Beyond One Million Genomes

D7.1

Catalogue of met and unmet use case WGs requirements 1v0.

Project Title (grant agreement No)	Beyond One Million Genomes (B1MG) Grant Agreement 951724		
Project Acronym	B1MG		
WP No & Title	WP7 - Support for 1+MG Use Case Working Groups		
WP Leaders	Serena Scollen (ELIXIR Hub), Marco Tartaglia (OPBG), Giovanni Tonon (HSR/ACC), Andres Metspalu (UT), Katja Kivinen (FIMM)		
Deliverable Lead Beneficiary	01 - EMBL/ELIXIR		
Deliverable	D7.1 - Catalogue of met and unmet use case WGs requirements 1v0.		
Contractual delivery date	31/1/2022	**Actual delivery date**	13/12/2022
Delayed	Yes		
Authors	Giselle Kerry (ELIXIR Hub), Serena Scollen (ELIXIR Hub), Juan Arenas (ELIXIR Hub)		
Contributors	N/A		
Acknowledgements (not grant participants)	1+MG Use Cases WG members had extensively contributed to this deliverable through the discussion in the WG meetings.		
Deliverable type	Report		
Dissemination level	Public		

Document History

Date	Mvm	Who	Description

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

02/11/2022	0v1	Giselle Kerry (ELIXIR Hub)	Initial draft
30/11/2022	0v2	Nikki Coutts (ELIXIR Hub)	WP comments addressed. Version circulated to B1MG-OG and B1MG-GB for feedback
13/12/2022	1v0	Giselle Kerry & Nikki Coutts (ELIXIR Hub)	B1MG-MB and B1MG-GB comments addressed. Version uploaded to the EC Portal

Table of Contents

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

1. Executive Summary

The 1+MG Use Case WGs leads and experts have contributed to the direction of the B1MG WPs activities since the beginning of the project. Their contributions are fundamental to determining the resulting infrastructure needs as well as analysing the final solutions and how they support the various scenarios.

This deliverable builds on previous informal and formal work, as well as contributions from other project and initiative activities:

- 1+MG use cases working group meetings
- Workshop to identify health care scenarios across 1+MG WGs
- Workshop to identify research scenarios across 1+MG WGs
- B1MG Operational group meetings (1+MG WGs & B1MG WP)
- Stakeholders forum outcomes
- 1+MG Group meetings

This deliverable gives the first formal snapshot of the requirements of the 1+MG use case WGs and their current implementation state, which has been collected as part of the B1MG project directly. This inventory of met and unmet needs of the use case WGs will subsequently be used to drive infrastructure development and prioritise future 1+MG infrastructure actions. Successful implementation of the requirements will be evaluated by 1+MG WGs, ensuring that the final 1+MG infrastructure is fit for purpose.

WP7 must still debate the implementation of the recently obtained EC recommendations to collect user requirements using particular software tracking techniques. The outcomes of those discussions will be incorporated into the future version of this deliverable.

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

2. Contribution towards project objectives

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

	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.	Yes
	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.	Yes
	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.	Yes
	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).	Yes
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	Yes
	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.	Yes
	7. Best practices in sharing and linking phenotypic and genetic data (WP3) by M12 & M24.	Yes
	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.	Yes

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 first overview of the 1+MG Use Cases WGs have been produced after:

1) Analysis of regular and ad hoc meetings and workshops including:
- 1+MG use cases working group meetings
- Workshop to identify health care scenarios across 1+MG WGs
- Workshop to identify research scenarios across 1+MG WGs
- B1MG Operational group meetings (1+MG WGs & B1MG WP)
- Stakeholder forum annual meeting outcomes
- 1+MG Group meetings In the case of WG9 the output of the monthly meetings with WP4/1+MG WG5 for the definition of PoC cancer has also been incorporated

2) Individual meetings with WG leaders
   a) 1+MG WG8 Rare Diseases
   b) 1+MG WG9 Cancer
   c) 1+MG WG10 Common and complex diseases
   d) 1+MG WG11 Infectious diseases

This report reflects the initial status of the WGs and establishes the baseline that will be re-evaluated and expanded in the remaining deliverables (D7.2 & D7.3). At this time, storage of requests for considerations from the individual WGs are maintained within a google spreadsheet [1] but are also displayed below for ease of reference:

Table 1 - An overview of the WGs identified needs and WP assignment for consideration

Working Group	**Description of Need**	**Technical Group**	**Current Status**
WG 8 (Rare Disease)	In RDs research and healthcare are often done in conjunction. Consider a single access path for RD diagnosis and novel pathogenic gene identification.	WP4	

[1] https://docs.google.com/spreadsheets/d/17s61hIL2tJOD8DqlCUCAn6AjjPzQ6vQY3vocsdvc6jc/edit#gid=0

WG 8 (Rare Disease)	The need for real time feedback should be considered when considering rare diseases	WP4	
WG 8 (Rare Disease)	Aside from the ability to search for probable matches in other databases via a link to MatchMaker Exchange, the architecture should provide notification when a MatchMaker Exchange query (e.g., submission to GeneMatcher) matches data currently stored in the B1MG infrastructure	WP4	
WG 8 (Rare Disease)	RD data is usually longitudinal - Evaluate how to address this from the ELSI and IT.	WP2 & WP4	
WG 8 (Rare Disease)	The current PoC should be extended to facilitate federated analysis	WP4	
WG 8 (Rare Disease)	The current PoC should be extended to facilitate federated learning	WP4	
WG 8 (Rare Disease)	Need for bi-directional data discovery mechanisms connection to other resources or IT networks	WP4	
WG 8 (Rare Disease)	Healthcare = research. As a result, rapid approval via a single method (at least for basic information – as in the case of a query directed to check a certain gene and later contact the submitters of records seemingly matching the query) should be the rule.	WP4	
WG9 (Cancer)	Beacon v1 allowed only basic queries on genomic data (single nucleotide variants), whereas in cancer it is desirable to be able to query for more complex features, like aberrant karyotypes, structural variants and copy number variations.	WP4	DONE - With V2 release of Beacon
WG9 (Cancer)	The proposed querying tool for rare diseases, Matchmaker Exchange, is not suited for cancer data. We encourage the usage of a standard electronic data capture tool together with a minimal data set to harmonise clinical data at source. Currently, we are evaluating the Cohort Genomic Platform (CGP) tool, developed by hospital San Raffaele and comparing different available minimal data set. Data can then be exported from CGP in JSON format, and be queried with the existing tools in the rare disease PoC.	WP4	
WG9 (Cancer)	The analyses that can be run on the queried data with GPAP are very specific for rare diseases. We suggest the implementation of a cancer-specific tool, like cbioportal, that can be run on a compartmentalised virtual machine at the storage node and provide the researcher with simple summary statistics.	WP4	
WG9 (Cancer)	Cancer has a longitudinal aspect i.e., multiple data points in time - therefore the need to be able to	WP2 & WP4	

	record proper measurements of disease progression and relapse definition is imperative.		
WG10 (Common & Complex Diseases)	WG10 does differ from the cancer and rare disease use cases since in WG10, the focus is on utilising polygenic risk scores consisting of many genetic loci rather than working with single SNPs.	WP4	
WG10 (Common & Complex Diseases)	Distributed analysis to different data hubs to generate PRS must be supported e.g., the capability to run common analysis tools on common input sets such as disease loci	WP4	
WG 11 (Infectious Diseases)	Host and pathogen data should be linked	WP4	

These requests for consideration will be shortly moved to an agreed tracking tool (after consultation with WP4), which can take into account the process from identification of the requirements through to implementation/resolution within the technical platform.

4. Description of work accomplished

4.1.1 [WG8 - Rare Disease]

At the end of 2021, WG8 successfully completed the Proof of Concept[2] (P.o.C.) for Rare Disease utilising a number of existing GA4GH Standards (Beacon, DUO,SAM,BAM, CRAM,VCF, Crypt4gh, htsget, phenopackets,Passports, TES & WES) and other existing services (Matchmaker Exchange, FEGA & ELIXIR/Life Science AAI).

[2] https://www.youtube.com/watch?v=6MtIJA4xXdU

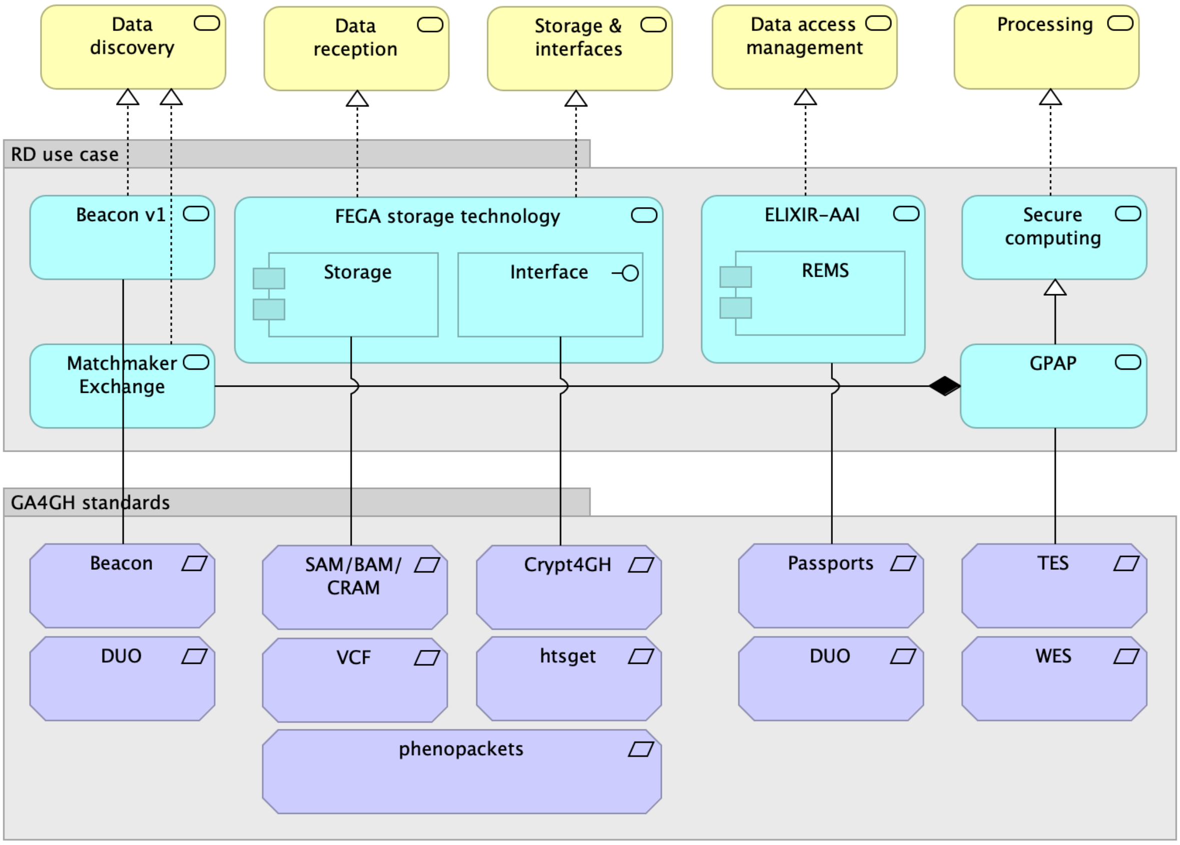

Figure 1 - Functionalities and Standards used in the Rare Disease P.o.C.

Whilst the P.o.C for Rare Disease was successful, there are still unmet requirements for the Rare Disease working group that will need to be addressed before the infrastructure can be scaled. At the time of writing the following areas have been highlighted as requirements that will need to be further considered;

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

- Two pathways were identified - healthcare (e.g., diagnosis) and research (e.g., identification of a new gene). These pathways are tightly imbricated when dealing with rare diseases i.e., Typically, the inquiry begins with a healthcare issue because it is directly tied to a specific patient who is awaiting diagnosis or treatment. In individuals suffering from a "new disease" (i.e., one that has yet to be characterised/identified/recognized), the purpose of diagnosis aligns with the "confirmation" of the discovery of a new disease gene or a previously unknown genotype-phenotype link. To put it another way, healthcare = research. As a result, rapid approval via a single method (at least for basic information – as in the case of a query directed to check a certain gene and later contact the submitters of records seemingly matching the query) should be the rule.
- The need for real time feedback should be considered when considering rare diseases (and other use cases).
- A registered user be allowed to have access to multiple queries - or have to reapply on a case by case basis
- Aside from the ability to search for probable matches in other databases via a link to MatchMaker Exchange, does the architecture provide for notification when a MatchMaker Exchange query (e.g., submission to GeneMatcher) matches data currently stored in the B1MG infrastructure?
- How are the ties between individual sequencing records and their respective submitters maintained over time (e.g., months/years)? This is also important in order to optimise the amount of data kept in the database.
- Large collections of WES - WES data standards need to be defined and consideration given on how to include them in 1+MG IT
- Need for bi-directional data discovery mechanisms connection to other resources or IT networks
- RD data is usually longitudinal - needs to be addressed from both ELSI and IT perspective
- Consider collection of other data types e.g., data from EHRs, MRIs, availability of biobank samples and how they should be included in the IT and recommend standards for these data types
- The current PoC should be extended to facilitate federated analysis
- The current PoC should be extended to facilitate federated learning

Whilst the P.o.C was successful, it is acknowledged that this is just the starting point for both the Rare Disease working group and also by which the other WG's can begin to benchmark their own use cases and make apparent their requirements to the relevant WP's accordingly.

4.1.2 [WG9 - Cancer]

WG9 analysed the Proof of Concept (P.o.C) designed for rare diseases by WG5/WP4 and evaluated how it fits with the specific needs for the Cancer use case . The overall modular structure of the existent P.o.C made it possible to break down the evaluation process. They ascertained that the storage solution and the management of authentication and authorization of users developed for rare diseases can be applied to the cancer use case as they are. However, the following modules need to be changed to create a functional PoC for Cancer:

- Beacon v1 allowed only basic queries on genomic data (single nucleotide variants), whereas in cancer it is desirable to be able to query for more complex features, like aberrant karyotypes, structural variants and copy number variations. The next implementation of the Beacon scheme (v2), released in Spring 2022, may solve this problem.
- Clinical/phenotypic data in cancer can be much more complex than in rare diseases, including longitudinal data (e.g. resequencing after acquired resistance), comorbidities, different therapies, etc. Therefore, the proposed querying tool for rare diseases, Matchmaker Exchange, is not suited for cancer data. We encourage the usage of a standard electronic data capture tool together with a minimal data set to harmonise clinical data at source. Currently, they are evaluating the Cohort Genomic Platform (CGP) tool, developed by hospital San Raffaele and comparing different available minimal data set. Data can then be exported from CGP in JSON format, and be queried with the existing tools in the rare disease PoC.
- The analyses that can be run on the queried data with GPAP are very specific for rare diseases. We suggest the implementation of a cancer-specific tool, like cbioportal, that can be run on a compartmentalised virtual machine at the storage node and provide the researcher with simple summary statistics.
- Cancer has a longitudinal aspect i.e., multiple data points in time - therefore the need to be able to record proper measurements of disease progression and relapse definition is imperative.

4.1.3 [WG10 - Common & Complex Disease]

At this time WG10 have not had the opportunity to run any simulations or confer with their stakeholders. However the following have been documented as already known unmet requirements

- WG10 does differ from the cancer and rare disease use cases since in WG10, the focus is on utilising polygenic risk scores consisting of many genetic loci rather than working with single SNPs.
- Distributed analysis to different data hubs to generate PRS must be supported e.g., the capability to run common analysis tools on common input sets such as disease loci

4.1.4 [WG11 - Infectious Disease]

WG11 have also not had the time to run any simulations or fully confer with their stakeholders. However, it is already noted that a key requirement for WG11 is that host and pathogen data should be linked. A meeting has been arranged with WG5 to discuss the Proof of Concept in June.

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

7. Conclusions

The Rare Disease P.o.C for Federated Data Access has been effectively realised through the use of existing standards and services, creating a substantial building block from which the remaining WG's can align and expand to support their particular use case.

By using this iterative approach of predominantly working with one WG at a time, services and standards may be modified and adapted to ensure interoperability is maintained as the overarching infrastructure expands to handle all essential use cases.

8. Next steps

- Evaluation of EC reviewers recommendations
- Discussion with technicals WGs and B1MG WPs to plan the implementation of the unmet use cases requirements

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