

# Contributions to secure Genomic Human Data sharing for precision medicine

Jordi Rambla i de Argila

---

TESI DOCTORAL UPF / 2022

Thesis supervisor

Dr. Arcadi Navarro i Cuartiellas

DEPARTMENT OF MEDICINE AND LIFE SCIENCES





*To my fathers and Núria, who always encouraged me to be in Science, indeed when I was hesitating.*

*To my daughters and Sònia, who are giving me a real life.*

*To Arcadi that envisioned my peculiarity as an asset.*

*And to my friend-colleagues that allowed me to generate this thesis and all other achievements we have reached together.*



## **In appreciation**

As mentioned in the main text, this thesis is the mentored work of a PhD student, but it strongly relies in communities. Communities close to me, like my current and previous teammates at Centre de Regulació Genòmica (CRG) and at the European Bioinformatics Institute (EBI), or the Barcelona Supercomputing Center (BSC), in particular the Life Sciences team, that coordinate ELIXIR Spain, and last but not least, the community at the Universitat Pompeu Fabra (UPF). International communities like the ELIXIR Europe Hub members, or the Global Alliance for Genomics and Health (GA4GH).

For each of these communities I can print a long list of names that contributed to make this work a reality and also a promise: all my team at the EGA at CRG, Alfonso, Angel, Gary, Michael, Peter, Roderic, Salvador, Serena, Thomas, Tony ... but it will take dozens of pages. As I'm writing this appreciation, more and more names come to my mind, therefore I'm solid that I would be unfair to many of them.

I want to specially mention my thesis supervisor, Arcadi, who gave me a lot of freedom, while opening doors that I can choose or not to cross. He sustained me in my mistakes, encouraged me when I was hesitating, and praised my successes – which are also his--, as a good and knowledgeable mentor knows how to do.



## **Abstract**

The advance of precision medicine depends on the availability of data that is continuously being generated in universities, research centres and hospitals world-wide. That data represents 1) the diversity of the human genotypes and phenotypes and 2) the diversity of ways diseases unfold in human beings. To make more data available it is important to improve the automation of the management of data, the discovery of data that could be reused, and the process of actually sharing the data. This thesis contributes to all these aspects focusing on both, data controllers and data consumers. To facilitate data management, EGA and Federated EGA are introduced as an example of standard integration. To facilitate data discovery, the Beacon version 2 protocol and an implementation of it are provided. To facilitate data sharing, harmonisation of the data use conditions and a categorization, simplification, of the levels of access is suggested.

## **Abstract in Catalan**

El progrés de la medicina de precisió depèn de la disponibilitat de dades que es produeixen a les universitats, centres de recerca i hospitals d'arreu del món. Aquestes dades representen 1) la diversitat de genotips i fenotips humans, i 2) la diversitat de formes en que les malalties es presenten. Per tal de que hi hagi més dades disponibles és important millorar-ne la gestió, localitzar (descobrir) dades que siguin d'interés, i simplificar el procés de compartir-les. Aquesta tèsi contribueix a aquests objectius i es centra tant en els generadors com en els consumidors de dades, Per facilitar la gestió de dades, s'introdueix EGA i la EGA Federada com exemples d'integració d'estàndars. Per facilitar la localització de dades, s'ofereix el protocol Beacon 2 i un programari que l'implementa. Per facilitar el compartir les dades es suggereix una harmonització de les condicions d'ús de les dades i les categories d'accés





## Preface or introduction

After the release of the Human Genome and the pioneer projects that aimed to assess the levels and patterns of human diversity (*e.g.*, the HapMap or the 1000 Genomes projects), it has become apparent that both, huge nucleotide variability in humans, with abundant rare variants, and the complexity of the interactions between heredity and environment will require of large sample sizes, which are difficult to obtain within any single study. As researchers tried to generate datasets by combining data produced by several studies, they were hampered, or indeed discouraged, by the difficulties in analysing together disparate data, in particular in what refers to the time and resources necessary to access and manage the data received; and to harmonising data for combined analysis.

Given the growing need, diverse initiatives were born to make all these processes smoother and more transparent both to the data producers and to the data consumers. Foremost among them, the Global Alliance for Genomics and Health or ELIXIR Europe was created in January 2013.

This thesis describes the contribution of the author, together with collaborators and colleagues all over the World, in advancing in all these fronts by providing ideas for simplification, protocols and tools that implement these ideas and real examples of its application.



## Table of contents

Abstract .....	vii
Abstract in Catalan.....	vii
Preface or introduction .....	ix
List of figures.....	xii
1. Introduction .....	13
1.1 The importance of data sharing for precision medicine.....	13
1.2 Harmonisation is essential.....	16
1.3 Data management, discovery, and sharing .....	17
1.4 The role of standards for sharing inside or outside of a federation.....	20
1.5 ELIXIR, the Federated EGA and the Beacon Network.....	23
2. Objectives .....	25
3. Results .....	27
Chapter 1.....	29
Chapter 2.....	39
Chapter 3.....	51
Chapter 4.....	55
4. Discussion and Conclusions .....	63
4.1 Data Management .....	63
4.2 Data Discovery .....	67
4.3 Data Sharing.....	80
4.5 Summary .....	83
5. Bibliography .....	85
6. Annex 1 .....	91
6.1 Details on the GA4GH Beacon version 2 specification .....	91
7. Annex 2.....	97
Annex Paper 1 .....	99
Annex Paper 2.....	107
Annex Paper 3.....	121
Annex Paper 4.....	133
Annex Paper 5.....	147

## List of figures

Figure 1 - Federated access to European Genomes .....	66
Figure 2 - The GA4GH Beacon v2 Model .....	70

# 1. INTRODUCTION

## 1.1 The importance of data sharing for precision medicine

Precision medicine (PM) is the application of medicine and scientific knowledge to tailor the diagnosis and treatment of an individual to its particular circumstances: genetic, phenotypic, lifestyle, environment, etc. While there is no universally accepted definition, the EU Health Ministers in their [Council conclusions on personalised medicine for patients](#), published in December 2015, defined personalised medicine as:

*“A medical model using characterization of individuals’ phenotypes and genotypes (e.g., molecular profiling, medical imaging, lifestyle data) for tailoring the right therapeutic strategy for the right person at the right time, and/ or to determine the predisposition to disease and/ or to deliver timely and targeted prevention.”*

If Precision Medicine must provide different preventive, prognostic, or therapeutic strategies depending on the characterization of a given individual, PM practitioner’s would need a corpus of knowledge for each of these dimensions, and also for their combinations, so they can assign each individual to a specific subpopulation. Not only that, they would also need information on which strategies have proven successful or unsuccessful for that subpopulation, so they can offer the successful ones to the target patient. As an example: “Which would be the right preventive strategy for a person living in Barcelona downtown, female, aged 40, with an Iberian genetic background, following a vegetarian diet, moderate exercise lifestyle, with a familiar history of breast cancer and with a given set of mutations in BRCA1 and BRCA2 genes?” A different subpopulation would be women older than 50, or following a different diet, having a different genetic background, or all of that together.

It is easy to understand that this explosion of combinations is impossible to evaluate in any single study, clinical trial or indeed longitudinal cohort. Therefore, researchers interested in a given subpopulation or combination of subpopulations would have two different options: 1) to recruit a new cohort of people that matches the criteria for being enlisted as a case or a control, or 2) to explore if similar data is already available from previous studies or in actual healthcare and focus on reanalysing it.

This problem is not circumscribed to small set of diseases, many biomedical domains related to disease research are facing these explosion of combinations:

Oncology confronts the huge variability in cancer genomes, the arising clonality during tumour evolution, the different rates of success of treatments and the complex and heterogeneous clinical journey of the patients.

- Complex or common diseases require huge sample sizes to disentangle the small contributions of multiple factors, sometimes negative, sometimes protective inherent to the nature of such diseases or conditions.
- Rare diseases studies are limited by the small number of cases available, the plethora of confounding phenotypic conditions, the diverse way of describing them, and the sensitivity of such personal data which adds strong limitations to the sharing of detailed data.
- Infectious diseases, especially for pandemics, which require a quick turnaround of information that circulates swiftly among interested parties.

The study of such diseases involves different types of data:

1. Phenotypic, that describes the conditions and physical characteristics of an individual
2. Clinical, including diagnosis, visits, treatments, interventions, biochemical tests, etc.
3. Genetic or genomic, from the individual, a sample obtained from her/him, from a given population, etc.
4. Images like MRI, electrocardiograms...
5. Metabolic, e.g., drug dynamics
6. Demographics that brings information about the economic and environment
7. Life style or behaviour that, e.g., provides information about diet and exercise

8. Cognitive for domains like neurological diseases or mental health
9. etc,

Researchers aiming to leverage existing data will need to:

1. Find existing datasets, which could be done browsing the bibliography or the Internet
2. Explore the datasets to confirm that they cover an interesting population
3. Request access to the data
4. Harmonise the data if gathered from different places or having a format not appropriate for the desired analysis

#### a) The role of data controllers

The process of sharing involves, at least, two parties: the consumer of the data (the researcher) and the data controller, which controls the governance and access to data they have produced.

Building large single cohorts from data obtained in healthcare or research facilities requires some elements to be in place *a priori*, (in no specific order):

1. Making data collected in research or healthcare available. In the case of healthcare, make them available for secondary use in research given that the appropriate legal and ethical aspects allow it.
2. Ideally, prepare the data (*e.g.*, harmonise the data) using popular standards for formats and vocabularies.
3. Providing a discovery mechanism for researchers that describes the data included in the data collection.

4. Establish the appropriate legal and ethical environment for data sharing.
5. Facilitate the process for requesting and granting access to the data.
6. Providing tools for moving the data to the places where it would be analysed (places that could be very close to the data or abroad).

In many cases, these aspects are delegated or centralised in facilities that specialise in providing the service to the data producer and controller communities and the data consumers.

Each of the aspects above is complex and there is extensive literature about them.

This Doctoral Thesis has mainly contributed to the discovery and harmonisation aspects of the process, while also contributing to the modelling of specialised infrastructures.

## **1.2 Harmonisation is essential**

Every research project or healthcare system has a preferred or organic set of rules to gather data from people into their data capture tools or in the core EHRs. This principle applies to basic aspects like the language used in data gathering (*e.g.*, English, Catalan, Turkish), the metric system units (*e.g.*, centimetres *vs.* inches) or the chosen standards (if any) for identifying phenotypic traits, person classification, medical conditions, etc. (*e.g.*, WHO ICD-10, HPO, ORDO, SNOMED).

Analyses that aggregate data from different sources need to address data heterogeneity before any analyses can be carried-out. The process of transforming data sets to make them comparable is called *data harmonisation*. Harmonisation includes different aspects like data conversion (*e.g.*, between different units), data mapping (*e.g.*, among different categories of two similar variables) or data reshaping (*e.g.*, from one database schema to a different one).

Data harmonisation can be done at the destination (where the analysis would be performed) or at the origin (at the source of the data). Harmonisation of the data at origin is preferable, because it is performed closer to the generation of data, allegedly the place that has a better understanding of such data and, hence, the place able to do the more



informed harmonisation. Harmonisation at origin removes the need of harmonising, time and again, at every destination, which could introduce errors and misinterpretations.

Harmonisation at the origin requires a previous selection or agreement on which would be the shape of the data once it is harmonised. Therefore, the data controller should have a good understanding of which will be the final form of the harmonised data, otherwise said: how the data will look once harmonised. In consequence, if data is harmonised at origin, the capacity of the source institution to provide a number of different conversions or harmonisations would be quite limited, as every harmonisation requires time and resources. The conversion to popular standards (models, unit systems, languages) is, then, the usual approach to solving that issue.

### **1.3 Data management, discovery, and sharing**

The aggregation of data obtained and shared in different places implies several phases: finding the data, getting it, managing it as provider or consumer, and analysing the data.

- ***Data Management*** is the set of processes by which data is captured or received, organised, stored and made ready for reuse.
- ***Data Discovery*** is the set of processes and tools that allows users to understand which data is available, where it is stored, and to get some information or details about it.
- ***Data Sharing*** is the set of processes and tools by which actual access to the data is provided and, when applicable, the intermediate steps to request and receive granted access are cleared.

Proper Data Management, including conditions for data reuse, is relatively easy for some data, specially data that could be shared openly without any restriction, but it requires consistent efforts when it comes to sensitive and/or large-scale datasets, like the ones required for Precision Medicine.

a) Models of data management, discovery and sharing

The set of processes for data management, discovery and sharing requires that the data provider sets up several components:

A place where the data is stored and available on demand.

- A solution for publishing the metadata (the information about the data), so the potential users can discover it.
- A mechanism for reviewing data access requests, or to track data usage if the data is publicly accessible.
- A system of user support to answer questions about the data or to handle reports on issues.

Similarly, the user should prepare for:

- Look for (discover) data that is available and suitable for the analysis.
- Apply to request access to data that is not publicly and anonymously available.
- Providing computing resources capable of performing the analysis.
- Transferring the data to the place where it would be analysed.
- Manage the data as agreed with the data provider, e.g., by limiting access to named authorised people or to destroy the data once the authorised usage period expires.

The data controller could organise each of the components listed above on premise or could delegate it to a third party. Examples of these delegations are moving the data to a cloud for data storage (like Amazon S3<sup>1</sup>), and listing the datasets in a public catalogue like the COVID-19 Data Portal<sup>2</sup> for data discovery. For data access requests, delegating usually

---

<sup>1</sup> <https://docs.aws.amazon.com/AmazonS3/latest/userguide/Welcome.html>

<sup>2</sup> <https://www.covid19dataportal.org/>

happens only when the delegated party is part of the same institution (an ethics committee, for instance) or when the delegated institution sits higher in the data controllership chain (e.g., a hospital delegating into the national health system). Delegating user support depends largely on how good are both, data quality and metadata (documentation); the better the quality and the documentation, the less support would be required.

## b) Data Analysis

Once the user has discovered one or more datasets and has gained access to them, the next step is to perform the desired analysis. Human genomics and health data analysis are increasingly becoming large-scale, complex operations, with two main features: the high sensitivity and the large volume of data. The fact that data is highly sensitive requires access and disclosure to be the minimal necessary. For some institutions, the large volume of data makes it infeasible both to transfer to and to store data in their local premises.

These two features incentivize institutions hosting sensitive data to try to provide some processing infrastructure to allow analysis close to data; or to arrange that both the data and the computing power are hosted at a professional service provider, hence delegating to them these aspects.

Therefore, data management, sharing and analysis are evolving from getting all the data sets on premise for analysis, to what is known as a *distributed analysis*: performing the analysis in several secured environments where each data set is located.

## c) Federation

Usually, the data providers don't have the infrastructure required for distributed analysis happening at their premises. It is a common approach to overcome this limitation by delegating to a third party, like a commercial cloud provider or a publicly funded computing facility (e.g., a supercomputing centre).

Organisations hosting similar data sets that anticipate being part of many distributed analyses could assemble themselves to address common challenges, and also to offer a uniform set of tools and procedures that makes distributed analysis easier for their users. This is like an *harmonisation of infrastructures* instead of the data. Organisations looking for an

economy of scale can also join forces to share the cost of data sharing operations among them. This type of collective solution is called a *Federation*.

The federation approach could be applied to different aspects of the data sharing process: managing user identities, storage or computing power, catalogues, etc.

The Beacon protocol, that will be introduced later, is suggesting a harmonisation of discovery infrastructures that provides options for harmonisation of data at origin,

## **1.4 The role of standards for sharing inside or outside of a federation**

As mentioned above, the success of analysing large cohorts relies on harmonisation: harmonisation of discovery, harmonisation of data and harmonisation of infrastructures. Harmonisation is based on agreements in the syntax and semantics of the data and the processes that became standards, *de iure* or *de facto*. Standards *de facto* are such products that became popular and used extensively. Standards *de iure* are usually born from a community need. The community organises –organically or leveraging existing bodies, entities or societies– and suggests a solution for that need. Examples of such organisations are: the World Wide Web Consortium<sup>3</sup> (W3C), the International Standards Organisation<sup>4</sup> (ISO) or the Global Alliance for Genomics and Health<sup>5</sup> (GA4GH).

### a) The Global Alliance for Genomics and Health

The Global Alliance for Genomics and Health (GA4GH) is a worldwide organisation with the goal of advancing precision medicine by enabling the responsible sharing of clinical and genomic data. The GA4GH was established in 2013 and, at the time of writing, has more than 600 organisational members and more than 2,000 subscribers from more than 90 countries.

The GA4GH model is to embrace or create “products” that are specifications or reference Ethical, Legal and Societal (ELSI) documents. The GA4GH is, nowadays, the organisation

---

<sup>3</sup> <https://www.w3.org/>

<sup>4</sup> <https://www.iso.org/>

<sup>5</sup> <https://www.ga4gh.org/>

in charge of maintaining (*de facto*) file format standards for bioinformatics like BED<sup>6</sup>, SAM/BAM<sup>7</sup>/CRAM<sup>8</sup> and VCF<sup>9</sup>.

The specifications for file formats provide a basic harmonisation on the structure (the syntax) of the shared data. The specifications for protocols describe the dialog that clients and servers could use to share the data. With some exceptions, the GA4GH has not been active in suggesting dictionaries or ontologies that could be used inside the file formats or specifications, which is the semantic component of communication.

The ELSI documents, like templates and guidelines, contribute to enabling data sharing by harmonising the conditions for such a process. If the conditions for data sharing are similar between different organisations, for new data controllers that want to share data it will become easier to rely on these examples and share data with a higher feeling of safety both for the data donors and for the hosting institutions.

The GA4GH organises its activities in several thematic work streams (WS) that rely, and are endorsed by, real world projects, called Driver Projects (DP). The Driver Projects contribute by explaining their needs or requirements, helping in the definition of the solutions to such requirements and piloting the implementations of such solutions.

In the context of this PhD thesis, we have contributed to GA4GH from two different Driver Projects: ELIXIR Beacon and the European Genome-phenome Archive (EGA). As part of these DP, we have participated in Work Streams on Discovery, Clinical & Phenotypic Data Capture (Clin-Pheno) and Data Use & Researcher Identities (DURI).

The Discovery WS<sup>10</sup> focuses on finding and querying datasets that could be relevant for a given researcher or clinician. The products designed by the Discovery WS take into consideration that the type and details of the shared data must be adjusted to the sensitivity of the data and the preferences of the data controller.

---

<sup>6</sup> <https://github.com/samtools/hts-specs/blob/master/BEDv1.pdf>

<sup>7</sup> <https://samtools.github.io/hts-specs/SAMv1.pdf>

<sup>8</sup> <https://samtools.github.io/hts-specs/CRAMv3.pdf>

<sup>9</sup> <https://samtools.github.io/hts-specs/VCFv4.3.pdf>

<sup>10</sup> <https://ga4gh-discovery.github.io/>

The Clinical & Phenotypic Data Capture WS<sup>11</sup> focuses on designing models for clinical and phenotypic data, including aspects like pedigree descriptions. It provides standard models like Phenopackets version 2 (*Jacobsen et al. 2022*), which includes some recommendations on the ontologies to be used.

The Data Use & Researcher Identities<sup>12</sup> (DURI) WS focuses on handling the identities and the authorizations of the users accessing services, i.e. the GA4GH Passports and Visas (*Voisin et al. 2021*) and also focuses on describing the conditions for using the available data, i.e. the Data Use Ontology (DUO) (*Lawson et al. 2021*).

## b) GA4GH Standards: Beacon version 1

The GA4GH Beacon version 1 (*Fiume et al. 2019*) is a protocol for discovering genomic variations. It is a basic REST API (*Fielding 2000*) (*Representational State Transfer - Application Programming Interface*) that allows querying a Beacon instance about the presence of a given genomic variation in the data managed by that Beacon. The answer to the query is simply a “Yes” or a “No”.

Beacon version 1 was conceived as a social experiment to test the actual feasibility for sharing data from organisations willing to do so. The rationale was to lower the technical complexity to a minimum so that the non-technical barriers will be more visible. The non-technical barriers range from legal limitations to psychological ones, like the fear of doing something wrong when sharing sensitive or potentially sensitive data.

Beacon version 1 has been quite popular, in particular when DNASTack<sup>13</sup>, a Canadian company, provided the Beacon Network<sup>14</sup> which allows querying several Beacons at once, aggregating the responses received from them. At the time of writing this thesis, the DNASTack Beacon Network links around 80 Beacons, but is no longer under development.

Beacon version 1 focused on demonstrating that sharing sensitive data is possible. During the conception of Beacon version 1 it was accepted that it would not be sufficient for its

---

<sup>11</sup> <https://ga4gh-cp.github.io/>

<sup>12</sup> <https://ga4gh-duri.github.io/>

<sup>13</sup> <https://dnastack.com/>

<sup>14</sup> <https://beacon-network.org/>

actual use in research nor in precision medicine. It was clear that, were Beacon v1 prove successful, another version should address the requirements of real world discovery. This has been the goal of GA4GH Beacon version 2.

## **1.5 ELIXIR, the Federated EGA and the Beacon Network**

ELIXIR<sup>15</sup> is an European virtual infrastructure that coordinates bioinformatics resources and helps to align infrastructures and resources that already exist in country members. ELIXIR focus in managing and keeping safe the increasing volume of data being generated by publicly funded research. ELIXIR starts as a community of disconnected resources that aims to become less heterogeneous by moving towards a federation of services that provide an homogeneous experience to its users. Therefore, leveraging standards and in particular those related with bioinformatics and sensitive human data is critical for ELIXIR. ELIXIR supported the inception and development of the Federated EGA through EXCELERATE<sup>16</sup> and CONVERGE<sup>17</sup> European Commission funded projects.

ELIXIR is also providing a Beacon Network of its own, which links to around a dozen Beacons. ELIXIR's Beacon Network<sup>18</sup> is currently evolving to include Beacons based on the new GA4GH Beacon specification: Beacon version 2.

---

<sup>15</sup> <https://elixir-europe.org/>

<sup>16</sup> <https://elixir-europe.org/excelerate>

<sup>17</sup> <https://elixir-europe.org/about-us/how-funded/eu-projects/converge>

<sup>18</sup> <https://beacon-network.elixir-europe.org/>





## 2. OBJECTIVES

The goal of this work is to contribute to secure sharing of sensitive human genomic data for precision medicine. For that goal to be successful, there is a clear requirement: that the contributions are significant enough and bring positive impacts to the community and to society at large. To tag the contributions as *significant*, suggested solutions must be adopted by a broad community (becoming a standard *de facto*) and/or the endorsement of a standards organisation must be obtained (becoming a standard *de iure*). Both options require the contribution of other people or institutions (a community) that provide requirements, that test the solution and return feedback; and also time for the solution to prove effective.

While some of the goals above are being achieved, the time element is out of reach for the duration of a PhD thesis and, therefore, its goals must be limited to provide a set of solutions and suggestions based on the community participation and feedback. Thus, the specific objectives of this thesis are:

- 1) To facilitate data management (as an enabler of the other objectives).
- 2) To facilitate data discovery by means of:
  - a. Allowing for discovering donors or cohorts with a given set of characteristics.
  - b. Allowing for centralised or distributed discovery.
  - c. Helping on the harmonisation of the query and the results.
- 3) To facilitate data sharing



### 3. RESULTS

In this PhD, contributions are provided for each of the three main goals:

- 1) Data Management, see papers: *Freeberg et al. 2021; Thorogood et al. 2021*
- 2) Data Discovery see papers: *Fiume et al. 2019; Rambla et al. 2022; Rueda et al. 2022*
- 3) Data Sharing see papers: *Dyke et al. 2018; Freeberg et al. 2021; Harrow et al. 2021; Lawson et al. 2021; Saunders et al. 2019; Thorogood et al. 2021*

The papers provide further details on each aspect. I include four of these papers as chapters of these thesis, using as criteria relevance for each of the three main goals and level of contribution. For Data Management, I led the work in the paper on the EGA update for 2021 (Chapter 1). For Data Discovery, I led the work for the two beacon papers that are included in Chapters 2 and 3. Finally, for Data Sharing, I contributed significantly to several papers, but include here one in Chapter 4. The rest of papers, while also relevant and related to the thesis, are included in an Annex (Annex 2).

The Discussion and Conclusions section includes additional background and considerations.



## Chapter 1

Freeberg, M. A., Fromont, L. A., D'Altri, T., Romero, A. F., Ciges, J., Jene, A., Kerry, G., Moldes, M., Ariosa, R., Bahena, S., Barrowdale, D., Barbero, M., Fernandez-Orth, D., Garcia-Linares, C., Garcia-Rios, E., Haziza, F., Juhasz, B., Llobet, O., Milla, G., . . . **Rambla, J.** (2021). [The European Genome-phenome Archive in 2021](#). *Nucleic Acids Research*, 50(D1), D980–D987.



# The European Genome-phenome Archive in 2021

Mallory Ann Freeberg<sup>1,†</sup>, Lauren A. Fromont<sup>1,†</sup>, Teresa D’Altri<sup>2</sup>, Anna Foix Romero<sup>1</sup>, Jorge Izquierdo Ciges<sup>1</sup>, Aina Jene<sup>2</sup>, Giselle Kerry<sup>1</sup>, Mauricio Moldes<sup>2</sup>, Roberto Ariosa<sup>2</sup>, Silvia Bahena<sup>1</sup>, Daniel Barrowdale<sup>1</sup>, Marcos Casado Barbero<sup>1</sup>, Dietmar Fernandez-Orth<sup>2</sup>, Carles Garcia-Linares<sup>1</sup>, Emilio Garcia-Rios<sup>1</sup>, Frédéric Haziza<sup>2</sup>, Bela Juhasz<sup>1</sup>, Oscar Martinez Llobet<sup>2</sup>, Gemma Milla<sup>2</sup>, Anand Mohan<sup>1</sup>, Manuel Rueda<sup>2</sup>, Aravind Sankar<sup>1</sup>, Dona Shaju<sup>1</sup>, Ashutosh Shimpi<sup>1</sup>, Babita Singh<sup>2</sup>, Coline Thomas<sup>1</sup>, Sabela de la Torre<sup>2</sup>, Umuthan Uyan<sup>2</sup>, Claudia Vasallo<sup>2</sup>, Paul Flicek<sup>1</sup>, Roderic Guigo<sup>2</sup>, Arcadi Navarro<sup>2</sup>, Helen Parkinson<sup>1</sup>, Thomas Keane<sup>1,\*</sup> and Jordi Rambla<sup>2,\*</sup>

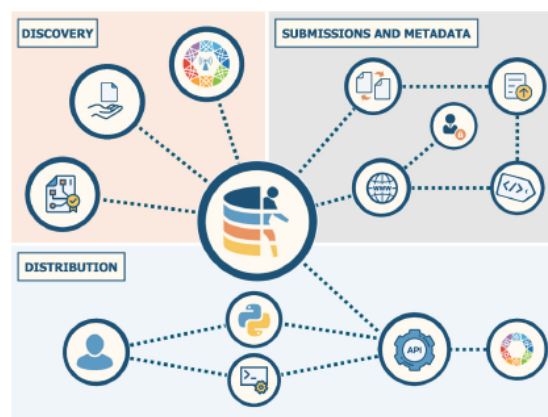
<sup>1</sup>European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Hinxton CB10 1SD, UK and <sup>2</sup>Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr Aiguader 88, Barcelona 08003, Spain

Received September 03, 2021; Revised October 08, 2021; Editorial Decision October 14, 2021; Accepted October 22, 2021

## ABSTRACT

The European Genome-phenome Archive (EGA - <https://ega-archive.org/>) is a resource for long term secure archiving of all types of potentially identifiable genetic, phenotypic, and clinical data resulting from biomedical research projects. Its mission is to foster hosted data reuse, enable reproducibility, and accelerate biomedical and translational research in line with the FAIR principles. Launched in 2008, the EGA has grown quickly, currently archiving over 4,500 studies from nearly one thousand institutions. The EGA operates a distributed data access model in which requests are made to the data controller, not to the EGA, therefore, the submitter keeps control on who has access to the data and under which conditions. Given the size and value of data hosted, the EGA is constantly improving its value chain, that is, how the EGA can contribute to enhancing the value of human health data by facilitating its submission, discovery, access, and distribution, as well as leading the design and implementation of standards and methods necessary to deliver the value chain. The EGA has become a key GA4GH Driver Project, leading multiple development efforts and implementing new standards and tools, and has been appointed as an ELIXIR Core Data Resource.

## GRAPHICAL ABSTRACT

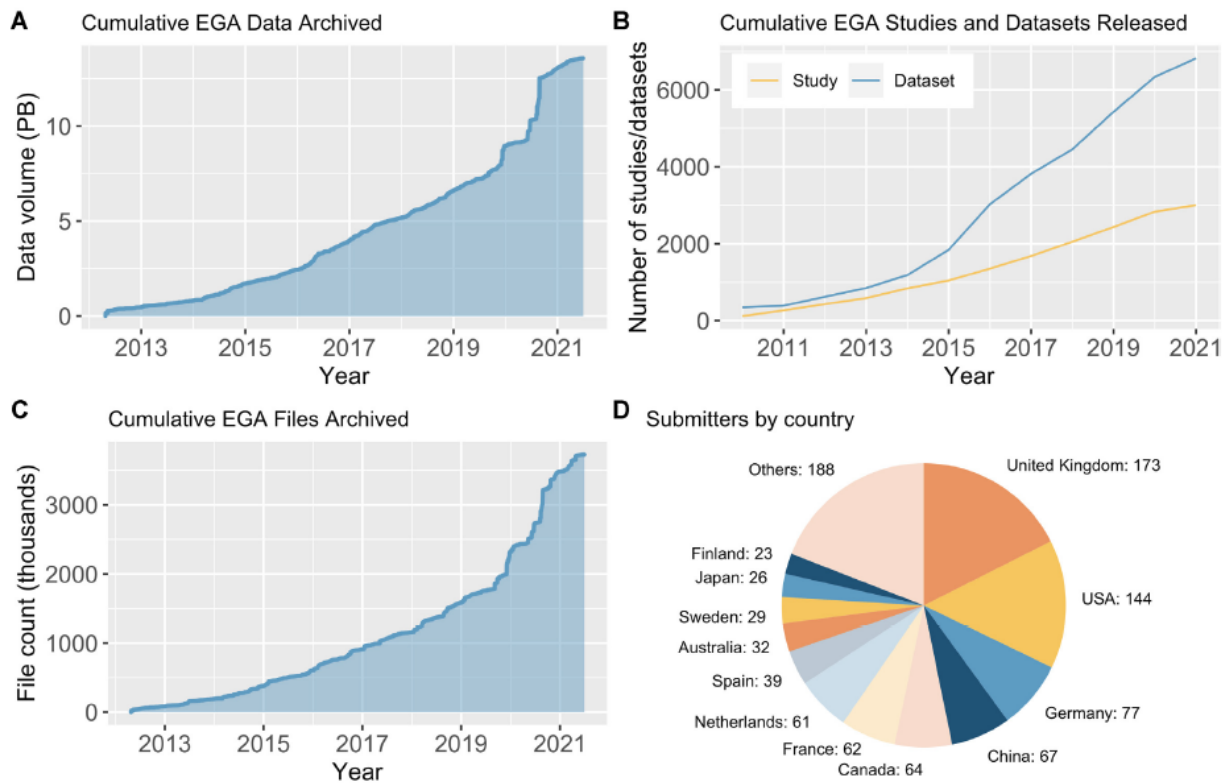


## INTRODUCTION

The European Genome-phenome Archive (EGA) is a resource for permanent secure archiving and sharing of all types of potentially identifiable genetic, phenotypic, and clinical data resulting from biomedical research projects (1). This data is subject to participant consent agreements, so sharing is restricted to bona fide researchers for specific research purposes. In recent years, governments world-wide have enacted data privacy protection laws and regulations to protect the rights of their citizens, further restricting how personal data is shared (2). In this environment, services for securely archiving and sharing sensitive human data for research are more important than ever. The EGA’s mission

\*To whom correspondence should be addressed. Email: [jordi.rambla@crg.eu](mailto:jordi.rambla@crg.eu)  
Correspondence may also be addressed to Thomas Keane. Email: [tk2@ebi.ac.uk](mailto:tk2@ebi.ac.uk)

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.



**Figure 1.** Data archived at EGA between 2013–2021. Cumulative size of data (A), number of studies and datasets (B), and number of files (C) archived and available for download from EGA per year. (D) Number of institutes per country that have archived data at the EGA.

is to foster data reuse, enable reproducibility, and accelerate biomedical and translational research in line with the FAIR (Findable, Accessible, Interoperable, and Reusable) principles (3).

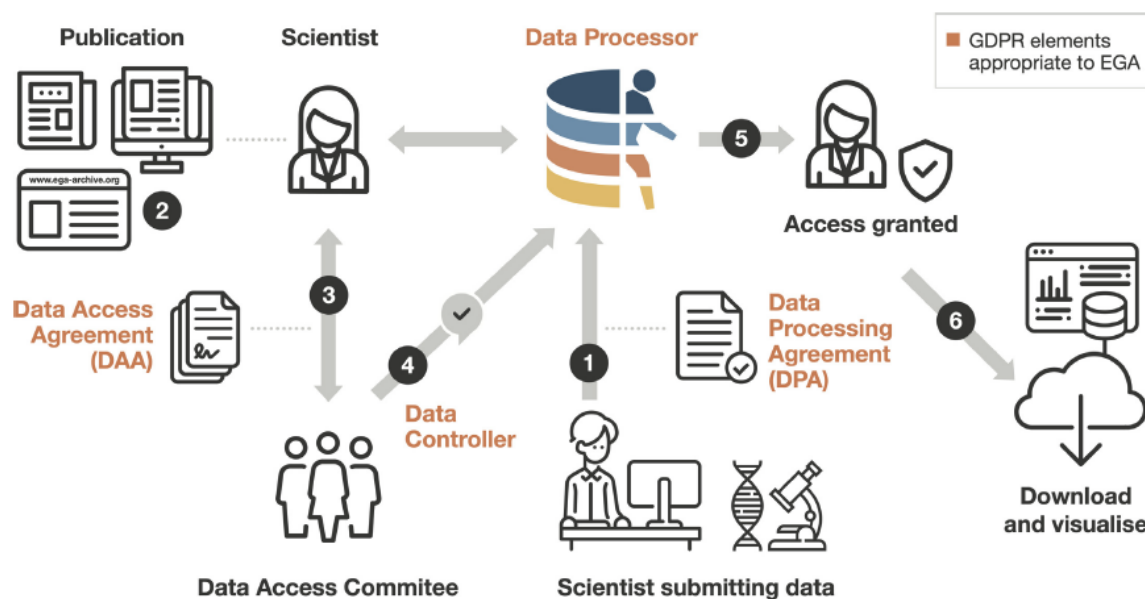
Since its launch in 2008, the EGA has experienced rapid growth, archiving over 4500 studies comprising 6800 datasets made up of nearly 15 PB of sensitive human data (Figure 1A–C). Studies archived in the EGA represent a variety of research fields (e.g. cancer, rare diseases, infectious diseases, common/chronic diseases), data types (e.g. genetic/genomic, phenotypic, clinical) and technologies (e.g. whole genome/exome sequencing, bulk and single cell RNA sequencing, DNA methylation-sensitive sequencing) from researchers around the world (Figure 1D). Since the inception of the Global Alliance for Genomics and Health (GA4GH), the EGA has been a founding partner and Driver Project, leading multiple workstream development efforts and piloting new standards and tools. To promote data discovery, the EGA co-leads the Beacon project (<https://beacon-project.io>) that will allow for the browsing of datasets that contain specific genomic information of interest. The EGA is also a core contributor to the GA4GH Researcher Passport standard, which can be used to reliably authenticate a researcher's digital identity and automate their access to a requested genomic dataset, and provided one of the first production level deployments.

To improve the FAIRness of human research data, EGA services include data submission, discovery, and access to

the global research community (Supplementary Table S1). For data submitters, the EGA offers a web-based Submitter Portal to guide users through the submission process, including assembling and validating metadata. Submitters are provided stable, globally unique identifiers to enable reference of datasets in publications and across genomics infrastructures. The EGA provides search options for discovering relevant datasets by keywords, data use conditions, variants, and accessions. To allow data controllers to manage data access permissions, the EGA offers a web-based portal and an API. Finally, the EGA has greatly expanded its data access services including support for downloading specific genomic regions, real-time visualisation in a genome browser, and more efficient file encryption approaches.

Data sharing and reuse is vital for advancing clinical and genomics research. A notable example of EGA data reusability is genotyping data from the UK Biobank, a large-scale biomedical research resource of in-depth genetic and health information (4). Released in 2017, this dataset contains directly genotyped and imputed data for all 500 000 UK Biobank participants and has been downloaded from the EGA by >600 researchers. Another example is the Wellcome Trust Case Control Consortium study (5). This study was released in 2007 and contains genome-wide case-control association data from over 5000 individuals to study seven major diseases in the British population. The data has been downloaded from the EGA by more than 2600 researchers and the study cited over 6000 times. Given the size





**Figure 2.** EGA facilitates the submission, discovery, access, and distribution of sensitive human data. A researcher submits controlled access human genetic, phenotypic and clinical data to EGA after signing a Data Processing Agreement (1). EGA processes, archives, and releases the dataset to be findable. Another researcher discovers data of interest at the EGA (2). They contact the Data Access Committee for the data of interest and agree to the terms of data reuse by signing a Data Access Agreement (3). The Data Access Committee informs EGA that access is approved (4). The EGA grants access to the requesting researcher (5) who can then download and visualise the data (6). *GDPR: General Data Protection Regulation.*

and value of data hosted at the EGA, it is important to consider how to improve the archive's value chain—that is, how the EGA can contribute to enhancing the value of human health data by facilitating its submission, discovery, access, and distribution, as well as leading the design and implementation of standards and methods necessary to deliver the value chain (Figure 2). These aspects will be addressed in this article.

### DEPOSITING DATA AT THE EGA

The start of the EGA value chain is the deposition of data. The submission process includes raw or processed data (or both) and metadata. Data correspond to the set of files produced by researchers from an experiment or data analysis and must be encrypted before submission using strong compression algorithms (e.g. AES256, Crypt4GH (6)). Metadata describe the data files and include information about the study, the samples from which the data were generated, and the process by which data were generated and analysed. The EGA receives studies of different sizes and complexity which can make submitting metadata challenging. The EGA offers a web-based, interactive Submitter Portal where users can enter and organise metadata manually. For large-scale or highly complex projects, the EGA provides an API to submit programmatically. The metadata model is based on the International Nucleotide Sequence Database Collaboration (7). The EGA actively contributes to developing additional models, for example the GA4GH Phenopackets standard (<http://phenopackets.org/>) for interoperable sharing of phenotype descriptions linked to disease, patient, and genetic information.

The submission process requires the signature of the Data Processing Agreement (DPA). The DPA states the conditions and responsibilities of data processing as well as the relationship between the data controller (Data Access Committee, DAC) and the data processor (EGA) (Figure 2). By signing this agreement, data controllers can ensure sensitive data are being handled according to data protection regulations and with security protections in place to prevent unauthorised access.

### DATA DISCOVERY AT THE EGA

The next step of the EGA value chain is providing users ways to discover EGA data relevant to their specific research aims. The EGA website ([www.ega-archive.org](http://www.ega-archive.org)) is the main entry point for data discovery, and in recent years this and other EGA services have been updated with new features.

#### Discovery by publication

Scientific publications are a common way for researchers to discover datasets that are relevant to their research. The EGA website displays links to associated publications for each study, enabling researchers to quickly find additional information about the original study and subsequent studies that have reused the data. To date, the EGA links to over 3000 publications, many of which are provided by submitters during the submission process. Additionally, the EGA continuously mines Europe PubMed Central (8) for EGA study and dataset accessions and adds links to these publications on the EGA website.

### Discovery by variants

The EGA Beacon API implements the GA4GH Beacon standard (9) and enables querying for genomic variants in datasets that have consented to be part of the EGA Beacon. In this way, dataset with variants of interest can be discovered by researchers prior to them applying for approval to access the entire dataset.

### Discovery by public metadata

The EGA website enables discovery of datasets by searching public metadata in different ways including by free text, controlled vocabularies, accessions, and other features (<https://ega-archive.org/howtosearch>). The search engine accounts for common spelling mistakes, capitalisation and most punctuation, and also suggests similar search term combinations with a higher number of results to increase the usefulness of the search. Researchers can perform similar searches over public metadata programmatically using the EGA Metadata API (<https://ega-archive.org/metadata/how-to-use-the-api>).

### Discovery by data use ontology

Human subject datasets often have use conditions such as ‘only available for cancer use’ or ‘only available for the study of pediatric diseases’ based on the original participant consent, which must be respected when sharing and studying these datasets. Working with the GA4GH Data Use and Researcher Identities workflow, the EGA has adopted the Data Use Ontology (DUO) (10,11) to describe these conditions using a standard vocabulary. DUO terms allow data controllers to semantically tag datasets with usage conditions, allowing the datasets to be automatically discoverable based on authorisation level or intended use. DUO terms are displayed on EGA dataset webpages and can be searched for using the textual search functionality.

### Discovery by data quality

High-quality data standards are essential to ensure the quality and credibility of archived data. The File Quality Control (QC) Report service (<https://ega-archive.org/about/quality-control-reports>) was developed to provide generic quality control reports for FASTQ, SAM/BAM/CRAM and VCF files deposited at EGA. QC Report allows anonymous EGA website users to view summary-level information regarding the files within a specific dataset, such as quality of reads, alignment quality, number and type of variants, and other features. Researchers benefit from being able to assess the quality of data prior to the data access decision, increasing the reusability of data.

### Discovery through linked resources

To broaden data discoverability, the EGA has established links with other public resources. For example, EGA samples are accessioned by BioSamples (12) which stores information about biological samples used in research. Within BioSamples, researchers can link samples from the same

study even if the data generated from those samples are in different archives. By linking samples, researchers can discover, for example, viral sequences archived at the European Nucleotide Archive (13) that have corresponding host genomic data archived at the EGA. In response to the COVID-19 pandemic, the European COVID-19 Data Portal (<https://www.covid19dataportal.org>) was established to accelerate COVID-19 research through data sharing. COVID-19 and SARS-CoV-2 studies archived at the EGA are indexed and displayed in the COVID-19 Data Portal, providing an additional route by which EGA data can be discovered by researchers. Finally, through daily synchronization with a metadata exchange server, the EGA provides summary information for and links to studies archived at the database of Genotypes and Phenotypes (dbGaP, <https://www.ncbi.nlm.nih.gov/gap/>). In this way, the EGA serves as a global hub for discovery of human data under access control.

## ACCESSING DATA AT THE EGA

### Data access model

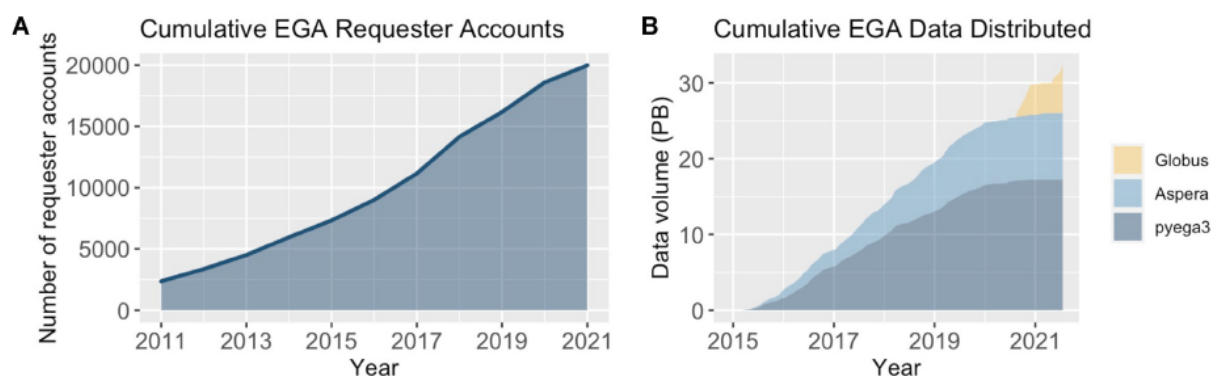
The next step in the value chain is providing data access to approved requestors. Given the complexity, scale, and diversity of global submitters and studies (Figure 3), the EGA operates a distributed data access model in which requests are made to the data controller, not to the EGA. The data controller comprises one or more individuals in a DAC that reviews access requests and approves or rejects them based on intended data use. Terms and conditions are specified in a Data Access Agreement (DAA) that an individual agrees to before being granted access. Such agreements include data management and security policies, terms for publication or embargoes, and restrictions on data use or sharing.

Once a researcher has identified datasets of interest, they contact the appropriate DAC to request access. If approved, an EGA account is created for the data requester. EGA accounts are individual: if more than one person from a research group or consortium wants access, everyone must be approved by the DAC. Sensitive human data resources can contain hundreds or thousands of datasets, each with its own controlling DAC and data use conditions. In fact, the EGA manages datasets for over 1,500 different DACs. By operating a distributed data access model, the EGA provides the infrastructure and services for secure data archive and distribution so that DACs can focus their efforts on reviewing data access requests.

This model has been extremely beneficial to promote data reuse: 624 of the studies deposited at the EGA have been used in other studies at least once. The METABRIC microRNA landscape study (14), which identified miRNAs that potentially play a role in breast cancer progression, has been re-used 25 times, generating scientific progress and accumulating over 675 citations to date.

### Authentication and authorisation

Authentication and authorisation infrastructure (AAI) management is key for operating the EGA. Authentication is verifying the identity of a user, while authorisation is confirming a user has access rights to specific information. The



**Figure 3.** EGA data distribution to approved researchers between 2011 and 2021. (A) Number of EGA data requester accounts created over time. (B) Amount of data distributed to approved researchers over time.

ability to manage and audit who has access to what data is required for preventing malicious or accidental unauthorised data access. The EGA's AAI implementation is compatible with the GA4GH AAI standard, ensuring that data access can be managed interoperably with other GA4GH AAI-compatible resources. Users can interact with multiple services that EGA has built on top of the AAI. To access sensitive metadata, a researcher signs in to the EGA website where their credentials are authenticated to verify their identity. They navigate to a dataset of interest and request download of the sensitive metadata, triggering the EGA to validate their request against the permissions assigned to their identity. If the researcher has permission to access the dataset, the request is authorised and they can download the metadata.

With the increasing number of data resources managing and analysing sensitive human data, two key needs have arisen: users want one set of credentials for multiple resources, and resources need to manage user identities and permissions in an interoperable way. The EGA has implemented solutions to address these needs. First, the EGA supports linking of EGA user identities to identities issued by the ELIXIR (15) AAI service (16,17). Once linked, ELIXIR credentials can be used with EGA services. Second, the EGA supports interoperable identities and permissions by conforming to the GA4GH Passports standard ([https://github.com/ga4gh-duri/ga4gh-duri.github.io/tree/master/researcher\\_ids](https://github.com/ga4gh-duri/ga4gh-duri.github.io/tree/master/researcher_ids)). A Passport is a machine-readable digital identity that contains information about what data someone is approved to access. A data requester can use the EGA Permissions API to retrieve a list of datasets they have access to at the EGA, while a DAC can use the API to add and remove permissions according to their data use policy. An updated web-based portal is under development as a service for DACs to manage permissions for their EGA datasets.

### Data distribution

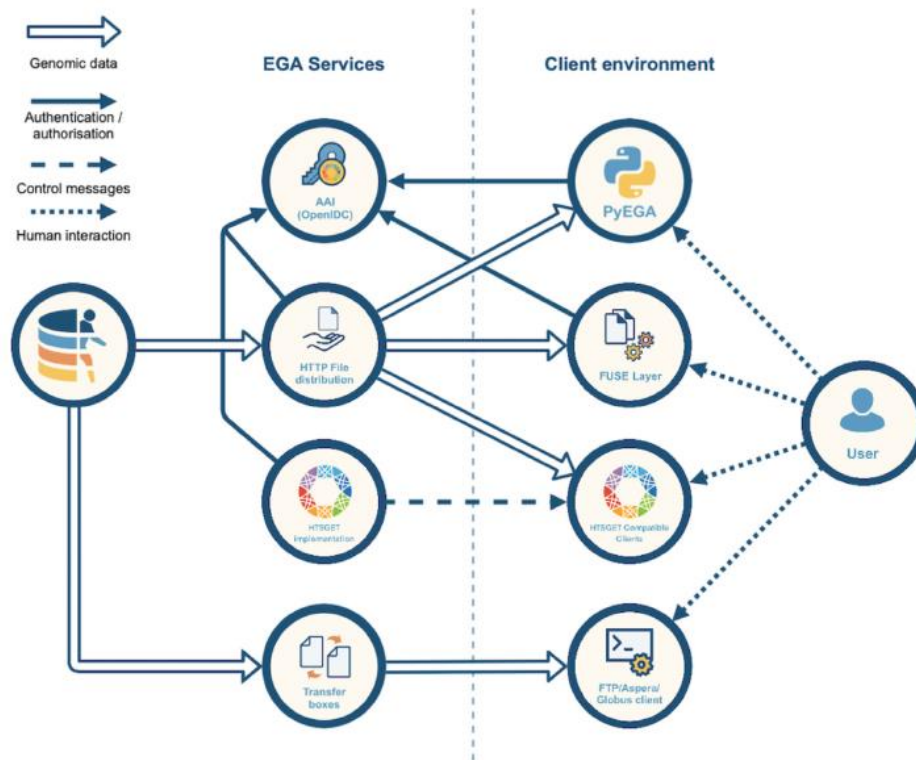
Genomic and phenotypic datasets archived at the EGA can be composed of a few files to hundreds of thousands of files ranging in size from very small to quite large. Importantly, these files must be encrypted at rest and during download

over secure channels to prevent unauthorised access. EGA brings value to this process by offering users a diversity of options based on their needs (Figure 4).

Large genomic data files pose challenges for researchers who have limited space to store files or network bandwidth to download them. Research questions can often be answered by looking at a specific region of the genome, for example a gene locus or chromosome. The EGA collaborated with the GA4GH to develop the *htsget* secure streaming protocol (18) for enabling real-time random access by genomic coordinates for sequencing read (e.g. CRAM) and variant (e.g. VCF) data. Specifying a genomic region using *htsget* results in a smaller file that can be downloaded more quickly. As an example, *htsget* was deployed in the RD-Connect Genome-Phenome Analysis Platform (<https://platform.rd-connect.eu/>) to enable rare disease researchers to inspect supporting read data in real-time through the Integrative Genomics Viewer browser (19).

Previous methods for retrieving data from the EGA were not parallelisable, required installing and running two tools (one to download, one to decrypt), and were prone to interruptions when downloading large files over an unstable connection. A new tool, PyEGA3 (<https://github.com/EGA-archive/ega-download-client>), offers enhanced features to support more efficient and robust data download. Files are securely delivered unencrypted to a user's local environment, removing the need for a separate decryption step. Download automatically restarts from where it left off, avoiding the need to start from scratch if the connection is interrupted, and users can specify the number of connections to enable parallel downloads. Finally, PyEGA3 implements *htsget* to support retrieval of specific genomic regions.

The EGA offers Filesystem in Userspace (FUSE) layer software solutions to allow users access to EGA files as transparently as if they were local files. The EGA FUSE client (<https://github.com/EGA-archive/ega-fuse-client>), after authenticating a user, mounts a virtual filesystem displaying all the files available to them. When they want to access these files, the streaming API decodes the byte stream and sends the data to the user over a secure channel.



**Figure 4.** The EGA offers a variety of secure data access and download services to meet user needs, many of which implement GA4GH standards. *FUSE: Filesystem in Userspace*. *AAI: Authentication and Authorization Infrastructure*. *OpenIDC: OpenID Connect, an open standard and decentralized authentication protocol*.

Further, the EGA is piloting the GA4GH Crypt4GH standard (6) in which file decryption occurs on the client side, reducing the stress on EGA servers. This approach enables researchers to inspect data files faster: specific parts of a data file of interest are decrypted on-the-fly, without having to decrypt the whole file, and the information is available in real-time.

## DISCUSSION

We are seeing the emergence of many human data resources across the globe including national biobanks, disease specific portals, clinical variants, and genetic association resources. Interoperable standards between the EGA and other human data repositories are instrumental to develop personalised medicine strategies. Active engagement by the EGA with international standards bodies, for example the GA4GH, Biobanking and Biomolecular Resources Research Infrastructure (20), and ELIXIR (15), is essential to further EGA interoperability. The EGA actively engages with GA4GH to develop and implement standards in areas such as genomic data formats (e.g. CRAM), secure streaming (e.g. htsget), and harmonisation of data access standards (e.g. researcher IDs, AAI interoperability, phenotype exchange formats). The EGA has been appointed an ELIXIR core data resource and partners with other ELIXIR human data infrastructures (e.g. RD-Connect,

Dutch Center for Translational Molecular Medicine) to provide implementations of GA4GH standards.

Molecular medicine is undergoing a paradigm shift as advances in high throughput DNA sequencing technology make it feasible to use genomics in clinical practice. The mission of the EGA is to enable sharing of human genetic data for research, acknowledging that in the future much of this data is likely to come from healthcare. However, as healthcare institutions are a national competence, data generated there is unlikely to be shared as freely as research data. Many countries, in Europe and beyond, are trying to address the interplay between using research generated data for personalized medicine and using healthcare generated data for secondary analysis in research, thus creating a virtuous circle between healthcare and research. Most of these countries are still in the planning, funding or organizing phases and, consequently, many aspects are still to be decided. It is clear, however, that all of them plan for a federated model, where the data is not leaving the corresponding jurisdiction and the control about who is accessing the data is kept locally. The EGA Strategic Committee started to plan for such a new scenario in the context of the ELIXIR EXCELERATE project, back in 2016. The EGA is currently transitioning from a centralised resource managed by EMBL-EBI (Hinxton, UK) and CRG (Barcelona, ELIXIR Spain, with key support of the Barcelona Supercomputing Centre) to a federated node model. The Federated EGA is designed to support national

data management requirements for genomic and clinical data collected from their citizens as part of healthcare or biomedical research projects. We have engaged with representatives from 14 ELIXIR nodes through the ELIXIR Federated Human Data community (<https://elixir-europe.org/communities/human-data>), Beyond 1 Million Genomes (B1MG), ELIXIR CONVERGE (<https://elixir-europe.org/about-us/how-funded/eu-projects/converge>), and 1 + Million Genomes (2) projects over the past 18 months to develop the federation model together. Our shared vision is that the Federated EGA will provide the cross border data sharing infrastructure and standards to enable secondary reuse of healthcare derived genetic data in Europe and beyond.

## DATA AVAILABILITY

The European Genome-phenome Archive can be accessed via: <https://ega-archive.org/>. Content is distributed under the EMBL-EBI Terms of Use available at <https://www.ebi.ac.uk/about/terms-of-use> and the CRG Terms of Use available at <https://www.crg.eu/en/content/legal-notice-privacy-policy>.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

The authors would like to thank former EGA team members for their contributions to the EGA: Sergi Aguilo, Mario Alberich, Jeff Almeida-King, Pablo Arce, Minjie Ding, Alfred Gil, Cristina Yenyxe Gonzalez Garcia, Jag Kandasamy, Vasudev Kumanduri, Ilkka Lapalainen, Audald Lloret, Sira Martinez, Dietmar Orth, Justin Paschall, Saif Ur Rehman, Gary Saunders, Alexander Senf, Marc Sities, Thomas Smith, Dylan Spalding, Nino Spataro, Alexander Vikhorev, Matthieu Vizuete-Forster, and Zahra Waheed. The authors thank graphic designer Frederike Werkmeister for designing Figure 2 of the present article. Additionally, the authors would like to acknowledge members of the ELIXIR Human Data Communities, the Global Alliance for Genomics and Health, the Galaxy Project, and all former and current project collaborators whose input and work has contributed to improving EGA services. We acknowledge support of the Spanish Ministry of Science and Innovation to the EMBL partnership, the Centro de Excelencia Severo Ochoa and the CERCA Programme / Generalitat de Catalunya. Finally, the authors would like to thank the Barcelona Supercomputing Centre and the EMBL-EBI Technical Services Cluster for technical services and support essential for operating the EGA.

## FUNDING

Horizon 2020 Programme of the European Union [CORBEL [654248], ELIXIR-EXCELERATE [676559], Solve-RD [779257], EASI-Genomics [824110], EJP-RD [825575], CINECA [825775], EuCanCan [825835], EUCanshare [825903], ELIXIR-CONVERGE [871075]];

Wellcome Trust Global Alliance for Genomics and Health [201535/Z/16/Z]; UK Biobank; Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation [2017-171304 (5022)]; European Molecular Biology Laboratory (EMBL); LaCaixa Foundation [004745/008034]; [LCF/PR/CE20/50740008]. Funding for open access charge: LaCaixa Foundation [LCF/PR/CE20/50740008].

**Conflict of interest statement.** Paul Flicek is a member of the Scientific Advisory Boards of Fabric Genomics, Inc. and Eagle Genomics, Ltd.

## REFERENCES

- Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J.D., Ur-Rehman, S., Saunders, G., Kandasamy, J., Caccamo, M., Leinonen, R. *et al.* (2015) The European Genome-phenome Archive of human data consented for biomedical research. *Nat. Genet.*, **47**, 692–695.
- Saunders, G., Baudis, M., Becker, R., Beltran, S., Bérout, C., Birney, E., Brooksbank, C., Brunak, S., Van den Bulcke, M., Drysdale, R. *et al.* (2019) Leveraging European infrastructures to access 1 million human genomes by 2022. *Nat. Rev. Genet.*, **20**, 693–701.
- Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, **3**, 160018.
- Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M. *et al.* (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.*, **12**, e1001779.
- Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
- Senf, A., Davies, R., Haziza, F., Marshall, J., Troncoso-Pastoriza, J., Hofmann, O. and Keane, T.M. (2021) Crypt4GH: a file format standard enabling native access to encrypted data. *Bioinformatics*, **37**, 2753–2754.
- Arita, M., Karsch-Mizrachi, I. and Cochrane, G. (2021) The international nucleotide sequence database collaboration. *Nucleic Acids Res.*, **49**, D121–D124.
- Ferguson, C., Araújo, D., Faulk, L., Gou, Y., Hamelers, A., Huang, Z., Ide-Smith, M., Levchenko, M., Marinos, N., Nambiar, R. *et al.* (2021) Europe PMC in 2020. *Nucleic Acids Res.*, **49**, D1507–D1514.
- Fiume, M., Cupak, M., Keenan, S., Rambla, J., de la Torre, S., Dyke, S.O.M., Brookes, A.J., Carey, K., Lloyd, D., Goodhand, P. *et al.* (2019) Federated discovery and sharing of genomic data using Beacons. *Nat. Biotechnol.*, **37**, 220–224.
- Woolley, J.P., Kirby, E., Leslie, J., Jeanson, F., Cabili, M.N., Rushton, G., Hazard, J.G., Ladas, V., Veal, C.D., Gibson, S.J. *et al.* (2018) Responsible sharing of biomedical data and biospecimens via the ‘Automatable Discovery and Access Matrix’ (ADA-M). *NPJ Genom. Med.*, **3**, 17.
- Dyke, S.O.M., Philippakis, A.A., Rambla De Argila, J., Paltoo, D.N., Luetkemeier, E.S., Knoppers, B.M., Brookes, A.J., Spalding, J.D., Thompson, M., Roos, M. *et al.* (2016) Consent codes: upholding standard data use conditions. *PLoS Genet.*, **12**, e1005772.
- Courtot, M., Cherubin, L., Faulconbridge, A., Vaughan, D., Green, M., Richardson, D., Harrison, P., Whetzel, P.L., Parkinson, H. and Burdett, T. (2019) BioSamples database: an updated sample metadata hub. *Nucleic Acids Res.*, **47**, D1172–D1178.
- Harrison, P.W., Ahamed, A., Aslam, R., Alako, B.T.F., Burgin, J., Buso, N., Courtot, M., Fan, J., Gupta, D., Haseeb, M. *et al.* (2021) The European Nucleotide Archive in 2020. *Nucleic Acids Res.*, **49**, D82–D85.
- Leivonen, S.-K., Sahlberg, K.K., Mäkelä, R., Due, E.U., Kallioniemi, O., Børresen-Dale, A.-L. and Perälä, M. (2014) High-throughput screens identify microRNAs essential for HER2 positive breast cancer cell growth. *Mol. Oncol.*, **8**, 93–104.
- Harrow, J., Drysdale, R., Smith, A., Repo, S., Lanfear, J. and Blomberg, N. (2021) ELIXIR: providing a sustainable infrastructure

- for life science data at European Scale. *Bioinformatics*, <https://doi.org/10.1093/bioinformatics/btab481>.
16. Linden,M., Prochazka,M., Lappalainen,I., Bucik,D., Vyskocil,P., Kuba,M., Silén,S., Belmann,P., Sczyrba,A., Newhouse,S. *et al.* (2018) Common ELIXIR Service for Researcher Authentication and Authorisation. *F1000Res.*, 7, 1199.
  17. Harrow,J., Hancock,J., ELIXIR-EXCELERATE,Community and Blomberg,N. (2021) ELIXIR-EXCELERATE: establishing Europe's data infrastructure for the life science research of the future. *EMBO J.*, 40, e107409.
  18. Kelleher,J., Lin,M., Albach,C.H., Birney,E., Davies,R., Gourtovaia,M., Glazer,D., Gonzalez,C.Y., Jackson,D.K., Kemp,A. *et al.* (2019) htsget: a protocol for securely streaming genomic data. *Bioinformatics*, 35, 119–121.
  19. Robinson,J.T., Thorvaldsdóttir,H., Turner,D. and Mesirov,J.P. (2020) igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). bioRxiv doi: <https://doi.org/10.1101/2020.05.03.075499>, 05 May 2020, preprint: not peer reviewed.
  20. Holub,P., Swertz,M., Reihls,R., van Enckevort,D., Müller,H. and Litton,J.-E. (2016) BBMRI-ERIC directory: 515 biobanks with over 60 million biological samples. *Biopreserv. Biobank.*, 14, 559–562.













## Chapter 2

Rambla, J., Baudis, M., Ariosa, R., Beck, T., Fromont, L. A., Navarro, A., Paloots, R., Rueda, M., Saunders, G., Singh, B., Spalding, J. D., Törnroos, J., Vasallo, C., Veal, C. D., & Brookes, A. J. (2022). [Beacon v2 and Beacon networks: A “lingua franca” for federated data discovery in biomedical genomics, and beyond](#). *Human Mutation*, 43, 791– 799.





# Beacon v2 and Beacon networks: A “lingua franca” for federated data discovery in biomedical genomics, and beyond

Jordi Rambla<sup>1,2</sup>  | Michael Baudis<sup>3</sup>  | Roberto Ariosa<sup>1</sup>  | Tim Beck<sup>4</sup> |  
Lauren A. Fromont<sup>1</sup> | Arcadi Navarro<sup>1,5,6,7</sup>  | Rahel Paloots<sup>3</sup>  |  
Manuel Rueda<sup>1</sup>  | Gary Saunders<sup>8</sup>  | Babita Singh<sup>1</sup>  | John D. Spalding<sup>9</sup> |  
Juha Törnroos<sup>9</sup>  | Claudia Vasallo<sup>1</sup>  | Colin D. Veal<sup>4</sup>  | Anthony J. Brookes<sup>4</sup> 

<sup>1</sup>Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain

<sup>2</sup>Department of Experimental and Health Sciences, Universitat Pompeu Fabra (UPF), PRBB, Barcelona, Spain

<sup>3</sup>Department of Molecular Life Sciences, University of Zurich and Swiss Institute of Bioinformatics, Zurich, Switzerland

<sup>4</sup>Department of Genetics & Genome Biology, University of Leicester, Leicester, UK

<sup>5</sup>Department of Experimental and Health Sciences, IBE, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, PRBB, Barcelona, Spain

<sup>6</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA), Universitat Pompeu Fabra, Barcelona, Spain

<sup>7</sup>Barcelona Beta Brain Research Center, Pasqual Maragall Foundation, Barcelona, Spain

<sup>8</sup>European Infrastructure for Translational Medicine, EATRIS, Amsterdam, The Netherlands

<sup>9</sup>ELIXIR Finland; CSC - IT Center for Science Ltd, Espoo, Finland

## Correspondence

Jordi Rambla, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain.

Email: [jordi.rambla@crgeu](mailto:jordi.rambla@crgeu)

Michael Baudis, University of Zurich and Swiss Institute of Bioinformatics, Zurich, Switzerland.

Email: [michael.baudis@uzh.ch](mailto:michael.baudis@uzh.ch)

## Funding information

Health Data Research UK,

Grant/Award Number: MR/S003703/1;

European Commission, Grant/Award Number:

EJP RD COFUND-EJP #825575; LaCaixa

Foundation, Grant/Award Number: 004745/

008034; ELIXIR Europe

## Abstract

Beacon is a basic data discovery protocol issued by the Global Alliance for Genomics and Health (GA4GH). The main goal addressed by version 1 of the Beacon protocol was to test the feasibility of broadly sharing human genomic data, through providing simple “yes” or “no” responses to queries about the presence of a given variant in datasets hosted by Beacon providers. The popularity of this concept has fostered the design of a version 2, that better serves real-world requirements and addresses the needs of clinical genomics research and healthcare, as assessed by several contributing projects and organizations. Particularly, rare disease genetics and cancer research will benefit from new case level and genomic variant level requests and the enabling of richer phenotype and clinical queries as well as support for fuzzy searches. Beacon is designed as a “lingua franca” to bridge data collections hosted in software solutions with different and rich interfaces. Beacon version 2 works alongside popular standards like Phenopackets, OMOP, or FHIR, allowing implementing consortia to return matches in beacon responses and provide a handover to their preferred data exchange format. The protocol is being explored by other research domains and is being tested in several international projects.

## KEYWORDS

Beacon, clinical genomics, data discovery, data sharing, GA4GH, REST API

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Human Mutation* published by Wiley Periodicals LLC.

## 1 | INTRODUCTION

The Global Alliance for Genomics and Health (GA4GH) (e.g., Rehm et al., 2021) was created in 2013 on the core mission to create a federated ecosystem for sharing genomic data and associated clinical information within a human rights framework. At its foundational meeting, the concept of a genomics “Beacon” was presented by Jim Ostell, as a means to engage and connect genomic data providers and developers as well as researchers with interest to access genomic variation data. The concept of this Beacon system was purposefully simple: Design a data access API, which allows users to query genomic data collections for the existence of a specific genomic variation and responds with a “Yes” or “No” answer. The name “Beacon” referred to the hope that such a system would be simple enough to engage willing participants, thereby lighting up the so far dark landscape of genomic data sharing.

While kept simple to encourage broad participation, the Beacon concept was aimed to trigger potential issues—for example, related to institutional policies or general regulatory issues regarding genomic information—but also to demonstrate the power of such a simple data sharing concept especially if implemented through a federated model, distributing Beacon queries to a large number of international nodes and providing an aggregation of the individual responses. While it was clear that such a Beacon network could become more useful through complex queries and richer responses, such extensions were proposed for a “version 2,” following the successful establishment of a working implementation of the original concept.

With the Beacon Project becoming one of the original GA4GH Driver Projects, it was enthusiastically adopted by members of the GA4GH developer community and genomic resource providers alike. By 2016, more than 35 organizations from all over the world had “beaconized” over 90 genomics datasets, many of which were connected to the network aggregator provided by DNASTack ([beacon-network.org](https://beacon-network.org)). At this point ELIXIR - the European bioinformatics infrastructure organization - joined the further development of the community project toward a standard specification, with improved usability and the goal of future use throughout biomedical genomics.

In 2018, the Beacon v1.0 protocol was accepted as official GA4GH standard (Fiume et al., 2019) following a formal review process. While this version and its updates introduced some improvements over the earlier editions, such as limited support for structural variant queries, some quantitative responses as well as the “handover” to external protocols, overall Beacon v1 stayed with the original “variant query and aggregate response” concept. However, at this point, it had become clear that further expansions of the protocol such as requested—especially for clinical applications in rare diseases and cancer genomics—required a re-design of the Beacon protocol to serve a wide range of use cases and leading to initiation of the “Beacon v2” design process.

### 1.1 | Designing a “clinical Beacon”

The initial concept of the Beacon protocol of returning a simple boolean response focuses more toward technical implementers and scientific researchers than clinicians. For broader use in clinical settings, each allele or mutation-specific query should ideally offer options to query and retrieve associated phenotypic data, metadata (e.g., age at disease onset, genotypic sex), associated diagnoses, therapeutic interventions, pedigree information, and so on. The success of the Beacon v1 concept and its enthusiastic adoption by genomics and rare disease communities has provided a strong argument to expand its usability toward a more general use in healthcare environments, while building on its conceptual simplicity.

With these objectives in mind, the GA4GH Beacon group engaged with GA4GH Driver Projects and with ELIXIR partners to identify the consensus requirements for the next generation Beacon. The following Driver projects were interviewed: Autism Speaks (<https://www.autismspeaks.org>), BRCA Exchange (Cline et al., 2018), CanDIG (Rehm et al., 2021), EGA (Freeberg et al., 2022)/ENA (Harrison et al., 2021)/EVA (Cezard et al., 2022), EuCanCan (<https://eucan.com>), European Joint Programme on Rare Diseases (EJP-RD, <https://www.ejprarediseases.org>), H3Africa (e.g., Mulder et al., 2018), GEM Japan ([https://www.amed.go.jp/en/aboutus/collaboration/ga4gh\\_gem\\_japan.html#anc-2](https://www.amed.go.jp/en/aboutus/collaboration/ga4gh_gem_japan.html#anc-2)), Genomics England (Koepli et al., 2015), Matchmaker Exchange (Pilipakis et al., 2015), Swiss Variant Interpretation Platform (SVIP, <https://svip.ch/SwissPersonalizedHealthNetwork> (SPHN, <https://sphn.ch/fr/en/home/>), and Variant Interpretation for Cancer Consortium (VICC). Some ELIXIR partners and communities such as Café Variome (Lancaster et al., 2015), hCNV community (<https://elixir-europe.org/communities/hcnv>), Fundación Progreso y Salud (<https://www.clinbioinfospa.es/>), RD-Connect (Thompson et al., 2014), CINECA (<https://www.cineca-project.eu>), and DisGeNET (Piñero et al., 2020), were also interviewed. In addition, members of the Beacon team detailed (a) requirements for specific diagnoses, including rare diseases, (b) a support for the clinical center to build their own Beacon, and (c) a procedure to gather them under clinical network of Beacon installations, via tight collaborations with hospitals in Catalunya, Spain, as well as Cancer Core Europe (<https://cancercoreeurope.eu/>) and Health-RI (<https://www.health-ri.nl/>), Netherlands. Finally, to address specific aspects in the Beacon development, working groups (“Scouts”) are working regularly on different aspects of the Beacon protocol such as security, filters, genomic variants, and protocol documentation.

### 1.2 | Requirements for a clinical Beacon

The health sector is increasingly seeking to use genetic/genomic tests as an integral part of the diagnostic process or in the selection of therapeutic procedures. Typical genomic data generation and sharing in healthcare includes five types of roles: the patient, the clinicians, the genetic analysts, intramural contributing researchers, and

external partners involved in associated research projects or technical aspects of diagnosis and data handling. Each contributor has a set of needs regarding data discovery. Citizens are both at the start and the end (as ultimate beneficiaries) of the data cycle: subject to different rules in different countries, they can consent to the use of their data for healthcare or secondary research use as long as their privacy and identity are protected. The general data protection regulation (GDPR) adds additional responsibilities on the operator of a Beacon to protect the privacy and rights of individuals whose data exists within a beacon, therefore data security (see section on security aspects) is therefore a key component to discovery tools such as Beacon. Finally, the new Beacon model should provide context for the genomic variant finding, including information about the biosample and molecular analysis procedures as well as observations and measurements describing the phenotypic state of individuals. Local researchers are responsible for structuring the data so it can be queried. As it is usually a manual process, different studies from the same institution often use different tools and select different information, making cross-querying or reuse of data difficult. For this reason, the Beacon team is committed to train internal researchers on data structuring for Beacon. External partners in large-scale European projects (e.g., CINECA or B1MG) are proponents of the Beacon v2. Understanding the partner needs has driven the addition of new features like cohorts, which are of high relevance for the scientific community. Projects also provide a platform to enable new Beacon features to be more visible (Fromont, 2021), and facilitate the use of these additions via training events. The ELIXIR Beacon project aims to disseminate Beacon and the Beacon network. In 2021, nine, preliminary version 2 and newly developed, Beacons were implemented across the ELIXIR Nodes in the ELIXIR Beacon network prototype (<https://beacon-network.elixir-europe.org>). Once the Beacon v2 specification stabilized, it was straightforward to write parsers to connect and translate to the various backends used in these implementations.

### 1.3 | Rare diseases use case

Even though many candidate variants can be identified in rare disease patients, a reliable genetic diagnosis cannot be achieved in at least 50% of cases (Zurek et al., 2021). Pinpointing causal genetic variants in these cases can be greatly assisted by increasing sample size(s) and finding other patients with similar phenotype profiles: an approach called "matchmaking."

"Matchmaking" cases—or patients with particular sets of disease phenotypes—are made difficult by the huge phenotypic and genotypic diversity of rare diseases patients, and the correspondingly large and variable way(s) in which rare diseases related data(sets) are collected into registries, biobanks, and sample catalogs. Each data set typically has its own access rules and gateways, making it difficult to connect data and deduce meaningful insights across resources for in depth genome analysis.

Previous projects such as matchmaker exchange (MME—Philippakis et al., 2015) and RD-connect genome-phenome analysis platform (GPAP—Matalonga et al., 2021) were successful in different ways in tackling these problems. MME is a small network of large databases that can interoperate to "matchmake" patients. However, the process requires the supply and transmission of patient profiles, with limited control over how a match is defined. This model comprises many data sharing policies, and therefore dissuades many potential users. The RD-connect GPAP approach utilizes a combination of phenotype, genotype, and biobank data to allow users to find subjects of interest. However, it is based on a multi-site centralized platform requiring the submission of data to the GPAP environment. Both systems also require authorized access, which further limits the availability of these data discovery solutions. Additionally, many smaller rare disease patient registries exist, along with sample catalogs and biobanks that operate their own systems. There is limited interoperability between any of these and the current larger solutions.

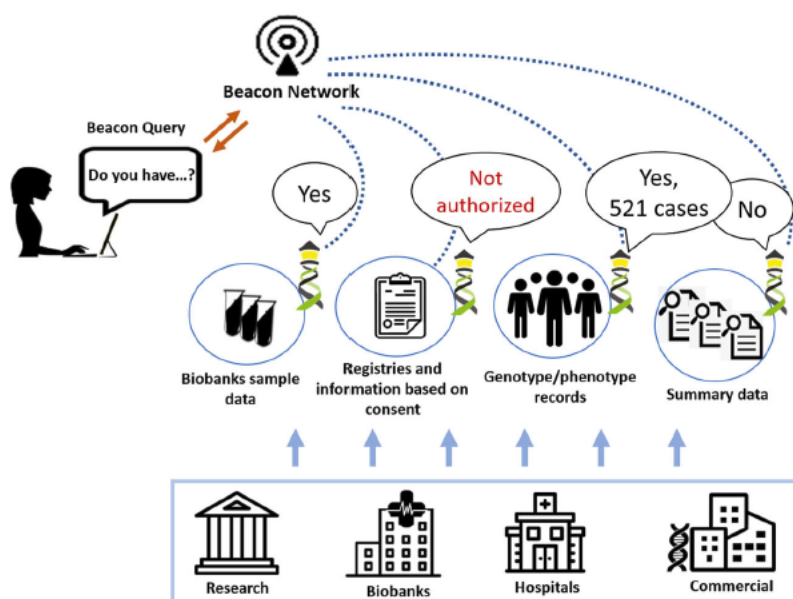
To improve on the above there needs to be a standardized way to interconnect diverse resources to provide safe, federated, flexible but powerful discovery queries, to (see Figure 1):

- Find a suitable registry based on summary data—response could be a yes/no;
- Find a suitable registry based on genotypes/phenotypes on a record level—response could be yes/no, counts;
- Find a biobank based on the sample data—response could be yes/no or counts of samples;
- Find a collection of subjects across many registries—response could be counts per registry;
- Find registries and biobanks based on consent and use conditions that apply to the assets in those resources.

The Beacon v2 API can help with these goals, as it is separated from the software and data models employed at any potential query target. The API provides a standardized way to send a query itself, and receive a standardized response containing yes/no, counts or data. The Beacon 2 models provide a description and attributes of the target type, an initial set of models covering query targets such as "individual" or "data set" are part of the API, however, this is extensible and new models can be made that fit a particular use case.

### 1.4 | Scope of Beacon v2 model

The original Beacon protocol did not specify an explicit data model but rather limited itself to reference genome mapped genome variations and simple, boolean responses. In contrast, and driven by the requirements detailed above, the Beacon v2 protocol allows for an extensible data model on top of its flexible framework (see our website <https://beacon-project.io/> for details). The Beacon v2 provides default support for a data model serving the needs of



**FIGURE 1** Beacon queries could be sent to Beacon instances directly or via Beacon networks. The response could be yes/no, counts or details if the user is properly authorized

biomedical genomics but also accommodating simple boolean responses.

The complete v2 data model (Figure 2) implements default entities such as "individual," "biosample," "observed genome variation," and "variant information" and their logical relations, as well as additional technical concepts. Here, the Beacon model follows common concepts such as those established through the GA4GH Data Working Group and compatible with, for example, the Phenopackets standard (Jacobsen et al., 2021) and the variant representation standard (VRS; Wagner et al., 2021). For example, Beacons can support queries combining phenotype parameters of an individual with genomic variation parameters ("are there individuals with phenotype X and variant Y"); or retrieve information from cancer samples of a certain histology that contain a mutation in a specified oncogene. Additionally, the "Variant Annotation" schema type can provide rich information about matched variants, but also can serve as the core of genomic knowledge resources for aggregated data about clinically actionable variants. Also, the Beacon v2 model supports the use of grouping concepts such as "data set" and "cohort," for example, to query data particular to a certain resource within a larger Beacon instance or a set of individuals from a given study cohort.

In summary, where Beacon v1 by design was limited to positional requests for genomic variations in specified datasets, v2 leverages common biomedical entity models for query and response. While default models and examples support the simple alignment across implementations and thereby empower federated Beacon queries, the extensibility of the model allows to tailor specific solutions for example, in the healthcare context.

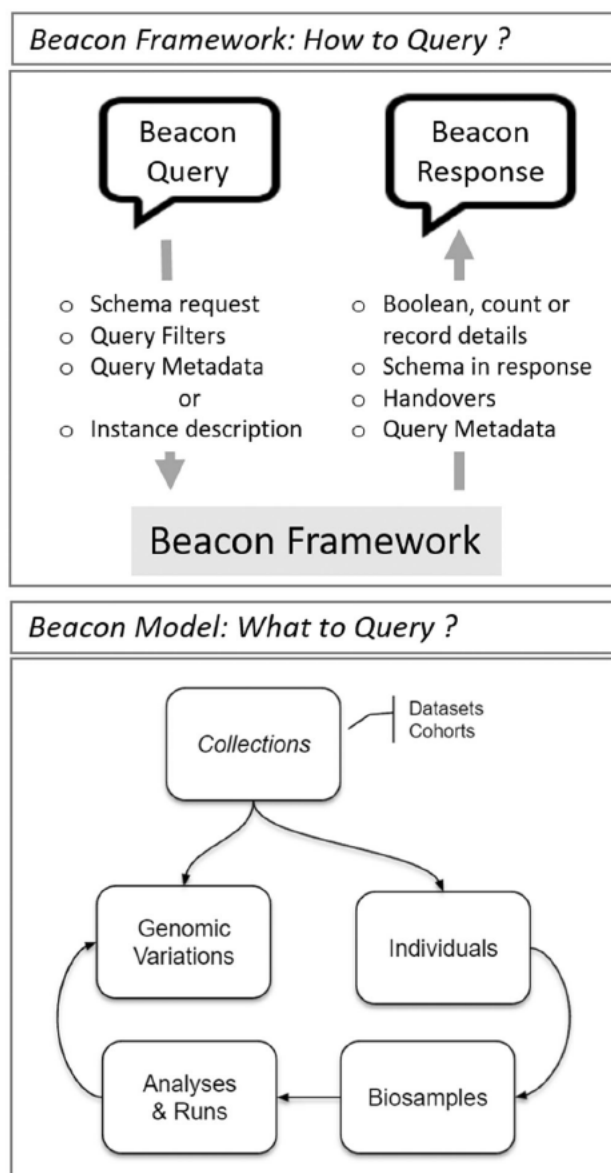
## 1.5 | Implementing Beacon over existing solutions

Beacon v2 is organized in two main blocks: the Beacon Framework and the Beacon Models. The Framework is agnostic to the knowledge

domain and includes the features related to Beacon instance description (metadata), query requests, query responses, filters, handovers, and so on. The Beacon Models describe the domain entities and the relationships between them. To obtain interoperability between Beacon instances, Beacon v2 includes a recommended model for clinical genomics diagnoses and research, as it is described in a previous section. Separating the Framework from the Model allows other disciplines to adopt the Beacon concept without departing from the standard itself and without any servitude of implementing a model that is not relevant to their domain.

We can also refer to the architecture of the Beacon design and to the architecture of the Beacon implementations or instances. Implementations of the previous version of Beacon (v1) existed in two different flavors: (1) solutions developed from scratch, that is by "beaconizing" a pure variant collection without existing data interface; and (2) Beacons created on top of existing solutions, for example: Cafe Variome (Lancaster et al., 2015), OpenCGA (<http://opencb.org/>), Progenetix (Huang et al., 2021), or RD-Connect GPAP. Beacon v1 was designed to be minimalistic, with a minimal implementation effort. Given that Beacon v2 has a broader scope and the effort to build a solution from scratch is significantly higher, Beacon v2 has been designed as a REST API *façade* for existing solutions. These could be professional and popular solutions or homemade, basic ones; in any case, the design principle was to make the implementation simple and as less intrusive as possible, lowering the barrier for implementation for existing solutions.

The Beacon designers envisioned a scenario in which genomics solution providers would be able to implement a Beacon on top of existing solutions, with a minimum of development effort and resources. To not be intrusive, the Beacon suggests an harmonization approach at query time. For example, a given backend could store the "male gender" concept as a "M" in the column "gender" of the "Person" table, while another Beacon could represent it as "male" in



**FIGURE 2** Beacon v2 is composed of two parts: the Beacon Framework and the Beacon Model. The former describes the request and response protocol, the latter describes the common entities in the clinical research domain although other models could be used

the column "sex" of the "cohort-members" table. Beacon v2 suggests using an ontology term for "male," for example, PATO:0000384, at query time. The Beacon instances will receive that term as part of the query and they would need to translate it, map it, to the corresponding internal representation to solve the query. The results are mapped to the model suggested by Beacon, obtaining, therefore, an harmonization at query and at response time. Importantly, such concepts can be implemented gradually, without the need to change the underlying data model ab initio.

Many Beacon instances will be part of networks, although a Beacon can be instantiated as a stand-alone solution. The Beacon

design includes several features aimed to be consumed by Beacon network aggregators. For example, a Beacon endpoint declares which entities are implemented in that particular instance, which are the ontology terms supported or the URL endpoints where different elements could be found.

In summary, Beacon v2 has been designed to be domain agnostic, but suggesting a model for clinical genomics as well as to act as an interoperability interface on top of existing resources, to enable their utilization as part of networked and federated data discovery solutions.

### 1.6 | How does it relate to other clinical research standards?

An ecosystem of clinical data standards enables healthcare and research systems to interoperate to unambiguously describe, store, exchange, and analyze health data on an international scale. Beacon works alongside established standards to provide a flexible data discovery solution with optional clinical applications. Beacon supports the use of semantic standards used to describe clinical concepts and it is compatible with open syntactic standards for harmonized clinical data storage and communication.

Controlled vocabularies and ontologies standardize the labeling of concepts for various biomedical domains, for example, SNOMED for diagnoses, HPO for phenotypic abnormalities, and LOINC for laboratory results. Beacon v2 "filters" support the discovery of patients and biosamples using ontology terms. Beacon does not limit which systems may participate by being agnostic to the semantic standards used by a data source. Beacons which use controlled vocabularies and ontologies declare this by providing an informational filters endpoint that is defined in the Beacon framework. Beacon reuses Phenopackets v2 specifications for describing ontologies and representing ontology classes. Phenopackets is a GA4GH approved standard for sharing disease and phenotypic information and has been adopted by the rare diseases research community for consistent characterization and representation of disease manifestations (Rubinstein et al., 2020). Beacon individual and biosample schemas compatible with Phenopackets v2 architecture are available to help streamline implementations of Beacon discovery when Phenopackets are used by a data provider.

The openEHR, HL7 fast healthcare interoperability resources (FHIR), and Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) standards are concerned with the storage of clinical data for healthcare, clinical data exchange, and the storage of clinical data for research, respectively. The Beacon architecture supports patient discovery across these three standards. The Beacon "individuals" model can be tailored and mapped to the components of these standards that store patient health information.

The openEHR specification defines the structure and function of electronic health record (EHR) systems. *Archetypes* are core concepts of the openEHR specification and are comprehensive, machine-interpretable, and reusable discrete models of health information,

such as observations of body mass index and arterial blood pressure (Bosca et al., 2015). The complementary FHIR standard is used by the healthcare industry to exchange EHR data. FHIR uses components called *resources* to access and perform operations on patient data. The resources define generic common health care concepts with clearly defined scope such as observation and condition (Ayaz et al., 2021). Dedicated Beacon model schemas can be used to accommodate implementations of openEHR archetypes and serializations of FHIR resources. Patients can be discovered by filtering on the patient characteristics captured by the clinical coding used by implementations, with the Beacon response accommodating a handover to the native data exchange format.

The OMOP CDM supports research by harmonizing healthcare data from diverse sources in a consistent and standardized way (Hripcsak et al., 2015). The CDM can be adapted to accommodate specialized medical research use cases, for example, the storage and analysis of rare disease patient data are enabled by including dedicated rare disease terminologies within the CDM (Zoch et al., 2021). The OMOP CDM consists of a collection of *table schemas* where each schema represents a specific OMOP domain such as observation and measurement. A community effort organized through Biohackathon Europe (described below) has mapped the Beacon model to OMOP table schemas to demonstrate the Beacon discovery of patients using OMOP vocabularies.

The annual Biohackathon Europe event hosted by ELIXIR brings together bioinformaticians, software engineers, data providers, and consumers to work on life sciences data challenges. During Beacon-focused projects at the event over consecutive years (in 2020 and 2021), we aimed to demonstrate the reality of Beacon discovery alongside existing clinical data standards. In 2020 we devised a proof of concept (POC) Beacon implementation to enable patient discovery of individuals described in the OMOP CDM. Serializations of SNOMED-coded synthetic FHIR resources were transformed and loaded into the OMOP CDM, and the POC Beacon was mapped to OMOP table schemas to enable individuals to be discovered using SNOMED ontology filters. This capability has been extended in 2021 to discover synthetic patients from EHRs and our ambition is to deliver a POC Beacon to demonstrate and support Beacon adoption alongside, and complementary to, established health information systems implementing these open standards.

## 1.7 | Security aspects

The Beacon uses a 3-tiered access model - anonymous, registered, and controlled access. A Beacon that supports anonymous access responds to queries irrespective of the source of the query. For a Beacon to respond to a query at the registered tier, the user must identify themselves to the Beacon, for example by using an ELIXIR identity. ELIXIR identities are controlled by the ELIXIR Authorization and Authentication Infrastructure (ELIXIR AAI; <https://elixir-europe.org/services/compute/aaai>). The ELIXIR Authentication and Authorization Infrastructure (AAI) enables researchers to use their home

organization credentials, community, or commercial identities (e.g., ORCID, LinkedIn) to sign in and access data and the services that they need. For a Beacon to respond to a controlled-access query, the user must have applied for and been granted access to, the Beacon (or data derived from one or more individuals within the Beacon) before sending the query. Note that a Beacon may contain datasets (or collections of individuals) whose data is only accessible at specified tiers within the Beacon. This tiered access model allows the owner of a Beacon to determine which responses are returned to whom depending on the query itself and the user who is making the request, for example, to ensure the response respects the consent or legal basis under which the data were collected, or to support requirements in different legal jurisdictions, for example, the data minimization or purpose limitation principles within GDPR (European Parliament and Council, 2016). As an example, the ELIXIR Beacon Network supports Beacons which respond at different tiers, for example, only Beacons which have a response for anonymous queries need respond to an anonymous request. A security document (ELIXIR, 2021) has been written to describe security best practice for users interested in deploying or running a Beacon or users who govern data hosted within a Beacon, and the requirements for adding the Beacon to the ELIXIR Beacon Network. Additionally, as Beacon implements a GA4GH approved standard it must go through the GA4GH approval process, which means the standard must be approved by both the Regulatory and Ethics, and Data Security foundational workstreams. As the Beacon standard extends in V2 toward supporting phenotype and range queries, the tiered access model becomes more important to ensure the Beacon response is appropriate to the underlying data.

All the measures described should allow a Beacon administrator to configure the access to the hosted data according to their sensitivity, ranging from total openness for allele frequencies in population studies to fully protected in particular diseases, therefore minimizing the risk of undesired re-identification (Bernier et al., 2022).

As a Beacon is designed to support data discoverability of controlled access datasets, it is recommended that synthetic or artificial data is used for testing and initial deployment of Beacon instances. The use of synthetic data for testing is important in that it ensures that the full functionality of a Beacon can be tested and/or demonstrated without risk of exposing data from individuals. In addition to testing or demonstrating a deployment, synthetic data should be used for development, for example adding new features. An example data set that contains chromosome specific vcf files is hosted at EGA under data set accession EGAD00001006673. This data set is accessible via EGA's test user and does not require obtaining separate credentials.

## 1.8 | Toward an "Internet of genomics"?

Since the inception of Beacon v2 idea in October 2018, many projects and initiatives have shown interest in the Beacon concept and its possibilities. The scope of projects is broad: resource discovery

(e.g., biobanks or registries), cohort discovery and description, proteomics, viral genomics (e.g., SARS-CoV-2 in Viral Beacon - <https://covid19beacon.crg.eu/>), plants, and so on. Some flagship projects like the European Joint Program on Rare Diseases (EJP-RD), the 1+ Million Genomes initiative (European Commission, 2021) and its supporting Beyond One Million Genomes project, or Horizon 2020 funded projects like BY-COVID, CINECA, CONVERGE, or EUCANCan.

The goals of these projects are diverse, ranging from sharing data in domains where there are no established standard solutions, to allow total control on the granularity of sensitive data sharing (from boolean answers to complete details, depending on the level of trust and if the audience is intramural or external). All of them share the vision that the future of sensitive data is federated discovery, query, and analysis and that only pragmatic approaches would make that possible. These pragmatic approaches translate into control, flexibility, simplicity, and capability to deal with heterogeneity. All of them are attributes that Beacon v2 has included in their design, therefore, these projects have looked at it as a solution to observe. Several of the mentioned projects have implemented Beacon v1 instances and tested the preliminary versions of Beacon v2.

Beacon v2 is designed to be an interface on top of existing solutions, however, the clinical genomics research facilities are, in many cases, facing a more basic issue: the lack of a solution to manage the genomic data and its relationship with the clinical care associated data (phenotypes and clinical journey). This need has led to the concept of the EGA Community Platform. The European Genome-phenome Archive (EGA) Community Platform is a proposal to combine existing solutions for genomic and metadata data management, with existing analysis solutions, all topped with a Beacon v2 interface. The clinical research facility could choose among the already tested solutions or add any of their preference, the only requirement is that it must implement a Beacon v2 interface. The aim of this concept is to facilitate the reuse of existing data, initially inside the institution, while paving the way for sharing with the community the generated knowledge in a safe and controlled way.

Beacons lighted independently, through EGA Community packages or by any other means, could be integrated in Beacon networks. Beacon networks could be internal to a hospital campus, a consortium, a region or country, or be organized by topic, one example being the ELIXIR Beacon Network (<https://beacon-network.elixir-europe.org/>), whose goal is to trigger the discovery of Beacons and to showcase the utility of such networks.

#### ACKNOWLEDGMENTS

The Beacon v2 project has received continued support from both GA4GH and ELIXIR. As a prominent GA4GH Discovery Work Stream product, the Beacon protocol and associated team has benefitted from administrative support in product submission and meeting management. We would like to acknowledge the support and contribution from the GA4GH Driver projects and other partners mentioned in the text. And last, but not least all the volunteers of the

Beacon Scouts and the enthusiastic Beacon v2 implementers. This study was funded by ELIXIR, the research infrastructure for life-science data and also by LaCaixa Foundation (grant number 004745/008034). Tim Beck was supported by a UKRI Innovation Fellowship at Health Data Research UK (MR/S003703/1). Anthony J. Brookes and Jordi Rambla were supported, in part, by the European Union's Horizon 2020 research and innovation program under the EJP RD COFUND-EJP #825575. Michael Baudis acknowledges funding under the BioMedIT Network project of Swiss Institute of Bioinformatics (SIB) and Swiss Personalized Health Network (SPHN).

#### CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

#### ORCID

Jordi Rambla  <http://orcid.org/0000-0001-9091-257X>  
 Michael Baudis  <https://orcid.org/0000-0002-9903-4248>  
 Roberto Ariosa  <https://orcid.org/0000-0001-8348-2524>  
 Arcadi Navarro  <https://orcid.org/0000-0003-2162-8246>  
 Rahel Paloots  <https://orcid.org/0000-0003-1239-1689>  
 Manuel Rueda  <https://orcid.org/0000-0001-9280-058X>  
 Gary Saunders  <https://orcid.org/0000-0002-7468-0008>  
 Babita Singh  <https://orcid.org/0000-0002-7989-9084>  
 Juha Tömroos  <https://orcid.org/0000-0001-9216-0455>  
 Claudia Vasallo  <https://orcid.org/0000-0002-0043-0882>  
 Colin D. Veal  <https://orcid.org/0000-0002-9840-2512>  
 Anthony J. Brookes  <https://orcid.org/0000-0001-8686-0017>

#### REFERENCES

- Ayaz, M., Pasha, M. F., Alzahrani, M. Y., Budiarto, R., & Stiawan, D. (2021). The fast health interoperability resources (FHIR) standard: Systematic literature review of implementations, applications, challenges and opportunities. *JMIR Medical Informatics*, 9(7), e21929. <https://doi.org/10.2196/21929>
- Bernier, A., Liu, H., & Knoppers, B. M. (2022). Computational tools for genomic data de-identification: Facilitating data protection law compliance. *Nature Communications*, 13, 391. <https://doi.org/10.1038/s41467-021-27890-5>
- Bosca, D., Moner, D., Maldonado, J. A., & Robles, M. (2015). Combining archetypes with fast health interoperability resources in future-proof health information systems. *Studies in Health Technology and Informatics*, 210, 180–184.
- Cezard, T., Cunningham, F., Hunt, S. E., Koylass, B., Kumar, N., Saunders, G., Shen, A., Silva, A. F., Tsukanov, K., Venkataraman, S., Flicek, P., Parkinson, H., & Keane, T. M. (2022). The European Variation Archive: A FAIR resource of genomic variation for all species. *Nucleic Acids Research*, 50, gkab960. <https://doi.org/10.1093/nar/gkab960>
- Cline, M. S., Liao, R. G., Parsons, M. T., Paten, B., Alquaddoomi, F., Antoniou, A., Baxter, S., Brody, L., Cook-Deegan, R., Coffin, A., Couch, F. J., Craft, B., Currie, R., Dlott, C. C., Dolman, L., den Dunnen, J. T., Dyke, S., Domchek, S. M., Easton, D., ... Spurdle, A. B. (2018). BRCA Challenge: BRCA Exchange as a global resource for variants in BRCA1 and BRCA2. *PLoS Genetics*, 14, e1007752. <https://doi.org/10.1371/journal.pgen.1007752>

- ELIXIR. (2021). *ELIXIR Beacon 2019–21 Deliverable D3.3*. [https://docs.google.com/document/d/1q7XuUB-24A\\_DogWT1AVrvkp\\_qHWWtbblCcxokHup\\_tts/edit](https://docs.google.com/document/d/1q7XuUB-24A_DogWT1AVrvkp_qHWWtbblCcxokHup_tts/edit)
- European Commission. (2021). *1+ Million genomes. Shaping Europe's digital future*. <https://digital-strategy.ec.europa.eu/en/policies/1-million-genomes>
- European Parliament and Council. (2016). *General Data Protection Regulations (GDPR): Regulation (EU) 2016/679; Article 5*. <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679%26from=EN#d1e1807-1-1>
- Fiume, M., Cupak, M., Keenan, S., Rambla, J., de la Torre, S., Dyke, S., Brookes, A. J., Carey, K., Lloyd, D., Goodhand, P., Haeussler, M., Baudis, M., Stockinger, H., Dolman, L., Lappalainen, I., Törnroos, J., Linden, M., Spalding, J. D., Ur-Rehman, S., ... Scollen, S. (2019). Federated discovery and sharing of genomic data using Beacons. *Nature Biotechnology*, 37(3), 220–224. <https://doi.org/10.1038/s41587-019-0046-x>
- Freeberg, M. A., Fromont, L. A., D'altri, T., Romero, A. F., Ciges, J. I., Jene, A., Kerry, G., Moldes, M., Ariosa, R., Bahena, S., Barrowdale, D., Barbero, M. C., Fernandez-Orth, D., Garcia-Linares, C., Garcia-Rios, E., Haziza, F., Juhasz, B., Llobet, O. M., Milla, G., ... Rambla, J. (2022). The European Genome-phenome Archive in 2021. *Nucleic Acids Research*, 50, gkab1059. <https://doi.org/10.1093/nar/gkab1059>
- Fromont, L. A. (2021). *Beacon cohorts: A model for cohort discovery in CINECA and beyond—CINECA—Common Infrastructure for National Cohorts in Europe, Canada, and Africa*. CINECA. <https://www.cineca-project.eu/blog-all/beacon-cohorts-a-model-for-cohort-discovery-in-cineca-and-beyond>
- Harrison, P. W., Ahamed, A., Aslam, R., Alako, B. T. F., Burgin, J., Buso, N., Courtot, M., Fan, J., Gupta, D., Haseeb, M., Holt, S., Ibrahim, T., Ivanov, E., Jayathilaka, S., Balavenkataraman Kadhivelu, V., Kumar, M., Lopez, R., Kay, S., Leinonen, R., ... Cochrane, G. (2021). The European Nucleotide Archive in 2020. *Nucleic Acids Research*, 49(D1), D82–D85. <https://doi.org/10.1093/nar/gkaa1028>
- Hripscak, G., Duke, J. D., Shah, N. H., Reich, C. G., Huser, V., Schuemie, M. J., & Ryan, P. B. (2015). Observational health data sciences and informatics (OHDSI): Opportunities for observational researchers. *Studies in Health Technology and Informatics*, 216, 574–578.
- Huang, Q., Carrio-Cordo, P., Gao, B., Paloots, R., & Baudis, M. (2021). The Progenetix oncogenomic resource in 2021. *Database: The Journal of Biological Databases and Curation*, 2021, baab043. <https://doi.org/10.1093/database/baab043>
- Jacobsen, J. O. B., Baudis, M., Baynam, G. S., Beckmann, J. S., Beltran, S., Callahan, T. J., Chute, C. G., Courtot, M., Danis, D., Elemento, O., Freimuth, R. R., Gargano, M. A., Groza, T., Hamosh, A., Harris, N. L., Kaliyaperumal, R., Khalifa, A., Krawitz, P. M., Köhler, S., ... Robinson, P. N. (2021). The GA4GH Phenopacket schema: A computable representation of clinical data for precision medicine. *medRxiv*, 11(27), 21266944. <https://doi.org/10.1101/2021.11.27.21266944>
- Koepfli, K. P., Paten, B., Genome 10K Community of Scientists O'Brien, S. J. (2015). The Genome 10K Project: A way forward. *Annual Review of Animal Biosciences*, 3, 57–111. <https://doi.org/10.1146/annurev-animal-090414-014900>
- Lancaster, O., Beck, T., Atlan, D., Swertz, M., Thangavelu, D., Veal, C., Dalgleish, R., & Brookes, A. J. (2015). Cafe Variome: General-purpose software for making genotype-phenotype data discoverable in restricted or open access contexts. *Human Mutation*, 36(10), 957–964. <https://doi.org/10.1002/humu.22841>
- Matalonga, L., Hernández-Ferrer, C., Piscia, D., Solve-RD SNV-indel working, g, Schüle, R., Synofzik, M., Töpf, A., Vissers, L., de Voer, R., Solve-Rd, D., Solve-Rd, D., Solve-Rd, D., Solve-Rd, D., Tonda, R., Laurie, S., Fernandez-Callejo, M., Picó, D., Garcia-Linares, C., Papakonstantinou, A., ... Solve-RD, C. (2021). Solving patients with rare diseases through programmatic reanalysis of genome-phenome data. *European Journal of Human Genetics*, 29(9), 1337–1347. <https://doi.org/10.1038/s41431-021-00852-7>
- Mulder, N., Abimiku, A., Adebamowo, S. N., de Vries, J., Matimba, A., Olowoyo, P., Ramsay, M., Skelton, M., & Stein, D. J. (2018). H3Africa: Current perspectives. *Pharmacogenomics and Personalized Medicine*, 11, 59–66. <https://doi.org/10.2147/PGPM.S141546>
- Philippakis, A. A., Azzariti, D. R., Beltran, S., Brookes, A. J., Brownstein, C. A., Brudno, M., Brunner, H. G., Buske, O. J., Carey, K., Doll, C., Dumitriu, S., Dyke, S. O., den Dunnen, J. T., Firth, H. V., Gibbs, R. A., Girdea, M., Gonzalez, M., Haendel, M. A., Hamosh, A., ... Rehm, H. L. (2015). The Matchmaker Exchange: A platform for rare disease gene discovery. *Human Mutation*, 36(10), 915–921. <https://doi.org/10.1002/humu.22858>
- Piñero, J., Ramírez-Anguaita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., & Furlong, L. I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48(D1), D845–D855. <https://doi.org/10.1093/nar/gkz1021>
- Rehm, H. L., Page, A. J. H., Smith, L., Adams, J. B., Alterovitz, G., Babb, L. J., Barkley, M. P., Baudis, M., Beauvais, M., Beck, T., Beckmann, J. S., Beltran, S., Bernick, D., Bernier, A., Bonfield, J. K., Boughtwood, T. F., Bourque, G., Bowers, S. R., Brookes, A. J., ... Rodarmer, K. W. (2021a). CanDIG: Federated network across Canada for multi-omic and health data discovery and analysis. *Cell Genomics*, 1(2), 100033. <https://doi.org/10.1016/j.xgen.2021.100033>
- Rehm, H. L., Page, A. J. H., Smith, L., Adams, J. B., Alterovitz, G., Babb, L. J., Barkley, M. P., Baudis, M., Beauvais, M. J. S., Beck, T., Beckmann, J. S., Beltran, S., Bernick, D., Bernier, A., Bonfield, J. K., Boughtwood, T. F., Bourque, G., Bowers, S. R., Brookes, A. J., ... Rodarmer, K. W. (2021b). GA4GH: International policies and standards for data sharing across genomic research and healthcare. *Cell Genomics*, 1(2), 100029. <https://doi.org/10.1016/j.xgen.2021.100029>
- Rubinstein, Y. R., Robinson, P. N., Gahl, W. A., Avillach, P., Baynam, G., Cederroth, H., Goodwin, R. M., Groft, S. C., Hansson, M. G., Harris, N. L., Huser, V., Mascalzoni, D., McMurry, J. A., Might, M., Nellaker, C., Mons, B., Paltoo, D. N., Pevsner, J., Posada, M., ... Haendel, M. A. (2020). The case for open science: Rare diseases. *JAMIA Open*, 3(3), 472–486. <https://doi.org/10.1093/jamiaopen/ooaa030>
- Thompson, R., Johnston, L., Taruscio, D., Monaco, L., Bérout, C., Gut, I. G., Hansson, M. G., t Hoen, P.-B. A., Patrinos, G. P., Dawkins, H., Ensini, M., Zatloukal, K., Koubi, D., Heslop, E., Paschall, J. E., Posada, M., Robinson, P. N., Bushby, K., & Lochmüller, H. (2014). RD-Connect: An integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. *Journal of General Internal Medicine*, 29, 780–787. <https://doi.org/10.1007/s11606-014-2908-8>
- Wagner, A. H., Babb, L., Alterovitz, G., Baudis, M., Brush, M., Cameron, D. L., Cline, M., Griffith, M., Griffith, O. L., Hunt, S. E., Kreda, D., Lee, J. M., Li, S., Lopez, J., Moyer, E., Nelson, T., Patel, R. Y., Riehle, K., Robinson, P. N., ... Hart, R. K. (2021). The GA4GH Variation Representation Specification: A computational framework for variation representation and federated identification. *Cell Genomics*, 1(2), 100027. <https://doi.org/10.1016/j.xgen.2021.100027>



- Zoch, M., Gierschner, C., Peng, Y., Gruhl, M., Leutner, L. A., Sedlmayr, M., & Bathelt, F. (2021). Adaption of the OMOP CDM for rare diseases. *Studies in Health Technology and Informatics*, 281, 138–142. <https://doi.org/10.3233/SHTI210136>
- Zurek, B., Ellwanger, K., Vissers, L. E. L. M., Schüle, R., Synofzik, M., Töpf, A., de Voer, R. M., Laurie, S., Matalonga, L., Gilissen, C., Ossowski, S., 't Hoen, P., Vitobello, A., Schulze-Hentrich, J. M., Riess, O., Brunner, H. G., Brookes, A. J., Rath, A., Bonne, G., ... Solve-RD, c (2021). Solve-RD: Systematic pan-European data sharing and collaborative analysis to solve rare diseases. *European Journal of Human Genetics*, 29, 1325–1331. <https://doi.org/10.1038/s41431-021-00859-0>

How to cite this article: Rambla, J., Baudis, M., Ariosa, R., Beck, T., Fromont, L. A., Navarro, A., Paloots, R., Rueda, M., Saunders, G., Singh, B., Spalding, J. D., Törnroos, J., Vasallo, C., Veal, C. D., & Brookes, A. J. (2022). Beacon v2 and Beacon networks: A “lingua franca” for federated data discovery in biomedical genomics, and beyond. *Human Mutation*, 43, 791–799. <https://doi.org/10.1002/humu.24369>



## Chapter 3

Rueda, M., Ariosa, R., Moldes, M. & Rambla, J. (2022). [Beacon v2 Reference Implementation: a toolkit to enable federated discovery of genomic and phenotypic data](#). *Bioinformatics*, Volume 38, Issue 19, 1 October 2022, Pages 4656–4657



*Bioinformatics*, YYYY, 0–0

doi: 10.1093/bioinformatics/xxxxx

Advance Access Publication Date: DD Month YYYY

Applications Note

## Databases and Ontologies

# Beacon v2 Reference Implementation: a toolkit to enable federated sharing of genomic and phenotypic data

Manuel Rueda<sup>1,†,\*</sup>, Roberto Ariosa<sup>1,†</sup>, Mauricio Moldes<sup>1,†</sup> and Jordi Rambla<sup>1</sup><sup>1</sup>European Genome-phenome Archive (EGA) in the Centre for Genomic Regulation (CRG), the Barcelona Institute of Science and Technology Dr. Aiguader 88, Barcelona, 08003 Spain.

\*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: XXXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

### Abstract

**Summary:** Beacon v2 is an API specification established by the Global Alliance for Genomics and Health initiative (GA4GH) that defines a standard for federated discovery of genomic and phenotypic data. Here we present the Beacon v2 Reference Implementation (B2RI), a set of open-source software tools that allow lighting up a local Beacon instance “out-of-the-box”. Along with the software, we have created detailed “Read the Docs” documentation that includes information on deployment and installation.

**Availability:** The B2RI is released under GNU General Public License v3.0 and Apache License v2.0. Documentation and source code is available at: <https://b2ri-documentation.readthedocs.io>

**Contact:** [manuel.rueda@crgeu](mailto:manuel.rueda@crgeu)

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

The Global Alliance for Genomics and Health (GA4GH) released in April 2022 the v2 of the Beacon specification, which defines an open standard for secure federated discovery of genomic and phenotypic data in biomedical research and clinical applications (Rambla, et al., 2022). Beacon v2 specification consists of two components, the *Framework* and the *Models*. The Framework defines the format for the requests and responses, whereas the Models define the structure of the biological data response (see Sup. Data ST1 and SF1). The overall function of these components is to provide the instructions to design a REST API.

Implementing a Beacon v2 API directly from the specification can be challenging for centers not having trained personnel. To demonstrate Beacon v2 capabilities and to facilitate the adoption, at the Centre for Genomic Regulation (CRG) we have developed the Beacon v2 Reference Implementation (B2RI), an open source Linux-based software toolkit that allows lighting up a local instance of Beacon “out-of-the-box”. In this communication, we will describe the software and summarize how its components work together to enable “beaconization” of biological data.

## 2 Methods and implementation

Overall, two basic elements are needed to implement a local instance of Beacon v2: i) an internal database (where the biological data are stored), and ii) a REST API that provides a standardized way to receive requests and send responses. The B2RI provides these basic elements, as well as a set of tools to transform biological data to the internal database format. The B2RI consists of four components:

- A set of tools for extraction, transformation and loading of metadata (e.g., sequencing methodology, bioinformatics tools), phenotypic data and genomic variants into a database.
- The database (an instance of MongoDB) (<https://www.mongodb.com>).
- The Beacon v2 query engine (i.e., a REST API)
- An example dataset consisting of synthetic data (CINECA synthetic cohort EUROPE UK1) (see Sup. Text ST6).

The software is available for download from Docker Hub ([https://hub.docker.com/r/beacon2ri/beacon\\_reference\\_implementation](https://hub.docker.com/r/beacon2ri/beacon_reference_implementation)) or through GitHub repositories (see Sup. Text ST7) and must be deployed on a local workstation/server. Hence, some security aspects of data access, such as external IP access, rely on the “jurisdiction” of each research centre. The software is written in *Python*, *Perl* and *Bash* and functions with a command-line interface (CLI) for control and operation.

We will now describe how the components work together to enable data conversion and access through the REST API.

## 2.1. Data ingestion

The data ingestion consists of three steps:

### 2.1.1 Transforming metadata and phenotypic data

Researchers/clinicians store metadata and phenotypic data in a wide variety of sources/formats (e.g., text files, CSV, Excel, databases, Electronic Health Records, PDF, etc.). The idea is that B2RI will facilitate converting data in those formats to the hierarchical structure of the Beacon v2 Models. The Models are a set of seven (*analyses*, *biosamples*, *cohorts*, *datasets*, *genomicVariations*, *individuals* and *runs*) entities (entry types in Beacon v2 specification) created to provide uniformity for the biological data responses (see Sup. Fig. SF2). The entry types are defined using JSON Schema and consist of multiple properties (or terms). As *input*, we provide an Excel template (see [https://github.com/EGA-archive/beacon2-ri-tools/tree/main/utills/bff\\_validator](https://github.com/EGA-archive/beacon2-ri-tools/tree/main/utills/bff_validator)) consisting of all Models properties “flattened-out” and separated into seven sheets (one per entry type). Note that it is not necessary to fill out all the sheets to light up a Beacon v2 instance. The user is responsible for filling out the Excel according to the entities and terms they want to share. Ontologies are defined at this level, but we are not enforcing the use of any particular ones, as ontologies depend on the domain of study (in any case, we provide examples in the documentation). Once the sheets are filled out, the B2RI comes with a utility that validates the Excel file against the Models JSON Schemas, and, if successful, it creates a set of JSON text files (JSON arrays) as an output that will be later loaded into the database.

### 2.1.2 Transforming genomic variations

For genomic data, the B2RI comes with a tool (see <https://github.com/EGA-archive/beacon2-ri-tools>) that takes as *input* a VCF (Danecek, et al., 2011) file (from DNaseq) and uses *BCFtools* (Narasimhan, et al., 2016), *SnpEff* (Cingolani, et al., 2012) and *SnpSift* (Cingolani, et al., 2012) (with data from *dbNSFP* (Liu, et al., 2020) (see Sup. Text ST8) and *ClinVar* (Landrum, et al., 2016) to annotate each VCF. Once annotated, the tool transforms VCF data to the *genomicVariations* entry type and serializes it as a JSON file.

### 2.1.3 Load data into MongoDB

Once transformed, the set of seven JSON files define what we call the *Beacon Friendly Format* (BFF) (see online documentation). The same tool used to process the VCF (see above) also enables loading BFF files into a MongoDB instance. We have chosen MongoDB as a *de facto* database as it works directly with JSON files. This way, we can store the data directly in the database according to the Beacon v2 Models and provide responses (Beacon v2 compliant) without the need of re-mapping

the data at the API level (see Sup. Text ST2). Once loaded into the database, the entry types will be referred to as MongoDB *collections*.

## 2.2 REST API

### 2.2.1 Queries

The API (see <https://github.com/EGA-archive/beacon2-ri-api>) follows REST principles and queries are carried out by sending requests (using either *GET* or *POST* HTTP methods) to Beacon v2 API endpoints (see Sup. Text ST2). Queries are performed using *request parameters* to map the API’s vocabulary to MongoDB collections. Queries can be further refined by using *filtering terms*. There exist four types of filtering terms *Bio-ontology*, *Custom*, *Numeric* and *Alphanumeric* (see Sup. Text ST4). Please see examples of API requests and responses in the Sup. Text ST5 and in the online documentation.

### 2.2.2 Security

The API can be configured according to different *security* and *granularity* levels. Three security levels (public, registered and controlled) can be set to grant differential external access and another three (boolean, counts and records) can be set for the granularity of the response (see Sup. Text ST3).

## Acknowledgements

We would like to thank Dietmar Fernández-Orth, Sabela de La Torre and Toshiaki Katayama for their contribution to previous versions of the software, and, to Prof. Michael Baudis (UZH) and EGA members for their comments. We also thank all early testers of the software and the referees for their valuable feedback.

## Funding

This study was funded by ELIXIR, the research infrastructure for life-science data (ELIXIR Beacon Implementation Studies 2019-2021 and 2022-2023).

*Conflict of Interest:* none declared.

## References

- Cingolani, P., et al. (2012) Using *Drosophila melanogaster* as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift, *Frontiers in genetics*, 3, 35.
- Cingolani, P., et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3, *Fly*, 6, 80-92.
- Danecek, P., et al. (2011) The variant call format and VCFtools, *Bioinformatics*, 27, 2156-2158.
- Landrum, M.J., et al. (2016) ClinVar: public archive of interpretations of clinically relevant variants, *Nucleic acids research*, 44, D862-868.
- Liu, X., et al. (2020) dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs, *Genome medicine*, 12, 103.
- Narasimhan, V., et al. (2016) BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data, *Bioinformatics*, 32, 1749-1751.
- Rambla, J., et al. (2022) Beacon v2 and Beacon Networks: a “lingua franca” for federated data discovery in biomedical genomics, and beyond, *Human mutation*.

## Chapter 4

Thorogood, A., Rehm, H. L., Goodhand, P., Page, A. J., Joly, Y., Baudis, M., **Rambla, J.**, Navarro, A., Nyronen, T. H., Linden, M., Dove, E. S., Fiume, M., Brudno, M., Cline, M. S., & Birney, E. (2021). [International federation of genomic medicine databases using GA4GH standards](#). Cell Genomics, 1(2), 100032.





## Commentary

# International federation of genomic medicine databases using GA4GH standards

Adrian Thorogood,<sup>1,2,\*</sup> Heidi L. Rehm,<sup>3,4</sup> Peter Goodhand,<sup>5,6</sup> Angela J.H. Page,<sup>4,5</sup> Yann Joly,<sup>2</sup> Michael Baudis,<sup>7</sup> Jordi Rambla,<sup>8,9</sup> Arcadi Navarro,<sup>8,10,11,12</sup> Tommi H. Nyronen,<sup>13,14</sup> Mikael Linden,<sup>13,14</sup> Edward S. Dove,<sup>15</sup> Marc Fiume,<sup>16</sup> Michael Brudno,<sup>17</sup> Melissa S. Cline,<sup>18</sup> and Ewan Birney<sup>19</sup>

<sup>1</sup>ELIXIR-Luxembourg and Biocore, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Belvaux, Luxembourg

<sup>2</sup>Centre of Genomics and Policy, Department of Human Genetics, McGill University, Montreal, QC, Canada

<sup>3</sup>Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA

<sup>4</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA

<sup>5</sup>Global Alliance for Genomics and Health, Toronto, ON, Canada

<sup>6</sup>Ontario Institute for Cancer Research, Toronto, ON, Canada

<sup>7</sup>University of Zurich and Swiss Institute of Bioinformatics, Zurich, Switzerland

<sup>8</sup>Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain

<sup>9</sup>Universitat Pompeu Fabra, Barcelona, Spain

<sup>10</sup>Institute of Evolutionary Biology (UPF-CSIC), Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain

<sup>11</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

<sup>12</sup>Barcelonaβeta Brain Research Center (BBRC), Pasqual Maragall Foundation, Barcelona, Spain

<sup>13</sup>CSC - IT Center for Science, Life Science Center, Espoo, Finland

<sup>14</sup>ELIXIR-Europe (Finland), Wellcome Genome Campus, Hinxton, Cambridgeshire, UK

<sup>15</sup>School of Law, University of Edinburgh, Edinburgh, UK

<sup>16</sup>DNASTack, Toronto, ON, Canada

<sup>17</sup>Department of Computer Science, University of Toronto and University Health Network, Toronto, ON, Canada

<sup>18</sup>UC Santa Cruz Genomics Institute, Mail Stop: Genomics, 1156 High Street, Santa Cruz, CA 95064, USA

<sup>19</sup>European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridgeshire, UK

\*Correspondence: [adrian.thorogood@uni.lu](mailto:adrian.thorogood@uni.lu)

<https://doi.org/10.1016/j.xgen.2021.100032>

We promote a shared vision and guide for how and when to federate genomic and health-related data sharing, enabling connections and insights across independent, secure databases. The GA4GH encourages a federated approach wherein data providers have the mandate and resources to share, but where data cannot move for legal or technical reasons. We recommend a federated approach to connect national genomics initiatives into a global network and precision medicine resource.

## Introduction

National-scale genomic sequencing initiatives are emerging worldwide to promote personalized healthcare and innovation. These national initiatives will generate genomic datasets for tens of millions of individual people as part of routine healthcare.<sup>1</sup> Connecting this wealth of data internationally offers great potential to advance our understanding of and our ability to address disease. Genomic and health-related data are sensitive, however, implicating the privacy of sequenced individuals and their families and typically attracting legal restrictions on disclosure and potentially also international transfer. The Global Alliance for Genomics and Health (GA4GH) is a standards-setting body established to promote the international sharing of genomic and health-related data.<sup>1</sup> It supports diverse models for sharing genomic

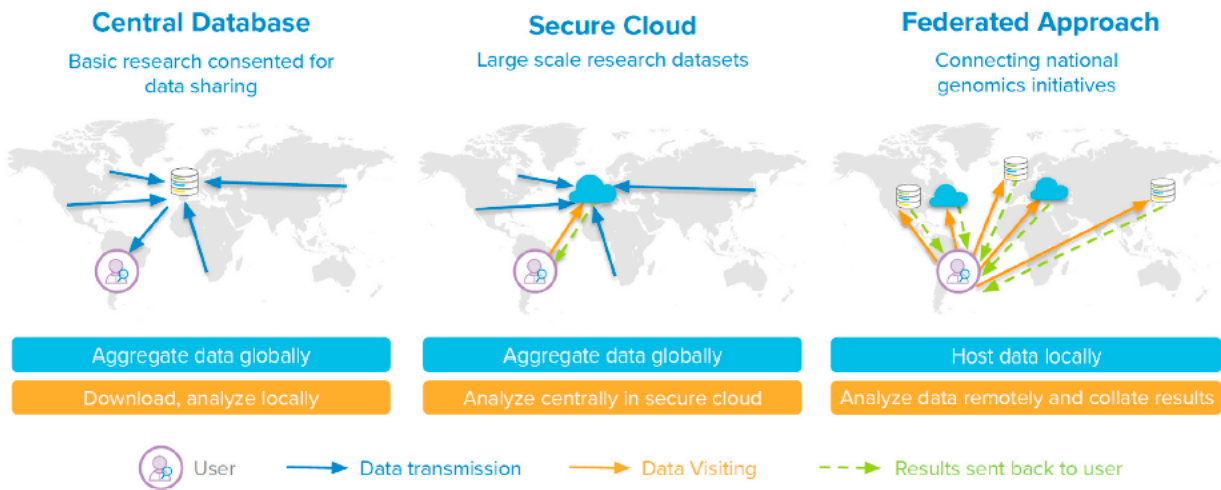
and health-related data with authorized users while also protecting competing interests. These models span central databases to networks of distributed databases connected by common infrastructure.<sup>2</sup> Data can be hosted in the cloud—along with methods, workflows, and computing resources—to facilitate secure, international access and large-scale analysis.<sup>3</sup>

A federated approach to data sharing is an alternative in which independent data providers maintain their own secure database. A data provider is any organization hosting a database of genomic and associated health data willing to share the data with data users—individuals and organizations who seek to analyze data. By adopting data and technical standards, they enable users to analyze data across multiple databases and combine the results. Each data provider maintains full

control over its data and access management in a secure computing environment. Data providers may choose to voluntarily align on common access policies and infrastructure to streamline user experience (Figure 1).<sup>4,5</sup> Federated approaches are highly attractive in principle, offering data providers more control without sacrificing opportunities for collaboration and openness. The concept is also flexible and can be adapted to different contexts. This flexibility can, however, lead to disagreement over what federated data sharing means in practice, stymying implementation.

In this commentary, we promote a shared vision for how and when to federate genomic and health-related databases. We review central considerations for developing these federated systems, including key design choices and trade-offs, and





**Figure 1. Data sharing approaches: Central database, secure cloud, and federated**

Central database: Data from multiple sources are pooled in a central database. Researchers download copies of data and analyze them in their own computing environment.

Secure cloud: Data from multiple sources are pooled in a central cloud environment. Researchers remotely visit data and run their analyses in the cloud and download the result.

Federation: Data remain within locally controlled databases and computing environments, which may be cloud environments. Researchers remotely visit data, run their analyses at each site, and receive a local result, which can then be aggregated.

how to incorporate GA4GH standards and frameworks. Federated approaches are justified over alternatives only where data cannot be pooled or transmitted for legal or technical reasons. Success is only likely where data providers have significant resources and a clear mandate to share. Federated approaches can involve different levels of organizational independence and security, with consequences for legal compliance, incentives, and costs. Data and technical standards—key enablers for data sharing generally—are especially vital for federated approaches, ensuring that data are FAIR (findable, accessible, interoperable, and re-useable) so as to enable analysis at scale.<sup>6</sup> Standard-setting bodies like the GA4GH are needed to bring together networks of independent data providers to drive adoption of these standards.

We recommend a federated system to connect national genomics initiatives into a global precision medicine resource. Connecting these resources would provide an opportunity for research on an unprecedented scale. A federated approach is necessary in this context. These initiatives face important security, sovereignty, and trust concerns that militate against pooling data in centralized environments. National initiatives are increasingly integrated with healthcare systems, which

tend to impose stricter rules around confidentiality and secondary use for research (though this depends very much on context). The sheer size of population-scale genomic databases makes them technically difficult to manage and transfer. Nations also expect their investments in large-scale genomic medicine initiatives to serve (competitive) national scientific, health, and wealth goals, with international research agendas being secondary. In light of all these concerns, trust across diverse countries and actors can be hard to establish. A federated approach is also feasible for national initiatives, who have the mandate to share and resources to make data and technical infrastructure—following GA4GH standards—available to the research community. This international use case, if successful, can provide a blueprint for expanding federated approaches to rich, real-time genomic data across national networks of hospitals and laboratories.

#### Key design choices and trade-offs

Federated approaches to data sharing allow data providers to preserve control, security, and accountability while (under the right conditions) still enabling data users to run analyses at scale. The level of data provider independence and the level of security varies across federated

approaches, with important implications for legal compliance, incentives, and costs. The following design considerations and trade-offs, drawn in part from experience in artificial intelligence and digital health contexts,<sup>7,8</sup> provide a guide for the genomics and health community.

#### Control over data

Federated data sharing approaches emphasize the independence of the participating data providers. The Oxford English Dictionary (third edition, 2015) defines federation as a “body . . . formed from a number of separate organizations . . . each retaining control of its own internal affairs.” A federated approach to data sharing typically means that data providers retain control over their own data, hosted in their own secure computing environment. Data providers also retain control over access management, i.e., who can access the data, for what purposes, and under what conditions. Greater control is meant to give data providers the confidence to make richer datasets available to a broader range of users, assuming they have the mandate and resources to do so.<sup>5</sup> The degree of individual organizational independence and control varies across federated approaches. At the most independent and loosely defined end, federation may simply be a group of independent data providers who voluntarily adopt a basic set of data

and technical standards. In this approach, there is no global data access committee, and data providers can independently establish their own data access policies. This approach is lightweight for data providers, but it requires data users to make separate access applications for each database and to navigate different access criteria. Although users face more paperwork, they are still able to access and analyze multiple databases separately and then integrate the results.

In more coordinated models of federation, data providers actively collaborate to align data standards and streamline user access. They may even agree to common access rules or to coordinate their access processes through a central data access portal or committee. Sharing sovereignty constrains independence over access management, though data providers still maintain direct control over data. This gives them greater flexibility to withdraw (certain kinds of) access at a later time, if conditions become less favorable. Users benefit from being able to access multiple resources with a single application and to trust their analyses will run reliably in different environments on interoperable datasets.

#### **Data utility**

On the one hand, federated approaches can enhance data utility. They provide a means to combine datasets into a virtual cohort, enabling analyses on datasets of larger scale and statistical power. Because data providers keep tight control over their datasets, they may be more willing and able to share richer, more routinely updated data. De-identification does not need to be as rigorous, as data are not disclosed, preserving utility. On the other hand, the utility of the datasets depends on the adoption of data and technical standards by data providers who require significant resources and expertise. Some data quality issues like record de-duplication can only be addressed collaboratively across data providers. This may be done securely through privacy-preserving record linkage. Users with limited access to data are unable to assess data quality or compare data across sources, exacerbating general data science challenges. They are more reliant on data providers to assist with data curation, analysis, and interpretation. Pooling and direct exchange of data has

long been a catalyst for the standardization of data elements, models, and quality. With no central repository to foster comparison, a federation of independent data providers may need compensating measures to actively drive standardization, such as standard-setting bodies, certifications, or trusted third-party curation services. These challenges can be facilitated by APIs (application programming interfaces) and containers. APIs are interfaces that allow users to query databases even with different underlying data formats. Containers are tools that bundle together software pipelines and their dependencies so they can run reliably in different computing environments.

#### **Security**

In federated data sharing models, each data provider grants authorized users remote access to data in its own computing environment. Access may be direct or indirect.<sup>9</sup> Users granted direct access may analyze each database separately, taking only summary statistics with them when they leave. This limits copying and transmission of data, reducing security risks and allowing continuous monitoring of user activity. The workflow is similar to contexts in which data are pooled centrally, in which users still need to segregate datasets for analytical reasons (e.g., applying different covariates and making independent estimates of significance). For even greater security, users can be limited to indirect access to data. Data remain hidden at all times behind secure firewalls. Users submit algorithms or queries, which are vetted and executed by the data provider, who returns summary or performance statistics.<sup>4</sup> Federated analysis means running the same analysis across multiple hidden databases. This has been demonstrated in artificial intelligence contexts, where models are trained across hospitals<sup>10</sup> or personal smart phones. Only in an idealized vision is federated analysis perfectly seamless for users; data providers may very well insist on their independence to control access to their own data and computing environments. Ultimately, greater data security has trade-offs. It constrains users' ability to interact with data. Data and technical standards become all-important to

ensure interoperability. Most importantly, the significant costs of both standardization and security fall to the data providers.

Federated data sharing models also introduce new security risks. Data providers face IT security risks when external users, or their software, are introduced into local computing environments. These risks can be alleviated through careful monitoring of user activity and airlocks to control introduction of external software (at additional cost). Federated approaches can also create security risks for users, who expose their research questions or code to a network of data providers. Where risks to users' queries and code are serious, they can be reduced through encryption and secure computing approaches in which data providers execute hidden code.<sup>7</sup>

#### **Legal compliance and ethics**

Federated approaches can alleviate legal and ethical concerns raised by data sharing, though they are not a panacea. The European Union General Data Protection Regulation 2016/679 (GDPR) has set a global standard for robust protection of personal data, which includes mandating limitations on international transfers of personal data outside the EU/EEA. It has also triggered a strong shift toward federated approaches for large scientific data infrastructure, in projects like the European Genome-Phenome Archive, European Open Science Cloud, the European 1+ Million Genomes Initiative, and the European Health Data Space. Secure local data hosting can improve accountability, trust, and individuals' ability to exercise rights like withdrawal of consent to further use or sharing of their data. Robust safeguards provide strong assurances of data protection, even when data are accessed by international researchers. International access within a European data center is still an international transfer, however. Clear legal pathways and privacy-enhancing technologies must be further developed before access can be extended outside Europe.<sup>11</sup> Even where data do not move, appropriate informed consent and ongoing transparency are still generally required for data sharing. Data subjects need to know who is accessing data and for what purposes. Research ethics oversight may also be a greater challenge for federated approaches than alternatives,

as data are analyzed across many different institutions and countries. To address this challenge, the GA4GH Ethics Review Equivalency Policy promotes international standards for ethics review, alongside cross-border coordination and recognition mechanisms.<sup>1</sup>

#### **Incentives**

A lack of incentives to provide data is a well-known barrier to data sharing. While federated approaches do not resolve this barrier, they do give data providers increased control and security, which may increase their willingness to share. Ongoing control may also mean data providers have more leverage to negotiate active collaboration, appropriate scientific recognition, or a share in commercial outputs. More conditions and transaction costs, however, discourage re-use of data, especially as they stack up across data providers. Indirect benefits to data providers include opportunities to develop local capacity and expertise in data infrastructure, management, and analysis. Ultimately, however, incentives must continue to be addressed through broader policy initiatives, investment in infrastructure, and cultural change.

#### **Sustainability**

The most important consideration for data providers considering a federated approach is cost. Data providers incur significant security, data management, and computing costs, including those related to adopting and maintaining standards. These costs are likely to be duplicated across data providers and thus higher overall in comparison to central databases. Federated approaches do spread these costs more evenly across data providers. One way to mitigate expense is through optimal network design. An international federation of genomic databases is enabled by pooling data on a national level. National pooling may raise fewer legal and trust issues, while also providing efficiencies.

#### **Enabling standards**

A key challenge for federated approaches is driving the adoption and maintenance of data and technical standards across numerous, independent organizations. Relying on voluntary adoption of community guidelines is likely to be too weak. Establishing formal partnership agreements could be too strong. The GA4GH, as an open standards-setting body, provides a

middle way. It offers a flexible and participatory model to drive the international adoption of consensus standards, collaborating with a network of Driver Projects and member organizations across the global genomics community.

The GA4GH develops and endorses data and technical standards that can be used to enable data sharing generally and federated approaches specifically (see Rehm et al. in this issue for details on these standards<sup>1</sup>). Data and metadata standards are key enablers for any discovery and re-use of data. Standard file formats provide standard structures for genomic data. The Phenotype Ontology provides a semantic ontology for expressing phenotypic data. Federated approaches additionally require technical standards to ensure the interoperability of distributed databases and computing environments. The GA4GH Beacon and Data Connect APIs allow researchers to find individuals with relevant genotypes or phenotypes in a database. Search interfaces can accept structured queries as input and release structured search results as output. Federated search is where users submit a single query that is run on and answered by multiple, independent databases, even where underlying structures differ. Each organization can determine the specificity of the search results (e.g., a simple yes/no, summary statistics, minimal health information associated with the variants) and its own access controls and security safeguards. Federated search has already been successfully demonstrated with GA4GH APIs.<sup>12</sup>

Authentication and authorization standards are needed to coordinate user access to multiple databases. OAuth 2.0 and OpenID Connect are useful tools to assist data providers in confirming the user seeking access is the person who has received approval to do so. Even where data providers retain independent control over access decisions, they may agree to coordinate user authentication protocols. CanDIG, a GA4GH Driver Project, uses an authentication scheme based on OpenID Connect, where each data provider authenticates the identity of its own employees, and that authentication is in turn accepted by the other participating nodes.<sup>13</sup> Each data provider continues to make its own authorization decisions based on local policy. Even

so, federated approaches are facilitated where data providers express their local data access and use credentials in a standard way. GA4GH Passports build on authentication standards to allow data providers to confirm a user has standard credentials.<sup>14</sup> The Data Use Ontology (DUO) allows data providers to ensure access requests match to standard data use conditions.<sup>15</sup> Federated analysis in particular requires interoperability between computing environments, because workflows are executed on behalf of data users on hidden databases. Federated analysis can be assisted by the GA4GH Cloud APIs, interfaces that allow users to look up data and tools and to execute portable workflows, driving larger-scale and more powerful analyses. The GA4GH Federated Analysis Systems Project (FASP) brings all these pieces together into end-to-end test scenarios, aiming to simulate how a researcher would search, access, and analyze genomic data across a network of real-world projects.<sup>1</sup>

#### **Conclusion**

Federated approaches to data sharing are flexible, involving design choices about data provider independence and secure access mechanisms. These choices influence data accessibility, data utility, legal compliance, and cost. The GA4GH encourages federated approaches where data providers have the will and resources to share but where data cannot flow because of legal, technical, or institutional policy reasons. Federated approaches come with costs and limitations, but they also provide opportunities to improve privacy protection, accessibility, and interoperability. Advancing federated approaches in genomics will also align the field with data sharing practices in digital health and artificial intelligence.

Creative mechanisms are needed to drive adoption of data and technical standards across networks of independent data providers. As a standards-setting body, the GA4GH is uniquely positioned to assist the genomics community to meet these challenges and bring the vision of a federated approach to genomics and human biomedical data sharing into reality, so as to realize the right of everyone to benefit from the progress of science.

**SUPPLEMENTAL INFORMATION**

Supplemental information can be found online at <https://doi.org/10.1016/j.xgen.2021.100032>.

**WEB RESOURCES**

European Genome-Phenome Archive, Federated EGA, <https://ega-archive.org/federated>  
 European Commission, European 1+ Million Genomes Initiative, <https://digital-strategy.ec.europa.eu/en/news/eu-countries-will-co-operate-linking-genomic-databases-across-borders>  
 European Commission, European Health Data Space, [https://ec.europa.eu/health/ehealth/dataspace\\_en](https://ec.europa.eu/health/ehealth/dataspace_en)  
 European Open Science Cloud, <https://eosc-portal.eu/>  
 Genomics England, Airlock Policy, Version 2.0, [https://www.genomicsengland.co.uk/about-gecip-for-gecip-members/documents/GA4GH-Federated-Analysis-Systems-Project-\(FASP\)](https://www.genomicsengland.co.uk/about-gecip-for-gecip-members/documents/GA4GH-Federated-Analysis-Systems-Project-(FASP)), <https://www.ga4gh.org/genomic-data-toolkit/2020-connection-demos/GA4GH-Ethics-Review-Recognition-Policy>  
 Google AI Blog, Federated Learning: Collaborative Machine Learning without Centralized Training Data, <https://ai.googleblog.com/2017/04/federated-learning-collaborative.html>  
 IRDIRC, Technology Primer: Overview of Technological Solutions to Support Privacy-Preserving Record Linkage, <https://www.irdirc.org/wp-content/uploads/2018/03/PPRL-Technical-Primer-V4-2.pdf>  
 OAuth 2.0 Authorization Framework, <https://datatracker.ietf.org/doc/html/rfc6749>  
 OpenID Connect Core 1.0 incorporating errata set 1, [https://openid.net/specs/openid-connect-core-1\\_0.html](https://openid.net/specs/openid-connect-core-1_0.html)

**ACKNOWLEDGMENTS**

We acknowledge the encouragement of the Global Alliance for Genomics and Health Steering Committee and the input of its members into the concept and content of the paper. We would like to thank Stephanie Li for assistance with preparing the manuscript graphics. A.T. acknowledges funding support from Genome Canada, Genome Quebec, and the Canadian Institutes of Health Research. H.L.R. and A.J.H.P. acknowledge funding under NIH U41HG006834 and U24HG011025. M. Baudis acknowledges funding under the BioMedIT Network project of Swiss Institute of Bioinformatics (SIB) and Swiss Personalized Health

Network (SPHN). M.L. acknowledges funding from the CINECA project (H2020 No 825775). M.S.C. acknowledges funding under NIH/NCI U01CA242954, NIH/NHLBI Fellowship 5118777. M. Brudno is a CIFAR Canada AI Chair. P.G. acknowledges funding from CIHR, Genome Canada, Wellcome Trust, and NIH. T.H.N. was funded by the Academy of Finland grant no 319968 and ELIXIR Europe 2019–2023 program.

**AUTHOR CONTRIBUTIONS**

A.T.: Conceptualization; Writing (Original Draft). H.L.R., P.G., E.B.: Conceptualization; Supervision; Writing (Review & Editing). A.J.H.P., Y.J., M. Baudis, J.R., A.N., T.H.N., M.L., E.S.D., M.F., M. Brudno, and M.S.C.: Writing (Review & Editing)

**DECLARATION OF INTERESTS**

M. Brudno holds financial interest in PhenoTips. E.B. is a consultant to Oxford Nanopore Technologies and Dovetail Inc. and a member of the *Cell Genomics* advisory board. H.L.R. is a member of the *Cell Genomics* advisory board. All other authors have no interests to declare.

**REFERENCES**

1. Rehm, H.L., Page, A.J.H., Smith, L., Adams, J.B., Alterovitz, G., Babb, L.J., Barkley, M.P., Baudis, M., Beauvais, M.J.S., Beck, T., et al. (2021). GA4GH: international policies and standards for data sharing across genomic research and healthcare. *Cell Genomics* 1, 100029-1-100029-33.
2. Contreras, J.L., and Reichman, J.H. (2015). Sharing by design: Data and decentralized commons. *Science* 350, 1312-1314.
3. Grossman, R.L. (2019). Data lakes, clouds, and commons: A review of platforms for analyzing and sharing genomic data. *Trends Genet.* 35, 223-234.
4. Popovic, J.R. (2017). Distributed data networks: a blueprint for Big Data sharing and healthcare analytics. *Ann. N Y Acad. Sci.* 1387, 105-111.
5. World Economic Forum (2019). Federated Data Systems: Balancing Innovation and Trust in the Use of Sensitive Data. <https://www.weforum.org/whitepapers/federated-data-systems-balancing-innovation-and-trust-in-the-use-of-sensitive-data>.
6. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B.,

- Boume, P.E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 160018.
7. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., et al. (2020). The future of digital health with federated learning. *npj Digital Medicine* 3, 119.
8. Kaissis, G.A., Makowski, M.R., Rückert, D., and Braren, R.F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* 2, 305-311.
9. Suver, C., Thorogood, A., Doerr, M., Wilbanks, J., and Knoppers, B. (2020). Bringing Code to Data: Do Not Forget Governance. *J. Med. Internet Res.* 22, e18087.
10. Jochems, A., Deist, T.M., van Soest, J., Eble, M., Bulens, P., Coucke, P., Dries, W., Lambin, P., and Dekker, A. (2016). Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital - A real life proof of concept. *Radiother. Oncol.* 121, 459-467.
11. ALLEA (2021). EASAC and FEAM joint initiative, International Sharing of Personal Health Data for Research. <https://easac.eu/publications/details/international-sharing-of-personal-health-data-for-research>.
12. Fiume, M., Cupak, M., Keenan, S., Rambla, J., de la Torre, S., Dyke, S.O.M., Brookes, A.J., Carey, K., Lloyd, D., Goodhand, P., et al. (2019). Federated discovery and sharing of genomic data using Beacons. *Nat. Biotechnol.* 37, 220-224.
13. Dursi, L.J., Bozoky, Z., de Borja, R., Li, H., Bujold, D., Lipski, A., Rashid, S.F., Sethi, A., Memon, N., Naidoo, D., et al. (2021). CanDIG: Federated network across Canada for multi-omic and health data discovery and analysis. *Cell Genomics* 1, 100033-1-100033-7.
14. Voisin, C., Linden, M., Dyke, S.O.M., Bowers, S.R., Reinold, K., Lawson, J., Li, S., Ota Wang, V., Barkley, M.P., Bemick, D., et al. (2021). GA4GH Passport standard for digital identity and access permissions. *Cell Genomics* 1, 100030-1-100030-12.
15. Lawson, J., Cabili, M.N., Kerry, G., Boughtwood, T., Thorogood, A., Alper, P., Bowers, S.R., Boyles, R.R., Brookes, A.J., Brush, M., et al. (2021). The Data Use Ontology to streamline responsible access to human biomedical datasets. *Cell Genomics* 1, 100028-1-100028-9.



## 4. DISCUSSION AND CONCLUSIONS

### 4.1 Data Management

#### a) Using standards in federated approaches

Thorogood *et al.* (2021) explain why federated approaches are attractive for the management of sensitive data, and discuss where and how federation of genomic data can be useful. They mention key aspects of data sharing: having a clear mandate and enough resources. Proper data management could help in the second aspect, as a proper management could reduce the amount of resources required.

The paper recommends establishing a federation that connects national resources for genomics, paving the way for research at a larger scale. This is especially relevant when these national resources leverage data coming from healthcare and they should preferably not leave the borders of any given country.

According to this paper there are several design considerations and trade-offs for choosing and setting up a federation of genomics and health data:

- 1) **Control over data.** Keeping the data inside controlled premises should provide control over who can access the data and how much, when and how the data is used. This feeling of control could be fictitious if the expertise, the governance procedures and resources are not appropriate. One listed benefit of the feeling of control is an increased capacity to share more and better data.
- 2) **Data utility.** If the principle that federation fosters more and better data sharing is met, the goal of building larger cohorts by composition of independent datasets could be easier to reach. On the other hand, the user could need additional help from the data providers to better understand the data. Again, this consequence highlights the importance of harmonisation and the contribution of standards to this end.
- 3) **Security.** Users' access to data could be direct or indirect. In direct access users are given access to the data so it can be moved to their premises. In indirect access, users must provide analysis algorithms and only get the results back. Both

approaches have positive and negative aspects that slide in the axis of confidentiality vs. flexibility. Again, the use of standards could alleviate the issues of sending algorithms that would run in a distant and automated system.

- 4) **Legal compliance and ethics.** This is the most obvious benefit of the participants in a federation of country-centric initiatives, as every local node should be compliant with only a limited set of rules. Centralised approaches are agnostic to such particularities.
- 5) **Sustainability.** The federation approach increases the cost of every node as each one needs to provide the infrastructure and human resources to keep the system running. Centralisation usually implies that a significant part of the cost is shared with others or indeed free.

The paper's conclusions are that federation approaches have more benefits than drawbacks for the sharing and management of sensitive data, and that standards are key for enabling the actual reuse of the data for its expected purpose: to contribute to generate new knowledge that is returned as benefits to society.

## b) ELIXIR

Harrow *et al.* (2021) introduce ELIXIR and ELIXIR-EXCELERATE, one of the projects that funded the kick-off of ELIXIR as an infrastructure. ELIXIR is a virtual infrastructure that coordinates and helps to align infrastructures and resources that already exist in country members. ELIXIR starts as a community of disconnected resources that aims to become less heterogeneous by moving towards a federation of services that provide an homogeneous experience to its users. EXCELERATE ran from 2015 to 2019 and was the seed of the Federated EGA and the Federated Human Data ELIXIR Community. This paper describes the value and need for current biology of analysing large amounts of data, the barriers and the reason why ELIXIR was funded. The sections about human data go into more details for that specific domain.

The paper explains why data management and sharing is crucial, and the efforts of ELIXIR in promoting proper data management as the foundation for data sharing and reuse. It describes the scenario in domains like crop research, where repositories do not exist yet and where privacy (for economic reasons) and fragmentation hinders progress.



Harrow *et al.* (2021) state that the first step to analyse heterogeneous data coming from different sources is making it easy to be found and reused by researchers. The lack of sustainability of the projects that generate data reinforces the need of centrally managed repositories, although currently the vast majority of them do not have long term funding and, hence, lack sustainability. Clearly, sustainability is a hidden key component of real world data management.

This paper also describes the role of EGA, Federated EGA and Local EGA in the overall ELIXIR context. ELIXIR CONVERGE, a project focused in data management tooling and capacity building is mentioned by the end of the paper. After EXCELERATE, another funded project, ELIXIR CONVERGE, has contributed to build the Federated EGA. ELIXIR strategic connections to GA4GH are mentioned, and it has been the seed of ELIXIR's contribution to the GA4GH Beacon project.

### c) The EGA, the Local EGA and the Federated EGA

Freeberg *et al.* (2021) describe a success model for managing sensitive human data. The paper assumes the principle that these valuable data should be preserved safely and securely for a long period and that most institutions would prefer, or be requested by journals and funders, to deposit the data into a service that is stable in the long term, the EGA in this case. Being the EGA a centralised service, data controllers actually delegate the operations of data management (and sharing) to the EGA, although they retain the ownership and the decision on received data access requests.

In the paper's discussion, it is mentioned that the EGA is transitioning from a centralised model to a federated one, where additional nodes join the EGA, to create a federation: the Federated EGA (FEGA). Each node will play the same role as the EGA but at a local scale. After the paper was published the transition has moved forward by the addition of five nodes: Finland, Germany, Norway, Sweden and Spain.

Due to limitations in length, and to its focus on the central EGA progress, the paper by Freeberg *et al.* (2021) omits the availability of a software solution for data management: the Local EGA<sup>19</sup>. This software solution was created to quickly set up a new Federated EGA

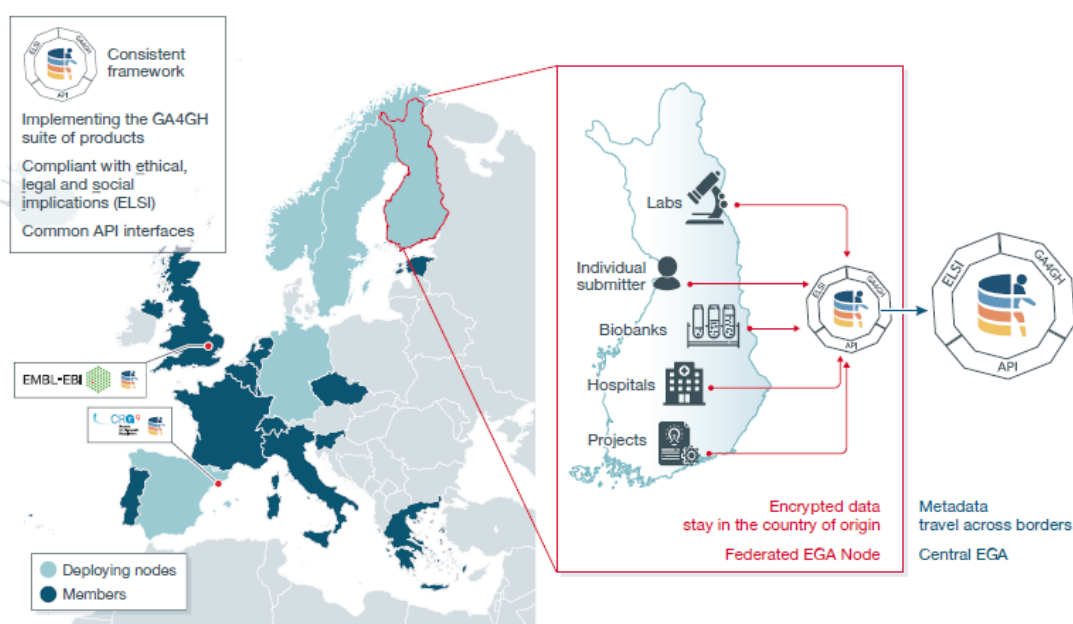
---

<sup>19</sup> <https://github.com/EGA-archive/LocalEGA>

node. However, given that the Local EGA solution has been designed to facilitate data management and sharing at a local level, with some planned modifications it could be leveraged to deploy a data management solution in a given institution to manage their own sensitive data. A paper about the Local EGA is in preparation.

With the addition of Local EGA, the more popular options for data management have been covered:

- 1) Locally on premises, by using the Local EGA solution
- 2) Delegating to another institution in the same jurisdiction, by depositing the data at a Federated EGA node
- 3) Delegating to an international service, by depositing at the Central EGA



**Figure 1 - Federated access to European Genomes.** On the left, 17 of the 23 Nodes are members of the ELIXIR Federated Human Data Community. EMBL-EBI (Cambridge, UK) and CRG (Barcelona, Spain) are specifically highlighted as these are the host institutes of the central EGA. Five Federated EGA deploying Nodes (Finland, Germany, Norway, Spain and Sweden) are also highlighted.

These implemented the Federated EGA framework in the first wave to manage archival, access and analysis of sensitive human data. On the right, a schematic view of the ELIXIR Finland Federated EGA deployment. Sensitive human data generated at laboratories, biobanks and hospitals, and/or by individual projects and submitters, are stored in encrypted format within the countries' jurisdiction. This sensitive data never leaves the Finnish borders. Metadata to describe the datasets is shared with the central EGA, which enables findability of these data. Authorised users are able to access these sensitive data remotely thanks to the suite of interoperable GA4GH standards.

Source: Harrow *et al.* 2021 - Figure 2

## 4.2 Data Discovery

### a) GA4GH Beacon version 2 specification

The formal approval of Beacon v1 by the GA4GH Steering Committee took place in October 2018. Fiume *et al.* (2018) describe the Beacon concept. Beacon v1 got the approval on that date because that was when the formal process of approval was inaugurated, and GA4GH leveraged that event to approve products that had been already available for some time. In the case of Beacon, it had been evolving from v0.2 to v1.0 between 2014 and 2018. The success, role and limitations of Beacon v1 were clear at the time of approval. Indeed, before the formal approval of Beacon v1, the Discovery WS team, owner of the Beacon specification, started to interview Driver Projects, ELIXIR Europe partners and other interested parties to gather the requirements of a potential new version of the Beacon protocol that could overcome the limitations of version 1. Fiume *et al.* (2018) describe that Beacon v1 could be used for discovery genomic and clinical data and that Beacon responses can convey more info than the basic “Yes” or “No”. The caveat is that this information is not queryable by itself and would be returned in a different format by each Beacon instance, requiring further harmonisation by the user or the Beacon client.

The Beacon working group, led by Professor Anthony Brookes from the University of Leicester, Professor Michael Baudis from the University of Zurich and by the author of this thesis from the Centre for Genomic Regulation, their respective teams and other contributors, started to design the new version according to the requirements expressed by the GA4GH Driver Projects. Rambla *et al.* (2022) describe that process and the resulting standard. This work contributes to goals 2a (discovering data), 2b (centralised or

distributed discovery), 2c (helping in harmonisation) and 3 (data sharing) of this doctoral thesis. In the following sections, a deeper discussion of some of the Beacon features could be found, for the basics, refer to Rambla *et al.* (2022). Annex 1 is extending information with technical aspects of the solutions provided to the challenges.

Some of the requirements, detailed in the following sections, where:

#### **Contributions to Goal 2c** (helping in harmonisation)

- 1) Providing a data model, broader than the Beacon v1 one, that could be leveraged as harmonised data.

#### **Contributions to Goal 2a** (discovering data)

- 2) Allowing for basic queries about the characteristics of the donors.
- 3) Allowing for querying about genomic variations from a gene.
- 4) Keeping the simplicity of the queries and the responses.

#### **Contributions to Goal 3** (data sharing)

- 5) Keeping the optional identification (authentication) and authorization of users.
- 6) Providing information on the conditions for data usage and access for the Beacon datasets.

#### **Contributions to Goal 2b** (centralised or distributed discovery)

- 7) Allowing for a smooth integration in networks of Beacons.

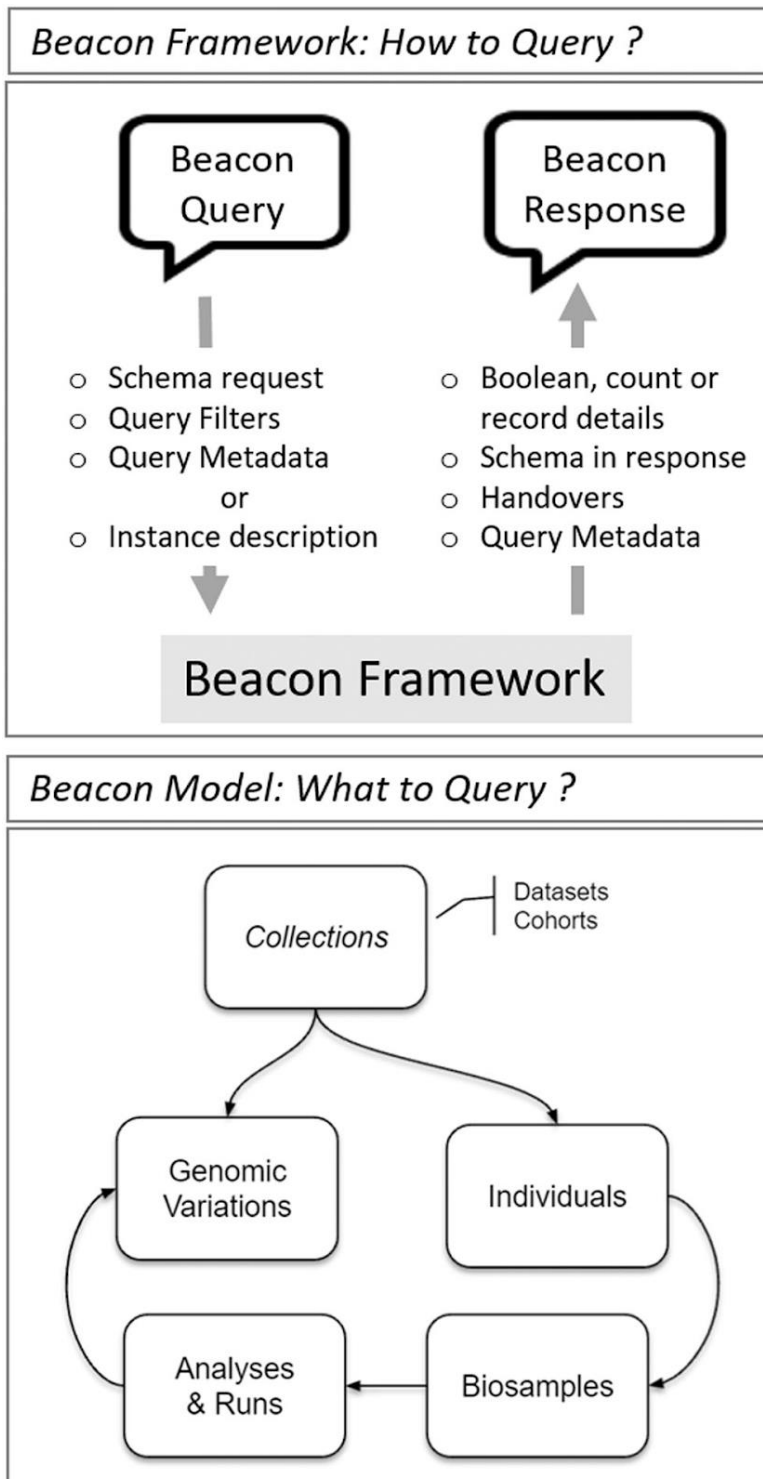
#### b) Providing a data model

Beacon v1 included a basic model for genomic variations. However, it lacked support for annotations about the variations or details about the samples and donors that originate them. For instance, BRCA Exchange, one of the GA4GH DP allows downloading data

about variants in the BRCA1 and BRCA2 genes available at some public repositories, processing and harmonising them. This operation is done periodically, hence, the data available at BRCA Exchange web site is not always up to date. All the repositories used by BRCA Exchange have a different model for describing the variants and their annotations. The BRCA Exchange lead team suggested that if all of these repositories could implement the same model, BRCA Exchange would benefit from harmonised and up to date data. This suggestion is crucial, as it links Beacon v2 with the concept of harmonising data at the origin (described in the Introduction), hence clients of the Beacons are not required to do the harmonisation themselves.

The Beacon team also interviewed partners closer to the clinical genomic diagnose domain. They suggested including information about the laboratory and bioinformatics analyses, as many times they compare different diagnoses in the context of the protocols applied to obtain the data. Another popular request was to Include information about the cohorts, as a means to understand the landscape of the data included in a given dataset.

The combination of all these requests led to the following model:



**Figure 2 - The GA4GH Beacon v2 Model.** Source: Rambla et al. 2022 - Figure 2

The model includes the following entry types:

- Cohorts: describe a formal or *ad hoc* cohort.

- Datasets: is a collection of other entry types, like genomic variations.
- Individuals: include details about sample donors of and their pedigrees.
- Biosamples: details information about samples taken from donors.
- Genomic variations: includes information about the genomic variations, both knowledge base level and case level.
- Runs: describes at high-level the procedures used to process the samples in the wet lab.
- Analysis: describes at high-level the procedures used to process sequencing data in the dry lab to determine the variations on them.

### c) Beacon v2 allows for other models and for additional schemas

The Beacon v2 specification has been designed to make the query and response protocol independent from the model that is implemented. It is possible, then, to improve the current version 2 model without changing the protocol specification. It is also possible to create new models for other domains, like imaging or viral genomics, or even biobanks who keep catalogues of their samples. These communities can leverage a protocol that is already very popular without the need to create a new one, and stack their model on top of it.

Additionally, although by default Beacon's responses are formatted using the model described above, the specification allows for other models or schemas for returning the information. As an example: a laboratory prefers to get variation data formatted using the ClinVar schema, as most of their processes understand that schema; if the queried Beacons have implemented that schema in their instances, the Beacon client in the lab could request that schema instead of the Beacon v2 default one.

These two features allow for an easier on boarding of new communities and avoid unnecessary transformations to the Beacon default model in cases where another model is already standard but a discovery protocol is not in place.

#### d) Richer query options

Genomic variations in Beacons belong to samples obtained from human donors. These donors could be part of a cohort planned in advance, like the ones in clinical trials, or simply being collected because they suffer a given disease or are part of the patient's family. It is important that Beacon allows for basic queries about the characteristics of donors, namely: gender, age range, ancestry, focus disease (when applicable), tagged as being affected by the focus disease or not, and some basic phenotypic attributes.

The requirement is for basic, simple queries, like “return only variants belonging to tumour samples from men above 30 years old diagnosed for melanoma”. In contrast, complex queries will include aspects like querying for combined events happening at different moments or dependencies between attributes, like “genomic variations from men older than 50 years having melanoma or colorectal cancer, but not present in women having melanoma”.

This requirement was refined by indicating that, preferably, all query terms should be expressed by ontology terms, not plain text. Including ontology terms in the Beacon specification has been embraced as a contribution to doctoral goal 2c.

A further refinement was to allow for “fuzzy” queries. “Fuzzy” queries are those that include a similarity option. By default, all queries imply an exact match, hence a search for “NCIT:C3510 Cutaneous Melanoma” must only return samples or donors tagged with such exact term and its descendants, like “NCIT:C54662 Nevroid Melanoma”. However, if similarity is set to “high” could return also the ascendant or siblings of the queried term, like “NCIT:C2920 Malignant Skin Neoplasm” (parent term) or “NCIT:C171101 Cutaneous Lymphoma” (sibling term). The capacity of looking for similar terms allows disparate but related annotations to be considered synonyms for the query purpose and, thus, they contribute to harmonisation (goal 2c) by alleviating the need of converting the hosted data to include the same exact terms

In summary, Beacon v2 allows for querying not only about the genomic variations but also about the samples or the donors that carry such variations. Additionally, Beacon v2 allows for querying about the cohort or cohorts hosted by a Beacon, which is another way that the requirement has been expressed. These features contribute to goal 2a.



#### e) Querying on genomic features or attributes

Beacon v1 only allows for queries on genomic positions, *e.g.*, searching for “an A in chromosome X position 1,123,101 in reference genome GRCh37”. Many clinically oriented users are more familiar with amino acid substitutions, irrespectively of the nucleotide change that is the cause of such substitution. For example, queries like “V600E in BRAF”, meaning looking for a glutamic acid (E) instead of a valine (V) in the position 600 of the BRAF protein, is common in clinically oriented environments. Other popular queries are genomic variations that are homozygous in a given sample or that produce a truncation of the protein or that has been reported as causative for a given condition. Including support for additional ways to query the hosted data contributes to goal 2a, as they facilitate the discovery of data from the user point of view.

The Beacon v2 specification addresses these requests by including sections in the genomic variation model for annotations about genomic features like HGVS gene name, amino acid substitution, diseases that have been associated with the variation, etc. It also includes a section about how that variation is found in individuals, like the zygosity. Additionally, information about frequency in populations is possible, allowing for queries that filter by minor allele frequencies (MAF) in specific populations.

#### f) Keep the simplicity of Beacon queries

A key value proposition of Beacon was the simplicity of Beacon v1 queries: having just one query endpoint and half a dozen parameters, understanding how to perform queries was very easy. Beacon v2 has been also designed to keep simplicity as a key value, although taking into account that it must cover a much broader set of queries and concepts.

Beacon v2 relies in two principles for keeping the simplicity:

- 1) Using the REST API (*REpresentational State Transfer Application Programming Interface*) approach instead of an *ad hoc* API, as REST brings consistency when querying diverse endpoints
- 2) Maintaining the complexity of the expressions to a minimum by avoiding to define a syntax for them.

### g) Maintaining the complexity of the expressions to a minimum

Query languages, like SQL, are very powerful because they provide a rich set of commands (or clauses) and because they support advanced expressions, with different boolean operators, a set of functions to handle strings, dates, numbers, etc.

While gathering requirements for Beacon v2, it was clear that the needs of Beacon users for querying were simple. The queries they described were like: “*show me all the men, older than 50 years, diagnosed with psoriasis*” or “*give me all individuals that have the V600E variant in BRAF and have African ancestry*”. Analysing these queries, we observed that the vast majority of conditions could be combined *just* using the AND boolean operator.

By limiting the available boolean operators in Beacon v2 to AND, however, it would not be possible to *directly* perform queries like “*show me all the men, older than 50 years, or the women younger than 35 years, diagnosed with psoriasis*”. Note that for solving expressions like this one, it must be clear if “diagnosed with psoriasis” refers to both men and women or only to women.

### h) Keep the simplicity of Beacon responses

Beacon v1 response was very easy to build and understand: simply a “Yes” or “No”, This is also one of Beacon v1 weaknesses: the lack of context to the response... “How many times has the variant been observed?”, “in somatic cells only or also in germline cells?”. Beacon v2 tries to solve the context issue by allowing richer queries, and by providing a much richer response.

However, sharing more details in the response introduces two concerns: 1) some Beacon instances are not allowed or prefer not to share more details and keep the “Yes/No” response, and 2) a richer response implies more software development effort to implement such a response.

As many interviewed projects declared that they are fine also with sharing a count in addition to the boolean response, we end up with the scenario of supporting three different levels of response in Beacon v2: boolean, count and details. For such purpose, Beacon v2 specification includes an attribute in the requests and in the responses: “granularity”.

A Beacon instance can support all three levels of granularity and apply one or another to every received request, depending on the sensitivity of the endpoint (e.g., cohorts vs individuals) and/or on the user being authenticated and/or authorised.

Every Beacon instance declares the default level of granularity, the one to be applied if no specific granularity is requested, in their **/configuration** endpoint. Every Beacon request includes the requested granularity and every Beacon response includes the returned granularity. Everything being allowed, the granularity response would match the granularity requested. However, if a Beacon instance receives a request for details (details granularity) and that request is not allowed because the user is not authenticated or authorised or that Beacon only supports boolean answers, the response would be boolean. The returned granularity would be stated in the response header; thus the Beacon client can check if the received granularity matches the requested one or if it has been adjusted by the queried Beacon.

#### i) Security aspects

Beacons can host data that is publicly and anonymously accessible, but also data that is considered sensitive by the data controllers. Some Beacons are allowed only to answer “Yes/no” queries, others are able to share summary or aggregated data, *e.g.*, counts of the results of a query, while others host anonymous data that could be shared openly. This difference of granularity in the response was added as a new requirement for Beacon v2.

Data controllers usually request that the people accessing the data identify themselves or, indeed, that they are explicitly authorised to access it. Beacon v2 must keep the security attributes in Beacon v1, including the identification (authentication) and authorization of users.

Some interviewed institutions said that they plan to use Beacon to share data among internal services. Users having access to their Beacon would have been identified and authorised previously elsewhere and, therefore, users reaching that Beacon should be considered fully trusted.

The Beacon v2 configuration declares which security access levels are supported. The options are public, registered and controlled (see Dyke *et al.* 2018 for further details), and a Beacon can support all or some of them.

The Beacon concept has been challenged by several authors (*Ayoş et al., 2021; Shringarpure & Bustamante 2015; von Thenen et al., 2018*) because, having the genome of a given person at hand, it would be possible with a systematic set of queries to a Beacon to determine if the person is part of that dataset. These privacy attack approaches are based on querying for alleles with a very low frequency and estimating the probability that such a combination of alleles is present in another person. In summary, the approach is similar to the microsatellite technique used in forensics to identify individuals with a very high probability.

Some authors (*Raisaro et al. 2017; Wan et al., 2017*) have suggested different mechanisms to alleviate that issue, for instance masking low frequency alleles or granting a number of queries to each user or removing a donor from the queryable set once a given number of positive hits for that donor has been reached.

Most of these techniques are complex to implement and only offer relative safety as there are mechanisms to “avoid” such countermeasures, *e.g.*, like querying from different computers to multiply the available budget. Additionally, the sensitivity of every allele is relative to the population it is compared with, so to say, the frequency in Europeans could be 0.1%, while the frequency in Icelandic population could be very different (higher or lower). In domains like rare diseases, the sensitivity of the information is very high but also the requirements of having as much details as possible about a case. On the contrary, in oncology there are many passenger variations and, many times, the ones of interest are the ones that are common enough to enable available treatments.

The Beacon v2 specification doesn't need to include any of the suggested countermeasures, as the approaches that have been used in real world Beacons are simple: Public Beacons tend to have only aggregated data or knowledge base information (*e.g.*, annotations on variations) and Beacons with sensitive information require that the user identifies itself and to sign a declaration of fair behaviour. The assumption is that the features provided in the specification should allow Beacon implementers to adjust the security to their specific audience and sensitivity of data.

j) Data use conditions

Once a dataset is discovered, the next question is “Is my project, institution or myself entitled to access the source dataset?” Including details about the conditions for data access by using a common set of terms is also a common request. GA4GH Data Use Ontology (Lawson *et al.* 2021) could fulfil this request. Beacon v1 included a previous solution named Consent Codes (Dyke *et al.* 2016), which has been superseded by DUO. Beacon v2 focuses on DUO for describing the data use conditions of hosted datasets. Lawson *et al.* (2021) describe the DUO standard and a further discussion could be found in the corresponding section below.

### k) Integration in networks

If Beacon v1 success was based on its simplicity, another crucial element was the existence of Beacon networks that allow querying many Beacons at once from a single point. Many users implicitly include the idea of Beacon networks when they show interest in hosting a Beacon (to be part of one or more networks) or when they show interest in using Beacons (by efficiently querying them through a network). Hence, any new version of the Beacon protocol must include the network aspect as part of the requirements.

The Beacon v2 specification includes many features that are conceived to make it easy for version 2 Beacons to integrate in networks, while allowing for the flexibility described in the previous sections. Examples of such features are the endpoints **/configuration** that describes aspects like the entry types (entities) implemented in a given Beacon instance, **/map** that describes the relationship between these entry types and the URLs to use for querying them, and **/filtering\_terms** that list the terms supported for filtering the results to provide.

These endpoints are designed to provide details about a Beacon instance to any client, which could be a web or a command line or, also, a Beacon Network service. That Beacon Network service could use the information provided by these endpoints to understand each Beacon and react accordingly to that.

As an example, let’s imagine a Beacon Network exists that is linked to 3 different Beacons:

- Beacon A only supports boolean queries about genomic variations, like a Beacon v1

- Beacon B supports detailed information about individuals, samples and genomic variations
- Beacon C accepts count requests about individuals but doesn't include data about genomic variations

Let's suppose that a user is interested in individuals having macular degeneration and posts the following query to the Beacon network: *"Tell me how many individuals have been diagnosed with macular degeneration"*.

As the Beacon Network knows that only Beacon B and C support individual information, and that both accept count requests, it could send that query just to them, avoiding Beacon A.

This simple example could be extended for authentication, support for a given ontology, etc. The Beacon service in the example could also know if a Beacon instance requires user authentication, and avoid sending them requests from anonymous users. Also the Beacon Network service can check if a term like "macular degeneration" is among the ones understood by Beacon B and C and skip those Beacons that will not understand that term.

Intelligence in Beacon Network services could go further and they could act as on-the-fly harmonisation services. Let's assume that both Beacon B and Beacon C understand the term "age-related macular degeneration", but Beacon B uses the MONDO Ontology (MONDO:0005150<sup>20</sup>) while Beacon C implements the NCIT ontology (NCIT:C84391<sup>21</sup>). Given that Beacon v2 specification recommends using ontology terms for querying, using MONDO:0005150 or NCIT:C84391 as query would only return results from Beacon B or C, respectively. However, as the Beacon Network service can harvest supported ontology terms from each Beacon instance on the network, it could leverage a service like EBI's Ontology Lookup Service<sup>22</sup> (OLS) to find synonyms among the lists (hence discovering

---

<sup>20</sup>

[https://www.ebi.ac.uk/ols/ontologies/mondo/terms?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FMONDO\\_0005150&lang=en&viewMode=PreferredRoots&siblings=false](https://www.ebi.ac.uk/ols/ontologies/mondo/terms?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FMONDO_0005150&lang=en&viewMode=PreferredRoots&siblings=false)

<sup>21</sup>

[https://www.ebi.ac.uk/ols/ontologies/ncit/terms?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FNCIT\\_C84391](https://www.ebi.ac.uk/ols/ontologies/ncit/terms?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FNCIT_C84391)

<sup>22</sup> [Ontology Lookup Service < EMBL-EBI](#)

that both ontology terms above are considered exact matches) and using this list to send the appropriate query to the corresponding Beacon without asking the user to provide all synonyms itself. In the example, the Beacon Network service, upon receiving a request on MONDO:0005150, will send that to Beacon B, but it will use the synonym (NCIT:C84391) for requesting data to Beacon C.

## l) The Beacon version 2 Reference Implementation

Rueda *et al* (2022) describe the Beacon version 2 Reference Implementation (B2RI) features, explaining how the B2RI contributes to goals 2a, 2b, 2c and 3. The main contribution of Beacon to this doctoral thesis goals is the specification approved by the GA4GH. However, the success of Beacon depends on actual Beacon deployments happening, and, for that, having software solutions readily available, easy to deploy and to configure is key. The Beacon version 2 Reference Implementation aims to be the first of such solutions.

The B2RI contributes to goal 2a (discovering donors and cohorts) by implementing the `/cohorts` and the `/individuals` endpoints described in the Bv2 specification. The discovery of donors and cohorts could also start from the `/biosamples` and the `/genomic_variations` endpoints, and then looking for the individuals related to such biosamples or variations. Both of them are also implemented in the B2RI.

The contribution to goal 2b (centralised or distributed discovery) is inherent to the implementation itself. For centralised discovery, several data providers could share their data through a common Beacon and keep the data separated by leveraging the cohorts or the datasets elements in Beacon.

Distributed discovery is based on independent Beacons being part of a network of Beacons that dispatches the request to all of them. B2RI implements all the endpoints that could be leveraged in a network: `/info`, `/service-info`, `/configuration`, `/map`, `/entry_types` and `/filtering_terms`.

The first version of the B2RI implements the whole Beacon v2 Model, hence contributing to harmonisation, but is limited in its contribution to harmonising by using several ontologies for filtering terms. This feature would be part of a future release.

### 4.3 Data Sharing

#### a) The Global Alliance for Genomics and Health

Rehm *et al.* (2021) introduce the Global Alliance for Genomics and Health (GA4GH) as an organisation that was created in 2013 to bring the benefits of genomics in health to all of humanity. It states that without the GA4GH or a similar initiative these benefits could eventually reach most people, but in a slower, more expensive and fragmented way. GA4GH would catalyse the process by harmonised data aggregation and federated approaches.

Rehm *et al.* (2021) also describe the GA4GH organisation in Workstreams and Driver Projects and the technical standards process until their approval. This is relevant for the Beacon v1 and Beacon v2 topics discussed in this doctoral thesis. GA4GH aims to be aligned with other standard organisations, in particular with those focused on the biomedical and healthcare domains, like HL7, and the aim to support and be interoperable with popular ontologies, like OMOP or the Human Phenotype Ontology (HPO), in order to avoid increasing unnecessarily the number of options for the community. Rehm *et al.* (2021) remind us that an excess of very similar standards does not help the community but, on the contrary, makes interoperability (or harmonisation) harder by increasing the fragmentation.

The lack of standardised and rich phenoclinical (phenotype and clinical course) data in "research genomes" is a clear limitation for reusing such data for genomic medicine. Hence, there is an increased need to make genomic data generated in healthcare available to trusted partners and the community. This limitation has been one of the motivations for moving from Beacon v1 to Beacon v2 as described in another section of this doctoral thesis.

Rehm *et al.* (2021) detail the challenges for the four typical disease areas (rare diseases, oncology, common or chronic diseases and infectious diseases) in genomic healthcare and how GA4GH products (like Beacon v2 and DUO) could help. While describing it, the importance of data sharing for each area is highlighted. By citing some products like Beacon in all the areas, Rehm is implicitly saying that these products are transversal and not limited to some of the areas.



Other challenges exist in the healthcare practice to make data generated in healthcare to be analysed, shared and returned back as practice improvements. Many data are not properly structured, are hosted in disparate systems, are written in multiple languages, etc. Ethical, Legal and Societal Issues (ELSI) must be included in the process as the data sharing process starts by getting the approval of data owners or donors for bringing such data in secondary use. The contributions of this doctoral thesis to the secure sharing of genomic human data for precision medicine does not address these later aspects, but focuses on posterior steps of the data sharing process.

## b) The Data Sharing process

*Data Sharing* is the set of processes and tools by which actual access to the data is provided and, when applicable, the intermediate steps to request and receive granted access are cleared.

Data sharing of human sensitive data is, in most cases, a long and paperwork-based process, which can become a very complex issue for many researchers. Typically, researchers must apply to get access to the data. They must provide details about themselves and their hosting institution, a description of the project, and must abide by some conditions on authorship sharing, data destruction after approval expiry, limitations in sharing with team members, and potential embargoes. Many of these requirements are due to the clauses included in the consent form signed by data donors.

The success or failure of these data access applications depends on the purpose and context of the project being among the ones permitted by the data owner and accepted by the data donors. In many cases, the conditions are not met and the application is denied. For researchers, getting and understanding the application documentation, compiling all the required information and signed documents (like the IRB approval or institutional endorsement) and following all the required steps could be very frustrating if the application is denied because a simple condition is not met. Having in advance a clear understanding of the data use conditions for a given dataset could avoid engaging in probably unsuccessful applications. An example of clear conditions is the Creative Commons approach<sup>23</sup> which has a limited set of options, identified by acronyms and logos;

---

<sup>23</sup> <https://creativecommons.org/licenses/>

once the users are familiar with them, it is very easy to understand the terms of usage for assets like images, sounds, etc.

### c) Data Use Ontology

Lawson *et al.* (2018) describe the context under which Data Use Ontology (DUO) was conceived (summarised in the previous paragraphs), the design, the places and counts of datasets where it is implemented and use cases. According to Lawson *et al.* (2018), the DUO is applied in three use cases:

- 1) In repositories, as metadata associated to the datasets to describe the data use conditions
- 2) To partially automate the data access request review process. By describing the application in DUO terms, they could be more easily compared with the DUO terms that should be honoured. That approach implies that the applicant describes their context in DUO terms or that a user interface guides the applicant through the process.
- 3) Including the DUO terms from the beginning of the process; in the consent form signed by the donors. This solution avoids a posterior translation of the textual consent form clauses to the DUO terms, bringing consistency to the whole process.

DUO is machine readable but less comprehensive than other approaches like Automatable Discovery and Access Matrix (ADA-M). This is an acceptable trade-off that brings simplicity at the cost of comprehensiveness or expressivity.

### d) Registered access level

Another characteristic of human sensitive data sharing is that many projects share some summary information (like allele frequencies in broad populations) but require a complete application to get access to any further detail. Dyke *et al.* (2018) describe that scenario: *open access (public)* and *controlled (restricted)* access as the main options, and introduces a third level – *registered access*– which, similarly to DUO, has the virtue of broadening the options while keeping it simple.

Most human sensitive data require a tight control on to whom that sensitive data is shared, in particular if details about the individual donors would be disclosed. Most projects, however, could share very useful information for the community (like allele frequencies in specific populations or cohorts) without the need to disclose details at individual level. Registered access aims to define a model where, in addition to *anonymous* or *identified and specifically authorised* users, there is a third homogenous level: identified users without specific authorization. Instead of opening the scenario to many different combinations of requirements to get access to some additional information, registered access reduces the problem for the data controllers to determine which data could be shared to this simply identified user.

## 4.5 Summary

As mentioned in the Objectives section, the goal of this thesis is to contribute to secure sharing of sensitive human genomic data for precision medicine. Given the nature of this work, it requires the participation of a broad international community that eventually demonstrates that our contributions are effective, and so the involvement in organisations that drive the design of the solutions has been a key component of this work.

To facilitate data management, our contribution is in the context of EGA, ELIXIR and GA4GH. EGA has evolved into the Federated EGA, with the support of ELIXIR and many of their members (the ELIXIR Nodes) and by adopting GA4GH products, which we have helped to design and adopt.

To facilitate data discovery and harmonisation, we proposed the Beacon specification, that is network aware, and also provided the Beacon Reference Implementation to facilitate a quick start.

To facilitate data sharing, in the initial steps of this doctoral thesis we have contributed to refine the three layered access level model (*i.e.*, public, registered and controlled access) and the Data Use Ontology (DUO) both of which help the data consumers to understand the context for obtaining access to the data.

The concepts, specifications and software developed for this thesis are open, and they now belong to the international scientific community, which, hopefully, will make them evolve so they keep being useful in the future

## 5. BIBLIOGRAPHY

(Ayoz *et al.*, 2021) Ayoz, K., Ayday, E. & Cicek, A. E. (2021). **Genome Reconstruction Attacks Against Genomic Data-Sharing Beacons.**

*Proceedings on Privacy Enhancing Technologies*, 2021(3), 28-48.

<https://doi.org/10.2478/popets-2021-0036>

(Dyke *et al.* 2016) Dyke, S. O. M., Philippakis, A. A., Rambla De Argila, J., Paltoo, D. N., Luetkemeier, E. S., Knoppers, B. M., Brookes, A. J., Spalding, J. D., Thompson, M., Roos, M., Boycott, K. M., Brudno, M., Hurles, M., Rehm, H. L., Matern, A., Fiume, M., & Sherry, S. T. (2016). **Consent Codes: Upholding standard data use conditions.** *PLOS Genetics*, 12(1), e1005772.

<https://doi.org/10.1371/journal.pgen.1005772>

(Dyke *et al.* 2018) Dyke, S. O. M., Linden, M., Lappalainen, I., de Argila, J. R., Carey, K., Lloyd, D., Spalding, J. D., Cabili, M. N., Kerry, G., Foreman, J., Cutts, T., Shabani, M., Rodriguez, L. L., Haeussler, M., Walsh, B., Jiang, X., Wang, S., Perrett, D., Boughtwood, T., . . . Flicek, P. (2018). **Registered access: authorizing data access.** *European Journal of Human Genetics*, 26(12), 1721–

1731. <https://doi.org/10.1038/s41431-018-0219-y>

(Fernández-Orth *et al.* 2022) Fernández-Orth, D., Rueda, M., Singh, B., Moldes, M., Jene, A., Ferri, M., Vasallo, C., Fromont, L. A., Navarro, A., & Rambla, J.

(2022). **A quality control portal for sequencing data deposited at the European genome–phenome archive.** *Briefings in Bioinformatics*, 23(3).

<https://doi.org/10.1093/bib/bbac136>

(Fielding 2000) Fielding, R. T. (2000). **Architectural Styles and the Design of Network-based Software Architectures** [Doctoral dissertation]. University of California, Irvine.

(Fiume *et al.* 2019) Fiume, M., Cupak, M., Keenan, S., Rambla, J., de la Torre, S., Dyke, S. O. M., Brookes, A. J., Carey, K., Lloyd, D., Goodhand, P., Haeussler, M., Baudis, M., Stockinger, H., Dolman, L., Lappalainen, I., Törnroos, J., Linden, M., Spalding, J. D., Ur-Rehman, S., . . . Scollen, S. (2019).

**Federated discovery and sharing of genomic data using Beacons.** *Nature Biotechnology*, 37(3), 220–224. <https://doi.org/10.1038/s41587-019-0046-x>

(Freeberg et al. 2021) Freeberg, M. A., Fromont, L. A., D’Altri, T., Romero, A. F., Ciges, J., Jene, A., Kerry, G., Moldes, M., Ariosa, R., Bahena, S., Barrowdale, D., Barbero, M., Fernandez-Orth, D., Garcia-Linares, C., Garcia-Rios, E., Haziza, F., Juhasz, B., Llobet, O., Milla, G., . . . Rambla, J. (2021). **The European Genome-phenome Archive in 2021.** *Nucleic Acids Research*, 50(D1), D980–D987. <https://doi.org/10.1093/nar/gkab1059>

(Harrow et al. 2021) Harrow, J., Hancock, J., Blomberg, N., Blomberg, N., Brunak, S., Capella-Gutierrez, S., Durinx, C., Evelo, C. T., Goble, C., Gut, I., Ison, J., Keane, T., Leskošek, B., Matyska, L., McEntyre, J., Miguel, C., Navarro, A., Newhouse, S., Nyrönen, T., Rambla, J., . . . Harrow, J. (2021). **ELIXIR-EXCELERATE: establishing Europe’s data infrastructure for the life science research of the future.** *The EMBO Journal*, 40(6). <https://doi.org/10.15252/emj.2020107409>

(Jacobsen et al. 2022) Jacobsen, J. O. B., Baudis, M., Baynam, G. S., Beckmann, J. S., Beltran, S., Buske, O. J., Callahan, T. J., Chute, C. G., Courtot, M., Danis, D., Elemento, O., Essenwanger, A., Freimuth, R. R., Gargano, M. A., Groza, T., Hamosh, A., Harris, N. L., Kaliyaperumal, R., Lloyd, K. C. K., . . . Robinson, P. N. (2022). **The GA4GH Phenopacket schema defines a computable representation of clinical data.** *Nature Biotechnology*, 40(6), 817-820. <https://doi.org/10.1038/s41587-022-01357-4>

(Lappalainen et al. 2015) Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J. D., ur-Rehman, S., Saunders, G., Kandasamy, J., Caccamo, M., Leinonen, R., Vaughan, B., Laurent, T., Rowland, F., Marin-Garcia, P., Barker, J., Jokinen, P., Torres, A. C., de Argila, J. R., Llobet, O. M., . . . Flicek, P. (2015). **The European Genome-phenome Archive of human data consented for biomedical research.** *Nature Genetics*, 47(7), 692–695. <https://doi.org/10.1038/ng.3312>

(Laurie et al. 2022) Laurie, S., Piscia, D., Matalonga, L., Corvo, A., Garcia, C., Fernandez-Callejo, M., Hernandez, C., Luengo, C., Ntalis, A. P., Protassio, J.,

Martinez, I., Pico, D., Thompson, R., Tonda, R., Bayes, M., Bullich, G., Camps, J., Paramonov, I., Trotta, J., . . . Beltran, S. (2022). **The RD-Connect Genome-Phenome Analysis Platform: Accelerating diagnosis, research, and gene discovery for rare diseases.** *Human Mutation*.

<https://doi.org/10.1002/humu.24353>

(Lawson et al. 2021) Lawson, J., Cabili, M. N., Kerry, G., Boughtwood, T., Thorogood, A., Alper, P., Bowers, S. R., Boyles, R. R., Brookes, A. J., Brush, M., Burdett, T., Clissold, H., Donnelly, S., Dyke, S. O., Freeberg, M. A., Haendel, M. A., Hata, C., Holub, P., Jeanson, F., Rambla, J., . . . Courtot, M. (2021). **The Data Use Ontology to streamline responsible access to human biomedical datasets.** *Cell Genomics*, 1(2), 100028.

<https://doi.org/10.1016/j.xgen.2021.100028>

(Raisaro et al. 2017) Raisaro, J. L., Tramèr, F., Ji, Z., Bu, D., Zhao, Y., Carey, K., Lloyd, D., Sofia, H., Baker, D., Flicek, P., Shringarpure, S., Bustamante, C., Wang, S., Jiang, X., Ohno-Machado, L., Tang, H., Wang, X. & Hubaux, J. P. (2017). **Addressing Beacon re-identification attacks: quantification and mitigation of privacy risks.** *Journal of the American Medical Informatics Association*, 24(4), 799-805. <https://doi.org/10.1093/jamia/ocw167>

(Rambla et al. 2022) Rambla, J., Baudis, M., Ariosa, R., Beck, T., Fromont, L. A., Navarro, A., Paloots, R., Rueda, M., Saunders, G., Singh, B., Spalding, J. D., Törnroos, J., Vasallo, C., Veal, C. D., & Brookes, A. J. (2022). **Beacon v2 and Beacon networks: A “lingua franca” for federated data discovery in biomedical genomics, and beyond.** *Human Mutation*, 43, 791-799.

<https://doi.org/10.1002/humu.24369>

(Rehm et al. 2021) Rehm, H. L., Page, A. J., Smith, L., Adams, J. B., Alterovitz, G., Babb, L. J., Barkley, M. P., Baudis, M., Beauvais, M. J., Beck, T., Beckmann, J. S., Beltran, S., Bernick, D., Bernier, A., Bonfield, J. K., Boughtwood, T. F., Bourque, G., Bowers, S. R., Brookes, A. J., . . . Birney, E. (2021). **GA4GH: International policies and standards for data sharing across genomic research and healthcare.** *Cell Genomics*, 1(2), 100029.

<https://doi.org/10.1016/j.xgen.2021.100029>

(Rueda et al. 2022) Rueda, M., Ariosa, R., Moldes, M. & Rambla, J. (2022). **Beacon v2 Reference Implementation: a toolkit to enable federated discovery of genomic and phenotypic data.** *Bioinformatics*, Volume 38, Issue 19, 1 October 2022, Pages 4656–4657, <https://doi.org/10.1093/bioinformatics/btac568>

(Saunders et al. 2019) Saunders, G., Baudis, M., Becker, R., Beltran, S., Bérout, C., Birney, E., Brooksbank, C., Brunak, S., van den Bulcke, M., Drysdale, R., Capella-Gutierrez, S., Flicek, P., Florindi, F., Goodhand, P., Gut, I., Heringa, J., Holub, P., Hooyberghs, J., Juty, N., . . . Scollen, S. (2019). **Leveraging European infrastructures to access 1 million human genomes by 2022.** *Nature Reviews Genetics*, 20(11), 693–701. <https://doi.org/10.1038/s41576-019-0156-9>

(Shringarpure & Bustamante 2015) Shringarpure, S. & Bustamante, C. (2015). **Privacy risks from genomic data-sharing Beacons.** *The American Journal of Human Genetics*, 97(5), 631-646. <https://doi.org/10.1016/j.ajhg.2015.09.010>

(Thorogood et al. 2021) Thorogood, A., Rehm, H. L., Goodhand, P., Page, A. J., Joly, Y., Baudis, M., Rambla, J., Navarro, A., Nyronen, T. H., Linden, M., Dove, E. S., Fiume, M., Brudno, M., Cline, M. S., & Birney, E. (2021). **International federation of genomic medicine databases using GA4GH standards.** *Cell Genomics*, 1(2), 100032. <https://doi.org/10.1016/j.xgen.2021.100032>

(Voisin et al. 2021) Voisin, C., Linden, M., Dyke, S. O., Bowers, S. R., Alper, P., Barkley, M. P., Bernick, D., Chao, J., Courtot, M., Jeanson, F., Konopko, M. A., Kuba, M., Lawson, J., Leinonen, J., Li, S., Ota Wang, V., Philippakis, A. A., Reinold, K., Rushton, G. A., . . . Nyrönen, T. H. (2021). **GA4GH Passport standard for digital identity and access permissions.** *Cell Genomics*, 1(2), 100030. <https://doi.org/10.1016/j.xgen.2021.100030>

(von Thenen et al., 2018) von Thenen, N., Ayday, E. & Cicek, A. E. (2018). **Re-identification of individuals in genomic data-sharing beacons via allele inference.** *Bioinformatics*, 35(3), 365-371. <https://doi.org/10.1093/bioinformatics/bty643>



(Wan et al., 2017) Wan, Z., Vorobeychik, Y., Kantarcioglu, M. & Malin, B. (2017). **Controlling the signal: Practical privacy protection of genomic data sharing through Beacon services.** *BMC Medical Genomics*, 10(S2).

<https://doi.org/10.1186/s12920-017-0282-1>



## 6. ANNEX 1

### 6.1 Details on the GA4GH Beacon version 2 specification

This annex is an extended alternative version of the section *GA4GH Beacon version 2 specification*, therefore some paragraphs between the additional details are almost identical to the ones that are part of that section.

#### a) Keep the simplicity of Beacon queries

The Beacon v1 specification describes an HTTP API, which is an API (*Application Programming Interface*) that relies on the HTTP (*HyperText Transfer Protocol*) protocol for sending the requests and returning the responses. An HTTP API offers a set of URLs (like **/search** or **/query**), or *endpoints*, that returns an HTTP document encoded in XML, JSON, or plain text as indicated in the HTTP request header. In Beacon v1, the response was encoded in JSON (*JavaScript Object Notation*).

HTTP requests accept parameters in the URL, like

**/search?term=parokeet&language=catalan**, which include conditions to consider when the server resolves the request (*e.g.*, the language for the returning results).

A key value proposition of Beacon was the simplicity of Beacon v1 queries: having just two API endpoints and half a dozen of parameters, understanding how to perform queries was very easy. Beacon v2 has been also designed to keep simplicity as a key value, although taking into account that it must cover a much broader set of queries and concepts.

Beacon v2 relies in two principles for keeping the simplicity:

- 1) Using the REST API (*REpresentational State Transfer Application Programming Interface*) approach instead of an *ad hoc* API, as REST brings consistency between queries
- 2) Maintaining the complexity of the expressions to a minimum by avoiding to define a syntax for them.

#### b) REST API

The REST approach is based on building the APIs endpoints according to the entities represented by a given solution. As an example, an API for a library could include the following endpoints: **/books**, **/users** or **/shelves**

The endpoint **/books** will return a list of all books in that library.

To get the list of books from a given author, parameters like **author** could be defined, *e.g.*,

**/books?author=James Joyce**

Getting the details of a given book, is as simple of adding the id of the book in the URL path, *e.g.*,

**/books/1234**

Similarly, related entities could be requested by using endpoints like,

**/books/1234/users**

That will return the list of users that have borrowed that book.

The previous examples show consistency in the mechanics, which allows users to infer which would be the way to query a given service.

Applying the principle to Beacon, getting the list of people in a Beacon would be possible by using the **/individuals** endpoint.

To get details on an specific individual (the one with id-123) we can use

**/individuals/123** and to get its samples **/individuals/123/biosamples**

As it is not mandatory that every Beacon v2 compatible instance implements *every* single endpoint defined in the Beacon v2 model, the Beacon includes an specific endpoint (**/entry\_types**) that informs the clients about which ones are available in that particular instance.

c) Maintaining the complexity of the expressions to a minimum

Query languages, like SQL, are very powerful because they provide a rich set of commands (or clauses) and because they support advanced expressions, with different boolean operators, a set of functions to handle strings, dates, numbers, etc.

While gathering requirements for Beacon v2, it was clear that Beacon users' needs for querying were simple. The queries they describe are like: “*show me all the men, older than 50 years, diagnosed with psoriasis*” or “*give me all individuals that have the V600E variant in BRAF and have African ancestry*”.

Analysing the queries above, we observed that the vast majority of conditions could be combined *just* using the AND boolean operator. For instance, for the first example written in SQL against an imaginary database:

```
SELECT * FROM INDIVIDUALS WHERE
gender="male" AND
age > 50 AND
diagnose = "psoriasis"
```

This simplification allows for leveraging the URL query parameters like:

```
/individuals?gender=male&age=>50&diagnose=psoriasis
```

By limiting the available boolean operators in Beacon v2 to AND, it would not be possible to *directly* perform queries like “*show me all the men, older than 50 years, or the women younger than 35 years, diagnosed with psoriasis*”. Note that for solving expressions like this one, it must be clear if “diagnosed with psoriasis” refers to both men and women or only to women. This is usually solved by using parentheses in the expression like:

```
SELECT * FROM INDIVIDUALS WHERE
((gender="male" AND
age > 50)OR (gender="female" AND
age < 35))
AND diagnose = "psoriasis"
```

Learning how to combine boolean operators and parentheses is not easy for many users, which leads to repeated queries until the result is the desired one. Many users, indeed, tend to translate spoken language into boolean expressions, which confuses them notably. For instance, “*return all individuals diagnosed with psoriasis and arthritis*” should be expressed with an OR operator:

```
SELECT * FROM INDIVIDUAL_DISEASES WHERE  
diagnose = "psoriasis" OR diagnose = "arthritis"
```

For Beacon v2, we considered that including non-basic queries could seriously tamper the “keep Beacon simple” goal, hence, another approach was preferred for non-basic queries, that rely on the user sending more than one query, if necessary, and combining the results on her/his own.

Beacon v2 is a *discovery* protocol, not an *analysis* tool, therefore it should not be used to get counts and compare the counts or distributions using statistical tests but to look carefully at the results and refine or extend the query accordingly. Using that consideration, we expect users looking for cases like the above example to perform two queries, one for males and one for females, which could be seen as more queries but also as an easier way to clearly disambiguate doubts such as if psoriasis applies only to women or both men and women. In the former case, the two queries would be:

```
/individuals?gender=male&age=>50&diagnose=psoriasis  
/individuals?gender=female&age=<35&diagnose=psoriasis
```

While in the latter they would be:

```
/individuals?gender=male&age=>50  
/individuals?gender=female&age=<35&diagnose=psoriasis
```

Of course, this approach is not optimal for all the use cases, in particular when the expressions could still be simple, such as “*return all individuals diagnosed with psoriasis and arthritis*”. In cases like that, we could leverage the flexibility of URL parameters that allow queries like

```
/individuals?diagnose=psoriasis,arthritis
```

The Beacon v2 specification itself is not detailing options like that, but it could be leveraged by implementations without breaking compatibility with the specification.

Queries like “show me the genomic variations from men older than 50 years having melanoma or colorectal cancer, but not present in women having melanoma” constitute another level of complexity. Indeed, in languages like SQL, a query like that is too difficult for any user that is not very experienced in the language.

#### d) Keep the simplicity of Beacon responses

Beacon v1 response was very easy to build and understand: simply a “Yes” or “No”, This is also one of Beacon v1 weaknesses: the lack of context to the response... “How many times has the variant been observed?”, “in somatic cells only or also in germline cells?”. Beacon v2 tries to solve the context issue by allowing richer queries, and by providing a much richer response.

However, sharing more details in the response introduces two concerns: 1) some Beacon instances are not allowed or prefer not to share more details and keep the “Yes/No” response, and 2) a richer response implies more software development effort to implement such a response.

As many interviewed projects declare that they are fine also with sharing a count in addition to the boolean response, we end up with the scenario of supporting three different levels of response in Beacon v2: boolean, count and details. For such purpose, Beacon v2 specification includes an attribute in the requests and in the responses: “granularity”.

A Beacon instance can support all three levels of granularity and apply one or another to every received request, depending on the sensitivity of the endpoint (e.g., cohorts vs individuals) and/or on the user being authenticated and/or authorised.

Every Beacon instance declares the default level of granularity, the one to be applied if no specific granularity is requested, in their **/configuration** endpoint. Every Beacon request includes the requested granularity and every Beacon response includes the returned granularity. Everything being allowed, the granularity response would match the granularity requested. However, if a Beacon instance receives a request for details (details granularity) and that request is not allowed because the user is not authenticated or authorised or that Beacon only supports boolean answers, the response would be boolean. Returned granularity would be stated in the response header, thus the Beacon client can check if the received granularity matches the requested one or if it has been adjusted by the queried Beacon.





## 7. ANNEX 2



## Annex Paper 1

Fiume, M., Cupak, M., Keenan, S., Rambla, J., de la Torre, S., Dyke, S. O. M., Brookes, A. J., Carey, K., Lloyd, D., Goodhand, P., Haeussler, M., Baudis, M., Stockinger, H., Dolman, L., Lappalainen, I., Törnroos, J., Linden, M., Spalding, J. D., Ur-Rehman, S., . . . Scollen, S. (2019). **Federated discovery and sharing of genomic data using Beacons**. *Nature Biotechnology*, 37(3), 220–224. <https://doi.org/10.1038/s41587-019-0046-x>



which should henceforth be encouraged. Indeed, the US congressional investigation on shadow pricing was spurred largely by coordinated action by patients made aware of pricing discrepancies through research published in a reputable medical journal<sup>9</sup>. Along these lines, we would argue that technological advancement in analytical tools and pricing analytics will be increasingly relevant under the current legal frameworks to establish direct communication and negotiation between payers and drug companies and to induce further transparency in the pricing process—as a major technology company sought to do last year<sup>14</sup>. □

Anurag S Rathore<sup>1</sup> and Faheem Shaheef  
Indian Institute of Technology Delhi, Department of  
Chemical Engineering, Hauz Khas, New Delhi, India.  
\*e-mail: asrathore@biotechcmz.com

Published online: 26 February 2019  
<https://doi.org/10.1038/s41587-019-0049-7>

#### References

1. Pharmaceutical Research and Manufacturers of America (PhRMA). Prescription medicines: costs in context. *Advocacy: Cost & Value of Medicines* <https://www.phrma.org/report/prescription-medicines-costs-in-context> (2016).
2. Greenwood, J. U.S. drug costs must be weighed against benefits. *Bloomberg News* <http://www.bloomberg.com/news/articles/2015-12-28/u-s-drug-costs-must-be-weighed-against-benefits> (2015).
3. Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. *Nat. Rev. Drug Discov.* 11, 191–200 (2012).
4. Brennan, Z. Congressmen investigate 'shadow pricing' of MS drugs. *Regulatory Focus* <https://www.raps.org/regulatory-focus/E2%84%A2/news-articles/2017/8/congressmen-investigate-shadow-pricing-of-ms-drugs> (2017).
5. Kelley, T. U.S. insulin prices rise as Sanofi, Novo await rivals. *Bloomberg News* <https://www.bloomberg.com/news/articles/2013-08-15/u-s-insulin-prices-rise-as-sanofi-novo-await-rivals> (2013).
6. Langreth, R. Hot drugs show sharp price hikes in shadow market. *Bloomberg News* <https://www.bloomberg.com/news/articles/2015-05-06/diabetes-drugs-compete-with-prices-that-rise-in-lockstep> (2015).
7. Barrett, P. & Langreth, R. The crazy math behind drug prices. *Bloomberg Businessweek* <https://www.bloomberg.com/news/articles/2017-06-29/the-crazy-math-behind-drug-prices> (2017).
8. Persistence Market Research. *Biosimilar Market: Pricing Analysis 2006–2017* (Persistence Market Research, New York, 2017).
9. Hartung, D. M., Bourdette, D. N., Ahmed, S. M. & Whitham, R. H. *Neurology* 84, 2185–2192 (2015).
10. Benmeleh, Y. FDA deals blow to Teva's defense plan with ruling on Bendeke. *Bloomberg News* <http://www.bloomberg.com/news/articles/2016-03-29/fda-deals-blow-to-teva-s-defense-plan-with-ruling-on-bendeke> (2016).
11. Serebrov, M. PBMs: The 'shadow' players in the drug pricing skirmish? *BioWorld* <http://www.bioworld.com/content/pbms-shadow-players-drug-pricing-skirmish-0> (2016).
12. Tufts Center for the Study of Drug Development (CSDD). *Cost of Developing a New Drug* (Tufts CSDD, Boston, 2014).
13. Loftus, P. U.S. investigates drugmaker contracts with pharmacy-benefit managers. *The Wall Street Journal* <https://www.wsj.com/articles/u-s-investigates-drugmaker-contracts-with-pharmacy-benefit-managers-1462895700> (2016).
14. Wingfield, N., Thomas, K. & Abelson, R. Amazon, Berkshire Hathaway and JPMorgan team up to try to disrupt health care. *The New York Times* <https://www.nytimes.com/2018/01/30/technology/amazon-berkshire-hathaway-jpmorgan-health-care.html> (2018).

#### Competing interests

The authors declare no competing interests.

Corrected: Publisher Correction

OPEN

# Federated discovery and sharing of genomic data using Beacons

**To the Editor** — The Beacon Project (<https://github.com/ga4gh-beacon/>) is a Global Alliance for Genomics & Health (GA4GH)<sup>1</sup> initiative that enables genomic and clinical data sharing across federated networks. The project is working toward developing regulatory, ethics and security guidance to ensure proportionate safeguards for distribution of data according to the GA4GH-developed “Framework for Responsible Sharing of Genomic and Health-Related Data”<sup>2</sup>. Here we describe the Beacon protocol and how it can be used as a model for the federated discovery and sharing of genomic data.

A Beacon is defined as a web-accessible service that can be queried for information about a specific allele. A user of a Beacon can pose queries of the form “Have you observed this nucleotide (e.g., C) at this genomic location (e.g., position 32,936,732 on chromosome 13)?” to which the Beacon responds with either “yes” or “no.” In this way, a Beacon allows allelic information of interest to be discovered by a remote searcher with no reference to a specific sample or patient, thereby mitigating privacy risks.

In principle, allelic information from any source (or species) can be distributed through a Beacon. For example, a Beacon may serve data from case-level observations, such as genetic variants identified from sequenced

samples, or from annotation resources such as variant–disease associations curated from scientific literature. Along with a “yes” response, a Beacon may optionally disclose metadata, including allele frequencies, pathogenicity scores and associated phenotypes, associated with the queried allele. Access to Beacons is securable through institutional systems for authentication and authorization (for example, ELIXIR AAI), allowing hosts to enforce proportionate safeguards for datasets that may be sensitive and consented for use only by trusted individuals and/or for specific purposes.

The Beacon Project is demonstrating the willingness of international organizations to work together to define standards for, and actively engage in, genomic data sharing. Several organizations have ‘lit’ (i.e., implemented) a Beacon, and these have been assembled into a single searchable network. In the years since the project's inception, over 100 Beacons have been lit by 40 organizations serving over 200 datasets. The datasets served through Beacons are searchable individually or in aggregate—for instance, via the Beacon Network (<https://beacon-network.org>), a federated search engine across the world's beacons.

Beacons are a general-purpose protocol for genomics data discovery and have been lit by both large and small organizations,

as well as by individuals. This has made available datasets collected from large-scale population sequencing efforts (for example, 1000 Genomes)<sup>3</sup>, clinical diagnostic settings, in silico predictions (for example, PolyPhen-2)<sup>4</sup>, expertly curated or crowd-sourced databases, scientific literature (for example, the Human Genome Mutation Database)<sup>5</sup> and variant curation efforts (for example, ClinVar)<sup>6</sup>. The International Cancer Genome Consortium<sup>7</sup> Beacon shares case-level somatic variant observations from over 60 cancer subtypes; the PhenomeCentral<sup>8</sup> Beacon shares observations from hundreds of clinical cases of undiagnosed and rare genetic diseases; and the BRCA Exchange (<https://brcaexchange.org/>) Beacon distributes consensus classifications for variants in *BRCA1* and *BRCA2* cataloged by the ENIGMA Consortium<sup>9</sup>, as well as variants collected from other resources as part of the GA4GH BRCA Exchange (<https://brcaexchange.org/>). The ELIXIR hub (<https://elixir-europe.org/>) is also integrating Beacon to connect geographically distributed data centers and unify their data access methodologies. This will enable aggregate sharing of allelic observations between sites, a feature that is not yet available through its services.

With continued adoption, Beacons will produce a large network of globally searchable genomics datasets that have the potential to unlock new genomics-derived discoveries and applications in medicine.

### Beacon protocol

Many former systems for genomic data sharing have followed a centralized model, wherein data generators deposit information into a single repository, such as the Sequence Read Archive (SRA)<sup>10</sup>. This model requires data generators to transfer whole copies of datasets over the internet, which will become inefficient and expensive as the rate of genomic data acquisition increases. An alternative, federated model for data sharing<sup>1</sup> requires organizations to host data independently and to interoperate via an agreed-upon technical language. This model removes the inefficiencies of large data transfers and gives host organizations more control over data privacy, security and representation.

For maximal interoperability, a Beacon is designed to be a communication layer that is compatible with any underlying representation of alleles or their annotations. For example, the GA4GH develops a data representation format for genomic variants and annotations, but in practice these data types may be stored in other formats as well (for example, VCF files or relational databases).

Sharing through Beacon is notably different from sharing fully descriptive data representations for genomic variants (for example, VCF) or annotations (for example, GFF). The Beacon protocol considers levels of data aggregation and obfuscation that can be added onto raw data representations (such as VCF) to convey useful information without explicitly referring to specific samples or individuals.

With these features in mind, the Beacon protocol was designed to be:

- Simple: Beacons can be implemented on top of any underlying variant or variant annotation data store.
- Federated: Beacons can be lit and maintained by individual organizations and assembled into a distributed network.
- General purpose: Beacons can be used to distribute any allelic dataset, including case-level observations or other annotations.
- Aggregative: Beacons provide a boolean answer to whether an allele was observed, possibly aggregated across an entire population, and therefore support deidentification in a way that sharing via VCF files does not.

- Securable: Beacon access can be restricted using institutional security protocols, and authorization schemes can be implemented to respect conditions consented to by patients and/or data owners.

The Beacon API (represented as a RESTful web application) provides a technical specification that a Beacon server must implement. The specification is open-source and available online at <https://github.com/ga4gh-beacon/specification>.

A Beacon has two available functions: the first lists information about the Beacon, including descriptions of the host organization and specific datasets that it serves; the second queries for the existence of information about specific alleles. Alleles are specified with chromosomal coordinates in addition to reference and alternate bases. Much as in their use in VCF, reference and alternative bases can be used together to specify exact matches for single nucleotide variants (SNVs) and small insertions or deletions. A Beacon responds either “yes” or “no” to signal whether the dataset(s) it serves have information about the queried allele. In the affirmative, a Beacon may optionally disclose metadata describing the observations or annotations associated with the queried allele. An example query and response is shown in Supplementary Fig. 1.

### Reference implementation

To simplify the process of lighting a Beacon, a free, open-source ‘reference implementation’ of the latest specification has been developed.

This implementation can create a public Beacon from a set of VCF files. It may be deployed locally or in a cloud-based environment maintained by a third-party provider (for example, Amazon, Google or Microsoft). Documentation and links to download and run the Beacon reference implementation are available (<https://github.com/ga4gh-beacon/>). Third-party organizations, such as Cafe Variome, DNASTack and the European Genome-phenome Archive (EGA), also support the ability to light Beacons from genetic variation datasets stored in those systems.

### Beacon security design

In principle, access to Beacons can be secured through any system of authentication or authorization, at the discretion of the host organization. The GA4GH is promoting different levels of data access (open, registered, and controlled) for convenience and for compatibility across its projects. Each so-called ‘access tier’ has distinct visibility and requirements for

authorization. For example, ‘open access’ Beacons are accessible to anonymous users of the internet, whereas ‘registered access’ Beacons are accessible to registered users (for example, bona fide researchers and clinicians) who have agreed to a set of conditions of data use<sup>11</sup>.

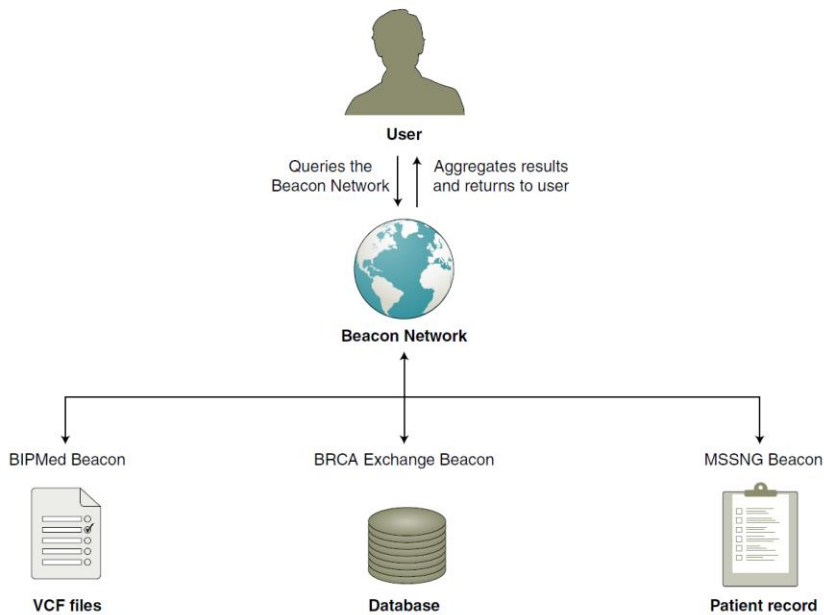
A Beacon may support one or more access tiers to provide progressive disclosure of increasingly sensitive information (for example, patient phenotypes and clinical information) as users pass through more stringent authentication and authorization checks. For example, tiered access makes it possible for organizations to allow anonymous users to discover the existence of an allelic observation, without the Beacon disclosing more information about it until users identify themselves. The ability for organizations to offer minimal data discovery up front can save substantial time and effort in data access applications when data might not contain relevant data points.

Beacon’s ability to reveal different information at specific access tiers affords genomic data stewards options for distributing allelic information, ranging from fully public to private. Access can be controlled using established authentication and authorization protocols (for example, OpenID Connect and OAuth2.0) to enforce proportionate safeguards for datasets that may be sensitive and/or consented for use only by trusted individuals for specific purposes.

### Attribute disclosure attacks and reidentification

The “yes” response from a Beacon signals the presence of an allele in a dataset comprising possibly many individuals’ genotypes, thereby mitigating risks associated with reidentifying specific individuals. Independent of their technical implementation, Beacon reidentification attempts require prior knowledge of genomic sequence data from the individual (or that of a close relative); they are arguably preceded by more harmful compromises to privacy. However, reidentification can pose additional risks if sensitive attributes about the individual can be inferred from Beacons (for example, HIV status or mental health condition). Such attacks have been characterized as “attribute disclosure attacks using DNA” (ADAD)<sup>12</sup>.

Querying a Beacon for many variants known to exist in a person’s genome could lead to confirmation of that person’s inclusion in a given database, potentially revealing sensitive information about that individual. The ability to reidentify individuals has been examined previously<sup>13</sup> and recently in the context of Beacons<sup>14</sup>. The power to reidentify an individual



**Fig. 1 | Beacon Network system architecture.** The user interacts with the Beacon Network system by asking for information about the existence of a particular genetic mutation. The Beacon Network federates the query across many Beacon instances serving various types of data, such as a variant database, VCF files or patient records. The Beacon Network collects the responses from Beacons and presents aggregated information to the user. BIPMed (<http://bipmed.org>), the Brazilian Initiative on Precision Medicine, is a population sequencing effort while MSSNG (<http://mss.ng>) collects sequence information from subjects with autism and their families.

whose genotypes are reflected through a Beacon depends on the number of individuals whose data is served, the allele frequency distribution of the pool, the scope of allowed queries (for example, exome versus genome), the type of DNA source (for example, normal tissue versus cancer sample) and the number of times a Beacon is queried. Models for population allele frequencies can be leveraged to reduce the number of queries required in such an attempt, but reidentification is still possible without using allele frequencies if a Beacon can be queried a large number of (for example, 10,000) times.

#### Risk mitigation schemes

User agreements, data use policies and technical enforcement of usage quotas can be established to limit the possibility of reidentification and ADAD through Beacons. Organizations are advised to specify terms of use that explicitly prohibit reidentification attempts through the service. When the risk of ADAD is considered too high for data to be distributed publicly, data stewards are encouraged to implement secured access. Compared with public-access tiers, secured-access tiers (either registered or controlled) impose extra social and/or

legal disincentives that can help prevent service misuse.

Beacon operators may further specify consent-based data use conditions from a structured set of Consent Codes to impose restrictions indicated by consent of research participants. These Consent Codes, which are general purpose and can be used by genomics data stewards, including Beacon operators, were designed with the purpose of supporting maximum data use and integration while respecting consent permissions<sup>15</sup>. The current set of Consent Codes is provided in Supplementary Table 1.

The ethical, legal and social status of health-related data that are typically considered sensitive in international policy and laws is being examined to provide guidance in aggregating Beacons and in implementing tiered protection of Beacon attributes based on sensitivity<sup>16</sup>. This guidance aims to enable consistent and proportionate provision of data protection for data that are considered more sensitive by individuals and society. Data stewards should consider the sensitivity of attributes used in describing their Beacons, as well as those in the data itself.

Technical provisions can also be used to reduce the statistical power of

reidentification attempts. Individual Beacons can be combined to form a single, aggregate Beacon, and direct access to participating Beacons can be blocked. Aggregate Beacons contain more data points than any of the individual Beacons while obscuring the origin of the data. As an example, a publicly accessible Beacon named Conglomerate has been lit as an aggregate of multiple independent Beacons.

An information budgeting approach can also be used to thwart reidentification attempts<sup>17</sup>, which rely on accumulating evidence from many queries for alleles carried by a specific individual. The power to reidentify an individual using this technique varies inversely with the frequency of the alleles being queried (i.e., very rare alleles are more revealing than common alleles). By metering the cumulative information disclosure for individuals, Beacons can be configured to restrict access before reidentification is possible within a desired level of statistical confidence.

Beacon is a general-purpose protocol for genomics data discovery, and as such can be used to distribute allelic information from various origins, including sequence observations from patients with known (for example, the International Cancer Genome Consortium)<sup>7</sup> or unknown (e.g., PhenomeCentral)<sup>8</sup> diseases, population studies (for example, 1000 Genomes)<sup>9</sup>, in silico predictions (for example, PolyPhen-2)<sup>4</sup>, expertly curated or crowd-sourced databases (for example, BRCA Exchange and ClinVar)<sup>6</sup>, and scientific literature (for example, the Human Genome Mutation Database)<sup>5</sup>. Additional Beacon implementations are ongoing in Europe, mainly through the ELIXIR Beacon project. The deployment of Beacons for select use cases is described below.

#### Matchmaking

A major obstacle to discovering the causes of rare diseases is sample size. A single affected family can be enough to identify one or more compelling candidate variants, but pinpointing causal genetic variants frequently requires examining unrelated cases with a variant in the same gene and similar phenotypic presentations. Recently, patient matchmaking has been formalized through efforts such as the Matchmaker Exchange (MME)<sup>18</sup>, in which users who contribute a case to a database within the federated network can find similar cases in other databases within the network.

MME is a secured-access system, requiring that only authorized databases and users can contribute and exchange patient profiles for matching. However, this inherently limits the discoverability of

the data, which may dissuade some users having candidate genes or variants they want to match. In addition to implementing the MME API<sup>19</sup> for patient matchmaking, several organizations within the MME have lit Beacons to serve aggregate views of their clinical datasets more publicly. This allows clinicians with candidate variants to quickly search for existing matches within the MME.

### Sequencing initiatives and archives

Large-scale sequencing initiatives, such as the 100,000 Genomes Project<sup>20</sup> conducted by Genomics England and the Precision Medicine Initiative<sup>21</sup>, promise to generate vast volumes of genotypic and associated health information. Data from these projects, once shared, help researchers make inferences on the genetic determinants of disease by way of comparative analysis and association studies.

The 1000 Genomes Project<sup>3</sup>, NHLBI Grand Opportunity Exome Sequence Project (<https://esp.gs.washington.edu/drupal/>), and Exome Aggregation Consortium<sup>22</sup> are exemplar large-scale initiatives that have shared genotypes from diverse populations through Beacons. As the number and scale of population sequencing efforts expand, a more accurate depiction of global sequence diversity will be available in aggregate through Beacons and the Beacon Network.

In addition, many of the largest genomic archives, such as dbGaP<sup>22</sup>, the European Genome-phenome Archive (<https://www.ebi.ac.uk/ega/home>) and the European Variation Archive (<http://www.ebi.ac.uk/eva>), have provided access to variation data through Beacons for some or all of their datasets. These Beacons collectively provide widespread discoverability across a large amount of data. Many of these resources are continually growing with new submissions and thus provide added value for data depositors by simplifying data distribution and unifying their consumption.

### Beacon Network

Beacon represents a simple protocol that, like internet protocols such as HTTP, describes a method for data discovery and exchange between distributed, collaborative systems. Toward developing an ‘internet for genomics’, it is useful to establish a network of protocol adopters and an efficient mechanism for searching across it.

The Beacon Network is a directory and search engine for Beacons. Although individual Beacons answer the question “Have you observed this allele?”, the Beacon Network answers the question “Who has observed this allele?”. The Beacon Network serves as a powerful, convenient and real-time genomic data distribution channel

through which users can discover the existence of alleles of interest and be directed to host organizations who have observed them. A schematic of the Beacon Network as a global federated network for genomic information discovery is shown in Fig. 1.

The Beacon Network is accessible either through its website or programmatically through an API, and enables fast, simultaneous search of hundreds of datasets from hundreds of thousands of individuals already served through Beacons worldwide.

Beacons can be freely registered to the Beacon Network and can be searched independently or in aggregate with other connected Beacons. The Beacon Network has received over 1.5 million queries in the three years since its launch. The value of datasets connected to the Beacon Network increases as more Beacons join, particularly for comparative applications like rare disease and donor matching.

### Conclusions and perspectives

The first version of the Beacon Project has validated the feasibility of a globally federated system for genomic data sharing. The conceptual and technical simplicity of the discovery question, “Have you observed this allele?”, enabled rapid and widespread adoption, and this has served to provide practical feedback for the GA4GH to continue to advance its best practices by holistically addressing regulatory, security and technical aspects of global genomics data sharing. However, the narrow focus of the initial Beacon question limits its utility to support other closely related use cases, and successive iterations of the protocol are planned to enable coverage of these.

Future extensions to the Beacon protocol may include the following:

- Support for discovering complex genomic alterations, including copy number variations (CNVs) and somatic copy number alterations (CNAs), which are major contributors to both inter-individual variation and disease susceptibility and prominent features of the oncogenomic mutation landscape;
- Integration of non-genomics data in queries, including the ability to discover similar cases on the basis of associated metadata;
- Support for quantitative attributes in responses (for example, allele frequencies) to facilitate statistical analyses that combine information disclosed through multiple Beacons;
- Handoff to services by which users may access additional information about a queried variant.

The development of data-rich extensions to the Beacon protocol will leverage the expertise of GA4GH members and stakeholders to iteratively design and evaluate the technical, privacy and security considerations in evolving Beacons to enable unprecedented access to genomics and clinical datasets through a global, federated ecosystem. □

*Editor’s note: This article has been peer-reviewed.*

Marc Fiume<sup>1\*</sup>, Miroslav Cupak<sup>1</sup>, Stephen Keenan<sup>2,3</sup>, Jordi Rambla<sup>4</sup>, Sabela de la Torre<sup>4</sup>, Stephanie O. M. Dyke<sup>5</sup>, Anthony J. Brookes<sup>6</sup>, Knox Carey<sup>7</sup>, David Lloyd<sup>8</sup>, Peter Goodhand<sup>2,9</sup>, Maximilian Haeussler<sup>10</sup>, Michael Baudis<sup>11,12</sup>, Heinz Stockinger<sup>12</sup>, Lena Dolman<sup>2,9</sup>, Ilkka Lappalainen<sup>3,13</sup>, Juha Törnroos<sup>13</sup>, Mikael Linden<sup>13</sup>, J. Dylan Spalding<sup>13</sup>, Saif Ur-Rehman<sup>3</sup>, Angela Page<sup>2,14</sup>, Paul Flicek<sup>15</sup>, Stephen Sherry<sup>16</sup>, David Haussler<sup>10</sup>, Susheel Varma<sup>15,8</sup>, Gary Saunders<sup>8</sup> and Serena Scollen<sup>8</sup>

<sup>1</sup>DNASTack, Toronto, Ontario, Canada. <sup>2</sup>Global Alliance for Genomics and Health, Toronto, Ontario, Canada. <sup>3</sup>European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK.

<sup>4</sup>Centre de Regulació Genòmica, Barcelona, Spain.

<sup>5</sup>Centre of Genomics and Policy, Department of Human Genetics, McGill University, Montreal, Quebec, Canada.

<sup>6</sup>Department of Genetics, University of Leicester, Leicester, UK.

<sup>7</sup>Genecloud, Sunnyvale, CA, USA.

<sup>8</sup>ELIXIR Hub, Wellcome Genome Campus, Hinxton, Cambridge, UK.

<sup>9</sup>Ontario Institute for Cancer Research, Toronto, Ontario, Canada.

<sup>10</sup>Genomics Institute, University of California at Santa Cruz, Santa Cruz, CA, USA.

<sup>11</sup>Department of Molecular Life Sciences, University of Zurich, Zurich, Switzerland.

<sup>12</sup>SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland.

<sup>13</sup>CSC – IT Center for Science Ltd, Espoo, Finland.

<sup>14</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA.

<sup>15</sup>Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK.

<sup>16</sup>National Center for Biotechnology Information, US National Library of Medicine, Bethesda, MD, USA.

\*e-mail: marc@dnastack.com

Published online: 4 March 2019

<https://doi.org/10.1038/s41587-019-0046-x>

### References

1. Global Alliance for Genomics and Health. *Science* 352, 1278–1280 (2016).
2. Knoppers, B. M. *HUGO J.* 8, 3 (2014).
3. 1000 Genomes Project Consortium. et al. *Nature* 526, 68–74 (2015).
4. Adzhubei, I. A. et al. *Nat. Methods* 7, 248–249 (2010).
5. Stenson, P. D. et al. *Hum. Genet.* 133, 1–9 (2014).
6. Landrum, M. J. et al. *Nucleic Acids Res.* 42, D980–D985 (2014).
7. International Cancer Genome Consortium. et al. *Nature* 464, 993–998 (2010).
8. Buske, O. J. et al. *Hum. Mutat.* 36, 931–940 (2015).
9. Spurdle, A. B. et al. *Hum. Mutat.* 33, 2–7 (2012).



10. Leinonen, R., Sugawara, H. & Shumway, M. & International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res* 39, D19–D21 (2011).
11. Dyke, S. O. M. et al. *Eur. J. Hum. Genet.* 24, 1676–1680 (2016).
12. Erlich, Y. & Narayanan, A. *Nat. Rev. Genet.* 15, 409–421 (2014).
13. Homer, N. et al. *PLoS Genet.* 4, e1000167 (2008).
14. Shringarpure, S. S. & Bustamante, C. D. *Am. J. Hum. Genet.* 97, 631–646 (2015).
15. Dyke, S. O. M. et al. *PLoS Genet.* 12, e1005772 (2016).
16. Dyke, S. O. M., Dove, E. S. & Knoppers, B. M. *Genomic Med.* 1, 16024 (2016).
17. Raisaro, J. L. et al. *J. Am. Med. Inform. Assoc.* 24, 799–805 (2017).
18. Philippakis, A. A. et al. *Hum. Mutat.* 36, 915–921 (2015).
19. Buske, O. J. et al. *Hum. Mutat.* 36, 922–927 (2015).
20. Peplow, M. *Br. Med. J.* 353, i1757 (2016).
21. Ashley, E. A. *J. Am. Med. Assoc.* 313, 2119–2120 (2015).
22. Mailman, M.D. et al. *Nat. Genet.* 39, 1181–1186 (2007).

#### Acknowledgements

J. Ostell conceived the project; Global Alliance for Genomics & Health provided substantial guidance and support. The Beacon Project team designed and developed the Beacon API. Members of various organizations implemented Beacons and contributed to its APIs. We are thankful for data contributors who elect to share their data. M.F. and S.O.M.D. are supported by Genome Quebec, Genome Canada, the Government of Canada, and

the Ministère de l'Économie, Innovation et Exportation du Québec (Can-SHARE grant 141210); S.O.M.D. is supported by the Canadian Institutes of Health Research (grants EP1-120608; EP2-120609) and the Canada Research Chair in Law and Medicine; M.H. is supported by BD2K NIH/NCI 5U54HG007990-02; S. Scollen, S.V., M.B., I.L., J.T., S.U.-R., S.d.I.T., M.L., H.S. and the EGA are supported by ELIXIR, the research infrastructure for life-science data. This work was supported by ELIXIR-EXCELERATE, funded by the European Commission within the Research Infrastructures programme of Horizon 2020, grant agreement number 676559 (J.D.S., I.L.), the Wellcome Trust grant numbers WT201535/Z/16/Z (P.F.) and WT098051 (S.K., D.L., P.F.), and the European Molecular Biology Laboratory (P.F., S.K., J.D.S., I.L.); A.J.B. is supported by the European Union FP7 Programme 'EMIF' IMI-JU grant no. 115372, and H2020 Programme 'GCOF' grant no. 643439.

#### Author contributions

M.F., S. Scollen, G.S., S.V., S.K., D.L., P.G., S. Sherry, M.B., I.L. and D.H. provided project leadership and management; M.C., J.R., S.d.I.T., J.T., K.C., A.J.B., M.H., M.B., H.S., M.L., J.D.S. and S.U.-R. designed and developed software; S.O.M.D. developed ethics and policy research; M.F. and M.C. designed and developed the Beacon Network;

P.F. provided security review; M.F. and L.D. wrote the manuscript with contributions from all other authors.

#### Competing interests

The authors declare no competing interests.

#### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41587-019-0046-x>.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

# CRISPResso2 provides accurate and rapid genome editing sequence analysis

**To the Editor** — The field of genome editing is advancing rapidly<sup>1</sup>, most recently exemplified by the advent of base editors that enable changing single nucleotides in a predictable manner<sup>2–4</sup>. For the validation and characterization of genome editing experiments, targeted amplicon sequencing has become the gold standard<sup>5</sup>. Here we present a substantially updated version of our CRISPResso tool<sup>6</sup> to facilitate the analysis of data that would be difficult to handle with existing tools<sup>6–9</sup>.

CRISPResso2 introduces five key innovations: first, comprehensive analysis of sequencing data from base editors; second, a batch mode for analyzing and comparing multiple editing experiments; third, allele-specific quantification of heterozygous or polymorphic references; fourth, a biologically informed alignment algorithm; and fifth, ultrafast processing time. We discuss each of these in turn below.

Our updated software allows users to readily quantify and visualize amplicon sequencing data from base-editing experiments. It takes as input raw FASTQ sequencing files and outputs reports describing frequencies and efficiencies of base editing activity, plots showing base substitutions across the entire amplicon region (Fig. 1a), and nucleotide substitution frequencies for a region specified by the user (Fig. 1b). Users can also specify the nucleotide

substitution (for example, C→T or A→G) that is relevant for the base editor used, and the software produces publication-quality plots for nucleotides of interest with heat maps showing conversion efficiency.

We also improved processing time and memory usage of CRISPResso2 to enable users to analyze, visualize and compare results from hundreds of genome editing experiments using batch functionality. This is particularly useful when many input FASTQ files must be aligned to the same amplicon or have the same guides, and the genome editing efficiencies and outcomes can be visualized together. In addition, CRISPResso2 generates intuitive plots to show the nucleotide frequencies and indel rates at each position in each sample. This allows users to easily visualize the results and extent of editing in their experiments for different enzymes (Fig. 1c).

In cases where the genome editing target contains more than one allele (for example, when heterozygous single nucleotide polymorphisms (SNPs) are present), genome editing on each allele must be quantified separately, even though reads from both alleles are amplified and mixed in the same input FASTQ file. Current strategies are not capable of analyzing multiple reference alleles and may lead to incorrect quantification. CRISPResso2 enables allele-specific quantification by

aligning individual reads to each allelic variant and assigning each read to the most closely aligned allele. Downstream processing is performed separately for each allele so that insertions, deletions or substitutions that distinguish each allele are not confounded with genome editing. To demonstrate the utility of our approach, we reanalyzed amplicon sequencing data from a mouse with a heterozygous SNP at the *Rho* gene in which an engineered SaCas9-KKH nuclease was directed to the P23H mutant allele<sup>10</sup>. CRISPResso2 deconvoluted reads, quantified insertions and deletions from each allele, and produced intuitive visualizations of experimental outcomes (Fig. 1d).

Existing amplicon sequencing analysis toolkits ignore the biological understanding of genome editing and instead optimize the alignment on the basis of sequence identity only. However, this can lead to incorrect quantification of indel events, especially in sequences with short repetitive subsequences where the location of indels may be ambiguous as a result of multiple alignments with the same best score. In such cases, it is reasonable to assume that indels should overlap with the predicted nuclease cleavage site. Our improved alignment algorithm extends the Needleman–Wunsch algorithm with a mechanism to incentivize the assignment of insertions or deletions to



## Annex Paper 2

Dyke, S. O. M., Linden, M., Lappalainen, I., de Argila, J. R., Carey, K., Lloyd, D., Spalding, J. D., Cabili, M. N., Kerry, G., Foreman, J., Cutts, T., Shabani, M., Rodriguez, L. L., Haeussler, M., Walsh, B., Jiang, X., Wang, S., Perrett, D., Boughtwood, T., . . . Flicek, P. (2018). **Registered access: authorizing data access.** *European Journal of Human Genetics*, 26(12), 1721–1731.

<https://doi.org/10.1038/s41431-018-0219-y>





## Registered access: authorizing data access

Stephanie O. M. Dyke<sup>1,35</sup> · Mikael Linden<sup>2,3</sup> · Ilkka Lappalainen<sup>2,3,4</sup> · Jordi Rambla De Argila<sup>5,6</sup> · Knox Carey · David Lloyd<sup>4,7</sup> · J. Dylan Spalding<sup>4</sup> · Moran N. Cabil<sup>8</sup> · Giselle Kerry<sup>4</sup> · Julia Foreman<sup>9</sup> · Tim Cutts<sup>9</sup> · Mahsa Shabani<sup>10</sup> · Laura L. Rodriguez<sup>11</sup> · Maximilian Haeussler<sup>12</sup> · Brian Walsh<sup>13</sup> · Xiaoqian Jiang<sup>14</sup> · Shuang Wang<sup>14</sup> · Daniel Perrett<sup>9</sup> · Tiffany Boughtwood<sup>15</sup> · Andreas Matern<sup>16</sup> · Anthony J. Brookes<sup>17</sup> · Miro Cupak<sup>18</sup> · Marc Fiume<sup>18</sup> · Ravi Pandya<sup>19</sup> · Ilia Tulchinsky<sup>20</sup> · Serena Scollen<sup>3</sup> · Juha Törnroos<sup>2</sup> · Samir Das<sup>21</sup> · Alan C. Evans<sup>21</sup> · Bradley A. Malin<sup>22</sup> · Stephan Beck<sup>23</sup> · Steven E. Brenner<sup>24</sup> · Tommi Nyrönen<sup>2,25</sup> · Niklas Blomberg<sup>3</sup> · Helen V. Firth<sup>9</sup> · Matthew Hurles<sup>9</sup> · Anthony A. Philippakis<sup>8</sup> · Gunnar Rättsch<sup>26</sup> · Michael Brudno<sup>27,28</sup> · Kym M. Boycott<sup>29</sup> · Heidi L. Rehm<sup>8,30</sup> · Michael Baudis<sup>31</sup> · Stephen T. Sherry<sup>32</sup> · Kazuto Kato<sup>33</sup> · Bartha M. Knoppers<sup>1</sup> · Dixie Baker<sup>34</sup> · Paul Flicek<sup>4</sup>

Received: 26 February 2018 / Revised: 8 May 2018 / Accepted: 20 June 2018 / Published online: 2 August 2018  
© The Author(s) 2018. This article is published with open access

### Abstract

The Global Alliance for Genomics and Health (GA4GH) proposes a data access policy model—“registered access”—to increase and improve access to data requiring an agreement to basic terms and conditions, such as the use of DNA sequence and health data in research. A registered access policy would enable a range of categories of users to gain access, starting with researchers and clinical care professionals. It would also facilitate general use and reuse of data but within the bounds of consent restrictions and other ethical obligations. In piloting registered access with the Scientific Demonstration data sharing projects of GA4GH, we provide additional ethics, policy and technical guidance to facilitate the implementation of this access model in an international setting.

### Introduction

As data sharing policies in genomics strive to keep pace with the state of data-intensive science [1, 2], current policies offer little choice for sharing genomic research data beyond the two established mechanisms of *open access*, when data are freely published on the World Wide Web, and *controlled access* (also called *managed* or *restricted access*), whereby qualified researchers apply for access on a project-by-project basis and their research plans are reviewed, often by a committee [3–5]. Both open and controlled access policy models have historically served the research community’s needs, scientific progress and clinical care. However, plans for greater integration of datasets and informatics platforms [6], along with ever greater sharing of health-related datasets and growing interest by clinicians and patients in also accessing genomic data, call for new streamlined models of data access that take greater

advantage of the richer access-control policies current technology is capable of enforcing. Access-control policies, and the technology that enforces them, must enable rapid and efficient access to data that is shared only for specific purposes to a wide range of users while effectively managing ethical and legal risks.

### The registered access policy model

Our proposals arise from discussions with a range of stakeholders engaging in international data sharing initiatives as members of the Global Alliance for Genomics and Health (GA4GH) [7]. GA4GH is an international coalition dedicated to improving human health by maximizing the potential of genomic medicine through effective and responsible data sharing, as founded on the *Framework for Responsible Sharing of Genomic and Health-Related Data* [8]. Our work has led us to conclude that there are specific datasets where existing consent agreements and ethical approval are compatible with a novel data access policy model called *registered access* [9]. This model would capitalize on the well-established role-based access control (RBAC) model for

✉ Stephanie O. M. Dyke  
stephanie.dyke@mcgill.ca

Extended author information available on the last page of the article

information technology security enforcement [10–14] and is based on the notion that potential users could be granted online access to data according to their roles (e.g., bona fide researcher or clinical care professional) and risk analysis, rather than on the basis of a specifically described project as is normally required in the controlled access models commonly implemented for research purposes. RBAC is widely implemented in government and industry throughout the world [15]. By capitalizing on RBAC-based access-control technologies, registered access could, in theory, provide access to all data shared in this way, following a unified general registration process and without the need for individualized data access committee review.

### Examples of registered access

Registration as a means to limit access to data to approved users—albeit with different approval processes—has already been used in several genomics projects. For example, the Wellcome Trust Case-Control Consortium required registration for access to summary allele frequency datasets once it was demonstrated that these data could potentially lead to the re-identification of study participants [16, 17]. While this risk was considered to be low, limiting access to consortium researchers seemed to be a reasonable mitigation strategy at the time and was judged by the Consortium Data Access Committee to be consistent with the participant consent agreements. More recently, the “Bravo” project requires a simple form of registration via logging in to access data. Recent policy recommendations based on risk assessment for such data aim to discriminate between a lower and higher risk of potential resulting harm in the case of re-identification, for example, limiting access to aggregate data according to whether data were associated with more sensitive health or demographic information (e.g., ethnicity information about small or vulnerable populations) [18, 19].

Another current example of a registration-based data access policy is the DatabasE of genomiC variation and Phenotype in Humans using Ensembl Resources (DECIPHER [20]). Users who have been approved by the project coordinator (a senior physician working at the center depositing the data) are granted registered access to that project data. DECIPHER projects can be linked to form a consortium, allowing intra-consortium sharing. PhenomeCentral is another example of a registered access policy for the identification of additional cases for ultra-rare disorders [21]. Along with DECIPHER, PhenomeCentral is part of the GA4GH Matchmaker Exchange (MME) initiative. PhenomeCentral users are required to be bona fide researchers or clinicians. This is validated through institutional email addresses, as well as through user-provided and publicly available information such as prior publications, scientific activity at conferences identified through web

searches, and mention on institutional websites. Users without a scientific track record (e.g., trainees) can be validated by a more senior colleague. Data entered into PhenomeCentral can then be shared either with chosen researchers or with pre-defined groups (consortia) who have leads responsible for approving membership.

In addition to these intra-consortia, coordinator-approved registration policies, several other current projects are providing, or plan to provide, registration-based access to the research community beyond their projects (see Box 1). These resources either grant an account following a review of an applicant’s credentials (based on submitted or public information) or following a simple registration of their identity. All involve online agreement to data use terms and conditions. CAGI, the Critical Assessment of Genome Interpretation, active since 2010, has several tiers of access according to the sensitivity of datasets [22], which are available to registered users ranging from unaffiliated researchers to trainees entering the field and individuals at companies to well-identified accomplished researchers. Vouching (e.g., of a mentor for a student) can also allow appropriate escalation of access.

### Implementation in GA4GH

Our model of registered access in the GA4GH context comprises a three-stage “Triple-A registration” process (Authentication, Attestation, and Authorization [9]), which aims to ensure both user identification and agreement to a standard set of general responsibilities while considerably simplifying the data access application process. Through the identification and authentication process, the individual provides “proof” that an asserted identity is their own. The attestation process establishes that the potential data user meets the requirements expected by the consent agreements and ethical approval of datasets in question and includes agreement to comply with the terms of data use required of registered users. Finally, authorization is the overall process by which users are granted access to data and permission to perform specific actions. We provide concrete examples of, and guidance for, each stage in the process based on three GA4GH Demonstration Projects with which we fleshed out standards that would be broadly applicable.

The Beacon Project (manuscript in press), the Matchmaker Exchange [23], and the BRCA Challenge (manuscript submitted) are among the initial demonstration projects that aimed to drive learning, identify requirements, assess value, and coordinate activity within the first phase of GA4GH. For each of these, we explored options for using registered access to improve and streamline access to data that had previously been available either through a controlled access application process and/or bound by protocol-specific restrictions.

**Box 1** Examples of current projects enabling registration-based access*Resource*

Critical Assessment of Genome Interpretation (CAGI)  
<https://genomeinterpretation.org>

Simons Foundation Autism Research Initiative (SFARI)  
<https://www.nextcode.com/ssc/>

mPower Public Researcher Portal [39]  
<http://sagebase.org/research-projects/mpower-researcher-portal/>

AACR Project GENIE  
<https://www.synapse.org/#!Synapse:syn7222066/wiki/410922>

Bravo (<http://bravo.sph.umich.edu>)

*Access requirements*

- Review of users
- Digital signing of data use agreement
- Review of users
- Online agreement to data use conditions
- Verification of user identity and training
- Online agreement to data use conditions
- Verification of user identity and training
- Online agreement to data use conditions
- Login with ID provider (Google ID) linked to work email address
- Online agreement to data use conditions

For the Beacon Project, which enables the discovery of genetic variants across multiple world-wide datasets, registered access is envisaged as a means to share more details than simple existence of genetic variants (e.g., that they are present in individuals with a specific health condition). In conjunction with Beacon partners ELIXIR (Europe's infrastructure for life science information) and NCBI (the US National Center for Biotechnology Information), registered access is being developed for access to appropriate metadata from controlled access datasets. Such metadata access is similar to current access protocols at dbGaP [25] and the European Genome-phenome Archive (EGA) [26]. Specifically, users with an eRA account (for the NIH Commons research grant system) are dbGaP registered users with access to some information about available controlled access datasets [24]. Similarly, EGA users who have obtained access to at least one EGA controlled access dataset have access to specific EGA information about available controlled access datasets after logging in with their EGA account.

The MME is a federated network connecting databases of genomic and phenotypic data using a common application programming interface to facilitate rare disease gene discovery, including from DECIPHER (open subset), the PhenomeCentral platform [21], GeneMatcher [27], MyGene2 [28], Patient Archive (*patientarchive.org*), and *matchbox*. In its current iteration, it requires two-sided inquiry (i.e., a search from two parties with a similar patient) and, in this way, connects two investigators looking for a match for the same candidate gene and disease. Each user must be registered in one of the databases in order for data to be deposited and queries made. Future iterations of MME will expand functionality and facilitate a one-sided inquiry, with bona fide investigators identified by a registered access process able to see details

of a matched case, including variants in a specific gene and high-level phenotypic information for their purposes as a scientific investigator working to understand the causes of rare diseases.

The goal of the BRCA Challenge is to translate the rapid expansion of sequencing capacity into useful knowledge and, in particular, learn how to rapidly interpret variant data to generate clinical utility. Its intent is to provide an umbrella under which many groups can collaborate and bring together data to improve the precision of assessing variants across both BRCA1 and BRCA2. While its main resource on BRCA variant interpretation is publicly available, overlaying registered access would allow enrichment of the dataset with data that cannot be shared openly: for example, patient data supporting clinical interpretations of variants may not be consented for open release but would be available to expert review teams, researchers, or clinicians.

To support these pilot implementations of registered access in GA4GH, we expand on our initial ethical–legal feasibility study and review of projects that are pioneering registration-based access policy (see Box 1) to describe plans for an international, unified approach that could lead to a standardized registration process allowing for access to a wide range of data resources. All three stages of registered access (authentication, attestation, and authorization) pose significant ethical–legal and technical challenges, which we attempt to address by providing policy and technical guidance.

## Authentication

A potential advantage of the registered access policy model is to efficiently provide data access to a relatively large

**Box 2** The “layered” registration system. Shows the main routes to user authentication for the categories of bona fide researcher and clinical care professional

*A person may receive bona fide researcher status if:*

1. Their home institution confirms they are researchers, OR
2. A person who satisfies condition (1) corroborates (“vouches for”) their researcher status (as a reference)

*A person may receive clinical care professional status if:*

1. Their home institution confirms they are clinical care professionals, OR
2. They have a physician or other clinical care professional license (ID/permit number)

number of authorised individuals and alleviate the considerable administrative burden on data custodians of managing controlled access requests. This model is premised on the trust that broad categories of registered users, such as researchers and clinical care professionals, will use the data accessed with the same appropriate care as they would manage controlled access data. Defining categories of users as bona fide researchers or clinical care professionals in this context rests largely on the information provided at the time of registration (user attributes) and the attestation they agree to. The attributes requested from users for the registration process, and particularly their verification, will have important implications for access to data protected by registered access authorization methods.

Based on an ethical–legal analysis of research ethics and other legal and administrative frameworks applicable to data sharing and access, it was previously proposed that several elements of controlled access review should be retained in registered access, including for how users might be authorized based on their “competence.” We considered whether it might be necessary to set a few differing levels of stringency for the registered access model (e.g., Registered, Registered+) to cater to different projects’ views of the requisite access and data sensitivity. However, we agreed that a minimal standard (basic registration criteria) could be established, thereby enabling mutual recognition between registration systems established in different parts of the world (e.g., ELIXIR and NCBI). This does not preclude policies that provide different levels of access to data to different categories of users. Indeed, such policies are enforceable using a combination of RBAC and attribute-based access control.

To qualify as either a bona fide researcher or clinical care professional, first of all, individuals will need to provide the following details of their identity and research/clinical activity: name; title; position; affiliation; and institutional email address, phone number, website, and mailing address. As these details may also be provided by an individual’s

organization, they are an important means of strengthening accountability and traceability of registered users and can be simply verified by web searches or calls to institutional switchboards.

We considered additional information that could demonstrate a research user’s professional status such as: researcher identity systems (e.g., ORCID or ISNI); PubMed publication IDs; and researcher accounts such as those with funding agencies (e.g., NIH Commons’ eRA), universities (email addresses or user accounts), and the major public archives (e.g., MyNCBI and PubMed Commons). Evidence of academic publication (in the context of a research position) is typically relied upon in the controlled access application process as an indication of researchers’ ability to use data [29]. However, concern was expressed regarding the value of journal publications and some researcher IDs as an indicator of professional activity, especially current activity. There was also concern about the rise in so-called “predatory” academic journals, leading to publications of dubious quality [30].

We eventually decided on a “layered” registration system whereby bona fide researchers or clinical care professionals could either demonstrate their status directly (by providing evidence of professional status, such as license numbers for clinical care professionals) or alternatively have their status “vouched for” by another registered user within their category (for researchers) or their employing institution (for researchers and clinical care professionals) (see Box 2). One use case for such a voucher approach would be for students or trainees who may have neither professional appointment nor publications, where the expectation would be for an advisor to support the registration.

### Responsibilities of institutions

Accountability of registered users is central to the registered access model. Within data access policy models, various approaches have been proposed to hold users accountable. One is co-signing of a data access agreement by the (home)



institution of the users and recognizing this institution as the ultimate responsible entity. Within this perspective, that institution can be legally held accountable if the researcher or clinician commits any wrongdoing. Although registered access does not require signing such an agreement between the home institution and the data custodians, one could argue that, if any wrongdoing happens, the users' institution will in all likelihood be contacted and asked to enforce administrative disciplinary measures in an analogous way as is currently done in some cases of scientific misconduct, such as plagiarism or publishing falsified data. In turn, in addition to the attestation registered users will have agreed to in registering for access to data, home institutions may require researchers and clinicians—who plan to use internal or external health data—to sign up to procedures and guidance documents such as a “Code of Conduct”, in order to bind them with the institutional rules and sanctions in this respect.

### Vouching

For the second route to registration for bona fide researchers, a person who has already been registered via their institution could corroborate another researcher's status, as a reference. The vouching researcher would need to confirm that they know and have identified the researcher they are registering. To promote accountability and community control, registered users would be able to see who has vouched for whom. This is akin to having a witness to one's competence and professional activity. The issue here is one of validation based solely on a personal statement and of potential liability for the researcher registering this way as they may not have institutional backup. “Vouchers” could also potentially be held liable. It is worth noting that a large-scale, successful community, the Debian community, maintains operating system software using a vouching approach based on Pretty Good Privacy (PGP) key signing. A member of the community must have their PGP public key signed by at least one existing member of the community before their key can be admitted into the Debian keyring (which then enables them to modify and upload software, participate in elections, etc.). There are strict guidelines on the level of proof required for signature—meeting in person, both parties show government photo ID, etc. There are also other similar prerequisites, such as accepting the social contract and advocacy by another member, and violations result in removal of access by the community.

By providing several routes to registration, we hope to enable access to as wide a group of potential data users as possible while maintaining a strong level of accountability. For clinical care professionals in the USA, the National Provider Identifier issued by the Centers for Medicare and Medicaid Services could be requested in addition to the

registered user's license number. As examples, in the UK, users could provide their General Medical Council licence number; in Germany, their *Lebenslange Arztnummer*; in France, their *numéro RPPS (répertoire partagé des professionnels de santé)*; in Australia, their Australian Health Practitioner Regulation Agency registration number; and in Canada, their Royal College of Physicians and Surgeons identification number. Registration for clinical care professionals will in most cases be linked to professional oversight and disciplinary governance frameworks.

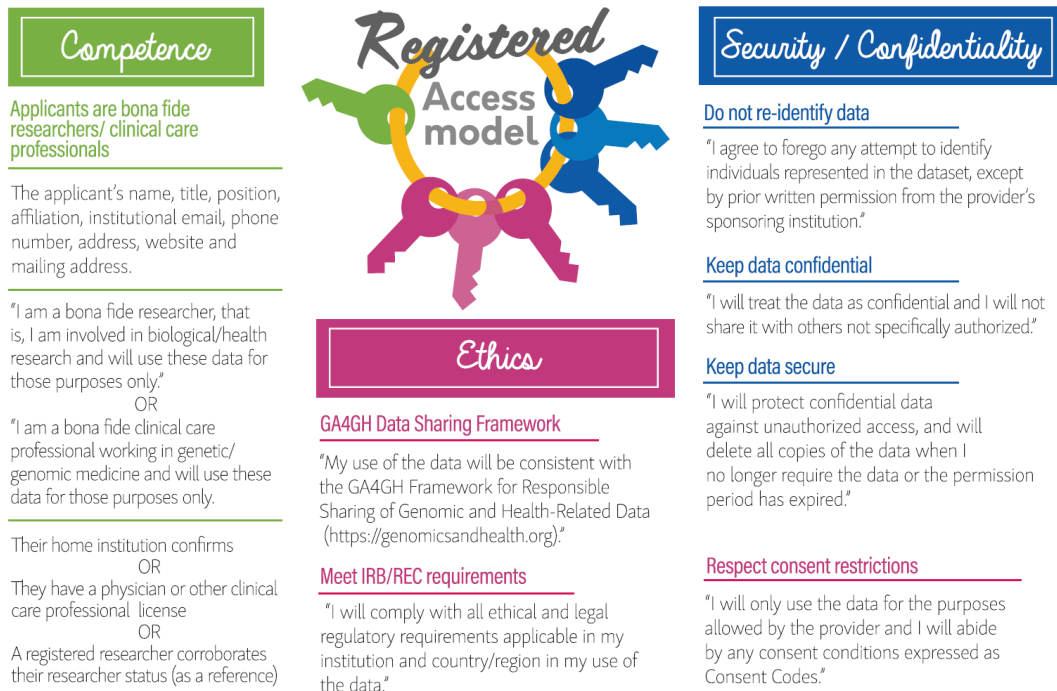
In case these routes did not allow registration of atypical potential users, as an additional route to registration, any individual would also be able to apply to a standard Data Access Committee (DAC), the committees that oversee access to controlled access data, to be assessed on a case-by-case basis for registered access status. DACs may also help register users whose organizations have yet to establish the organizational or technical protocols to facilitate registration (see discussion under “Accessibility” below).

### Attestation

Integral to the definitions of bona fide researcher and clinical care professional are the statements and agreements included in the attestation stage of the registration process (see Fig. 1). Indeed, controlling the purpose of data use is a key component of data protection principles and the European Union (EU) General Data Protection Regulation (GDPR) [31].

One of the attestation statements refers to respecting consent-based data use permissions and restrictions, which should ideally be expressed as Consent Codes [32]. The GA4GH Consent Codes are a structured way of recording consent permissions so they can be made clear to users and to enable maximum data aggregation (with the same or broader permissions). Another attestation statement prohibits any attempt to identify individuals based on combining shared data with other public or non-public data sources. It allows for exceptions to this condition in some circumstances, with “prior written permission of the provider's sponsoring institution.” This is to enable the recontact of participants if warranted (e.g., for the return of individual research results) or for permission to conduct research into privacy risks. We plan to provide general guidance for the attestation statement about keeping data secure, such as that the data should be kept encrypted at rest and in transit between systems, and that only authorized individuals have access to the keys (<http://genomicsandhealth.org/work-products-demonstration-projects/security-infrastructure>).

We also plan to include an educational module as part of the registration process. Ultimately, we aim to enable a single format for the registration process but support a model that



**Fig. 1** The registered access policy model. The figure shows the authentication and attestation requirements of the GA4GH registered access policy model for the user categories of bona fide researcher and

clinical care professional. The seven statements shown in quotation marks form the attestation stage of the process

would allow for additional attestation statements attaching extra conditions of use for some datasets or for data from some providers. For example, Australian Genomics is considering the model that researchers need to supply proof of Human Research Ethics Committee (HREC) approval (HREC number, title, etc.), which can then be easily verified by web search.

In agreeing on these definitions of bona fide researcher/clinical care professional and level of security, cross-federation becomes possible (e.g., to enable European bona fide researchers to present queries to US Beacons at the registered access level and vice versa). The current attributes chosen to define registered users in these categories are designed to cover most of the use cases. When exceptions arise, there would have to be a very strong need for the new definition to make everyone deploy it (and populate values to the existing users retrospectively). Importantly, such exceptions would only be considered valid if driven by informed consent requirements or national laws.

Although our model does not include a review and approval of the user's specific research or data use plans, we considered requesting abstracts of general planned data use in lay terms that would be published to enhance transparency. This may be reconsidered, especially to reinforce

registered users' commitments to using data for appropriate research or clinical care purposes and to further the aims of public transparency. Another interesting suggestion regarding transparency was to request and publish links to public researcher profiles for all registered researchers.

## Authorization

In agreeing on the proposed routes to registration, we have effectively delegated the authorization of registered users for the two categories described here to established professional employment, accreditation, or accomplishment. The data sharing environment is therefore assumed from individuals' bona fides (including work practices and the security aspect) along with the basic set of requirements set out in the registration attestation. Along with efforts to automate registered access, this potentially limits the amount of manual authorization that will be required.

Our pilot implementation of the first registration route for academic researchers (their home institution confirms they are researchers) is the simplest in terms of liability for the category of "bona fide researchers," and therefore the "safest" place to start. ELIXIR is piloting an approach

where an ELIXIR user authenticates their identity through their own research organization's account, and the organization confirms researcher's status. Organizational validation is assumed to improve the provenance of the researcher's professional status because home organizations are vetted by funders, are expected to know their researchers, and can also provide the authentication credentials securely to their researchers. A challenge will be to define the requirements an organization needs to meet to become trusted in a global GA4GH registered access system (e.g., for federated identity in research, the UK has minimal checks <https://www.ukfederation.org.uk/content/Documents/EligibleOrganisations>).

Our experience is that this is often a more controversial and difficult challenge than the verification of individuals' identity and role. For instance, institutions that might oversee clinicians and researchers wanting access to genetic data could include a range of clinical genetics centers (publicly funded/charitable/private); primary care centers, which treat certain inherited conditions and other contexts in which genetic testing may be commissioned or communicated without genetics specialists; and research institutions (university/other public/charitable/private). The challenge, therefore, may be to establish standards for those entities facilitating registration, including the institutions hosting registered users. A particularly crucial element of the institutional aspect of access control is the identification of accounts that no longer meet the access criteria. There needs to be well-defined, well-understood mechanisms for reviewing and revoking status, and registries of users will need to demonstrate that they successfully ensure sponsors do so in a timely manner. Examples of situations which access control workflows may need to account for include staff moving from one role to another (which may alter the user's clinical care professional vs. researcher category) or leaving the profession.

From a technical point of view, we split the registered access architecture into two components, which can be separated organizationally and geographically: a component that manages the individual's identity and attributes, and the party that relies upon this component to confirm identity and attributes. The OpenID Connect technical standard (<http://openid.net/connect/>) refers to these two components as the "OpenID Provider" and the "relying party" respectively. There may be several registries and relying parties managed by different organizations in different geographical locations.

The OpenID Provider is responsible for authenticating a registered user's identity and for sharing attributes that the relying party may use to authorize access (see Box 2 and Fig. 1). Given OpenID Connect's broad use worldwide, we suppose that organizations such as ELIXIR in Europe or

NCBI in the US could deploy the technology needed to operate as an OpenID Provider for their constituencies.

Registered access relying parties are the entities that consume OpenID authorizations and enforce access rights and privileges based on the registered access status and attributes of the users. To be able to use registered access claims, a relying party needs to trust one or several OpenID Providers. In order to establish a federation of OpenID Providers and relying parties, they need to agree on the exact semantics of registered access status and attributes; how credentials are verified by the OpenID Provider and expressed to the relying party; what technical protocols are used to share between the registry and the relying party; and how to protect the confidentiality, integrity, and availability of the communication.

User attributes and attestations are provided to relying parties through the standardized OpenID Connect protocol, which is based on OAuth 2.0 [33]. These standards provide a mechanism through which OpenID Providers may authenticate users and provide "claims"—data structures that encode various user attributes—that can be cryptographically validated by relying parties and used in mediating access to data. Once identity has been authenticated and registered access attributes shared, OAuth 2.0 will mediate the requested access based on the data holder's access policy.

The GA4GH is working to define a set of custom claims for registered access that all OpenID Providers and relying parties can adopt (Library Cards [34]) providing interoperability across the ecosystem of registered access adopters. Strong identity-proofing will be required within a unified identity framework, especially in the future, for registration that is independent of institutional listing or peer vouching. We plan to use existing guidelines [35] for how to establish and maintain trust in digital identities. These frameworks rank a spectrum of assurance levels, and relying parties can report (in claims) which of these levels was used to perform identity proofing.

Researcher attributes and registered access status count as personal, identifiable information, which is protected by privacy laws, including the new GDPR in the EU. To protect the privacy of researchers and respect data protection laws, it is proposed that OpenID Providers limit the amount of personal data shared with relying parties. This would mean communicating only a pseudonymous identifier of the researcher (i.e., an alphanumeric code, which is needed for thwarting re-identification and other attacks on multiple relying parties simultaneously) and their registered access status (which is needed for verifying the requestor's status), including its route and provenance, i.e., which registry delivered the status. Consent is one of the six lawful bases to process personal information in the GDPR [36]. Article 4(11) defines consent as: "any freely given, specific,

informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her". For the registration process, this would entail providing users with a way to consent to the sharing of their personal data for the purposes of gaining registered status, which ELIXIR has integrated into its pilot system.

Different datasets, even within an institution, may have different requirements, such as the Consent Codes associated with data. Such datasets may require additional Attestation statements, beyond those recommended by GA4GH, for access (see Fig. 1); a data steward [37] (or the data custodian or guardian as referred to in different locations) must specify and enable such Attestations, and they will usually be guided by research ethics committees and institutional review boards in these responsibilities. While access conditions must reflect the use permissions of the dataset, additional Attestations/restrictions may complicate or prevent the aggregation of data from many sources.

## Accessibility

In the interests of efficiency and alleviating administrative burden on data custodians—particularly given the number of potential registered users—efforts should be made to automate the registered access process. Additionally, from an information security perspective, self-asserted attributes provide little accountability and raise the possibility of identity theft. We therefore sought to incorporate automated (or delegated, e.g., institutional) checks of user attributes. As our plans for the processing of registered access attributes for bona fide researcher registration draw on pre-existing academic infrastructure, we envisage minimal investment from an institutional perspective, reducing barriers to adoption of this system. It will be important to install a comparable system for access by researchers in industry.

Since 2005, research and education institutions have been operating technical frameworks called identity federations that allow researchers to use their home institution's credentials (such as user accounts and passwords) to access services that are outside their home institutions. To register their bona fide researcher status and make the related attestations within such federations, a researcher would first need to log in at their home institution, which then delivers their fresh and validated role and affiliation information to the registration process. Currently, there is some form of national research and education identity federation in 72 countries (<https://refeds.org/federations>) using many different systems but usually the same

technology and that are bridged with a system called edu-GAIN (<https://www.edugain.org>; a sister service of eduroam, <https://www.eduroam.org/>). A benefit of using an identity federation for registered access is that the researcher status is not self-asserted by the researcher but instead claimed by the research institution employing the researcher. The home institution is also able to provide more fine-grained information on the person's affiliation (<http://software.internet2.edu/eduperson/internet2-mace-dir-eduperson-201602.html#eduPersonAffiliation>) than a simple institutional e-mail address check, which often does not differentiate between researchers, students, and administrative staff. Additional details that could support registered access through federated identity management would be the categorization of bio/health researchers or even "registered following GA4GH standards." A challenge of identity federation is that currently there is no widely deployed framework for the level of assurance of the identity and authentication of users. Data protection laws also make some institutions hesitate to release researchers' personal data to other jurisdictions. Collaborations such as the Federated Identity Management for Research Collaboration (FIM4R) aim to establish common standards that meet the needs of various research communities [38, 39].

## Conclusion

While there remain many challenges in implementing registered access, especially at scale and with respect to the legal and administrative tools to facilitate registration through the proposed range of routes, the GA4GH pilots have allowed us to flesh out various aspects and better understand its practical utility. The main goal of registered access is to streamline access to datasets that require acceptance of terms and conditions due to consent agreements or because of a level of ethical and legal risk, and to enable access to multiple datasets at once as well as to facilitate data discovery and use. We also envisage that the simplicity, and clarity, of the standard conditions of data access and use in registered access (the attestation) will both encourage greater use of the data and respect for its ethical use, as seen with licensing terms, such as GNU General Public License and Creative Commons.

The registered access model and services described above must correctly maintain protections that were agreed to by study participants as well as researchers and clinicians who wish to study their data in order to eventually advance biomedical knowledge and benefit society. The registered access policy model will then need to be recognized and supported by many stakeholders, including research ethics boards, such that the language used in consent forms and research agreements are

compatible with this access model. This will make a big difference in how “*silo-ed*” data continue to be. Ultimately, the confidence the research community will gain in the system will determine the extent of the resources it will ultimately provide.

Finally, while we have focused initially on registration criteria for researchers and clinical care professionals, many of whom have not generally had access through the controlled access system, we anticipate that data users will eventually include members of the public, including patients and citizen scientists (see e.g., mPower [40]), as well as other groups such as volunteer health-care providers and journalists. We plan to consider expanding registered access for these important and diverse groups in the near future, within the permissions of consent, and ethical standards, and with broad consultation with patient advocacy groups and research participants.

Another important aspect of improving data access is the development of ethics tools to support the assessment of data sensitivity and therefore the risk in data sharing to better determine proportionate levels of protection (e.g., open or registered). A coherent approach involves considering both the risk of re-identification of data and its sensitivity, along with the data sharing expectations of individuals and communities (Data Sharing Privacy Test [41]).

We expect registered access will inform and may even replace many controlled access mechanisms as the level of accountability that it can achieve is demonstrated over time. Data Access Committees may come to play new roles, such as deciding which data are suited to registered access, as well as reviewing applications of atypical potential users and handling other aspects of data governance (e.g., data use breaches or retractions).

We believe that it is ethically desirable to use less restrictive access controls, wherever suitable, to increase the chances of having the best research from the most people using the data that has been contributed. To needlessly reduce appropriate access likely undermines the intentions and desires of research participants as well as hindering the course of research progress.

**Acknowledgements** We thank Dr. David Kelsey (STFC Rutherford Appleton Laboratory) for helpful comments and discussion of registered access.

**Funding** SOMD is supported by the Canadian Institutes of Health Research (EP1-120608; EP1-120609; CEE-151618), Genome Quebec, Genome Canada, the Government of Canada, the Ministère de l'Économie, Innovation et Exportation du Québec (Can-SHARE grant 141210), and the Canada Research Chair in Law and Medicine. ML, IL, JT, and TN are supported by the ELIXIR, the research infrastructure for life-science data, and the H2020 ELIXIR-EXCELERATE grant 676559. IL and GK are supported by the European Molecular Biology Laboratory; MS by Research

Foundation Flanders (FWO); MH by NIH/NHGRI 5U41HG002371-15; SW by NIH/NHGRI R00HG008175; S Beck by the National Institute for Health Research UCLH Biomedical Research Centre (BRC369/CN/SB/101310); S Brenner by NIH/NHGRI U41HG007346; BMK by the Canada Research Chair in Law and Medicine; and PF by WT201535/Z/16/Z and the European Molecular Biology Laboratory.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Bobrow M. Funders must encourage scientists to share. *Nature*. 2015;522:129.
2. Auffray C, Balling R, Barroso I, Bencze L, Benson M, Bergeron J, et al. Making sense of big data in health research: towards an EU action plan. *Genome Med*. 2016;8:71.
3. Toronto International Data Release Workshop Authors, Birney E, Hudson TJ, Green ED, Gunter C, Eddy S, et al. Prepublication data sharing. *Nature*. 2009;461:168–70.
4. Ramos EM, Din-Lovinescu C, Bookman EB, McNeil LJ, Baker CC, Godynskiy G, et al. A mechanism for controlled access to GWAS data: experience of the GAIN Data Access Committee. *Am J Human Genet*. 2013;92:479–88.
5. Brenner SE. Be prepared for the big genome leak. *Nature*. 2013;498:139.
6. Manolio TA, Fowler DM, Starita LM, Haendel MA, MacArthur DG, Biesecker LG, et al. Bedside back to bench: building bridges between basic and clinical genomic research. *Cell*. 2017;169:6–12.
7. Global Alliance for Genomics and Health. GENOMICS. A federated ecosystem for sharing genomic, clinical data. *Science*. 2016;352:1278–80.
8. Knoppers BM. Framework for responsible sharing of genomic and health-related data. *Hugo J*. 2014;8:3.
9. Dyke SO, Kirby E, Shabani M, Thorogood A, Kato K, Knoppers BM. Registered access: a ‘Triple-A’ approach. *Eur J Human Genet*. 2016;24:1676–80.
10. Ferrairole D, Kuhn, R. Role-based access control. Proceedings of the 19th National Computer Security Conference. 1992.
11. Ferrairole D, Cugini, J, Kuhn, R. Role-based access control: features and motivations. Proceedings of the 11th Conference in Computer Security Applications. 1995.
12. Sandhu R, Coyne E, Feinstein H, Youman C. Role-based access control models. *IEEE Comput*. 1996;29:38–47.

13. Barkley J, Cincotta, A, Ferraiolo, D, Gavrilu, S, Kuhn, D. Role based access control for the world wide web. 20th National Computer Security Conference. 1997.
14. Ferraiolo D, Sandhu, R, Gavrilu, S, Kuhn, D, Chandramouli, R. A proposed standard for role-based access control. National Institute of Standards and Technology. 2000. ACM Transactions on Information and System Security, Vol. 4, No. 3, August 2001, pp. 224–274.
15. O'Connor AC, Loomis RJ. Economic analysis of role-based access control: final report. 2010.
16. Homer N, Szelling S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* 2008;4:e1000167.
17. Craig DW, Goor RM, Wang Z, Paschall J, Ostell J, Feolo M, et al. Assessing and managing risk when sharing aggregate genetic variant data. *Nat Rev Genet.* 2011;12:730.
18. National Human Genome Research Institute. Workshop on sharing aggregate genomic data report. 2016.
19. National Institutes of Health. Proposal to update data management of genomic summary results under the NIH genomic data sharing policy. 2017.
20. Chatzimichali EA, Brent S, Hutton B, Perrett D, Wright CF, Bevan AP, et al. Facilitating collaboration in rare genetic disorders through effective matchmaking in DECIPHER. *Hum Mutat.* 2015;36:941–9.
21. Buske OJ, Girdea M, Dumitriu S, Gallinger B, Hartley T, Trang H, et al. PhenomeCentral: a portal for phenotypic and genotypic matchmaking of patients with rare genetic diseases. *Hum Mutat.* 2015;36:931–40.
22. Hoskins RA, Repo S, Barsky D, Andreoletti G, Moul J, Brenner SE. Reports from CAGI: the Critical Assessment of Genome Interpretation. *Hum Mutat.* 2017;38:1039–41.
23. Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, et al. The matchmaker exchange: a platform for rare disease gene discovery. *Hum Mutat.* 2015;36:915–21.
24. Wong KM, Langlais K, Tobias GS, Fletcher-Hoppe C, Krasnewich D, Leeds HS, et al. The dbGaP data browser: a new tool for browsing dbGaP controlled-access genomic data. *Nucleic Acids Res.* 2016;45:D819–26.
25. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet.* 2007;39:1181–6.
26. Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, Ur-Rehman S, et al. The European Genome-phenome Archive of human data consented for biomedical research. *Nat Genet.* 2015;47:692–5.
27. Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum Mutat.* 2015;36:928–30.
28. Chong JX, Yu JH, Lorentzen P, Park KM, Jamal SM, Tabor HK, et al. Gene discovery for Mendelian conditions via social networking: de novo variants in KDM1A cause developmental delay and distinctive facial features. *Genet Med.* 2016;18:788–95.
29. Dyke SOM, Saulnier KM, Pastinen T, Bourque G, Joly Y. Evolving data access policy: the Canadian context. *FACETS.* 2016;1:138.
30. Moher D, Shamseer L, Cobey K. Stop this waste of people, animals and money. *Nature.* 2017;549:23–5.
31. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016.
32. Dyke SO, Philippakis AA, Rambla De Argila J, Paltoo DN, Luetkemeier ES, Knoppers BM, et al. Consent codes: upholding standard data use conditions. *PLoS Genet.* 2016;12:e1005772.
33. Internet Engineering Task Force. The OAuth 2.0 Authorization Framework, RFC6749.
34. Cabili MN, Carey K, Dyke SOM, Brookes AJ, Fiume M, Jeanson F, et al. Simplifying research access to genomics and health data with Library Cards. *Sci Data.* 2018;5:180039.
35. Grassi PA, Fenton JL. NIST Special Publication 800-63-3. Digital Identity Guidelines. National Institute of Standards and Technology. 2017. Available from <https://pages.nist.gov/800-63-3/sp800-63-3.html> (accessed 17 January 2018).
36. Article 29 Data Protection Working Party. Guidelines on consent under Regulation 2016/679. 2018. Report no. 17/EN WP259 rev.01.
37. Global Alliance for Genomics and Health. Data Sharing Lexicon. 2016.
38. Broeder DJB, Kelsey D, Kershaw P, Luders S, Lyall A, Nyronen T et al. Federated identity management for research collaborations. Report No. CERN-OPEN. 2012;006.
39. Atherton CJ, Barton T, Basney J, Broeder D, Costa A, van Daalen M, et al. Federated Identity Management for Research Collaborations (Version 2.0). Zenodo, 2018. <https://doi.org/10.5281/zenodo.1296031>.
40. Wilbanks J, Friend SH. First, design for data sharing. *Nat Biotechnol.* 2016;34:377–9.
41. Dyke SOM, Dove ES, Knoppers BM. Sharing health-related data: a privacy test? *NPJ Genom Med.* 2016;1:16024.

## Affiliations

Stephanie O. M. Dyke<sup>1,35</sup> · Mikael Linden<sup>2,3</sup> · Ilkka Lappalainen<sup>2,3,4</sup> · Jordi Rambla De Argila<sup>5,6</sup> · Knox Carey · David Lloyd<sup>4,7</sup> · J. Dylan Spalding<sup>4</sup> · Moran N. Cabili<sup>8</sup> · Giselle Kerry<sup>4</sup> · Julia Foreman<sup>9</sup> · Tim Cutts<sup>9</sup> · Mahsa Shabani<sup>10</sup> · Laura L. Rodriguez<sup>11</sup> · Maximilian Haeussler<sup>12</sup> · Brian Walsh<sup>13</sup> · Xiaoqian Jiang<sup>14</sup> · Shuang Wang<sup>14</sup> · Daniel Perrett<sup>9</sup> · Tiffany Boughtwood<sup>15</sup> · Andreas Matern<sup>16</sup> · Anthony J. Brookes<sup>17</sup> · Miro Cupak<sup>18</sup> · Marc Fiume<sup>18</sup> · Ravi Pandya<sup>19</sup> · Ilia Tulchinsky<sup>20</sup> · Serena Scollen<sup>3</sup> · Juha Törnroos<sup>2</sup> · Samir Das<sup>21</sup> · Alan C. Evans<sup>21</sup> · Bradley A. Malin<sup>22</sup> · Stephan Beck<sup>23</sup> · Steven E. Brenner<sup>24</sup> · Tommi Nyrönen<sup>2,25</sup> · Niklas Blomberg<sup>3</sup> · Helen V. Firth<sup>9</sup> · Matthew Hurler<sup>9</sup> · Anthony A. Philippakis<sup>8</sup> · Gunnar Rättsch<sup>26</sup> · Michael Brudno<sup>27,28</sup> · Kym M. Boycott<sup>29</sup> · Heidi L. Rehm<sup>8,30</sup> · Michael Baudis<sup>31</sup> · Stephen T. Sherry<sup>32</sup> · Kazuto Kato<sup>33</sup> · Bartha M. Knoppers<sup>1</sup> · Dixie Baker<sup>34</sup> · Paul Flicek<sup>4</sup>

- <sup>1</sup> Centre of Genomics and Policy, Faculty of Medicine, McGill University, Montreal, QC, Canada
- <sup>2</sup> CSC – IT Center for Science, Espoo, Finland
- <sup>3</sup> ELIXIR Hub, Wellcome Genome Campus, Hinxton, Cambridge, UK
- <sup>4</sup> European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK
- <sup>5</sup> Centre for Genomic Regulation, Barcelona, Spain
- <sup>6</sup> Universitat Pompeu Fabra, Barcelona, Spain
- <sup>7</sup> The Global Alliance for Genomics and Health, MaRS Centre, West Tower, 661 University Avenue, Suite 510, Toronto M5G 0A3 ON, Canada
- <sup>8</sup> Broad Institute of MIT and Harvard, Cambridge, MA, USA
- <sup>9</sup> Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK
- <sup>10</sup> Center for Biomedical Ethics and Law, Department of Public Health and Primary Care, University of Leuven, Leuven, Belgium
- <sup>11</sup> National Human Genome Research Institute, NIH, Bethesda, MD, USA
- <sup>12</sup> Genomics Institute, University of California at Santa Cruz, Santa Cruz, CA, USA
- <sup>13</sup> OHSU, Portland, OR, USA
- <sup>14</sup> Department of Biomedical Informatics, UC San Diego, La Jolla, CA, USA
- <sup>15</sup> Australian Genomics Health Alliance, 50 Flemington Road, Parkville, VIC 3052, Australia
- <sup>16</sup> Bioreference Laboratories, Inc., Elmwood Park, NJ, USA
- <sup>17</sup> Department of Genetics and Genome Biology, University of Leicester, Leicester, UK
- <sup>18</sup> DNASTack, Toronto, ON, Canada
- <sup>19</sup> Microsoft, Redmond, WA, USA
- <sup>20</sup> Google, Mountain View, CA, USA
- <sup>21</sup> McGill Centre for Integrative Neurosciences, Montreal Neurological Institute, McGill University, Montreal, QC, Canada
- <sup>22</sup> Vanderbilt University Medical Center, Nashville, TN, USA
- <sup>23</sup> UCL Cancer Institute, University College London, London, UK
- <sup>24</sup> Department of Plant & Microbial Biology, University of California, Berkeley, CA, USA
- <sup>25</sup> ELIXIR Compute Platform, ELIXIR, Wellcome Genome Campus, Hinxton, Cambridge, UK
- <sup>26</sup> Department of Computer Science, Biomedical Informatics, ETH Zurich, Zurich, Switzerland
- <sup>27</sup> Department of Computer Science, University of Toronto, Toronto, ON, Canada
- <sup>28</sup> Centre for Computational Medicine, Hospital for Sick Children, Toronto, ON, Canada
- <sup>29</sup> Children’s Hospital of Eastern Ontario Research Institute, University of Ottawa, Ottawa, ON, Canada
- <sup>30</sup> Department of Pathology, Brigham & Women’s Hospital & Harvard Medical School, Boston, MA, USA
- <sup>31</sup> University of Zurich & Swiss Institute of Bioinformatics, Zurich, Switzerland
- <sup>32</sup> National Centre for Biotechnology Information, US National Library of Medicine, Bethesda, MD, USA
- <sup>33</sup> Department of Biomedical Ethics and Public Policy, Graduate School of Medicine, Osaka University, Osaka, Japan
- <sup>34</sup> Martin, Blanck & Associates, Alexandria, VA, USA
- <sup>35</sup> Present address: Montreal Neurological Institute, Faculty of Medicine, McGill University, Montreal, QC, Canada








### **Annex Paper 3**

Harrow, J., Hancock, J., Blomberg, N., Blomberg, N., Brunak, S., Capella-Gutierrez, S., Durinx, C., Evelo, C. T., Goble, C., Gut, I., Ison, J., Keane, T., Leskošek, B., Matyska, L., McEntyre, J., Miguel, C., Navarro, A., Newhouse, S., Nyrönen, T., Rambla, J., . . . Harrow, J. (2021). **ELIXIR-EXCELERATE: establishing Europe's data infrastructure for the life science research of the future.** *The EMBO Journal*, 40(6).

<https://doi.org/10.15252/emj.2020107409>



# ELIXIR-EXCELERATE: establishing Europe's data infrastructure for the life science research of the future

Jennifer Harrow<sup>†</sup> , John Hancock<sup>†</sup> , ELIXIR-EXCELERATE Community<sup>‡</sup> & Niklas Blomberg<sup>\*</sup> 

Creating knowledge by connecting and analysing large amounts of life science data is transforming our society, allowing us to start addressing major scientific and societal challenges, such as adaptation to climate change or pathogen outbreaks in an interconnected world. Modern biology is dependent on the generation, sharing and integrated analysis of digital data at scale. A deeper understanding of biological systems is now becoming possible thanks to breakthroughs in technologies that study life systematically at different scales, from molecules and single-cell pathogens to complex animal or plant models and ecosystems as well as across temporal ranges spanning split-second reactions to multi-year clinical or agronomic trials, and beyond. The key to analyse and leverage this complex, fragmented and geographically dispersed life science data landscape is to ensure it is easy to find and reuse by researchers. This article comments on ELIXIR, an international organisation that brings together bioinformatics researchers and life science resources across Europe and integrates them into a single federated infrastructure.

## Addressing the data challenges of modern biology

At present, analysis often involves integrating large datasets from multiple sources. Life science archives are rapidly increasing in size and complexity; for example, the archives held by EMBL-EBI double in size approximately every 2 years (Cook *et al*,

2016) so that long-term data stewardship is vital. The chances of retrieving data from any given scientific publication may decline by as much as 17% per year (Vines *et al*, 2014). Data have been generated for different research purposes at thousands of facilities across the world and are captured and stored in diverse formats. This creates a significant barrier to data integration and reuse (Rigden *et al*, 2016), as well as necessitating a massive data storage and exchange burden (Cook *et al*, 2016). In addition, data need to remain accessible and be updated long term for future reuse. Over 1,000 data resources exist in Europe and over 5,000 worldwide (<https://bigd.big.ac.cn/databasecommons/>), but only a small fraction of these have institutional support and long-term funding commitments (Imker, 2020). The fact that the mid- and long-term sustainability of many crucial bioinformatics resources, such as UniProt (<https://www.uniprot.org/>), Ensembl (<https://www.ensembl.org/>), EGA (<https://ega-archive.org/>) and Silva (<https://www.arb-silva.de/>), is not guaranteed threatens the foundations of academic and industrial life science activities, risking the loss of an immense wealth of biological and biomedical information, and wasting those associated historical investments. To address these challenges, ELIXIR became operational in 2014. Intergovernmental by nature, it is funded by financial contributions from its member countries (each of which, along with EMBL-EBI, hosts an ELIXIR Node), alongside other grants. Here, we describe the progress made by ELIXIR as

a result of European Union's €19 million ELIXIR-EXCELERATE grant from 2015 to 2019. This funding was provided, following the ESFRI and European Council decision in 2014 to categorise ELIXIR as one of Europe's three priority research infrastructures. A broader description of the ELIXIR Infrastructure, platforms and communities can be found in J. Harrow, R. Drysdale, A. Smith, S. Repo, J. Lanfear, and N. Blomberg (submitted). Here, we focus on the developments that have direct impact on users of bioinformatics services built on the ELIXIR infrastructure, funded through ELIXIR-EXCELERATE.

ELIXIR has worked to meet its key challenges around data sharing, reuse and resource sustainability by consolidating Europe's national centres and bioinformatics resources into a coordinated infrastructure (both a technical network and a people network), operating as a distributed virtual organisation. Figure 1 highlights key milestones in its progress, such as the development of ELIXIR Communities, and partnership with the Global Alliance for Genomics and Health, as ELIXIR becomes established as a key European life science infrastructure and moves to a mature operational phase during its 2019–2023 scientific programme. ELIXIR's successful development is underlined by the fact that it now brings together more than 220 institutes within 23 members (22 countries plus EMBL-EBI), meeting the needs of over a half-million life scientists across Europe. ELIXIR increasingly ensures that users (individual scientists, companies, large consortia

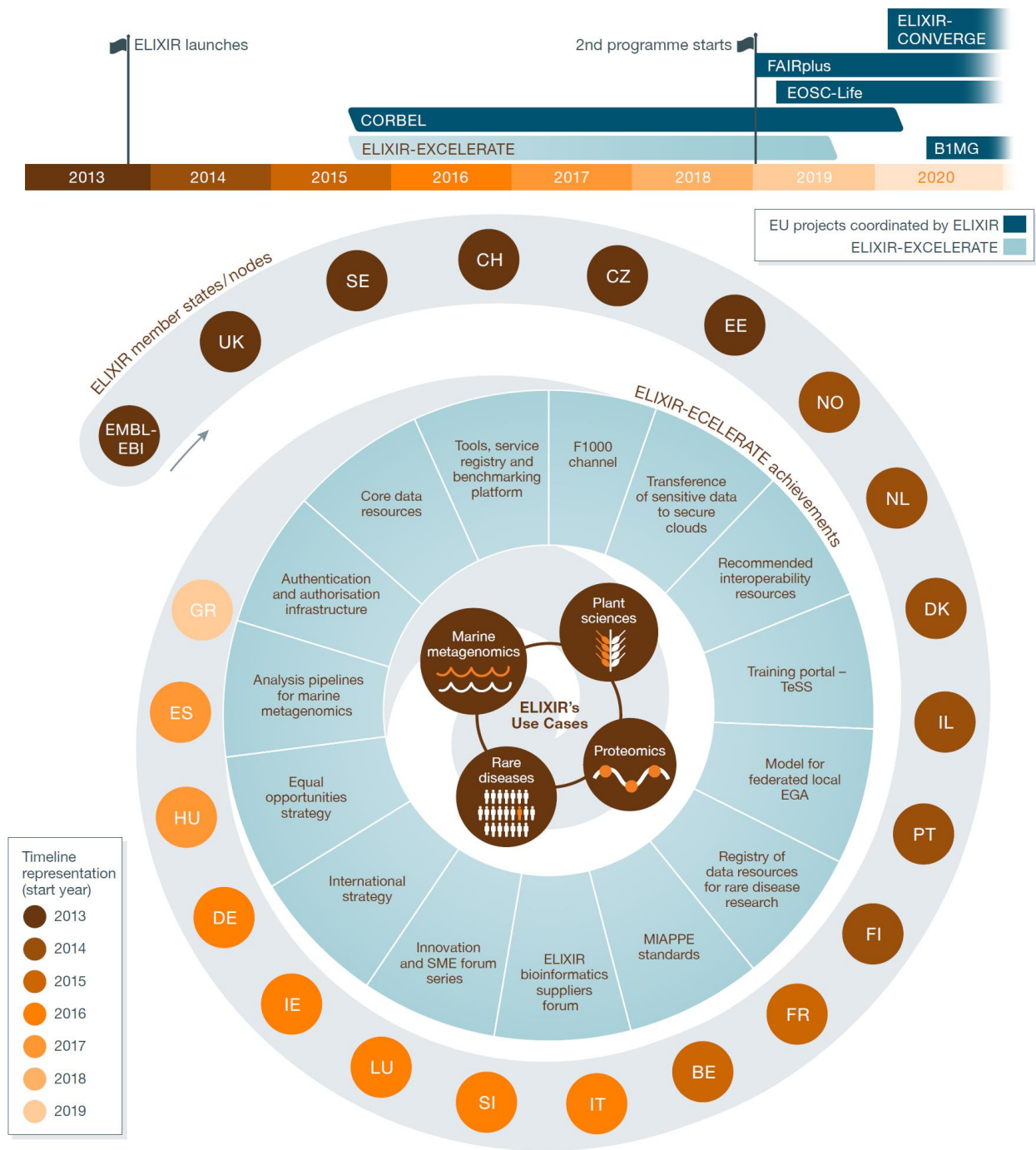
ELIXIR Hub, Wellcome Genome Campus, Hinxton, Cambridge, UK

\*Corresponding author. E-mail: [niklas.blomberg@elixir-europe.org](mailto:niklas.blomberg@elixir-europe.org)

<sup>†</sup>These authors contributed equally to this work

<sup>‡</sup>ELIXIR-EXCELERATE Community members are listed in Appendix 1.

DOI 10.15252/emboj.2020107409 | The EMBO Journal (2021) 40: e107409 | Published online 9 February 2021



**Figure 1. Development of ELIXIR as Europe's life science data infrastructure.**

Schematic overview of the establishment of ELIXIR, including the ELIXIR-EXCELERATE achievements, a timeline of when members joined and an overview of Use Cases established during the launch of the ELIXIR-EXCELERATE grant.

and other research infrastructures) can easily access data resources, built on strong community standards and safeguarded over

the long term. It has produced a concerted effort to connect national infrastructures that reach out to local and regional centres with

Europe-wide reference data resources and support services for data standards. By coordinating Europe's national and international

capabilities into a coherent infrastructure, our 500,000+ users will seamlessly navigate an ecosystem of life science data services.

In this publication, we summarise the development of ELIXIR (as depicted in Fig 1) in the context of the H2020 ELIXIR-EXCELERATE project. We focus on describing how ELIXIR has developed mechanisms and foundations to align its operations with the FAIR principles, and how this supports the effective use of open life science data. We further illustrate how the expertise from individual ELIXIR member institutions has worked together to align national and international services into a standards-based infrastructure, operating at a pan European scale.

### Building a stable and sustainable infrastructure for biological information across Europe with the aid of ELIXIR-EXCELERATE

Underlying the success of ELIXIR has been its ability to work with users in different domains of life science research. Researchers work on generic solutions that can also be applied in other communities; lessons learnt are then taken from a specific field and widen the uptake of those solutions to other unrelated fields, by sharing them across ELIXIR's members and user communities. ELIXIR provides its expertise via five technical domains of implementation called Platforms—Compute, Data, Interoperability, Tools and Training. The headline outputs of the ELIXIR Platforms during EXCELERATE are summarised in Table 1. For more details, see J. Harrow, R. Drysdale, A. Smith, S. Repo, J. Lanfear, and N. Blomberg (submitted) and the individual Platform web pages [<https://elixir-europe.org/platforms>].

Orthogonal to the Platforms is what have become known as the ELIXIR Communities (originally called Use Cases in ELIXIR-EXCELERATE), which bring together individuals from across the ELIXIR members to identify and address specific issues relevant to that area. The first four Use Cases in ELIXIR-EXCELERATE were Human Data, Rare Diseases, Marine Metagenomics and Plant Sciences, each with their own unique technical and legal challenges in addition for Human Data and Rare Diseases.

ELIXIR facilitates the optimal reuse of existing and future life science data by applying the FAIR principles (Wilkinson *et al*, 2016). Data must be findable, accessible, interoperable and, ultimately, reusable. The

**Table 1. ELIXIR Platforms and major outputs from the ELIXIR-EXCELERATE grant**

Platform	Headline outputs
Compute	Developed the ELIXIR Authentication and Authorisation Infrastructure (ELIXIR AAI) Demonstrated technology to transfer sensitive human data to secure clouds Container orchestration Hybrid cloud capacities Integrated solutions for ELIXIR Communities
Data	Defined criteria for and identified the ELIXIR Core Data Resources Contributed to the establishment of Global BioData Coalition Infrastructure to support community annotation and linking
Interoperability	Identified the set of ELIXIR Recommended Interoperability resources Bioschemas: schema.org submission and adoption for life sciences CWL: workflow interoperability and adoption Developed framework deploying interoperability services Bring your own data and capacity-building workshops
Tools	Developed ELIXIR Tools and Service registry (bio.tools) Developed EDAM ontology for the annotation of tools and services by the community Developed the ELIXIR benchmarking platform (OpenEBench) making DREAM challenges results available
Training	Delivered over 850 training events, to over 19,000 people across 60 countries Established TeSS, ELIXIR's Training Portal, a registry of training events and materials Consolidation and expansion of the network of training providers in Europe Established the ELIXIR Train-the-Trainer programme, E-learning platform and Virtual Coffee Room Developed the ELIXIR Training Toolkit

FAIR principles describe how data, including life science data, can be fully utilised by both humans and computer systems (see Box 1). Well-managed research data in the life sciences generate value in the research community, industry, education and society at large, far beyond the initial researcher's laboratory. For example, an impact report of the European Bioinformatics Institute showed that its value to researchers and funders was over 20 times its operational cost [<https://beagrie.com/static/resource/EBI-impact-report.pdf>]. In ELIXIR's vision, the FAIR principles apply not only to the data, but also to the tools and workflows used to analyse and interoperate them, the training resources needed to build capacity internationally to analyse and manage the data, and the compute infrastructure needed to access and analyse data at scale.

ELIXIR is an open infrastructure and does not “own” or operate data resources or other services. Rather, it provides a coordinated data and service backbone that allows partners (e.g. other Research Infrastructures [<https://www.esfri.eu/health-food>], national resources, institutional archives) to make use of existing resources and connect and interoperate their own resources, building on service levels guaranteed by the ELIXIR branding. Ensuring interoperability between resources and data enables long-term, cost-effective data management and drives

“standards as the default” across the life sciences. However, this also relies on the stability of key datasets that underlie data reuse in the life sciences.

In addition, ELIXIR-EXCELERATE enabled a series of capacity-building activities in emerging scientific areas, such as genome assembly and annotation, where six successful high-level workshops were given across Europe, and single-cell transcriptomics, later growing to form an ELIXIR user Community.

### Establishing an open data framework for European life science through ELIXIR infrastructure

Essential to FAIR data and optimal data reuse is a strategy for data management. A data management strategy defines how to handle the entire data lifecycle. ELIXIR underpins and drives good data management practice in the life sciences, and in future is committed to making data available within the framework of the European Open Science Cloud (EOSC) [<https://www.eosc-portal.eu/>]. ELIXIR promotes open and free data access to the maximum extent possible, since it is difficult, if not impossible, to interoperate and integrate data across a complex web of licences and contractual limitations—discoveries get lost in legal red tape. ELIXIR recognises, however, that restrictions are needed for some data types, such as

**Box 1. ELIXIR has published and bases its work on the following guiding principles for FAIR Data Management in the life sciences (ref <https://f1000research.com/documents/6-1857>):**

- i Open sharing of research data is a core principle for publicly funded research and ELIXIR encourages all funders to adopt Open Data mandates and aims to support those mandates.
- ii Data Management is a crucial part of good scientific practice and research excellence and is being followed up in the CONVERGE project.
- iii Whenever possible, biological research data should be submitted to the recommended community deposition databases.
- iv All data submitted to Open Data archives must be annotated in accordance with community-defined standards.
- v ELIXIR members facilitate the national implementation of a harmonised FAIR Data Management programme for the life sciences.
- vi FAIR Data Management requires professional skills, reusable tools, services and workflows, and adequate resources.
- vii Good research data management requires appropriate funding for data infrastructures.

personal data. Charging for, or restricting access to data, seriously limits the ability of research organisations, both public and private, to exploit and create additional value from collective public research investments.

In the following sections, we describe concrete implementations and standards developed by ELIXIR during the ELIXIR-EXCELERATE project, summarising the general principles that emerge from these examples and how they help both computer and bench scientists go about their work. This serves to underline the value of a large collaborative infrastructure in developing new services that can have direct benefit to any life scientist.

### Distributed search and access to plant phenotype datasets

The exploitation of modern genomics and phenotyping technologies is increasingly driving the development of new crops and commercial plant cultivars that are needed to address major challenges to be faced by agriculture such as adaptation to climate change, decreasing its environmental impacts and feeding the expanding population. Plant phenotype data are central to the development of new and improved crops and to identifying the genomic regions underlying particular traits. This type of data is difficult to find because there are no central repository and no plan to build one. Indeed, the heterogeneous nature of phenotype data led to the implementation of diverse infrastructures and experimental platforms backed by specialised data collection and management schemes, including dedicated ontologies used to describe phenotypes of interest

within specific plant communities. Nevertheless, it is important that these datasets become FAIR to allow improved reproducibility and reusability across the different communities working with plant phenotype data. To address this, the ELIXIR Plant Sciences Community played a major role in extending and further developing the MIAPPE standard [<https://www.miappe.org/>] to describe plant phenotyping experiments (making the phenotype data more readily interpretable) and the Breeding API (BrAPI) standard [<https://brapi.org/>], which allows machine-actionable access to disparate datasets. Based on these innovations, the Community has developed the FAIDARE BrAPI-based portal [<https://urgi.versailles.inra.fr/aidare/>], a data discovery service to search relevant, ontology-annotated datasets with linkage to the European Nucleotide Archive (ENA) [<https://www.ebi.ac.uk/ena/browser/home>]. The FAIDARE federation currently indexes eight resources, providing access to over 150,000 datasets (as of 3 September 2020). This work is led by the ELIXIR France Node, in collaboration with EMPHASIS, the plant phenotyping ESFRI, international groups such as the MIAPPE and BrAPI consortia, and with support from ELIXIR's Interoperability Platform.

### Anonymised prioritisation of disease-related genomic variants

Sharing sensitive human genomic data across borders is essential to gain an understanding of the genetic basis of diseases, especially in the case of rare diseases where large datasets are needed and single countries (particularly smaller countries with inherently smaller local datasets) do not

have sufficient infrastructure to archive and distribute these data. Human genetic data raise specific issues with regard to findability, as data must not be identifiable (i.e. traceable to a single individual). To address this, the ELIXIR Federated Human Data Community is working with the Global Alliance for Genomics and Health (GA4GH) use the Beacon discovery service for resources across ELIXIR. A Beacon is defined as a web-accessible service that can be queried for information about a specific allele, with no reference to a specific sample or patient (Fiume *et al*, 2019). Lightweight metadata provided by a data resource (a "Beacon") can be interrogated to ask "Have you observed this nucleotide at this specific chromosomal position?", and the query response is a "Yes" or "No" answer. A Beacon may serve data from case-level observations, such as genetic variants identified from sequenced samples, or from annotation resources, such as variant-disease associations curated from scientific literature. Beacons represent an important step towards collaborative, responsible sharing of human genomic data, compatible with sharing information about identifiable data and the European GDPR regulation.

Work within ELIXIR has driven the establishment of a network of ELIXIR Beacons via strategic partnering with national data owners to enable data flow to the Beacon service. Development of the Beacon infrastructure has involved strong interactions with the ELIXIR Training, Compute and Interoperability Platforms, reflecting recognition that Beacons represent a simple and useful mechanism for data discovery. Our work further aims to increase the integration of the Beacon API with human data resources throughout ELIXIR and extend its application to other data resources, and currently, there are 42 international organisations using this API to serve > 1,000,000 anonymised human samples across 200 datasets.

### Privacy-compliant access to human genomic data

The European Genome-Phenome Archive (EGA) is a database infrastructure for archiving and distributing sensitive human genomic and phenomic data that, by definition, require controlled access. As with Beacons, the key issue for the EGA is to protect sample confidentiality while

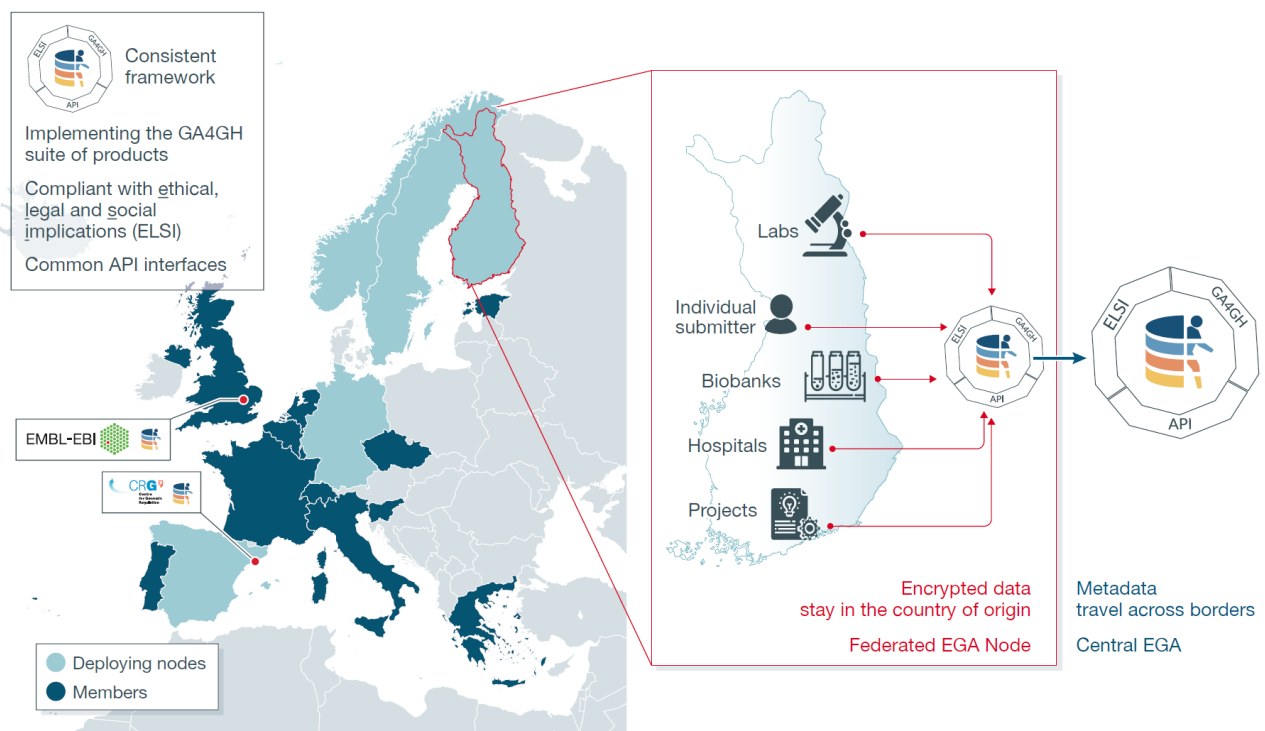
enabling research. The EGA complements Beacons by making confidential data accessible. EGA was founded in 2008, and in 2013, a collaboration was established to mirror the infrastructure at the Centre for Genomic Regulation (CRG) in Barcelona, part of ELIXIR—Spain. During ELIXIR-EXCELERATE, ELIXIR supported improvements in the submission process to the EGA and the development of local EGA, an infrastructure allowing easy, local installation of EGA to be used for the collection of genome and phenome information at the national, regional or even institutional level.

Local EGA [<https://ega-archive.org/federated>] allows deposition of sensitive human data locally (complying with national guidelines for storing that data), but enables data reuse across national boundaries. Local EGAs store metadata from the central EGA, allowing the use of the local EGA to search

both the main and local EGA. Search and retrieval of information from the local EGA is also possible using the local EGA API, allowing the building of local services based on the data available. Additionally, the central EGA gathers non-sensitive metadata (e.g. dataset descriptions) from all the data submitted to local EGAs, so a search at the central EGA allows data located across all local EGAs to be found. The local EGA activity links to the Compute, Training and Interoperability Platforms of ELIXIR. The collaborative work of local EGA instances with the central EGA is collectively known as the Federated EGA. Indeed, the Federated EGA (see Fig 2) is part of the European COVID-19 Data Platform [<https://www.covid19dataportal.org/>], which has been established to facilitate the sharing of national SARS-CoV-2 viral and associated host sequence datasets.

### ELIXIR infrastructure for rare disease research

The ELIXIR-EXCELERATE Rare Disease Community laid the foundation towards building a robust bioinformatics infrastructure at the European level for and with the rare diseases (RD) community. The catalogue of ELIXIR Rare Disease resources (<https://rare-diseases.bio.tools>) was developed in collaboration with the ELIXIR Tools Platform and currently contains 133 relevant referenced tools. Some of these tools are directly linked to OpenEBench (<https://openbench.bsc.es/dashboard>; part of the ELIXIR Tools Platform) where public reference datasets have been made available to the community for the benchmarking of genomic variant calling tools and pipelines. The Community also benchmarked and established a genomic variant calling



**Figure 2. Federated access to European Genomes.**

On the left, 17 of the 23 Nodes are members of the ELIXIR Federated Human Data Community. EMBL-EBI (Cambridge, UK) and CRG (Barcelona, Spain) are specifically highlighted as these are the host institutes of the central EGA. Five Federated EGA deploying Nodes (Finland, Germany, Norway, Spain and Sweden) are also highlighted. These implemented the Federated EGA framework in the first wave to manage archival, access and analysis of sensitive human data. On the right, a schematic view of the ELIXIR Finland Federated EGA deployment. Sensitive human data generated at laboratories, BioBanks and hospitals, and/or by individual projects and submitters, are stored in encrypted format within the countries' jurisdiction. These sensitive data never leave the Finnish borders. Metadata to describe the datasets is shared with the central EGA, which enables findability of these data. Authorised users are able to access these sensitive data remotely thanks to the suite of interoperable GA4GH standards.

pipeline that was integrated in the RD-Connect Genome-Phenome Analysis Platform (GPAP; <https://platform.rd-connect.eu>) and, in collaboration with the Tools and Interoperability Platform, was adapted to GA4GH [<https://www.ga4gh.org/>] and Common Workflow Language (CWL) [<https://www.commonwl.org/>] standards to run on the WES-TES GA4GH cloud [<https://www.ga4gh.org/news/ga4gh-wes-api-enable-s-portable-genomic-analysis/>].

The ELIXIR-EXCELERATE Rare Disease Community also contributed to the definition of the FAIR Guiding Principles (Wilkinson *et al*, 2016) and generic services such as the DCAT-based FAIR data point specification [<https://github.com/FAIRDataTeam/FAIRDataPoint-Spec>], Bioschemas [<https://bioschemas.org/>] extensions, ontology services and their tailoring for RD research, in collaboration with the Interoperability Platform. This data FAIRification process has subsequently been applied to several RD registries (e.g. Osteogenesis imperfecta in collaboration with the Rizzoli Institute in Bologna, Italy, and vascular anomalies in collaboration with the Radboud Medical Centre and their registry software provider Castor EDC, the Netherlands), and is being further developed and scaled up through the European Joint Program Rare Disease with the aim to develop sustainable FAIRification services and integration in routine RD workflows for establishing a FAIR-based virtual platform for rare disease multidisciplinary research.

### Processing and deposition data resources for biodiversity data from the marine metagenome

Metagenomic data are a relatively new source of genomic data derived from samples from a wide range of environments, ranging from marine and soil, to the human gut. Standards to process and deposit data, for assembly of metagenomic-assembled genomes (MAGs), and their deposition into appropriate databases, have been lacking. As a first step to addressing this, and to increase the amount of data available from the marine environment, the ELIXIR Marine Metagenomics Community (with major involvement from EMBL-EBI and ELIXIR—Norway, ELIXIR—France and ELIXIR—Italy) refined databases and tools specific to marine metagenomics and worked to provide better integration and compatibility

across those workflows and tools. In collaboration with the Interoperability and Compute Platforms, the Community drove the development and use of CWL for the description of metagenomic analysis pipelines to increase transparency and reproducibility (e.g. <https://github.com/EBI-MetaGenomics/pipeline-v5>). Furthermore, the use of these more formal tool and pipeline descriptions allows them to be rapidly repurposed to establish a transcriptome annotation pipeline (<https://github.com/EBI-MetaGenomics/workflow-is-cwl>), with the outputs forming the backbone of MetDB [<http://metdb.sb-roscoff.fr/metdb/>], a new micro-eukaryotic marine transcriptome database, which is being adopted within EOSC-Life. The Community published best practices (ten Hoopen *et al*, 2017) that serve as a foundation for a community standard to enable reproducibility and better sharing of metagenomic datasets. In future, the Marine Metagenomics Community is planning to broaden its scope to focus on the microbiome as a whole, enabling a larger community to benefit from the workflows and tools developed through ELIXIR-EXCELERATE.

### Principles emerging from the work of ELIXIR-EXCELERATE

ELIXIR provides small amounts of funding to support infrastructure elements relating, for example, to the needs of particular Communities. The ELIXIR-EXCELERATE use cases became the first members of the ELIXIR Communities. A number of projects funded by ELIXIR have worked to improve workflows to analyse data. As well as the work of the ELIXIR Marine Metagenomics Community, there is an ongoing effort to standardise workflows for fluxomics by the ELIXIR Metabolomics Community, a community that emerged within ELIXIR after the start of ELIXIR-EXCELERATE. Underlying this activity is the adoption of new standards and technical developments for workflow description. A trailblazer for this was the adoption of CWL, used to describe workflows by the ELIXIR Marine Metagenomics Community. Additionally, work is taking place under the auspices of the ELIXIR Galaxy Community to improve Galaxy's utility as a tool for reproducible analysis, including improving the use of software virtualisation using different container technologies.

Key to the reuse of data is ease of deposition into central databases, often Core Data

Resources (CDRs) (Durinx *et al*, 2017) such as the ENA, and the adoption of metadata standards (including ontologies) to describe data and make it more understandable and reusable. Many ELIXIR Communities have undertaken work to improve deposition of data into central databases, notably the Plant Science, Marine Metagenomics, Metabolomics and Proteomics Communities. These efforts improve submission to a range of databases; not only ENA, but also MetaboLights [<https://www.ebi.ac.uk/metabolights/>] and PRIDE [<https://www.ebi.ac.uk/pride/>], and by doing so, make the data accessible to the broader scientific community. Major databases have their own standards for data description but many have evolved in response to new requirements from ELIXIR Communities.

In some cases, data cannot be readily consolidated in a central database. The reasons for this differ. For human data, it may not be possible or desirable for data to cross-national borders for regulatory reasons. ELIXIR has addressed this by building on two interlocking solutions—the Federated EGA, linked to the central EGA archive, to allow local storage of data in a standard format combined with regulated sharing of metadata; and Beacons, which allow non-identifiable identification of potentially useful datasets. In the case of distributed datasets such as those handled by the Plant Science Community, the barriers are more technical in nature, reflecting the huge disparity of the data to be described. In this case, there has also been a need to gain adoption of metadata standards to describe the provenance of datasets. The adoption of BrAPI was predicated on the adoption of the MIAPPE metadata standard by databases that wished to be part of the FAIDARE network. This multi-layered approach provides an excellent example of how the development of a suite of standards can deliver reusable data to a community of researchers.

### How does ELIXIR'S work help the working scientist?

Much of ELIXIR'S work during the ELIXIR-EXCELERATE project was directed to developing guidelines and approaches for the FAIRification of data in different aspects. It increased the findability of human, plant and marine metagenome data using a variety of infrastructures, either to deposit data



in central databases, or to federate datasets when consolidation has not been a practical solution. Accessibility does not exclusively pertain to data, but also to other types of objects such as software tools, workflows and training materials. Therefore, ELIXIR has extended FAIRness of software resources via the bio.tools registry [<https://bio.tools/>], part of the ELIXIR Tools Platform, which makes descriptions of, and access to, research software resources easier and more standardised and provided the TeSS registry for training materials, training workflows and training events [<https://tess.elixir-europe.org/>] which enables scientists to find and access training resources easily. FAIRsharing [<https://fairsharing.org/>] provides curated resources on data and metadata standards, enabling interoperability of datasets and software, both via registries and specifications that can be applied *at source* (e.g. Bioschemas and the DCAT-based FAIR data point specification). This is the kind of work that is often invisible to many researchers in the life science arena, but it results in working data processing resources, and better described datasets that are more suitable for data reuse.

## Conclusions and future directions

The evolution of ELIXIR during the ELIXIR-EXCELERATE project resulted in a mature infrastructure that benefits the European life science community at a number of levels and was an essential learning phase for ELIXIR. At the highest level, it drove the evolution of national bioinformatics communities by the formation of the national ELIXIR Nodes. Various themes have emerged across the broad range of activities improving the FAIRness of data and the software resources used to process and analyse that data. A lot was learnt about how to ensure data findability and accessibility using a variety of mechanisms, and about the hard work needed to make data and software interoperable. To build on this, ELIXIR has initiated a tools ecosystem that will integrate diverse research software descriptions through its registries such as bio.tools and Biocontainers [<https://biocontainers.pro/>], benchmarking through OpenEBench and integration of the workflows through the WorkflowHub registry [<https://workflowhub.eu/>]. To improve data discoverability, a key future development will be to widen the uptake of the Bioschemas

standard which allows the discovery of datasets on the web, and via tailored tools.

Thanks to the developments during the ELIXIR-EXCELERATE project, ELIXIR was able to quickly respond to the 2019–2020 SARS-CoV-2 pandemic (Blomberg & Lauer, 2020). For example, the ELIXIR Galaxy Community, with close links to the Tools, Training and Compute Platforms, has played a key role in the European efforts to identify potential therapeutic small molecules against the SARS-CoV-2 Spike protein [<https://covid19.galaxyproject.org/cheminformatics/#background>]. In collaboration with the INSTRUCT-ERIC ESFRI and the UK Diamond Light Source, the Galaxy Community provided distributed compute infrastructure for the implementation of rapid, parallel workflows to prioritise potential small molecules. This made use of the recently implemented PULSAR network, which enables a job execution system distributed across several European centres, allowing the scaling of the computing power of Galaxy instances over different resources. The Galaxy Community has also driven the development of a European network of accessible Galaxy servers [<https://galaxyproject.eu/>]. More broadly, ELIXIR has supported research into SARS-CoV-2 across its many Platforms [<https://elixir-europe.org/services/covid-19>], including facilitating the development of the COVID-19 Disease Map [<https://covid19map.elixir-luxembourg.org/minerva/>].

A key objective of ELIXIR is the long-term sustainability of datasets and software. To stabilise datasets over the long term, a major aspect is to ensure stable funding of key databases and remove them from the usual funding cycle based on expectation of scientific innovation. Building on its development of its CDRs, a process developed during the ELIXIR-EXCELERATE project, ELIXIR contributed to the establishment of the Global BioData Coalition [<https://globalbiodata.org/>], whose aim is to coordinate national funders worldwide to support major data resources. For software, ELIXIR sees the European Open Science Cloud as a key infrastructure for maintaining widely usable workflows, making them accessible for any life science scientist to use and, in the context of infrastructures such as Galaxy, to modify workflows to support individual needs. ELIXIR coordinates the EOSC-Life project [<https://www.eosc-life.eu/>], which aims to facilitate access to life science data, tools and workflows in the

context of a hybrid cloud infrastructure, across a range of data types provided by the various life science ESFRI infrastructures.

ELIXIR's work on workflows leading to data deposition has surfaced the importance of pre-submission data management. To address this, ELIXIR coordinates the ELIXIR-CONVERGE project [<https://elixir-europe.org/about-us/how-funded/eu-projects/converge>], bringing together infrastructure and national expertise in data management across its members. Capacity building is a key output of ELIXIR-EXCELERATE, and its training activities, including its TeSS registry and standards for training courses, including post-training follow-up, continue to develop.

Human data remain a key priority for ELIXIR, which is achieved via its own technical developments, community coordination via its Human Data Communities and coordinating European engagement with initiatives such as GA4GH and B1MG (the Beyond One Million Genomes project) [<https://b1mg-project.eu/>]. More broadly, the ELIXIR Community structure, which brings together experts in particular technical and scientific areas with the potential to carry out small projects to develop infrastructural components, is a key way for ELIXIR to learn what needs to be done in the future, and expand the areas in which the ELIXIR infrastructure is usable by different stakeholders.

In conclusion, since 2014, ELIXIR has evolved into a dynamic yet well-developed infrastructure enabling state-of-the-art life science research. ELIXIR combines technical and coordination activities both across Europe and globally. Its vision for the future is shaped by its constituent communities, both formal and informal, and is focussed on building a technical infrastructure to provide FAIR data and software, structures that deliver capacity building within its Nodes, and sustainability of the data and tools ecosystem upon which life science scientists increasingly rely.

## Acknowledgements

JH, JH and NB guided the development, writing and final edits. We wish to acknowledge all those who have engaged in and supported ELIXIR activities who we were unable to cite explicitly in these publications due to citation limits. ELIXIR core funding is contributed by its Member States. ELIXIR implementation is supported by ELIXIR-EXCELERATE, funded by the European Commission within the Research

Infrastructures Programme of Horizon 2020, Grant Number 676559. The TeSS portal is also supported by BBSRC UK (Delivering ELIXIR-UK Grant, Grant Agreement Number BB/L005050/1).

### Author contributions

We confirm that the funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

### Conflict of interest

The authors declare that they have no conflict of interest.

### References

- Blomberg N, Lauer KB (2020) Connecting data, tools and people across Europe: ELIXIR's response to the COVID-19 pandemic. *Eur J Hum Genet* 28: 719–723
- Cook CE, Bergman MT, Finn RD, Cochrane G, Birney E, Apweiler R (2016) The European bioinformatics institute in 2016: data growth and integration. *Nucleic Acids Res* 44: D20–D26
- Durinx C, McEntyre J, Appel R, Apweiler R, Barlow M, Blomberg N, Cook C, Gasteiger E, Kim J-H, Lopez R et al (2017) Identifying ELIXIR core data resources. *F1000Res* 5: 2422
- Fiume M, Cupak M, Keenan S, Rambla J, de la Torre S, Dyke SOM, Brookes AJ, Carey K, Lloyd D, Goodhand P et al (2019) Federated discovery and sharing of genomic data using Beacons. *Nat Biotechnol* 37: 220–224
- ten Hoopen P, Finn RD, Bongo LA, Corre E, Fosso B, Meyer F, Mitchell A, Pelletier E, Pesole G, Santamaria M et al (2017) The metagenomic data life-cycle: standards and best practices. *GigaScience* 6: 1–11
- Imker HJ (2020) Who bears the burden of long-lived molecular biology databases? *Data Sci J* 19: 8
- Rigden DJ, Fernández-Suárez XM, Galperin MY (2016) The 2016 database issue of nucleic acids research and an updated molecular biology database collection. *Nucleic Acids Res* 44: D1–D6
- Vines TH, Albert AYK, Andrew RL, Débarre F, Bock DG, Franklin MT, Gilbert KJ, Moore J-S, Renault S, Rennison DJ (2014) The availability of research data declines rapidly with article age. *Curr Biol* 24: 94–97
- Wilkinson MD, Dumontier M, Ijz A, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE et al (2016) The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 3: 160018
-  **License:** This is an open access article under the terms of the Creative Commons Attribution 4.0 License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
- ### Appendix 1
- #### ELIXIR-EXCELERATE Community
- Niklas Blomberg [N.B.] (0000-0003-4155-5910), ELIXIR Hub, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.
- Søren Brunak [S.B.] (0000-0003-0316-5866), Technical University of Denmark & University of Copenhagen.
- Salvador Capella-Gutierrez [S.C.-G.] (0000-0002-0309-604X), Barcelona Supercomputing Center (BSC).
- Christine Durinx [C.D.] (0000-0003-4237-8899), SIB Swiss Institute of Bioinformatics.
- Chris T. Evelo [C.T.E.] (0000-0002-5301-3142) Department of Bioinformatics - BiGCaT, Research School NUTRIM, Maastricht University.
- Carole Goble [C.G.] (0000-0003-1219-2137; 0000-0001-7219-632X), The University of Manchester.
- Ivo Gut [I.G.] (0000-0001-7219-632X) Centro Nacional de Análisis Genómico (CNAG-CRG), Centre for Genomic Regulation, Barcelona Institute of Technology; Universitat Pompeu Fabra.
- Jon Ison [J.I.] (0000-0001-6666-1520) Technical University of Denmark.
- Thomas Keane [T.K.] (0000-0001-7532-6898), European Bioinformatics Institute.
- Brane Leskošek [B.L.] (0000-0001-5202-2349) University of Ljubljana, Faculty of Medicine, ELIXIR-SI.
- Luděk Matyska [L.M.] (0000-0001-6399-5453) Masaryk University; CESNET.
- Johanna McEntyre [J.M.] (0000-0002-1611-6935) European Bioinformatics Institute.
- Célia Miguel [C.M.] (0000-0002-1427-952X) Biosystems & Integrative Sciences Institute (BioISI), Faculdade de Ciências, Universidade de Lisboa, Portugal; iBET Instituto de Biologia Experimental e Tecnológica, Oeiras, Portugal.
- Arcadi Navarro [A.N.] (0000-0003-2162-8246) Institute of Evolutionary Biology (UPF-CSIC), Department of Experimental and Health Sciences, Pompeu Fabra University, Barcelona Biomedical Research Park, Carrer del Dr. Aiguader 88, 08003 Barcelona, Spain; Catalan Institution of Research and Advanced Studies (ICREA), Passeig de Lluís Companys 23, 08010 Barcelona, Spain; CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Carrer del Dr. Aiguader 88, 08003 Barcelona, Spain; Barcelonaβeta Brain Research Center (BBRC), Pasqual Maragall Foundation, Wellington 30, 08005, Barcelona, Spain.
- Steven Newhouse [S.N.] (0000-0003-1531-5198) European Bioinformatics Institute.
- Tommi Nyrönen [T.N.] (0000-0002-5569-5183) CSC - IT Center for Science Ltd.
- Patricia Palagi [P.P.] (0000-0001-9062-6303) SIB Swiss Institute of Bioinformatics.
- Bengt Persson [B.P.] (0000-0003-3165-5344) NBIS (National Bioinformatics Infrastructure Sweden) and ELIXIR-SE, Department of Cell and Molecular Biology, SciLifeLab, Uppsala University, Sweden.
- Cyril Pommier [C.P.] (0000-0002-9040-8733) Université Paris-Saclay, INRAE, URGI, 78026, Versailles, France; Université Paris-Saclay, INRAE, Bioinformatics, Plant bioinformatics facility, 78026, Versailles, France.
- Jordi Rambla [J.R.] (0000-0001-9091-257X) CRG, Centre for Genomic Regulation, Barcelona, Spain.
- Marco Roos [M.R.] (0000-0002-8691-772X) Leiden University Medical Center, Leiden, Netherlands.
- Gabriella Rustici [G.R.] (0000-0003-3085-1271), University of Cambridge, Cambridge, UK.
- Andrew Smith [A.S.] (0000-0003-3085-1271) ELIXIR Hub, Wellcome Genome Campus, Cambridge, UK.
- Alfonso Valencia [A.V.] (0000-0002-8937-6789) Barcelona Supercomputing Center (BSC), Barcelona, Spain; Institució Catalana de Recerca i Estudis Avançats (ICREA).
- Celia van Gelder [C.G.] (0000-0002-0223-2329) DTL Dutch Techcentre for Life Sciences.
- Jiri Vondrasek [J.V.] (0000-0002-6066-973X) IOCB AS CR.
- Nils Peder Willassen [N.P.W.] (0000-0002-4397-8020) UiT The Arctic University of Norway.
- Juan Arenas [J.A.] (0000-0001-5497-8045), ELIXIR Hub, Wellcome Genome Campus Hinxton, Cambridge, CB10 1SD, UK.
- Helen Parkinson [H.P.] (0000-0003-3035-4195) European Bioinformatics Institute.
- Robert D. Finn [R.D.F.] (0000-0001-8626-2148) European Bioinformatics Institute.

Sergi Beltran [S.B.] (0000-0002-2810-3445) Centro Nacional de Análisis Genómico (CNAG-CRG), Centre for Genomic Regulation, Barcelona Institute of Technology; Universitat Pompeu Fabra; Dep. Genetics, Microbiology and Statistics, Facultat de Biologia, Universitat de Barcelona.

Leslie Matalonga [L.M.] (0000-0003-0807-2570) Centro Nacional de Análisis Genómico (CNAG-CRG), Centre for Genomic Regulation, Barcelona Institute of Technology.

Hannah Hurst [H.H.] (0000-0002-1119-9321), ELIXIR Hub, South Building, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.

Paul Kersey [P.K.] (0000-0002-7054-800X) European Bioinformatics Institute.

Ilkka Lappalainen [I.L.] (0000-0001-5762-893X) CSC - IT Center for Science Ltd.

Pascal Kahlem [P.K.] (0000-0002-8810-444X), Scientific Network Management, Barcelona, Spain.

Gary Saunders [G.S.] (0000-0002-7468-0008), ELIXIR Hub, South Building, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.

Sirarat Sarntivijai [S.S.] (0000-0002-2548-641X) ELIXIR Hub, South Building, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.

Rachel Drysdale [R.D.] (0000-0003-3037-0216), ELIXIR Hub, South Building, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.

Jonathan Tedds [J.T.] (0000-0003-2829-4584) ELIXIR Hub, South Building, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK<sup>§</sup>.

Jeremy Lanfear [J.L.] (0000-0002-8007-5568) ELIXIR Hub, South Building, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.

Jennifer Harrow [J.H.] (0000-0003-0338-3070) ELIXIR Hub, South Building, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.

<sup>§</sup>Correction added on 15 March 2021, after first online publication: the author's name was changed from Johnathan Tedds to Jonathan Tedds.



## **Annex Paper 4**

Lawson, J., Cabili, M. N., Kerry, G., Boughtwood, T., Thorogood, A., Alper, P., Bowers, S. R., Boyles, R. R., Brookes, A. J., Brush, M., Burdett, T., Clissold, H., Donnelly, S., Dyke, S. O., Freeberg, M. A., Haendel, M. A., Hata, C., Holub, P., Jeanson, F., Rambla, J., . . . Courtot, M. (2021). **The Data Use Ontology to streamline responsible access to human biomedical datasets.** *Cell Genomics*, 1(2), 100028.

<https://doi.org/10.1016/j.xgen.2021.100028>



## Technology

# The Data Use Ontology to streamline responsible access to human biomedical datasets

Jonathan Lawson,<sup>1,37</sup> Moran N. Cabili,<sup>1,37</sup> Giselle Kerry,<sup>2</sup> Tiffany Boughtwood,<sup>3</sup> Adrian Thorogood,<sup>4,5</sup> Pinar Alper,<sup>5</sup> Sarion R. Bowers,<sup>6</sup> Rebecca R. Boyles,<sup>7</sup> Anthony J. Brookes,<sup>8</sup> Matthew Brush,<sup>9</sup> Tony Burdett,<sup>2</sup> Hayley Clissold,<sup>6</sup> Stacey Donnelly,<sup>1</sup> Stephanie O.M. Dyke,<sup>10</sup> Mallory A. Freeberg,<sup>2</sup> Melissa A. Haendel,<sup>9</sup> Chihiro Hata,<sup>11</sup> Petr Holub,<sup>12</sup> Francis Jeanson,<sup>13</sup> Aina Jene,<sup>14</sup> Minae Kawashima,<sup>15</sup> Shuichi Kawashima,<sup>16</sup> Melissa Konopko,<sup>17</sup> Irene Kyomugisha,<sup>18</sup> Haoyuan Li,<sup>19</sup> Mikael Linden,<sup>20</sup> Laura Lyman Rodriguez,<sup>21</sup> Mizuki Morita,<sup>22</sup> Nicola Mulder,<sup>23</sup> Jean Muller,<sup>24,25</sup>

(Author list continued on next page)

<sup>1</sup>Broad Institute of Harvard and the Massachusetts Institute of Technology, Cambridge, MA, USA

<sup>2</sup>European Molecular Biology Laboratory—European Bioinformatics Institute (EMBL-EBI), Hinxton, UK

<sup>3</sup>Australian Genomics, Murdoch Children's Research Institute, Parkville, VIC, Australia

<sup>4</sup>Centre of Genomics and Policy, Department of Human Genetics, McGill University, Montreal, QC, Canada

<sup>5</sup>ELIXIR-Luxembourg, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Belvaux, Luxembourg

<sup>6</sup>Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK

<sup>7</sup>RTI International, Research Triangle Park, NC, USA

<sup>8</sup>University of Leicester, Leicester, UK

<sup>9</sup>University of Colorado Anschutz Medical Campus, Aurora, CO, USA

<sup>10</sup>McGill Centre for Integrative Neuroscience, Montreal Neurological Institute, Department of Neurology & Neurosurgery, Faculty of Medicine, McGill University, Montreal, QC, Canada

<sup>11</sup>Bioinformatics and DDBJ Center, National Institute of Genetics, Mishima, Japan

<sup>12</sup>BBMRI-ERIC, AT and Masaryk University, Brno, Czech Republic

<sup>13</sup>University Health Network, Toronto, ON, Canada

<sup>14</sup>Centre de Regulació Genòmica (CRG), Barcelona, Spain

<sup>15</sup>National Bioscience Database Center, Japan Science and Technology Agency, Tokyo, Japan

<sup>16</sup>Database Center for Life Science, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, Kashiwa, Japan

(Affiliations continued on next page)

## SUMMARY

Human biomedical datasets that are critical for research and clinical studies to benefit human health also often contain sensitive or potentially identifying information of individual participants. Thus, care must be taken when they are processed and made available to comply with ethical and regulatory frameworks and informed consent data conditions. To enable and streamline data access for these biomedical datasets, the Global Alliance for Genomics and Health (GA4GH) Data Use and Researcher Identities (DURI) work stream developed and approved the Data Use Ontology (DUO) standard. DUO is a hierarchical vocabulary of human and machine-readable data use terms that consistently and unambiguously represents a dataset's allowable data uses. DUO has been implemented by major international stakeholders such as the Broad and Sanger Institutes and is currently used in annotation of over 200,000 datasets worldwide. Using DUO in data management and access facilitates researchers' discovery and access of relevant datasets. DUO annotations increase the FAIRness of datasets and support data linkages using common data use profiles when integrating the data for secondary analyses. DUO is implemented in the Web Ontology Language (OWL) and, to increase community awareness and engagement, hosted in an open, centralized GitHub repository. DUO, together with the GA4GH Passport standard, offers a new, efficient, and streamlined data authorization and access framework that has enabled increased sharing of biomedical datasets worldwide.

## INTRODUCTION

To address global scientific challenges in health, human biomedical data must be shared and integrated worldwide.<sup>1</sup> To promote discovery and improve healthcare, researchers and clinicians

need to be able to find, access, harmonize, and re-use data from diverse data sources. Data access for research is often facilitated by data repositories, and in a growing number of federated data environments<sup>2</sup> that aggregate datasets within or among themselves and make the results available to the



Satoshi Nagaie,<sup>26</sup> Jamal Nasir,<sup>27</sup> Soichi Ogishima,<sup>26</sup> Vivian Ota Wang,<sup>28</sup> Laura D. Paglione,<sup>29</sup> Ravi N. Pandya,<sup>30</sup> Helen Parkinson,<sup>2</sup> Anthony A. Philippakis,<sup>1</sup> Fabian Prasser,<sup>31</sup> Jordi Rambla,<sup>14</sup> Kathy Reinold,<sup>1</sup> Gregory A. Rushton,<sup>1</sup> Andrea Saltzman,<sup>1</sup> Gary Saunders,<sup>17</sup> Heidi J. Sofia,<sup>32</sup> John D. Spalding,<sup>2</sup> Morris A. Swertz,<sup>33</sup> Ilia Tulchinsky,<sup>34</sup> Esther J. van Enkevort,<sup>33</sup> Susheel Varma,<sup>35</sup> Craig Voisin,<sup>34</sup> Natsuko Yamamoto,<sup>36</sup> Chisato Yamasaki,<sup>36</sup> Lyndon Zass,<sup>23</sup> Jaime M. Guidry Auvil,<sup>28</sup> Tommi H. Nyrönen,<sup>20</sup> and Mélanie Courtot<sup>2,38,\*</sup>

<sup>17</sup>ELIXIR Hub, Wellcome Genome Campus, Hinxton, UK

<sup>18</sup>Division of Human Genetics, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa

<sup>19</sup>Canada's Michael Smith Genome Sciences Centre, Vancouver, BC, Canada

<sup>20</sup>ELIXIR-Finland, CSC - IT Center for Science Ltd, Espoo, Finland

<sup>21</sup>Patient-Centered Outcomes Research Institute, Washington, DC, USA

<sup>22</sup>Okayama University, Okayama, Japan

<sup>23</sup>Computational Biology Division, IDM, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa

<sup>24</sup>Laboratoire de Génétique Médicale, Institut de Génétique Médicale d'Alsace, INSERM U1112, Université de Strasbourg, Strasbourg, France

<sup>25</sup>Laboratoire de Diagnostic Génétique, Institut de Génétique Médicale d'Alsace, Hôpitaux Universitaires de Strasbourg, Strasbourg, France

<sup>26</sup>Tohoku Medical Megabank Organization (ToMMo), Tohoku University, Sendai, Japan

<sup>27</sup>Department of Life Sciences, University of Northampton, Northampton, UK

<sup>28</sup>Office of Data Sharing, National Cancer Institute, NIH, Rockville, MD, USA

<sup>29</sup>Spherical Cow Group, Rego Park, NY 11374, USA

<sup>30</sup>Microsoft Research, Redmond, WA 98052, USA

<sup>31</sup>Berlin Institute of Health at Charité—Universitätsmedizin Berlin, Berlin, Germany

<sup>32</sup>National Human Genome Research Institute, NIH, Bethesda, MD, USA

<sup>33</sup>Genomics Coordination Center, Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands

<sup>34</sup>Google Cloud, Kitchener, ON N2H 5G5, Canada

<sup>35</sup>Health Data Research UK, Gibbs Building, 215 Euston Road, London NW1 2BE, UK

<sup>36</sup>Osaka University, Osaka, Japan

<sup>37</sup>These authors contributed equally

<sup>38</sup>Lead contact

\*Correspondence: [mcourtot@gmail.com](mailto:mcourtot@gmail.com)

<https://doi.org/10.1016/j.xgen.2021.100028>

research community. Challenges arise in the aggregation of datasets with varying ethical or regulatory conditions on data reuse. Different conditions may stem from different applicable data protection laws (e.g., limits on allowable purposes of processing, transfers to third countries), informed consents (e.g., specific vs. broad), policies (e.g., IRB data release authorizations), or data sharing agreements (e.g., within consortia).<sup>3</sup> Due to this heterogeneity of re-use conditions, it can be difficult for researchers to search and find appropriate datasets, methods of requesting and accessing those datasets vary, and there is no shared understanding of the allowable uses and/or downstream analyses of the data once access is approved.

Current processes to access sensitive human biomedical data can be cumbersome, time and cost intensive, and variable between repositories. In typical workflows, Data Access Committees (DACs) manually review data use terms; this process can be delayed by the need to interpret data use terms often described in inconsistent and ambiguous language. There can also be inconsistency in access determinations across DACs, particularly for broadly defined data use terms, such as “permitted use for a disease and related conditions.” Similarly, language in a consent form prohibiting “commercial use” has been interpreted differently by DACs, ranging from not allowing commercial organizations access to the data to not allowing the data to be used for commercial purposes— independently of the organization type. Finally, these interpretations can shift over time, increasing the risk that data are used in a way that does not reflect what the research participant

originally agreed to and leading to inconsistent data sharing practices.

To address the needs for consistent terminology and reliable interpretations of allowable data uses, the GA4GH Data Use and Researcher Identities (DURI) work stream<sup>4</sup> developed a data authorization and access framework to streamline the process for granting researchers access to biomedical datasets based on their credentials and research purposes. A main component of this framework is the Data Use Ontology (DUO), a standard, machine-readable vocabulary of data use terms that enables direct matching between data use conditions and intended research use. DUO is complemented by the GA4GH Passport standard (see Voisin et al. in this issue),<sup>5</sup> which provides a machine-readable representation of a researcher's data access permissions. Together, the GA4GH DUO and Passport standards enable automating access by researchers to multiple datasets based on their authentication and authorization levels and has been deployed by various organizational members of the GA4GH DURI work stream. DUO is now the accepted GA4GH standard for data use terms, based on use cases from several GA4GH Driver Projects.<sup>6</sup> Australian Genomics, EGA, GENome Medical alliance (GEM) Japan, Human Heredity & Health in Africa (H3Africa), U.S. National Heart, Lung, and Blood Institute, BBMRI-ERIC, and U.S. National Cancer Institute have all contributed to the establishment and review of DUO terms, which are aligned with data use terms or phrasing of their respective consent forms. Over 200,000 datasets worldwide have been annotated with



**Table 1. Count of datasets annotated with DUO by data custodian as of February 2021**

Data custodian	Datasets annotated with DUO
Broad Institute	225
Sanger Institute	700
EGA	1,021
HDR UK	568
BBMRI-ERIC	In progress. Manual for data managers with guidance for DUO annotations released: <a href="https://doi.org/10.5281/zenodo.4427731">https://doi.org/10.5281/zenodo.4427731</a>
AMED Biobank Network (GEM Japan)	203,900
Australian Genomics	14
H3Africa	16

A census of datasets annotated with DUO in February 2021 highlights widespread adoption of the standard. Early implementers such as EGA are now requiring DUO annotation upon dataset submission. New partners such as BBMRI-ERIC are only starting the annotation process. AMED Biobank has made a very large number of DUO annotations, as they consider each sample to be its own dataset. An example implementation in the EGA is described in [supplemental information](#).

machine-readable DUO terms (Table 1). DUO has been successfully leveraged by software such as the Broad Data Use Oversight System (DUOS) to enable automated matching between access requests and DUO annotation on datasets (see Cabili et al. in this issue).<sup>7</sup>

In this study, we report on the DUO standard, describe the curated structured vocabulary and hierarchies, and review use cases and considerations in implementing DUO for the management and access of biomedical datasets. DUO has been successfully used to annotate genomics datasets worldwide, and its usage is being expanded to direct mapping into consent forms and automated matching of requests to permissions by DACs. Future uses of DUO include annotation to different data types such as samples and integration within GA4GH Passport visas.

## DESIGN

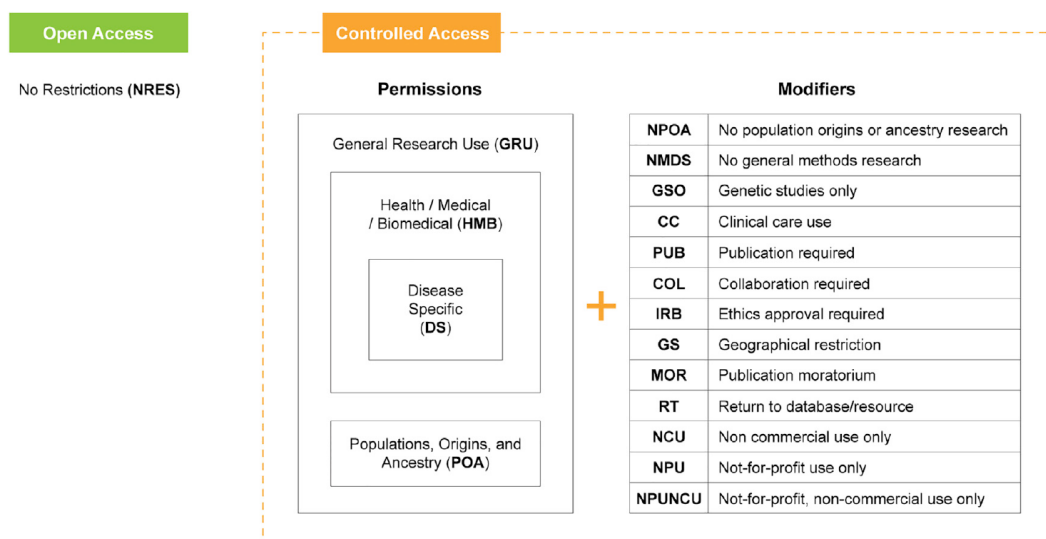
DUO is a structured vocabulary of standard human- and machine-readable data use terms. DUO's original list of terms was informed by review of common terminologies used by major international controlled-access genomic repositories (e.g., U.S. National Institutes of Health database for Genotypes and Phenotypes, NIH dbGaP,<sup>8</sup> and European Genome-Phenome Archive, EGA<sup>9</sup>), as well as policy tools developed by the GA4GH Regulatory and Ethics Work Stream (REWS).<sup>3,10</sup> Contributors from those efforts joined to form the Data Use group, which met regularly both through videoconferences and face-to-face meetings. External efforts such as the Informed Consent Ontology (ICO)<sup>11</sup> were additionally reviewed for interoperability and synergistic evolution; DUO has been directly imported in ICO to describe data use conditions instead of duplicating its content. The DUO terms are intended to be a

simple set of data use terms most often used or referenced in consent forms that include provisions for data sharing. DUO does not aim to represent all possible data use terms, consent phrases, or complex logical permutations of permissions, limitations, or requirements. Structurally, DUO contains 25 terms representing two types of data use terms, permissions and modifiers (Table S1):

- Permission terms include “general research use,” “health or medical or biomedical use,” “disease specific research,” and “population origins or ancestry research only” and are expressly permitted uses or focused areas of research.
- Modifier terms add requirements, limitations, or prohibitions within the permitted boundary (Figure 1).

DUO is use-case driven, and requests for new data use terms in DUO must be supported by specific use cases that promote and facilitate data sharing. Each DUO term was developed based on contributions and reviews from community experts and implementers. Contributions to DUO are public and created by raising GitHub issues;<sup>12</sup> anyone may submit a request to add a new term or comment on an existing request. Requests are discussed by the DUO work stream leads and driver project implementers on the tracker, on the DUO mailing list, and during periodic teleconferences. Once approved, changes are open to the public for further discussion over a comment period of 2 weeks, as per the DUO governance policy.<sup>13</sup>

DUO is implemented in the Web Ontology Language (OWL),<sup>14</sup> a World Wide Web Consortium standard. Development of DUO follows Open Biomedical Ontologies (OBO) development principles,<sup>15</sup> ensuring interoperability with other ontological resources, such as those describing disease entities.<sup>16</sup> As per OBO guidelines, DUO is built under the Basic Formal Ontology (BFO)<sup>17</sup> upper-level ontology. The DUO root terms “data use permission” and “data use modifier” are subclasses of “data item” (IAO:0000027), itself a type of “information artifact entity” (IAO:0000030) and “generically dependent continuant” (BFO:0000031). While BFO provides the framework for the DUO hierarchy, it proved confusing to use for most users. We consequently worked with the developers of the EMBL-EBI Ontology Lookup Service (OLS)<sup>18</sup> to design and implement a system allowing selection of suitable entry levels in the DUO hierarchy. The “preferred root” toggle shown in Figure 2 allows most users to browse only classes of interest, while expert ontologists can instead select the complete view. DUO terms are stable, with each DUO term having its unique Uniform Resource Identifier, which can be browsed using the OLS. Most importantly, the meaning associated with a specific DUO ID is permanent; this guarantees consistency through time of the data use terms. Different versions of DUO are available through the GitHub repository,<sup>19</sup> including an editors' version that captures ongoing development and stable, released versions. Released versions of DUO are associated with permanent URLs (PURLs) for sustainability:<sup>20</sup> the most recent release is always available from <http://purl.obolibrary.org/obo/duo.owl>, while previous versions can be accessed through their date-based PURL, providing choice for users who prefer to use a specific historical view of the ontology<sup>21,22</sup> for stability while transitioning to the latest version.



**Figure 1. Data Use Ontology permissions and modifiers**

DUO is a hierarchical vocabulary of data use terms most often used to denote secondary usage conditions for controlled access datasets. DUO does not aim to represent all possible data use terms, consent phrases, or complex logical permutations of permissions, limitations, or requirements. As of June 2021, DUO contains 25 terms representing two types of data use terms, permissions and modifiers. Permissions such as General Research Use (GRU), Health or Medical or Biomedical use (HMB), Disease Specific research (DS), and Population Origins and Ancestry research (POA) standardize allowed usage of the datasets. Modifiers are used to further qualify main categories of controlled access.

Terms are positioned in the DUO hierarchy, such that subclasses are more specific sets of instances than their parents. This allows for inference of new knowledge through description logic underpinning OWL reasoners.<sup>23</sup> For example, when searching for datasets for a “disease-specific” research use (Figure 2), a researcher would see query results of datasets matching this use term and its parents, “health and biomedical research” (direct superclass) and “general research” (indirect superclass). The initial structure of the repository was generated using the ontology development kit,<sup>24</sup> which provides a way of creating an ontology project ready for pushing to GitHub. Development of the ontology follows a modular approach for greater flexibility both by developers of DUO and its users. For example, the DUO Japanese translation is stored as a separate file from the main ontology. This file is merged in at release time via an automated script, allowing different files and features to remain independent until they are ready to be published and/or to be excluded at release time on demand—for example, for users who do not require translations from English. The same script also executes SPARQL<sup>25</sup> queries to render CSV versions, again for easy human browsing in the GitHub repository. Finally, the script merges relevant subsets of external ontologies imported through the MIREOT method<sup>26</sup> to promote ontology re-use and consistent identification of ontology terms across resources.

To increase community awareness and engagement, DUO is hosted under an open, centralized GitHub repository. This enables tagging of versions and continuous integration tests to be run at each iteration via the Travis CI software. After each

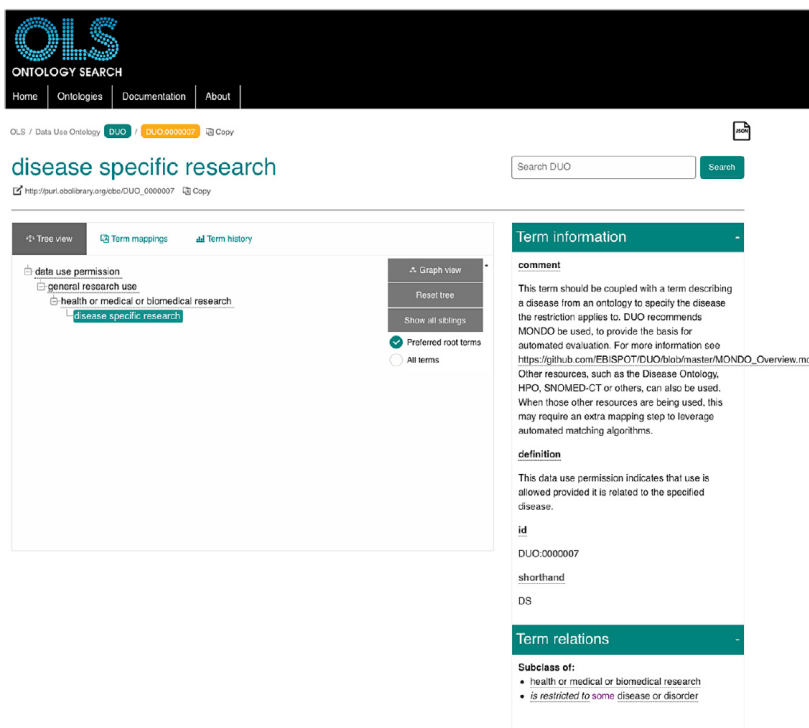
modification of the source file, the ELK reasoner<sup>27</sup> is run to ensure ongoing consistency of the ontology.

## RESULTS

To ensure trustworthiness and sustainability of its technical standards, the GA4GH applies an open and consistent development and product approval process.<sup>1</sup> In 2019, DUO was unanimously approved as a GA4GH standard by the GA4GH Steering Committee, joining other products in the GA4GH Genomic Toolkit suite.<sup>1</sup> Figure 3 displays the current implementers of DUO.

DUO has been incorporated in several central aspects of the data access request process (Box 1). First, DUO terms are applied as dataset metadata to be stored alongside the data they describe in a repository, making it easier for data custodians to manage their datasets compliantly and facilitate researchers’ querying of the datasets by their data use terms. Repositories can add DUO annotations to their dataset files, either retrospectively through curation of existing data or inter-actively at submission time. Users can search for datasets according to data use terms to determine what datasets are available for their purposes before requesting data access. This improved accessibility and interoperability of datasets increases their FAIRness:<sup>28</sup> 2.6% of data requesters who applied for access to Sanger’s Cancer Genome Project (CGP) datasets between April and October 2020 had used the EGA DUO search tool to find re-usable datasets compatible with their research purposes.

In a second use case, DUO terms have been leveraged by DACs to facilitate and, for the first time, automate parts of the



**Figure 2. Browsing the Data Use Ontology**

The DUO OWL file has been loaded in human-friendly browsers such as the Ontology Lookup Service (OLS). This enables interactive navigation through the hierarchy and display of additional properties such as definition, comment, or relations to other terms. For example, the “disease specific research” DUO term, [http://purl.obolibrary.org/obo/DUO\\_0000007](http://purl.obolibrary.org/obo/DUO_0000007), clarifies that it should be used in conjunction with a term from a disease ontology. The “Preferred root terms” button (middle, active green checkbox) guides display of the top classes to be displayed to the user instead of presenting the complex upper-level BFO hierarchy (accessible by selecting “All terms”)

## DISCUSSION

Since its approval as a GA4GH standard,<sup>1</sup> DUO has been widely implemented across diverse biomedical projects worldwide. Beyond requests for and comments on new data use terms, DUO standard implementers have contributed by proposing translations in other languages, such as Japanese, or in “plain language,” which has been shown to increase understanding and participation

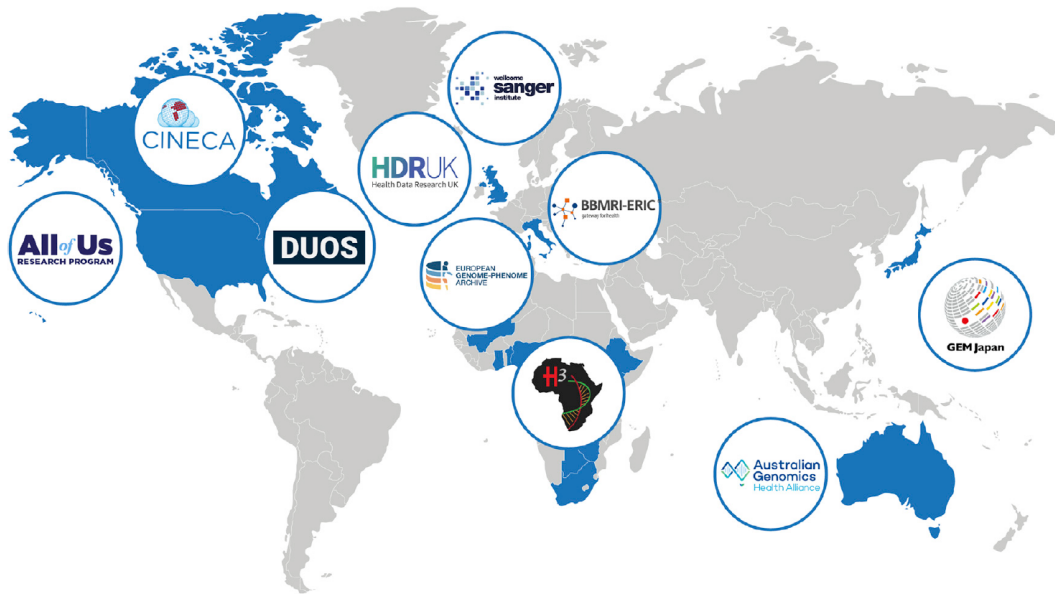
of research participants.<sup>32</sup> To this end, DUO was successfully extended for consent use as the Machine-Readable Consent Guidance described earlier, which was approved as a GA4GH standard in July 2020<sup>33</sup> and is being actively reviewed and implemented by IRBs and research studies. In addition, community members enthused by the success and simplicity of DUO aim to further extend its application beyond genomic datasets to resources such as biological specimens, imaging data, and public health data. The Finnish Institute for Health and Welfare biobank<sup>34</sup> has already implemented DUO in requiring sample depositors to describe sample/data use terms when depositing in their repositories. Indeed, nothing precludes developing applications or extensions of DUO for other scientific resources.

Successful external extensions of the standard can be fed back to GA4GH, allowing for continual improvement in utility and function for the community.

DUO terms can also be used in healthcare settings and alongside complementary standards. Health Level Seven International (HL7)’s Fast Healthcare Interoperability Resources (FHIR)<sup>35</sup> Consent resource,<sup>36</sup> as well as other tools or standards, such as the Automatable Discovery and Access Matrix (ADA-M) or OASIS’s LegalRuleML,<sup>37</sup> use logic for expressing more complex data use rules. The HL7 standard permits an implementer to adopt a default rule for a given use term (e.g., everything permitted by default, everything restricted by default) and then specify exceptions. LegalRuleML and ADA-M explicitly define if a rule for coded data use is a permission, prohibition, or condition. This approach requires users to “translate” their intuitive thinking into machine-based logic and can lead to complexity, confusion, and a greater risk of error.

As a third use case, DUO terms are incorporated into the data sharing language in consent forms written during the study inception.<sup>30,31</sup> Incorporating DUO terms at this early stage is important to enable more effective and consistent data use management. This addresses current challenges in the common use of informed consent language that does not fully capture the scope and issues related to data sharing and secondary research purposes, resulting in uncertainty for participants regarding research expectations as well as for data providers and data stewards or DACs in assessing how datasets can be distributed. The consent clauses in the Machine-Readable Consent Guidance are accompanied by explanations and guidance for consistency, and to ensure prospective capture as machine-readable data use terms. This is currently undergoing evaluation and validation by IRBs, and we anticipate this becoming a recommendation that could be more broadly followed.

data access request process. The use of DUO in electronic data access systems enables automated matching by software algorithm, leveraging the DUO hierarchy and logical structure. An implementation in automating data access requests has been piloted for NIH and the Broad Institute through DUOS<sup>7</sup> and is now being extended to other databases. The DUOS software platform performs automated DUO-based data use oversight and provides interfaces to simplify the work of DACs. An empirical evaluation of the results demonstrates that the DUO is broadly useful, matching ~96% of consent terms in examined datasets, and that using DUOS to automate the process streamlines the review process while maintaining efficacy and consistency.



**Figure 3. Current implementations of the Data Use Ontology**

DUO has been implemented to annotate genomics datasets worldwide. As of November 2021, implementers include repositories, databases, and projects in North America, Europe, Africa, Europe, Asia, and Australia.

### Limitations of the study

The GA4GH DUO standard represents the data use terms commonly used by data management professionals for sharing of biomedical datasets, while minimizing the complexity of logical permutations of data use terms, essential to global interoperability and data sharing.<sup>38</sup> For example, DUO adopts the term “not-for-profit use only” rather than decomposing “profit” and whether it is “allowed,” “forbidden,” or “restricted” in specific instances, thus not requiring users to mix and match terms with potentially opposing meanings; DUO is not built to capture the entire spectrum of possible data use combinations, as pursuing a vocabulary to describe all possible combinations of data use would likely lead to an infinitely complexifying model given the constant increase of possible terms and combination permutations. This intentional limitation of the DUO terminology space has been encouraged by researchers, in line with the DURi leadership’s vision for DUO as a concise standard to facilitate compatibility of terms.

Arguments to the contrary espouse DUO and the aspiration for a limited vocabulary as counter to the needs of specific participant communities. A red herring example often used to justify this contrary position is that rare disease research participants often believe that DUO’s limited scope would not be able to represent the unique, specific diseases they have, such as ataxia-telangiectasia or Diamond-Blackfan anemia. Yet this reflects an inversion of understanding, as permitting unique, edge-case-like types of research would be permissible via many of the existing DUO terms, particularly those such as General Research Use and Health/Medical/Biomedical Use. Annotating those datasets with more general DUO terms also increases the probability of researchers reaching those dis-

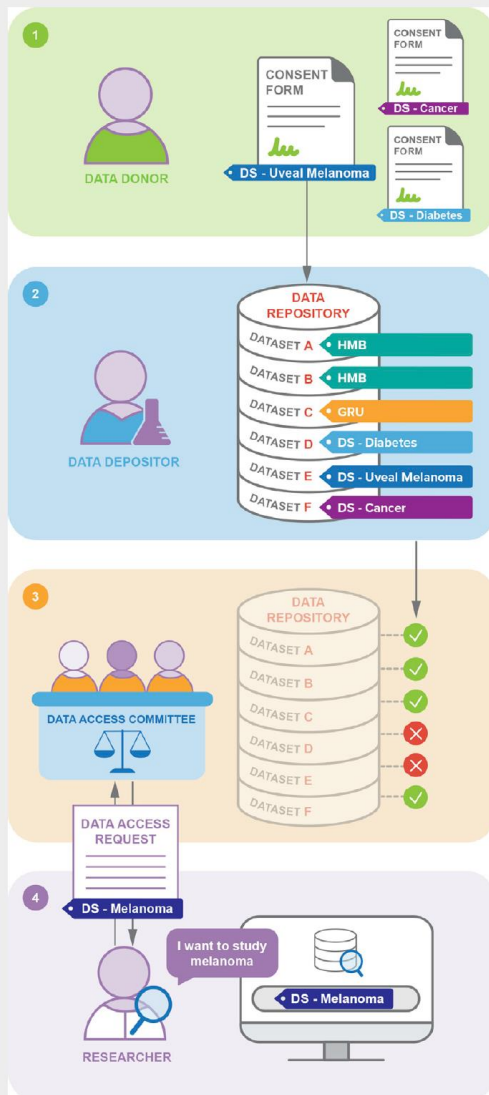
ease-specific findings, possibly impacting scientific discoveries to prevent and treat such diseases. Ultimately, after engaging with the DUO team, representatives of the RARE-X rare disease community became strong proponents of DUO and advocate for its use among other rare disease participant groups. To help clarify this to future adopters of DUO, the DURi work stream is actively developing DUO implementation guidance and is also evaluating whether it would be feasible to provide a DUO-based software service to aid groups in choosing DUO terms that fit their needs.

Currently, the implementation and use of DUO may be limited by the need to retrospectively translate consent form language into DUO terms. This limits the number of dataset annotations possible and potentially generates variability in the mapping of legacy consent form conditions to DUO terms. To prospectively mitigate this issue, we have finalized the Machine-Readable Consent Guidance<sup>29</sup> to propose a consent form already mapped into DUO terms. DUO also supports DACs and data custodians with workshops and trainings on how to translate consent forms to DUO terms.

### Conclusion

DUO has been adopted worldwide for use in annotation of over 200,000 datasets to describe data use conditions for human biomedical data (Table 1). The GA4GH DUO and Passport standards, part of a joint strategy to streamline access to data, have not yet been connected to enable a singular process. As a next step, the DURi working group of GA4GH is planning to integrate DUO terms into Passport visas, combined with advocating for policy shift in approving access to groups of datasets by data use profile rather than individualized datasets. This will allow

**Box 1. DUO at each step of the data access process**



**STEP 1: CONSENT FORM ANNOTATION**

Data donors—participants in trials and studies—agree to data use purposes described in consent forms. Consent forms are written by research teams in compliance with national, local, or institutional regulations and/or policies. To maintain stewardship and accessibility, these forms should adopt clear-language data use terms, and templates should be made publicly accessible. DUO standard data use terms can be embedded directly in the consent forms' clauses, following the GA4GH Machine-Readable Consent Guidance.<sup>29</sup> Organizations may add additional usage param-

eters beyond DUO, for example, to protect intellectual property.

**STEP 2: DATASET ANNOTATION**

Datasets hosted in controlled-access repositories are annotated with DUO terms denoting the data use terms that must be adhered to for approval for secondary data usage. The DUO terms can be added retrospectively by repository custodians for legacy datasets and/or prospectively by data depositors upon data submission.

**STEP 3: DATASET DISCOVERY**

A researcher can use DUO terms to search for datasets with relevant use conditions in a data repository. For example, they can search for all datasets consented for melanoma research. This returns only the list of datasets that would be permitted for use given this specific condition. Alternatively, the researcher can query a specific dataset for their use case, without needing to contact the DAC or other help resources. This process allows the researcher to streamline the process of identifying suitable datasets and avoid unnecessary data access request submissions.

**STEP 4: DATA ACCESS REQUEST**

A researcher requests access to relevant datasets and describes the research purpose using DUO terms. This enables efficient triaging by the DAC, either manually or using an automated matching algorithm.<sup>7</sup> The DAC reviews the access request to determine if the proposed research is consistent with the data use terms and if so, grants the researcher access to the datasets. The use of DUO terms facilitates a streamlined and standardized review by DACs.

authenticated researchers to automatically access new and existing datasets matching their DAC-approved data use profile after sign-in. Further streamlining the access process will minimize the need for multiple consecutive requests as new data are released either for a specific project or in a new repository. Such an approach also sets a precedent for establishing trust between DACs and enhanced alignment in the approval process: we envision users' data use profiles could be shared across DACs. As biomedical datasets are produced in greater numbers, across diverse settings, reliance on DUO-based mechanisms is critical to streamline data access to enable scientific collaborations.

**STAR★METHODS**

Detailed methods are provided in the online version of this paper and include the following:

- [KEY RESOURCES TABLE](#)
- [RESOURCE AVAILABILITY](#)

- Lead contact
- Materials availability
- Data and code availability
- **METHOD DETAILS**
- Evolution of DUO

### SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.cgen.2021.100028>.

### ACKNOWLEDGMENTS

The authors would like to acknowledge the GA4GH Data Use and Researcher Identities (DURI) work stream. We are also grateful for contributions from the developers of the EMBL-EBI Ontology Lookup Service (Simon Jupp, Nico Matentzoglou, and Henriette Harmse) who implemented the “preferred root” visualization to support DUO users. Thanks to Stephanie Li for her ongoing help with graphics and design for this manuscript and other DUO materials. The DUO logo was designed by Spencer Philips at EMBL-EBI. We are grateful to Angela Page, Michael Baudis, and Peter Goodhand (GA4GH), Ewan Birney (EMBL, GA4GH), Bartha Knoppers and Anne-Marie Tassé (McGill University), Ayad Aliomer (Optum), Michele Mattioni (Seven Bridges), Neil Otte (University of Buffalo), and Cooper Stansbury (University of Michigan) for their support. M.A.S. and E.J.v.E. acknowledge contributions from Jelmer Veen, Marije van der Geest, and Aneas Hodseldmans for BBMRI-NL Directory work.

G.K., M.A.F., H.P., J.D.S., and M.C. were funded by EMBL-EBI Core Funds and Wellcome Trust GA4GH award number 201535/Z/16/Z. T. Burdett was funded by EMBL-EBI Core Funds. T. Boughtwood was funded by NHMRC GNT111353, GNT200001, and the Australian MRFF. P.A. was funded by ELIXIR Luxembourg. S.D. and K.R. were funded by the Broad Institute. M.A.H. and M.B. received funding from NIH #5R24OD011883. M.L. and M.C. were funded by the CINECA project (H2020 No 825775). N.M. and L.Z. were funded by H3ABioNet, NIH grant number U24HG006941. S.O. and C.Y. received funding from the Japan Agency for Medical Research and Development (AMED) under grant numbers JP19kk020501 and JP18kk0205012. A.A.P. was funded by NHGRI AnVIL, award number U24HG010262. F.P. was supported, in part, by the European Union’s Horizon 2020 research and innovation program under the EJP RD COFUND-EJP #825575. M.A.S. and E.J.v.E. were funded by FAIR genomes (ZonMW #846003201) and EOSC-Life (H2020 #824087). S.V. was funded by the Industry Strategy Challenge Fund by the UK Government. G.S. was funded by ELIXIR, the research infrastructure for life science data.

### AUTHOR CONTRIBUTIONS

All authors contributed to investigation and writing—review and editing. J.L., M.N.C., M.C., J.D.S., and A.A.P. contributed to conceptualization. J.L., G.A.R., A.A.P., S.D., A.S., L.L.R., T.H.N., M.C., M.N.C., A.J.B., F.J., S.O.M.D., S.R.B., H.C., G.K., and T. Boughtwood contributed to validation. G.K., T. Boughtwood, H.C., S.R.B., N.M., S.O., M. Kawashima, M.M., N.Y., S.N., C.H., S.K., A.J., J.R., J.D.S., M.A.S., E.J.v.E., S.V., C.Y., and L.Z. contributed to data curation. M.C., J.L., A.T., T. Boughtwood, G.K., P.A., S.R.B., H.C., P.H., H.P., F.P., J.R., A.J., G.S., J.D.S., M.A.S., S.V., C.V., S.O., M. Kawashima, N.M., and C.Y. contributed to writing—original draft.

### DECLARATION OF INTERESTS

M.N.C. is an employee of Foundation Medicine and equity holder of Roche. A.A.P. is a venture partner at GV and an employee of Alphabet Corporation. He has received funding from MSFT, Verily, IBM, Intel, Bayer, and Novartis. The views expressed by L.L.R. are the author’s own and do not necessarily represent those of her organization.

### INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in science. One or more of the authors of this paper self-identifies as a member of the LGBTQ+ community.

Received: February 28, 2021

Revised: July 2, 2021

Accepted: August 9, 2021

Published: November 10, 2021

### REFERENCES

1. Rehm, H.L., Page, A.J.H., Smith, L., Adams, J.B., Alterovitz, G., Babb, L.J., Barkley, M.P., Baudis, M., Beauvais, M.J.S., Beck, T., et al. (2021). GA4GH: international policies and standards for data sharing across genomic research and healthcare. *Cell Genomics* 1, 100029-1-100029-33.
2. Thorogood, A., Rehm, H.L., Goodhand, P., Page, A.J.H., Joly, Y., Baudis, M., Rambla, J., Navarro, A., Nyronen, T.H., Linden, M., et al. (2021). International Federation of Genomic Medicine Databases Using GA4GH Standards. *Cell Genomics* 1, 100032-1-100032-5.
3. Woolley, J.P., Kirby, E., Leslie, J., Jeanson, F., Cabili, M.N., Rushton, G., Hazard, J.G., Ladas, V., Veal, C.D., Gibson, S.J., et al. (2018). Responsible sharing of biomedical data and biospecimens via the “Automatable Discovery and Access Matrix” (ADA-M). *npj Genomic Med.* 3, 1-6.
4. GA4GH Data Use and Researcher ID work stream. <https://ga4gh-duri.github.io/>.
5. Voisin, C., Linden, M., Dyke, S.O.M., Bowers, S.R., Reinold, K., Lawson, J., Li, S., Ota Wang, V., Barkley, M.P., Bernick, D., et al. (2021). GA4GH Passport standard for digital identity and access permissions. *Cell Genomics* 1, 100030-1-100030-12.
6. GA4GH Driver projects. <https://www.ga4gh.org/how-we-work/driver-projects/>.
7. Cabili, M.N., Lawson, J., Saltzman, A., Rushton, G., O’Rourke, P., Wilbanks, J., Rodriguez, L.L., Nyronen, T., Courtot, M., Donnelly, S., and Philippakis, A.A. (2021). Empirical validation of an automated approach to data use oversight. *Cell Genomics* 1, 100031-1-100031-6.
8. Tryka, K.A., Hao, L., Sturcke, A., Jin, Y., Wang, Z.Y., Ziyabari, L., Lee, M., Popova, N., Sharopova, N., Kimura, M., and Feolo, M. (2014). NCBI’s Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res.* 42, D975-D979.
9. Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J.D., Ur-Rehman, S., Saunders, G., Kandasamy, J., Caccamo, M., Leinonen, R., et al. (2015). The European Genome-phenome Archive of human data consented for biomedical research. *Nat. Genet.* 47, 692-695.
10. Dyke, S.O.M., Philippakis, A.A., Rambla De Argila, J., Paltoo, D.N., Luetkemeier, E.S., Knoppers, B.M., Brookes, A.J., Spalding, J.D., Thompson, M., Roos, M., et al. (2016). Consent Codes: Upholding Standard Data Use Conditions. *PLoS Genet.* 12, e1005772.
11. Lin, Y., Harris, M.R., Manion, F.J., Eisenhauer, E., Zhao, B., Shi, W., Karnovsky, A., and He, Y. (2014). Development of a BFO-Based Informed Consent Ontology (ICO). Proceedings of the 5th International Conference on Biomedical Ontologies (ICBO), Houston, Texas, USA.
12. DUO tracker. <https://github.com/EBISPOT/DUO/issues>.
13. DUO governance policy. <https://github.com/EBISPOT/DUO/blob/master/Governance2021.md>.
14. Web Ontology Language Reference. <https://www.w3.org/TR/owl-ref/>.
15. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., et al.; OBI Consortium (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25, 1251-1255.
16. Shefchek, K.A., Harris, N.L., Gargano, M., Matentzoglou, N., Unni, D., Brush, M., Keith, D., Conlin, T., Vasilevsky, N., Zhang, X.A., et al. (2020).



- The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.* 48 (D1), D704–D715.
17. Grenon, P., Smith, B., and Goldberg, L. (2004). Biodynamic ontology: applying BFO in the biomedical domain. *Stud. Health Technol. Inform.* 102, 20–38.
  18. Jupp, S., Burdett, T., Malone, J., Leroy, C., Pearce, M., McMurry, J., et al. (2015). A new Ontology Lookup Service at EMBL-EBI. *Proceedings of the 8th International Conference on Semantic Web Applications and Tools for Life Sciences*, Cambridge, UK., 118–119.
  19. DUO Github repository. <https://github.com/EBISPOT/DUO>.
  20. Overton, J.A., Cuffaro, M., and The OBO Foundry Operations Committee Technical Working Group; and Mungall, C.J. (2020). String of PURLs – frugal migration and maintenance of persistent identifiers. *Data Science.* 3, 3–13.
  21. DUO releases. <http://purl.obolibrary.org/obo/duo/releases/>.
  22. DUO October 2017 release. <http://purl.obolibrary.org/obo/duo/releases/2017-10-16/duo.owl>.
  23. Matentzoglou, N., Leo, J., Hudhra, V., Parsia, B., Sattler, U., et al. (2015). A survey of current, stand-alone OWL Reasoners. *Proceedings of the 4th International Workshop on OWL Reasoner Evaluation (ORE-2015)*. Athens, Greece, June 6, 2015.
  24. <https://douroucoulis.wordpress.com/2018/08/06/new-version-of-ontology-development-kit-now-with-docker-support/>
  25. Hancock, J.M. (2004). SPARQL (SPARQL Protocol and RDF Query Language). In *Dictionary of Bioinformatics and Computational Biology* (Wiley).
  26. Courtot, M., Gibson, F., Lister, A., Malone, J., Schober, D., Brinkman, R.R., Ruttenberg, A., et al. (2011). MIREOT: The minimum information to reference an external ontology term. *Applied Ontology* 6, 23–33.
  27. Kazakov, Y., Krötzsch, M., and Simančík, F. (2014). The Incredible ELK. *J. Autom. Reason.* 53, 1–61.
  28. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 160018.
  29. GA4GH Machine-readable consent guidance (2020). [https://www.ga4gh.org/work\\_stream/regulatory-ethics/](https://www.ga4gh.org/work_stream/regulatory-ethics/).
  30. Powell, K. (2021). The broken promise that undermines human genome research. *Nature* 590, 198–201.
  31. NIH Releases New Policy for Data Management and Sharing. <https://osp.od.nih.gov/2020/10/29/nih-releases-new-policy-data-management-and-sharing/>.
  32. Vasilevsky, N.A., Foster, E.D., Engelstad, M.E., Carmody, L., Might, M., Chambers, C., Dawkins, H.J.S., Lewis, J., Della Rocca, M.G., Snyder, M., et al. (2018). Plain-language medical vocabulary for precision diagnosis. *Nat. Genet.* 50, 474–476.
  33. GA4GH releases three new deliverables to support responsible genomic data sharing. <https://www.ga4gh.org/news/ga4gh-releases-three-new-deliverables-to-support-responsible-genomic-data-sharing/>.
  34. Finnish institute for health and welfare. <https://thl.fi/en/web/thl-biobank>.
  35. Bender, D., and Sartipi, K. (2013). HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, 326–331.
  36. Consent - FHIR v4.0.1 <https://www.hl7.org/fhir/consent.html#:~:text=A%20Privacy%20Consent%20Directive%20as,purposes%20and%20periods%20of%20time>
  37. OASIS LegalRuleML TC [https://www.oasis-open.org/committees/tc\\_home.php?wg\\_abbrev=legalruleml](https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=legalruleml)
  38. Thorogood, A. (2020). Policy-aware data lakes: a flexible approach to achieve legal interoperability for global research collaborations. *J. Law Biosci.* 7, a065.

## STAR★METHODS

### KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
ELK reasoner	Kazakov et al., 2014 <sup>27</sup>	<a href="https://www.korrekt.org/page/The_Incredible_ELK">https://www.korrekt.org/page/The_Incredible_ELK</a>
Ontology Lookup Service	Jupp et al., 2015 <sup>18</sup>	<a href="https://www.ebi.ac.uk/ols/index">https://www.ebi.ac.uk/ols/index</a>
Ontology Development Kit	<a href="https://douroucouli.wordpress.com/2018/08/06/new-version-of-ontology-development-kit-now-with-docker-support/">https://douroucouli.wordpress.com/2018/08/06/new-version-of-ontology-development-kit-now-with-docker-support/</a>	<a href="https://doi.org/10.5281/zenodo.4662066">https://doi.org/10.5281/zenodo.4662066</a>
DUO GitHub repository	This manuscript	<a href="http://purl.obolibrary.org/obo/duo">http://purl.obolibrary.org/obo/duo</a>
Released DUO file	This manuscript	<a href="http://purl.obolibrary.org/obo/duo.owl">http://purl.obolibrary.org/obo/duo.owl</a>

### RESOURCE AVAILABILITY

#### Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Mélanie Courtot ([mcourtot@gmail.com](mailto:mcourtot@gmail.com)).

#### Materials availability

This study did not generate new unique reagents.

#### Data and code availability

The Data Use Ontology source files, scripts and documentation are licensed under CC-BY 4.0 and available from the GitHub repository <http://purl.obolibrary.org/obo/duo>. This manuscript describes the 2021-02-23 release of DUO, permanently publicly available at <http://purl.obolibrary.org/obo/duo/releases/2021-02-23/duo.owl>

### METHOD DETAILS

The GA4GH community has been previously involved in the development of two main controlled vocabularies/"information models" that systematically capture data use restrictions on human genomics and health datasets: (1) Consent Codes<sup>10</sup> and (2) ADA-M.<sup>3</sup> Further details on the process by which these vocabularies were created is described elsewhere.<sup>3,10</sup> Preceding these efforts, guidance from the NIH's database of Genotype and Phenotype (dbGaP)<sup>8</sup> led to the organic creation of a data use restriction vocabulary by requesting data depositors to represent the conditions for secondary use of the deposited datasets using the dbGaP vocabulary. This dbGaP vocabulary included a set of a handful of nucleating terms that are often used (such as: "General Research Use" (GRU), "Health/Medical/Biomedical research only" (HMB)) and also allowed depositors to add new terms to the vocabulary if a suitable term didn't previously exist.

The goal in creating DUO was to create a human and Machine-Readable representation of these 3 vocabularies and to code and maintain it in a form of a versioned ontology that will allow automated computation of software systems (e.g., as needed by a search function) on the ontology terms. An ontology encodes the hierarchy between terms which is critical for machine based automated computation. Before attempting to create DUO we defined 5 main goals:

1. Generate an ontology that is easy to use for the end user and unambiguous.
2. Generate a lean ontology based on real life use cases; and evolve gradually.
3. Ontology categories could be used to represent Data Use Conditions and Research Purposes. Thus, definitions should be generalized accordingly.
4. Include categories to support piloting ADA-M and Consent Codes as a human interface to define data use restrictions and research purposes.
5. Ideally, support a matching algorithm that uses boolean logic.

To create DUO we conducted the following steps:

1. Consent code and ADA-M integration proposal: In early 2018, we reviewed the Consent Codes,<sup>10</sup> the NIH dbGaP data depositor guide and the ADA-M information model<sup>3</sup> and created a proposal of a set of data use restriction terms and their hierarchy as the basis for the DUO ontology.





2. DUO refinement: In the GA4GH 2018 spring in person meeting in Toronto we initiated a review process in which key potential DUO users trimmed down the set of terms to be included in the initial version of DUO and confirmed their hierarchy to ensure that common software-based use cases can be coded using DUO. These users included representatives from GA4GH driver projects (e.g, The All of Us research program, Australian Genomics, ANVIL), and representatives of data repositories that were seeking a Machine-Readable data use ontology (e.g, dbGaP, EGA, Sanger, The Broad Institute). These processes continued during the GA4GH bi-weekly DURl team video-conference meetings, where the team systematically discussed and approved terms, their definition and hierarchy in the ontology. Whenever a controversy arose the team relied on the guiding principles of creating (a) a lean ontology that (b) supports a real-life use case. In the absence of an immediate real life use case our team refrained from adding terms in favor of creating a lean ontology to begin with.
3. Ontology representation of DUO: Once a stable first version of DUO was agreed on, the ontology was implemented in the Web Ontology Language (OWL),<sup>14</sup> a World Wide Web Consortium standard. Development of DUO follows Open Biomedical Ontologies (OBO) development principles,<sup>15</sup> ensuring interoperability with other ontological resources, such as those describing disease entities.<sup>16</sup> As per OBO guidelines, DUO is built under the Basic Formal Ontology (BFO)<sup>17</sup> upper-level ontology. The DUO root terms “data use permission” and “data use modifier” are subclasses of “data item” (IAO:0000027), itself a type of “information artifact entity” (IAO:0000030) and “generically dependent continuant” (BFO:0000031). DUO terms are stable, with each DUO term having its unique Uniform Resource Identifier, which can be browsed using the OLS. Most importantly, the meaning associated with a specific DUO ID is permanent; this guarantees consistency through time of the data use terms. Different versions of DUO are available through the GitHub repository,<sup>19</sup> including an editors’ version which captures ongoing development, and stable, released versions. Released versions of DUO are associated with permanent URLs (PURLs) for sustainability:<sup>20</sup> the most recent release is always available from <http://purl.obolibrary.org/obo/duo.owl>, while previous versions can be accessed through their date-based PURL, providing choice for users who prefer to use a specific historical view of the ontology<sup>21,22</sup> for stability while transitioning to the latest version.
4. Pilot adoption: once the OWL version of DUO was available, the use of the ontology in live software systems was piloted. This included a pilot by the EGA and Sanger as well as a pilot by the Broad Institutes DUOS data repository<sup>7</sup> were working software systems in both data repositories were referencing the DUO OWL libraries to tag datasets in their system and underlie their search features. DUOS is used in the All-of-Us and ANVIL GA4GH driver projects.
5. GA4GH product approval: Once the use of DUO was demonstrated via GA4GH driver projects pilots, DUO was unanimously approved as a GA4GH standard, following the GA4GH official product review and approval process, by the GA4GH steering committee in Jan 2019.<sup>1</sup>

### Evolution of DUO

Contributions to DUO are public and created by raising GitHub issues,<sup>12</sup> anyone may submit a request to add a new term, or comment on an existing request. Requests are discussed by the DUO work stream and driver project implementers on the tracker, DUO mailing-list and during periodic teleconferences. Once approved, changes are open to the public for further discussion throughout a comment period of two weeks, as per the DUO governance policy.<sup>13</sup> External efforts such as the Informed Consent Ontology (ICO)<sup>11</sup> were additionally reviewed for interoperability and synergistic evolution; DUO has been directly imported in ICO to describe data use conditions instead of duplicating its content.



## **Annex Paper 5**

Saunders, G., Baudis, M., Becker, R., Beltran, S., Bérout, C., Birney, E., Brooksbank, C., Brunak, S., van den Bulcke, M., Drysdale, R., Capella-Gutierrez, S., Flicek, P., Florindi, F., Goodhand, P., Gut, I., Heringa, J., Holub, P., Hooyberghs, J., Juty, N., . . . Scollen, S. (2019). [Leveraging European infrastructures to access 1 million human genomes by 2022.](#)

*Nature Reviews Genetics*, 20(11), 693–701.



## Leveraging European infrastructures to access 1 million human genomes by 2022

Gary Saunders<sup>1</sup>, Michael Baudis<sup>2</sup>, Regina Becker<sup>3</sup>, Sergi Beltran<sup>4,5</sup>, Christophe Bérout<sup>6,7</sup>, Ewan Birney<sup>8</sup>, Cath Brooksbank<sup>8</sup>, Soren Brunak<sup>9,10</sup>, Marc Van den Bulcke<sup>11</sup>, Rachel Drysdale<sup>1</sup>, Salvador Capella-Gutierrez<sup>12</sup>, Paul Flicek<sup>8</sup>, Francesco Florindi<sup>13</sup>, Peter Goodhand<sup>14,15</sup>, Ivo Gut<sup>4,5</sup>, Jaap Heringa<sup>16</sup>, Petr Holub<sup>13</sup>, Jef Hooyberghs<sup>17</sup>, Nick Juty<sup>18</sup>, Thomas M. Keane<sup>8</sup>, Jan O. Korbel<sup>19</sup>, Ilkka Lappalainen<sup>20</sup>, Brane Leskosek<sup>21</sup>, Gert Matthijs<sup>22</sup>, Michaela Th. Mayrhofer<sup>13</sup>, Andres Metspalu<sup>23</sup>, Arcadi Navarro<sup>24,25,26</sup>, Steven Newhouse<sup>8</sup>, Tommi Nyrönen<sup>20</sup>, Angela Page<sup>15,27</sup>, Bengt Persson<sup>28</sup>, Aarno Palotie<sup>29</sup>, Helen Parkinson<sup>8</sup>, Jordi Rambla<sup>26</sup>, David Salgado<sup>6</sup>, Erik Steinfeldt<sup>13</sup>, Morris A. Swertz<sup>30</sup>, Alfonso Valencia<sup>12,31</sup>, Susheel Varma<sup>8</sup>, Niklas Blomberg<sup>1</sup> and Serena Scollen<sup>1</sup> \*

**Abstract** | Human genomics is undergoing a step change from being a predominantly research-driven activity to one driven through health care as many countries in Europe now have nascent precision medicine programmes. To maximize the value of the genomic data generated, these data will need to be shared between institutions and across countries. In recognition of this challenge, 21 European countries recently signed a declaration to transnationally share data on at least 1 million human genomes by 2022. In this Roadmap, we identify the challenges of data sharing across borders and demonstrate that European research infrastructures are well-positioned to support the rapid implementation of widespread genomic data access.

### Precision medicine

An approach for disease treatment and prevention that takes into account individual variability in genes, environment and lifestyle for each person.

Genomics has the potential to benefit overall health by ensuring that patients receive timely and effective diagnosis, information and treatment. For example, international collaborations that integrate genomic, phenotypic and clinical data have achieved new paradigms in the diagnosis and care of patients with rare diseases<sup>1</sup> (BOX 1). However, realizing the potential of precision medicine beyond rare diseases will require systematic access and integration of research and health-care data at a greater scale, for example, across countries<sup>2–4</sup>.

Across Europe, several national initiatives are being established to generate genomic data, most of which are disease agnostic, although some initiatives focus on cancer, infectious diseases and/or rare diseases (FIG. 1). Recently, representatives of 21 member states of the European Union (EU) signed a joint declaration to deliver cross-border access to human genomes by the end of 2022 (REF.<sup>5</sup>) (TABLE 1). Whole-genome sequencing data at this scale have the potential to transform our understanding of disease, leading to improved diagnostics and the development of effective prevention programmes and precision medicine treatments. However, handling data on a large, transnational scale does not come without challenges.

Researchers and clinicians will need remote access to sensitive human data across national boundaries to assemble and manage very large cohorts or identify individuals with rare phenotypes, with the governance and security necessary to interface with health-care systems. Currently, each European country sets its own regulatory framework for the processing of health and genetic data and to enable access to these data for research. Moreover, genetic and associated data generated through health care are not shared as widely as research data; given that health care is a national competence and subject to national laws, it is often problematic for health data from one country to be exported outside regional or national jurisdictions.

Transformation of the European life sciences and health data landscape will be possible only by aligning national and international initiatives, by connecting developments across projects and countries into a long-term, standards-based infrastructure operating at continental scale. It will also be essential to provide a procedural framework that will guarantee research participants' and patients' rights while allowing controlled access to data across borders. Despite the many challenges, enabling access to genomic data at this scale is possible by building on established European research infrastructures.

\*e-mail: serena.scollen@elixir-europe.org  
<https://doi.org/10.1038/s41576-019-0156-9>

**Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-ERIC).** A research infrastructure that brings together key stakeholders from the biobanking field to support biomedical research and facilitate the development of new therapies by offering management services, support with ethical, legal and societal issues, and a number of online tools and software solutions.

By implementing a Europe-wide framework of experts and long-term services, the [European Strategy Forum On Research Infrastructures \(ESFRIs\)](#), which includes the [Biobanking and Biomolecular Resources Research Infrastructure \(BBMRI-ERIC\)](#) and [ELIXIR \(ELIXIR Europe\)](#), aims to drive the coordination of efforts at both the national and international level. In this Roadmap, we present opportunities that will enable secure and compliant transnational access to controlled-access human genomic data that has been consented for secondary use. We consider key issues according to their priority, including data-sharing models, data discovery, data standards, computing, regulatory frameworks and training needs. By leveraging existing services to achieve this ambitious aim, Europe can be positioned as a global leader in this field.

#### Data access and management

Access and management of genomic data are now more of a challenge than the generation of the data themselves. To enable effective, cross-border access to data, a coordinated, secure, federated environment that enables population-scale genomic, phenotypic and

biomolecular data to be accessible across international borders will be required. Many national and European life-science research programmes as well as public-private partnerships, such as the [Innovative Medicines Initiative](#), have made and continue to make considerable investments in data and knowledge management infrastructure. However, efforts are mostly independent, resulting in fragmented and overlapping investments in data management.

One possible solution to facilitate access and manage human data across borders is to develop federated systems for data sharing (FIG. 2). Data are geographically dispersed but discoverable and/or accessible in such a way that data queries can be responded to as if they were deposited in a single database. For example, [Matchmaker Exchange](#)<sup>6</sup> is a federated data-sharing platform that successfully facilitates the matching of patients with rare diseases with similar phenotypic and genotypic profiles. The willingness of patients with a rare disease to share data has driven earlier implementation compared to models that are being established for data sharing and/or access beyond rare diseases. Nevertheless, two platforms in mature stages of development are moving towards use for case-driven implementation, the [European Genome-phenome Archive \(EGA\)](#); also known as the [European Nucleotide Archive or European Variation Archive](#)<sup>7</sup> and the [Personal Health Train \(PHT\)](#).

**European Genome-phenome Archive.** The EGA is a resource for the permanent archiving and sharing of controlled-access genetic and phenotypic human data that result from biomedical research projects. The central EGA, which is operated from the [European Bioinformatics Institute, UK](#), and the [Centre for Genomic Regulation, Spain](#), hosts over 1,700 studies that comprise more than 4,000 data sets from more than 900 data providers and has served data to over 10,000 requestors since 2008 (REF.<sup>7</sup>). The EGA is one of several ELIXIR core data resources and the recommended database for deposition of controlled-access human data<sup>8</sup>.

The EGA is now being extended to a federated model, which will enable local implementations at research institutes in different national ELIXIR Nodes. The overall goal is to provide secure, standardized, documented and interoperable services under the framework of the EGA. The fundamental principle of the EGA federated framework is that data sets remain within appropriate jurisdictional boundaries whereas metadata (that is, data set descriptions) are centralized and searchable through a common application programming interface (API). After data discovery, access to the data themselves can be requested from the source, for example, by applying to a data access committee, to establish agreements for data use. The EGA participates in the large-scale, funded projects [euCanSHare](#) and [EUCANCan](#), two European-Canadian cooperative projects aimed at facilitating genomic data analysis, sharing and management in cardiovascular and cancer research, respectively, as well as in the transcontinental [Common Infrastructure for National Cohorts in Europe, Canada and Africa \(CINECA\)](#) project. CINECA will encompass 18 organizations representing European, Canadian and African

#### Author addresses

<sup>1</sup>ELIXIR Hub, Wellcome Genome Campus, Hinxton, Cambridge, UK.

<sup>2</sup>University of Zurich, Zurich, Switzerland.

<sup>3</sup>Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Luxembourg, Luxembourg.

<sup>4</sup>CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain.

<sup>5</sup>Universitat Pompeu Fabra (UPF), Barcelona, Spain.

<sup>6</sup>Aix Marseille Univ, INSERM, MMG, Marseille, France.

<sup>7</sup>Département de Génétique Médicale et de Biologie Cellulaire, APHM, Hôpital d'Enfants de la Timone, Marseille, France.

<sup>8</sup>European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK.

<sup>9</sup>Department of Health Technology, Technical University of Denmark, Lyngby, Denmark.

<sup>10</sup>Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark.

<sup>11</sup>Cancer Centre, Epidemiology and Public Health, Sciensano, Ixelles, Belgium.

<sup>12</sup>Barcelona Supercomputing Centre (BSC), Barcelona, Spain.

<sup>13</sup>BBMRI-ERIC, Graz, Austria.

<sup>14</sup>Ontario Institute for Cancer Research, Toronto, Ontario, Canada.

<sup>15</sup>Global Alliance for Genomics and Health, Toronto, Ontario, Canada.

<sup>16</sup>Department of Computer Science, Vrije Universiteit, Amsterdam, Netherlands.

<sup>17</sup>Flemish Institute for Technological Research, VITO, Mol, Belgium.

<sup>18</sup>School of Computer Science, The University of Manchester, Manchester, UK.

<sup>19</sup>European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany.

<sup>20</sup>CSC — IT Center for Science, Espoo, Finland.

<sup>21</sup>IBMI, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia.

<sup>22</sup>Katholieke Universiteit Leuven, Leuven, Belgium.

<sup>23</sup>Estonian Genome Center, University of Tartu, Tartu, Estonia.

<sup>24</sup>Institute of Evolutionary Biology (UPF-CSIC), Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain.

<sup>25</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.

<sup>26</sup>Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain.

<sup>27</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA.

<sup>28</sup>Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala, Sweden.

<sup>29</sup>Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland.

<sup>30</sup>BBMRI-NL/University Medical Center Groningen, University of Groningen, Groningen, Netherlands

<sup>31</sup>ICREA, Pg., Barcelona, Spain.

**Box 1 | A coordinated infrastructure for the rare diseases research community**

Rare diseases are individually uncommon but are estimated to affect around 7% of the population or approximately 30 million people across Europe<sup>1</sup>. Over 80% of rare diseases are of genetic origin and, in general, only very few individuals in a single country are affected. Owing to the heterogeneity and low prevalence of each disease, it is difficult to gain access to a substantial number of cases with the same disease, which poses numerous technical and scientific challenges for research. Furthermore, as the commercial incentives to explore the underlying mechanism of these diseases are insufficient, very few drugs currently exist to treat rare diseases.

Coordinated access to genomic and phenotypic information across Europe is transforming rare disease research. The **ELIXIR Rare Diseases Community** promotes and funds activities between ELIXIR platforms and relevant rare disease research infrastructures and initiatives. This community provides a strong example of how a coordinated infrastructure can provide direct, tangible benefits to health-care systems and patients. For example, the RD-Connect platform<sup>1</sup> includes a biobank and registry finder, a sample catalogue (integrated with the Biobanking and Biomolecular Resources Research Infrastructure) and the genome–phenome analysis platform (GPAP). Genomic data available in GPAP are processed through a validated standard pipeline, and the raw data are deposited in the European Genome–phenome Archive<sup>7</sup> for long-term storage. GPAP is part of the International Rare Diseases Research Consortium, Global Alliance for Genomics and Health (GA4GH) Matchmaker Exchange, the GA4GH Beacon network and the GA4GH ‘Discovery’ work stream. GPAP is a scalable and interoperable system that enables genome discovery, access and analysis that could be easily deployed at national nodes to provide access to 1 million human genomes. In this sense, other local systems based on RD-Connect have already been deployed using containers, enabling full control of data discovery and access and allowing data to be kept within national boundaries (for example, *Proyecto Genoma 1000 Navarra*). GPAP is working towards providing tiered discoverability and data access between local instances based on user permissions.

**ELIXIR**

An intergovernmental organization that coordinates life science resources from across Europe, including databases, software tools, training materials, cloud storage and supercomputers, to form a single infrastructure that facilitates data sharing, exchange of expertise and best practice development. Ultimately, ELIXIR’s goal is to help researchers gain new insights into how living organisms work.

**Federated**

A term used to describe an architecture that allows information sharing between information technology systems and applications.

**ELIXIR Nodes**

One or more research institutes within a member country that run the resources and services that are part of ELIXIR; there are currently 23 ELIXIR Nodes.

**Application programming interface**

(API). An access point that enables applications to communicate with one another, for example, allowing an application to access a particular database.

cohorts to develop and apply the necessary international infrastructure to responsibly share and analyse data based on existing cohorts’ data, operating within existing consent and EU General Data Protection Regulation (GDPR) 2016/679 regulations.

**Personal Health Train.** Another possible solution being developed by consortia in the Netherlands and Germany is the PHT, which is a concept for the (re)use of personal data in health care, disease prevention and research. The key concept of the PHT is to share data in a federated manner — to bring algorithms to the data where they happen to be, rather than transmitting data to a central place. This approach is achievable using a suite of standardized computational interfaces and executable computational containers. The train metaphor explains the infrastructure: ‘stations’ with health-related data are connected by secure and monitored ‘tracks’ along which care professionals, researchers or citizens can run ‘trains’ that carry questions and return answers. Bringing questions to data rather than moving data is a key differentiator of the PHT, addressing scalability issues with data transmission and mitigating legal, ethical, societal and technical barriers associated with enabling (cross-border) physical data access.

**Data discoverability for reuse**

An essential element to unlock access for authorized researchers to 1 million human genomes across the EU is the awareness of the existence and location of these data. This requires the provision of metadata that characterizes the samples and genomes, such as their association with certain diseases, as well as their registration in a

searchable database that allows data to be found by both humans and computers. As demonstrated by the EGA, metadata can be shared and made searchable through a common interface even when data is hosted locally.

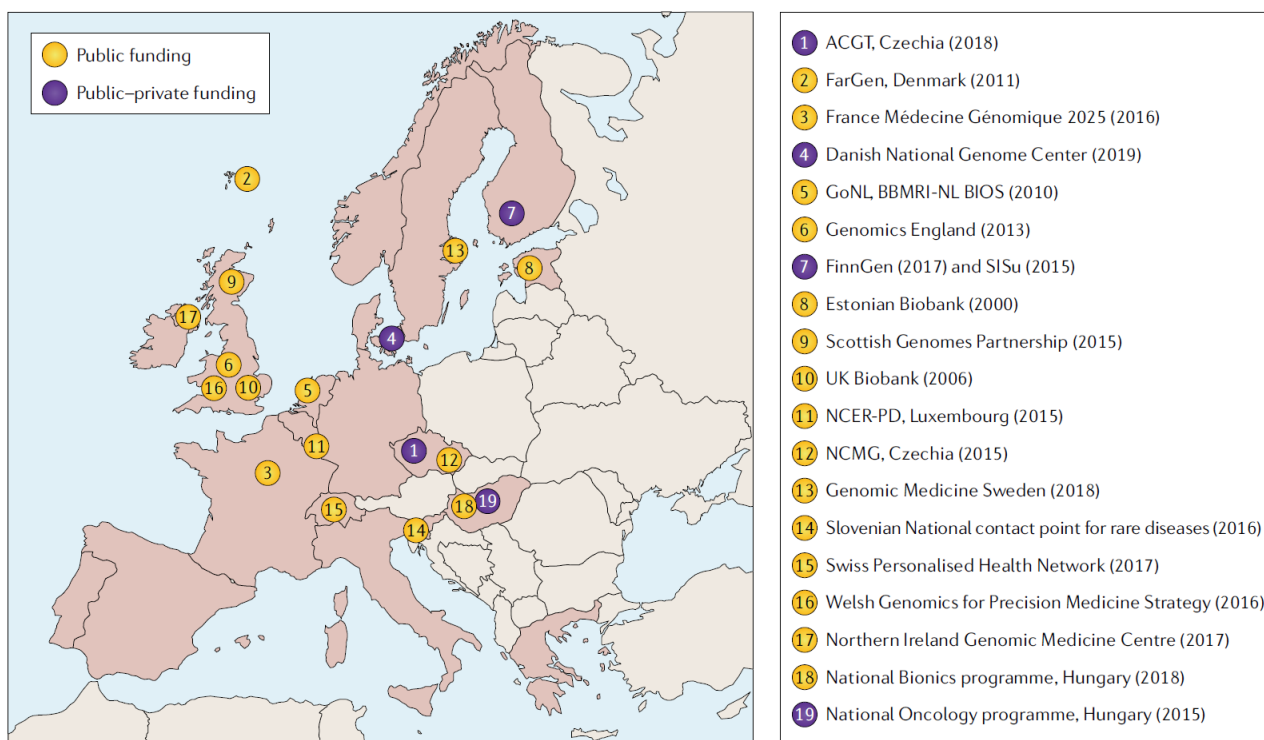
The discovery of genomic data can be enhanced further through the implementation of ‘beacons’, a federated data discovery protocol that allows users to find specified genetic variants across multiple data sets<sup>9</sup>. To maintain participant anonymization, only the presence or absence of the specified variant in data collections is reported. This information enables the researcher to contact the persons responsible for the respective data set, learn more about the data and to formally request access where these data are of interest. Beacon is an approved international standard of the policy-framing and standards-setting organization for genomics, GA4GH. Currently, nine ELIXIR member countries have launched national beacons.

A large part of the data and samples needed to sequence 1 million genomes is already stored in biobanks, and is searchable, for example, via the Directory of the BBMRI-ERIC, the European research infrastructure for biobanking<sup>10</sup>. BBMRI-ERIC facilitates access to high-quality samples and data by connecting more than 500 biobanks and sample collections across 21 EU countries. The **BBMRI-ERIC Directory** is a tool to share aggregated information about biobanks that are willing to collaborate and provide access to others. It forms the largest catalogue of biobanks in the world, with more than 100 million samples readily available for researchers<sup>11</sup>. The biobank information standard group, **Minimum Information About Biobank data Sharing (MIABIS) 2.0** (REF.<sup>12</sup>) and BBMRI-ERIC Interoperability Forum groups are working on developing a common API and common data exchange models for distributed search, whereby donor-level and sample-level information is kept stored in local biobanks but information on the availability of donors and samples matching search criteria is proffered. The **ELIXIR Scientific Programme (2019–2023)** will see the generation of the necessary interfaces and data models to allow biobanks to become interoperable with the beacon discovery protocol for the genetic data component. As described above, this protocol helps local biobanks to make their samples more findable but does not centralize collection and storage, which are maintained at the local or national level.

**Genomics data standards and reference data**

High-content phenotypic data are often heterogeneous and recorded using varied standards and ontologies. Communities working with these data need coordinated expert advice on which standards to adopt in order to enable federated data access. To facilitate reuse, data producers must have compatible (interoperable) interfaces and provide computational services that allow data integration. Going forward, the vast majority of human multi-omics data are expected to come from health care rather than research. Harmonized data governance architectures allow for broad spheres of responsible data access, enabling researchers to perform analysis on virtual cohorts of populations or the use of virtual analytical tools, without data movement.

# ROADMAP



**Fig. 1 | Examples of current health care-focused and genomics-based national initiative projects across ELIXIR members.** In many European countries (for example, Spain and Italy) health care is administered regionally and, until now, genomics-based projects have been linked to the regional health-care authorities. For brevity, these regional projects are not included. ACGT, Analysis of Czech Genome for Theranostics; BBMRI-NL, Biobanking and Biomolecular Resources Research Infrastructure – The Netherlands; BIOS, Biobank-based integrative omics study; FarGen, Faroe Genome Project; GoNL, Genome of the Netherlands; NCER-PD, National Centre of Excellence in Research on Parkinson’s disease; NCMG, National Center for Medical Genomics; SISu, Sequencing Initiative Suomi.

## General Data Protection Regulation (GDPR) 2016/679

A regulation in European Union (EU) law on data protection and privacy for all individuals within the EU and the European Economic Area. It also addresses the export of personal data outside the EU and European Economic Area.

## Containers

A system for building highly portable packages of bioinformatics software, containerization and virtualization technologies for isolating reusable execution environments for these packages and an integrated workflow system that automatically orchestrates the composition of these packages for entire pipelines.

## Biobanks

Biorepositories that store biological samples (usually human) for use in research.

## Collaboration with the Global Alliance for Genomics and Health.

GA4GH has a 5-year plan to provide standards upon which federated data sites (including those managed by research, health-care and commercial organizations as well as those run by individuals) use, analyse and store the data needed to drive precision medicine. To meet the aims of the EU declaration it will be necessary to establish coordinated European collaboration with GA4GH, for example, by building on existing collaborations between ELIXIR and GA4GH, the long-term goals of which are aligned.

Currently, ELIXIR contributes resources to the development and implementation of GA4GH standards via implementation studies and infrastructure projects that fund **GA4GH driver projects** — real-world genomic data initiatives that have signed on to help scope, develop and pilot GA4GH standards. For example, ELIXIR Beacon is a GA4GH driver project that actively contributes to four of the eight **GA4GH work streams**, including ‘Clinical and phenotypic data capture’, ‘Data use and researcher identities’, ‘Discovery’ and ‘Genomic knowledge standards’. Each GA4GH work stream is designed for the purpose of developing standards that overcome technical and regulatory hurdles to international genomic data sharing, and ELIXIR delegates co-lead four of these work streams (‘Discovery’, ‘Data

use and researcher identities’, ‘Genomic knowledge standards’ and ‘Large-scale genomics’).

As another example, the ELIXIR-linked GA4GH driver project EGA actively contributed to the **GA4GH ‘Data use and researcher identities’ work stream** by supporting the development, and now deployment, of the Data Use Ontology, an approved standard that provides a computable representation of data use requirements. This collaboration is a natural fit, as the encoding of data consent in machine-readable format is essential to the EGA’s goal of providing an archive for sensitive human data that has been consented for research, and to enable access to these sensitive data in a timely manner for approved researchers.

An extension to the collaboration between ELIXIR and GA4GH was announced in February 2019, which will take the form of a strategic partnership with specific efforts in cloud computing and identity and access management, building on the ELIXIR Authentication and Authorization Infrastructure (AAI). ELIXIR AAI allows service providers to control and manage the access rights of their users, while enabling researchers to sign in to access data and services. The vision for the extended collaboration between ELIXIR and GA4GH is to increase visibility of ELIXIR’s GA4GH-related work beyond that which



any single driver project or even a suite of individual ELIXIR-managed driver projects could provide alone. Thus, the intention is to coordinate and position ELIXIR to provide a gateway for GA4GH into Europe.

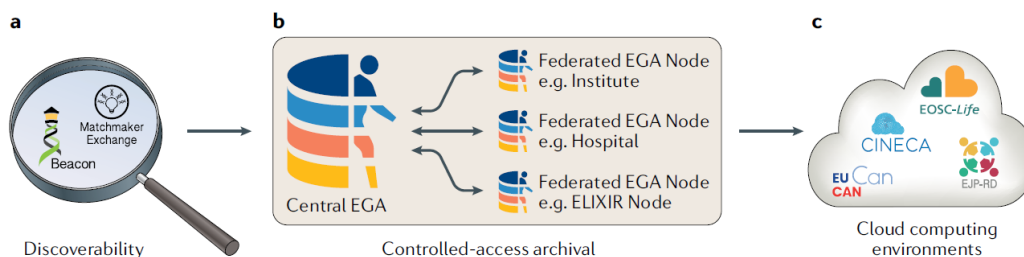
**Collaboration with the International Organization for Standardization.** BBMRI-ERIC provides quality management services to all its biobanks and contributes to the development of European and international standards. To ensure defined and computer-actionable information on the quality of the biological material and associated data, BBMRI-ERIC leads work within the International

Organization for Standardization Technical Committee 276, which holds responsibility for standardization in the field of biotechnology processes, on an interoperable provenance information model. The aim is to have a complete chain of provenance information from sample acquisition to data generation and processing, thereby allowing assessment of fitness of the data, including genetic and phenotype data for particular analyses. All BBMRI-ERIC biobanks abide by a 'partner charter' and 'access policy' that set a high bar for how these biobanks operate and collect and store samples. To make sure that samples and associated data are used effectively,

Table 1 | EU declaration signatory and membership status

Country	Declaration signatory <sup>a</sup>	BBMRI-ERIC status	ELIXIR status	EMBL status
Austria	Yes	Full Member	No	Full Member
Belgium	No	Full Member	Member	Full Member
Bulgaria	Yes	Full Member	No	No
Croatia	Yes	No	No	Full Member
Cyprus	Yes	Observer	Observer	No
Czech Republic	Yes	Full Member	Member	Full Member
Denmark	No	No	Member	Full Member
Estonia	Yes	Full Member	Member	Prospect Member
Finland	Yes	Full Member	Member	Full Member
France	No	Full Member	Member	Full Member
Germany	No	Full Member	Member	Full Member
Greece	Yes	Full Member	Member	Full Member
Hungary	Yes	No	Member	Full Member
Iceland	No	No	No	Full Member
Ireland	No	No	Member	Full Member
Israel	–	No	Member	Full Member
Italy	Yes	Full Member	Member	Full Member
Latvia	Yes	Full Member	No	No
Lithuania	Yes	No	No	Full Member
Luxembourg	Yes	No	Member	Full Member
Malta	Yes	Full Member	No	Full Member
Montenegro	–	No	No	Full Member
Netherlands	Yes	Full Member	Member	Full Member
Norway	Yes	Full Member	Member	Full Member
Poland	No	Full Member	No	Full Member
Portugal	Yes	No	Member	Full Member
Slovakia	No	No	No	Full Member
Slovenia	Yes	No	Member	No
Spain	Yes	No	Member	Full Member
Sweden	Yes	Full Member	Member	Full Member
Switzerland	No	Observer	Member	Full Member
Turkey	–	Observer	No	No
United Kingdom	Yes	Full Member	Member	Full Member

BBMRI-ERIC, Biobanking and Biomolecular Resources Research Infrastructure; EMBL, European Molecular Biology Laboratory; EU, European Union. A hyphen stands for 'not applicable'. <sup>a</sup>The initiative is also open to countries of the European Economic Area and the European Free Trade Association.



**Fig. 2 | The concept of EGA federation — from data discoverability to raw sensitive human data access.**

**a | Discoverability.** Metadata is shared from each of the sensitive data archives to a centralized database upon which query interfaces can be built; these can be project-specific portals or interfaces to query the metadata associated with all data sets across a federated network. The Global Alliance for Genomics and Health (GA4GH) driver projects ELIXIR Beacon and MatchMaker Exchange, for example, provide standards and interfaces to query such metadata in order to aid discoverability. **b | Controlled-access archival.** The GA4GH driver project European Genome–phenome Archive (EGA) provides interoperable programmatic interfaces that are required to enable metadata transfer and user authentication and authorization (provided by ELIXIR Authentication and Authorization Infrastructure, for example) across the federated network of controlled-access archives. **c | Cloud computing environments** that are, for example, community-curated workflows (such as those found in containers) able to be executed remotely and run locally at one or more sensitive data archives by implementing the standards from the GA4GH ‘Cloud’ and ‘Large-scale genomics’ work streams.

specifications for sample quality and data selection from designated samples should be defined. In doing so, it will be possible to avoid pitfalls and inefficiencies that arise when comparing data of different quality.

#### Computing resources to access genomics data

Many challenges remain to fully realize the potential of cloud computing services across Europe so that they can be used in seamless transnational workflows. Restrictions on the export of human genomic data derived from health care mean that we need to develop cloud computing models where researchers can bring their analysis to the data. Resource allocation and cost models must be developed to allow transnational access and collaborative projects, cloud interoperability standards need further development, and widespread adoption of cloud computing with harmonization of task and workflow execution systems is required. Furthermore, the GDPR allows individual EU member states to define their own safeguards to process health and genetic data (Article 9.4 GDPR)<sup>13</sup>. Therefore, security standards and user access protocols that encompass the diversity between individual countries must be established, with the necessary mutual recognition processes.

Ultimately, the vision is that national life-science clouds are compatible with life-science services and operate in a securely accessible cloud ecosystem that spans local private clouds, national community clouds, European research and innovation oriented clouds (for example, [European Open Science Cloud \(EOSC\)](#)), as well as commercial clouds (for example, Google Cloud, Microsoft Azure or Amazon Web Service), while simultaneously meeting full individual and national level identity and access requirements. Therefore, data could be organized as a federation, where data processors can access data sets, computational tools to process them and scalable computer resources, with a linked electronic identity provided by technologies such as ELIXIR AAI<sup>14</sup> or [BBMRI-ERIC AAI](#). Building on identity, security is a design principle for the integration of

infrastructure services, and this principle must encompass the whole integrated technical and software service process. Committing to an integrated security principle will help to build and maintain trust in the infrastructure for genomic data management. This also includes synchronizing terms of use and ensuring legal compliance, which will help prevent misuse of data, in turn increasing trust in the overall ecosystem.

Within the EOSC, the biomedical science research infrastructures aim to connect existing national cloud infrastructures associated with biomedical science research infrastructure nodes; adopt interoperable AAI services such as the ELIXIR AAI service; provide secure data transfers between biomedical science research infrastructures to facilitate sensitive data processing such as the reference data set distribution service ([GA4GH data repository service schemas](#)); and implement agreed standards for workflow and task execution such as the [GA4GH workflow execution service](#) and task execution service ([GA4GH task execution schemas](#)) standard APIs. Alignment with EOSC will thus drive federated computation via the implementation of standards to make clouds compatible both within the life sciences globally (for example, by using the GA4GH cloud standards) and with other science domains in EOSC.

National and regional capacities are actively developing the necessary software layers that enable genomics data management to leverage investments made in electronic infrastructures. For example, the [Trygve project](#) will invest €6 million from 2017 to 2020 to develop and facilitate access to secure electronic infrastructures for human data, suitable for hosting large-scale, cross-border biomedical research studies. Services will be based on key ELIXIR technologies such as the EGA, cloud capacities of the ELIXIR Nodes and the AAI. Another example is the [High Performance Computing Research Infrastructure Eastern Region project](#), which will invest €20 million into a secure national super-computing centre in Slovenia to support national and regional research infrastructures, including life-science

**Broad consent**

Consent for an unspecified range of future research subject to a few content and/or process restrictions.

ESFRIs with high-performance computing services. Services will be aligned with ELIXIR key technologies such as cloud and container capacities of the Nodes and federated AAL.

**Bioinformatics training**

Keeping pace with the constant development of new technologies and infrastructure services is difficult, particularly for early-career clinicians and researchers who are being exposed to big data analysis for the first time. Bioinformatics capacity and competence across Europe must improve to empower efficient and effective access and analyses of genomic data. This will rely on the establishment and dissemination of best practices in bioinformatics training, providing support to training providers across Europe in developing and delivering training events, and the provision of a sustainable training infrastructure.

Existing training and corresponding materials could be used; for example, the ELIXIR training platform, an interactive training community that spans all member states, offers a seamlessly integrated technical infrastructure, including its flagship [Training eSupport System \(TeSS\)](#). The TeSS is a training toolkit that can be adopted and implemented by all ELIXIR Nodes and contains guidelines, metrics and training descriptors, as well as a course portfolio to support the training needs of the ELIXIR community. Within the ELIXIR framework, a training programme developed by the European Bioinformatics Institute delivers world-leading training in bioinformatics and scientific service provision to the research community, empowering scientists at all career stages and across sectors to make the most of biological data and strengthening bioinformatics capacity across the globe.

Beyond bioinformatics, the European research infrastructures deliver innovative ‘business process’ training programmes for managers and operators of research infrastructures, such as the [Executive Master's in Management of Research Infrastructure](#) developed by the [RItrain](#) project. This programme enables managers of research infrastructures across all domains to gain expertise on compliance, data coding (for example, using Data Use Ontology), governance, organization, financial and staff management, funding, intellectual property, service provision, and outreach in an international context. Additionally, the Coordinated Research Infrastructures Building Enduring Life-science Services (CORBEL) project enables staff exchanges, short courses and webinars for technical operators of the research infrastructures. Such initiatives are critical to developing the human resources necessary to run research infrastructures and engage with patients and citizens as well as experts and are beginning to set Europe apart from the rest of the world.

**Regulatory issues**

For 1 million human genomes to be shared transnationally by 2022, regulatory issues will need to be resolved within the community and rules will be required to implement procedures that can be efficient and still privacy-preserving (for example, inclusion criteria for

participants, or how and what information is shared with participants). Intellectual property rights management needs to be agreed and regulatory differences between countries solved. In addition, training as well as competent guidance on practical issues of data exchange across Europe and internationally will be essential.

In May 2018, the European GDPR came into force with the aim to harmonize data protection law in the EU. However, the principle setup of the GDPR allows flexibility for scientific research purposes, which poses practical challenges<sup>15</sup>. For example, the GDPR allows broad consent as one possible legal basis for data processing provided that organizational and technical safeguards are in place to protect the rights and freedom of the data subjects in research. This condition increases the responsibility and accountability of the data controller, which leads to extensive documentation requirements.

Moreover, although the GDPR is directly applicable in all member states of the European Economic Area, it leaves a high degree of freedom to countries regarding the implementation of many research-relevant provisions<sup>16,17</sup>. According to [GDPR article 9\(4\)](#), each country is free to set its own rules for processing health and genetic data as well as for research exemptions. Not only does this affect the way such data must be handled but also offers the possibility to use an alternative legal basis to consent in order to comply with [GDPR articles 6 and 9](#). For example, in Ireland, the legislation to process genetic data for research requires that explicit consent be obtained<sup>18</sup>. In the Netherlands, explicit consent is required as well but can be waived if it is impossible to ask for explicit consent or if it requires a disproportionate effort<sup>19</sup>. By contrast, in Sweden, consent can be flexible under the condition that an ethics approval is obtained<sup>20</sup>. Such different requirements for processing the same data provide a major threat to scientific collaborations in the EU, as biomedical research needs clear policies and support for high-quality risk analysis for the storage, processing and access to sensitive human data.

The initiative and willingness of so many countries to share genomic data for research and health purposes now provides a great opportunity to enter a dialogue of harmonization between the countries at the governmental level. Activities are already in motion on the level of research infrastructures. Ethical and legal concerns for all infrastructures dealing with human health data are very similar with respect to, for example, privacy, consent, protection of personal data, differences in national legislation and their implementation. ELIXIR and [BBMRI-ERIC](#) have agreed to explore and develop the necessary regulatory frameworks and policies jointly, with expert input from representatives from both infrastructures. To this end, ELIXIR and [BBMRI-ERIC](#) are in the process of developing a collaboration strategy with the intent of establishing a long-term relationship and knowledge exchange concerning both legal and ethical requirements surrounding the use of sensitive data for research.

However, harmonization and collaboration on regulatory aspects, and in particular data protection issues, must go beyond these two infrastructures. Therefore,

## Box 2 | Summary of recommendations

A coordinated, secure, federated environment that enables population-scale genomic, phenotypic and biomolecular data to be accessible across international borders (see the table) will be required to enable the committed European Union (EU) member states to achieve their goal to access 1 million genomes and other health-related data.

Research infrastructures, such as ELIXIR and the Biobanking and Biomolecular Resources Research Infrastructure, already connect national centres across Europe. They have established groups for developing shared data models, state of the art data encryption processes and establishment of cross-boundary 'data use agreements'. Lessons learned and solutions developed can be used. It will be critical to ensure coordination and integration of national reference genomes and cohorts that allow for high-precision analysis of national populations and the establishment of national variant frequency databases based on whole-genome sequencing data. The EU must take the lead on policy framing and technical standards-setting on a global stage in collaboration with organizations such as the Global Alliance for Genomics and Health to enable data access to authorized researchers.

Necessary minimal infrastructure component	In development <sup>a</sup>	Implemented at scale <sup>a</sup>
Genomics data and clinical information standards geared towards specific disease communities	Yes	No
Common application programming interfaces to enable remote data discovery and access	Yes	Yes
Computational resources, including secure, federated cloud computing environments that offer secure access across national boundaries to raw data and interoperable results	Yes	Yes
Regulatory frameworks that enable access to and the processing of genomic data across borders, including the management of transnational user access and compliance	Yes	No
A repository of tools and services, including workflows to analyse deposited data while enabling these analysis workflows to operate on data across national borders. This will contribute towards data reproducibility and provenance, which are of high importance in both research and clinical practices	Yes	Yes
A training and capacity-building programme to develop the skills and workforce required for genomics and big data in health care as well as shift the culture towards openness and integration of research data across national boundaries	Yes	Yes

<sup>a</sup>'In development' and 'Implemented at scale' refer to locally defined status within ELIXIR and/or Biobanking and Biomolecular Resources Research Infrastructure.

BBMRI-ERIC coordinates the [GDPR Code of Conduct for Health Research Initiative](#), which brings together more than 130 individuals (such as legal and ethics experts, researchers, patient advocates, industry representatives and biomedical science research infrastructures) that represent more than 80 organizations in the field of health research. The aim of the code of conduct is to provide an instrument, following GDPR article 40, to give health research-specific guidance for data protection based on ethical and data protection principles. It takes into account the specific features of processing personal data in the area of health to find the right balance in enabling research while protecting the privacy of research participants and patients.

Additionally, BBMRI-ERIC supports the biobanking community by facilitating compliance with regulatory requirements and best practice standards through a common service on ethical, legal and social issues that

includes a helpdesk and knowledge base (Ethical, Legal and Social Issues in Biobanking)<sup>21</sup>. Within the CORBEL project — an initiative of 13 biomedical research infrastructures that aims to create a platform for harmonized user access to biological and medical technologies, biological samples and data services — these services have been broadened to support the broader biomedical science research infrastructure community and are set up to address the ethical, legal and societal challenges of genomic research.

### Conclusions

Our understanding of the human genome is recognized as a primary factor for improvement in health care. Initiatives on a national scale are being established to generate genomic data to realize the benefits of precision medicine. The most advanced — Genomics England in the UK — has now completed full genome sequencing for more than 100,000 participants<sup>22</sup> and has already demonstrated benefits by providing a diagnosis for one in four participants of the rare disease component of the initiative. No other national sequencing initiative has reached this scale, with most being currently at the stage of inception.

Data sharing knowledge and technologies sit mostly within the research sector where, to date, most data have been generated. As the majority of genomics data generation shifts to the health-care sector<sup>3</sup>, a sector that is not used to handling data at this scale, the knowledge that already exists should be leveraged. Providing access to sensitive human data to authorized researchers within one country is challenging in itself; providing access to 1 million human genomes cross-border by 2022 (as proposed by the EU declaration<sup>5</sup>) will be even more so. Beyond the technical capabilities, such a project needs to ensure that patients are satisfied and understand how their data are shared, or willingness to participate will dwindle and future benefits will not be realized.

Efficient management of genomics data from human participants, ensuring that the privacy of individuals is preserved, will be vital to meet current aims. To truly federate services for controlled-access human data we will need to identify, develop and disseminate global interoperable and reusable standards, and these standards must be persistent, stable and fit for purpose. We have described in this paper the infrastructure that exists to build upon for transnational-scale genomics data access and our minimal recommendations for an EU-wide infrastructure for accessing and analysing genomics data (BOX 2).

A strong and active collaboration between ESFRIs working under the CORBEL project (and beyond) is the best option to implement the EU declaration, with the support of all the signatories. The federated infrastructure needed to deliver access to genomic and health data at a transnational scale must be an open infrastructure: it will not 'own' all data resources in Europe; rather, it should operate as an 'interoperability backbone' that allows partners (for example, ESFRIs, international initiatives, national coordination units and institutional data centres) to make use of existing resources and connect and interoperate their resources. As such, the blueprint

we are outlining in this paper builds on a unique set of European research organizations that exist within the transnational regulatory and institutional framework of the EU. Distributed European research infrastructures such as BBMRI-ERIC and ELIXIR are unique, and in contrast to the more commonly formed research consortia and large-scale initiatives, for example, the Human Cell Atlas<sup>23</sup> or the NIH Big Data to Knowledge initiative<sup>24</sup>, they connect national infrastructures and resources via a permanent legal framework. Thus, we are outlining a strategy to overcome a major challenge in European research — that the assembly of large cohorts will require transnational collaboration and

pooling of data over international borders — by building on the established, strong European institutions. By building on global standards and maintaining active international collaborations, this infrastructure can serve as a template for a truly international federation. A sustainable infrastructure for users that manages data identifiers, secure data archiving and access, and ensures mappings between resources will enable long-term, cost-effective data management and drive standards as the default across the European life science and health data landscape.

Published online 27 August 2019

- Lochmüller, H. et al. RD-Connect, NeurOmics and EUREnOmics: collaborative European initiative for rare diseases. *Eur. J. Hum. Genet.* **26**, 778–785 (2018).
- Horgan, D. From here to 2025: personalised medicine and healthcare for an immediate future. *J. Cancer Policy* **16**, 6–21 (2018).
- Auffray, C. et al. Making sense of big data in health research: towards an EU action plan. *Genome Med.* **8**, 71 (2016).
- Birney, E., Vamathevan, J. & Goodhand, P. Genomics in healthcare: GA4GH looks to 2022. Preprint at *bioRxiv* <https://www.biorxiv.org/content/10.1101/203554v1> (2017).
- The European Commission. Declaration of cooperation: towards access to at least 1 million sequenced genomes in the European Union by 2022. *European Commission* [http://ec.europa.eu/newsroom/dae/document.cfm?doc\\_id=50964](http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50964) (2018).  
**This declaration from the European Commission posits the provision of transnational access to at least 1 million human genomes by 2022.**
- Philippakis, A. A. et al. The Matchmaker Exchange: a platform for rare disease gene discovery. *Hum. Mut.* **36**, 915–921 (2015).
- Lappalainen, I. et al. The European Genome-phenome Archive of human data consented for biomedical research. *Nat. Genet.* **47**, 692–695 (2015).
- Durinx, C. et al. Identifying ELIXIR core data resources. *Version 2. F1000Res.* **5**, 2422 (2016).
- Fiume, M. et al. Federated discovery and sharing of genomic data using Beacons. *Nat. Biotechnol.* **37**, 220–224 (2019).  
**The Beacon API protocol is an approved GA4GH to federated genomics data discoverability and has many implementations across ELIXIR.**
- Holub, P. et al. BBMRI-ERIC directory: 515 biobanks with over 60 million biological samples. *Biopreserv. Biobank.* **14**, 559–562 (2016).
- Litton, J. E. Launch of an infrastructure for health research: BBMRI-ERIC. *Biopreserv. Biobank.* **16**, 233–241 (2018).
- Merino-Martinez, R. et al. Toward global biobank integration by implementation of the minimum information about biobank data sharing (MIABIS 2.0 Core). *Biopreserv. Biobank.* **14**, 298–306 (2016).
- European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *EUR-Lex* <http://data.europa.eu/eli/reg/2016/679/oj> (2016).
- Linden, M. et al. Common ELIXIR service for researcher authentication and authorisation. *F1000Res.* **7**, 1199 (2018).
- Kaye, J. et al. Are requirements to deposit data in research repositories compatible with the European Union's General Data Protection Regulation? *Ann. Intern. Med.* **170**, 332–334 (2019).
- Dove, E. S. The EU General Data Protection Regulation: implications for international scientific research in the digital era. *J. Law Med. Ethics* **46**, 1013–1030 (2018).
- Shabani, M. & Borry, P. Rules for processing genetic data for research purposes in view of the new EU General Data Protection Regulation. *Eur. J. Hum. Genet.* **26**, 149–156 (2018).
- Harris, S. Data protection act 2018 (section 36(2)) (health research) regulations 2018. *eISB* <http://www.irishstatutebook.ie/eli/2018/si/314/made/en/pdf> (2018).
- Government of the Netherlands. Regels ter uitvoering van Verordening (EU) 2016/679 van het Europees Parlement en de Raad van 27 april 2016 [Dutch]. *Rijksverheid* <https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2017/12/08/tk-uitvoeringswet-algemene-verordening-gegevensbescherming-en-mvt-tbv-rvs-def/tk-uitvoeringswet-algemene-verordening-gegevensbescherming-en-mvt-tbv-rvs-def.pdf> (2017).
- Government Offices of Sweden. Lag (2003:460) om etikprövning av forskning som avser människor [Swedish]. *Regeringskansliet* <http://rkrattsbaser.gov.se/sfst?bet=2003:460> (2018).
- Mayrhofer, M. & Schlünder, I. Mind the gap: from tool to knowledge base. *Biopreserv. Biobank.* **16**, 458–462 (2018).
- Genomics England. The UK has sequenced 100,000 whole genomes in the NHS. *Genomics England* <https://www.genomicsengland.co.uk/the-uk-has-sequenced-100000-whole-genomes-in-the-nhs> (2018).
- Rozenblatt-Rosen, O. et al. The Human Cell Atlas: from vision to reality. *Nature* **550**, 451–453 (2017).
- Paten, B. et al. The NIH BD2K center for big data in translational genomics. *J. Am. Med. Assoc. Inform. Assoc.* **22**, 1143–1147 (2015).

#### Acknowledgements

The authors thank D. Lloyd (ELIXIR-Hub), U. Gerst-Talaz (ELIXIR-EE), A. Jene and J. Dopazo (ELIXIR-ES) for reviewing and commenting on this manuscript whilst in preparation. Additionally, the authors would like to acknowledge all members of the ELIXIR Federated Human Data, Rare Diseases, and Human Copy Number Variation Communities whose input and work has contributed to this manuscript and whose combined work in future under the banner of the ELIXIR Human Data Communities, along with the five ELIXIR Platforms (Compute, Data, Interoperability, Tools and Training), shall provide workable solutions to meet the aims of the EU Declaration to share at least 1 million genomes transnationally by 2022. Within this group the authors would like to specifically acknowledge V. Satagopam (ELIXIR-LU), N. Jareborg (ELIXIR-SE), M. Chiara (ELIXIR-IT), H. Peterson (ELIXIR-EE), A. Dimopoulos (ELIXIR-GR) and A. Ardashirdavani (ELIXIR-BE). The authors would like to thank all the contributors of BBMRI-ERIC Common Service IT.

#### Author contributions

G.S., E.B., S.Br., P.F., N.B. and S.S. researched the literature. G.S., E.B., S.Br., P.F., I.G., N.B. and S.S. provided substantial contributions to discussions of the content. G.S., R.B., S.Be., C.Bé., C.Br., M.V.d.B., S.C.-G., F.F., J.He., P.H., J.Ho., N.J., T.M.K., J.O.K., G.M., M.T.M., A.M., T.N., A.Pag., B.P., H.P., J.R., D.S., M.A.S., S.V., N.B. and S.S. wrote the article. G.S., M.B., R.B., S.Be., C.Bé., C.Br., M.V.d.B., R.D., S.C.-G., F.F., P.G., I.G., J.He., P.H., J.Ho., N.J., T.M.K., J.O.K., I.L., B.L., G.M., M.T.M., A.M., A.N., A.V., S.N., T.N., A.Pag., B.P., A.Pal., H.P., J.R., D.S., E.S., M.A.S., S.V., N.B. and S.S. reviewed and/or edited the manuscript before submission.

#### Competing interests

E.B. is a paid consultant to Oxford Nanopore, Glaxo-SmithKline and Dovetail Inc. S.Br. acknowledges funding from the Danish Agency for Science, Technology and Innovation (09–067306), Novo Nordisk Foundation (NNF14CC0001). P.F. is a member of the scientific advisory boards of Fabric Genomics, Inc., and Eagle Genomics, Ltd. The other authors declare no competing interests.

#### Peer review information

*Nature Reviews Genetics* thanks H. Rehm, B. Knoppers, E. Dove and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

#### Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

#### RELATED LINKS

BBMRI-ERIC: <http://www.bbMRI-eric.eu/>  
 BBMRI-ERIC AAI: <https://web.bbMRI-eric.eu/Policies/>  
 BBMRI-ERIC Directory: <https://directory.bbMRI-eric.eu/>  
 CINECA: <https://edukad.etag.eu/project/4011?lang=en>  
 CORBEL: <http://www.corbel-project.eu/about-corbel.html>  
 Data Use Ontology: <https://github.com/EBISPOT/DUO>  
 ELIXIR core data resources: <https://elixir-europe.org/platforms/data/core-data-resources>  
 ELIXIR Europe: <https://elixir-europe.org/>  
 ELIXIR Rare Diseases Community: <https://www.elixir-europe.org/communities/rare-diseases>  
 ELIXIR Scientific Programme (2019–2023): <https://elixir-europe.org/about-us/what-we-do/elixir-programme>  
 Ethical, Legal, and Social Issues in Biobanking: <http://www.bbMRI-eric.eu/services/common-service-elsi/>  
 EUCANCan: <https://eucanCan.com/>  
 euCanShare: <http://www.eucanShare.eu/>  
 European Genome-phenome Archive (EGA): <https://ega-archive.org/>  
 European Open Science Cloud (EOSC): <https://www.eosc-portal.eu/>  
 European Strategy Forum on Research Infrastructures (ESFRIs): <https://www.esfri.eu/roadmap-2018>  
 Executive Master's in Management of Research Infrastructure: <http://www.emmri.unimib.it/en/>  
 GA4GH data repository service schemas: <https://github.com/ga4gh/data-repository-service-schemas>  
 GA4GH 'data use and researcher identities' work stream: <https://ga4gh-duri.github.io>  
 GA4GH driver projects: <https://www.ga4gh.org/how-we-work/driver-projects/>  
 GA4GH task execution schemas: <https://github.com/ga4gh/task-execution-schemas>  
 GA4GH work streams: <https://www.ga4gh.org/how-we-work/workstreams/>  
 GA4GH workflow execution service schemas: <https://ga4gh.github.io/workflow-execution-service-schemas/>  
 GDPR article 9(4): <https://gdpr-info.eu/art-9-gdpr/>  
 GDPR Code of Conduct for Health Research initiative: <http://www.code-of-conduct-for-health-research.eu/>  
 Global Alliance for Genomics and Health (GA4GH): <https://www.ga4gh.org>  
 High Performance Computing Research Infrastructure Eastern Region: [https://www.hpc-rivrsi/home\\_en/](https://www.hpc-rivrsi/home_en/)  
 Innovative Medicines Initiative: <https://www.imi.europa.eu/>  
 Matchmaker Exchange: <https://www.matchmakerexchange.org/>  
 Minimum Information About Biobank data Sharing (MIABIS) 2.0: <http://www.bbMRI-eric.eu/services/miabis/>  
 Personal Health Train (PHT): <http://www.dtls.nl/fair-data/personal-health-train/>  
 Proyecto Genoma 1000 Navarra: <https://www.nagen1000navarra.es/en/home>  
 Rltrain: <http://rltrain.eu/>  
 Training eSupport System: <https://tess.elixir-europe.org/>  
 Tryggve project: <https://neic.no/tryggve/>

