Determine Data Reception Requirem...
  - Define GoE submission, QC, and acc...
  - Develop GoE submission pipelines

2023 Q3
2023 Q4
2024 Q1
2024 Q2
2024 Q3
2024 Q4
2025 Q1
2025 Q2
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Requirements analysis with SIMPL
Assess SIMPL test environment
Develop interoperability with SIMPL
Interact with the Data Spaces Support Centre
Requirements analysis of eWallet and linking to GA4GH Passports
Requirements analysis of the eID and extension from LS AAI
Feasibility plan / roadmap for migration to eID
Interaction with EHDS via Healthdata@EU
Determine Data Reception Requirements
Define GoE submission, QC, and access SOPs
Develop GoE submission pipelines

Beyond One Million Genomes
B1MG has received funding from the European Union’s Horizon 2020 Research and Innovation programme under grant agreement No 951724



Beyond One Million Genomes
B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

6. Discussion

This document outlines the updated roadmap for the deployment of a genomic data infrastructure enabling secure cross border access to genomic and phenotypic data according to FAIR principles. It is a significant update on the original roadmap, which has been informed by the feedback from different stakeholders and the evolving landscape of regulation and data spaces within the European Union. As such it tries to ensure that the proposed infrastructure can adapt to future changes by ensuring the different functionalities provided by the infrastructure are provided by applications or services that are linked by standards, so each application or service can evolve and adapt as required.

7. Conclusions

The roadmap presented takes account of the GDI timelines and defines the generic steps each node must take to deploy a GDI node, as well as outlines to core actors that GDI must collaborate and become interoperable with.

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724