Leveraging European infrastructures to access 1 million human genomes by 2022

*Gary Saunders[1], Michael Baudis[2], Regina Becker[3], Sergi Beltran[4,5], Christophe Béroud[6,7], Ewan Birney[8], Cath Brooksbank[8], Søren Brunak[9,10], Marc Van den Bulcke[11], Rachel Drysdale[1], Salvador Capella-Gutierrez[12], Paul Flicek[8], Francesco Florindi[13], Peter Goodhand[14,15], Ivo Gut[4,5], Jaap Heringa[16], Petr Holub[13], Jef Hooyberghs[17], Nick Juty[18], Thomas M. Keane[8], Jan O. Korbel[19], Ilkka Lappalainen[20], Brane Leskosek[21], Gert Matthijs[22], Michaela Th. Mayrhofer[13], Andres Metspalu[23], Arcadi Navarro[24,25,26], Steven Newhouse[8], Tommi Nyrönen[20], Angela Page[15,27], Bengt Persson[28], Aarno Palotie[29], Helen Parkinson[8], Jordi Rambla[26], David Salgado[6], Erik Steinfelder[13], Morris A. Swertz[30], Alfonso Valencia[12,31], Susheel Varma[8], Niklas Blomberg[1] and Serena Scollen[1]**

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[1] (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[2–4].

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.[5]) (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[6] 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)[7] 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.[7]). The EGA is one of several ELIXIR core data resources and the recommended database for deposition of controlled-access human data[8].

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

[1]ELIXIR Hub, Wellcome Genome Campus, Hinxton, Cambridge, UK.
[2]University of Zurich, Zurich, Switzerland.
[3]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Luxembourg, Luxembourg.
[4]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain.
[5]Universitat Pompeu Fabra (UPF), Barcelona, Spain.
[6]Aix Marseille Univ, INSERM, MMG, Marseille, France.
[7]Département de Génétique Médicale et de Biologie Cellulaire, APHM, Hôpital d'Enfants de la Timone, Marseille, France.
[8]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK.
[9]Department of Health Technology, Technical University of Denmark, Lyngby, Denmark.
[10]Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark.
[11]Cancer Centre, Epidemiology and Public Health, Sciensano, Ixelles, Belgium.
[12]Barcelona Supercomputing Centre (BSC), Barcelona, Spain.
[13]BBMRI-ERIC, Graz, Austria.
[14]Ontario Institute for Cancer Research, Toronto, Ontario, Canada.
[15]Global Alliance for Genomics and Health, Toronto, Ontario, Canada.
[16]Department of Computer Science, Vrije Universiteit, Amsterdam, Netherlands.
[17]Flemish Institute for Technological Research, VITO, Mol, Belgium.
[18]School of Computer Science, The University of Manchester, Manchester, UK.
[19]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany.
[20]CSC — IT Center for Science, Espoo, Finland.
[21]IBMI, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia.
[22]Katholieke Universiteit Leuven, Leuven, Belgium.
[23]Estonian Genome Center, University of Tartu, Tartu, Estonia.
[24]Institute of Evolutionary Biology (UPF-CSIC), Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain.
[25]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.
[26]Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain.
[27]Broad Institute of MIT and Harvard, Cambridge, MA, USA.
[28]Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala, Sweden.
[29]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland.
[30]BBMRI-NL/University Medical Center Groningen, University of Groningen, Groningen, Netherlands
[31]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[1]. 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[1] 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[7] 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[9]. 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[10]. 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[11]. The biobank information standard group, Minimum Information About BIobank data Sharing (MIABIS) 2.0 (REF.[12]) 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.

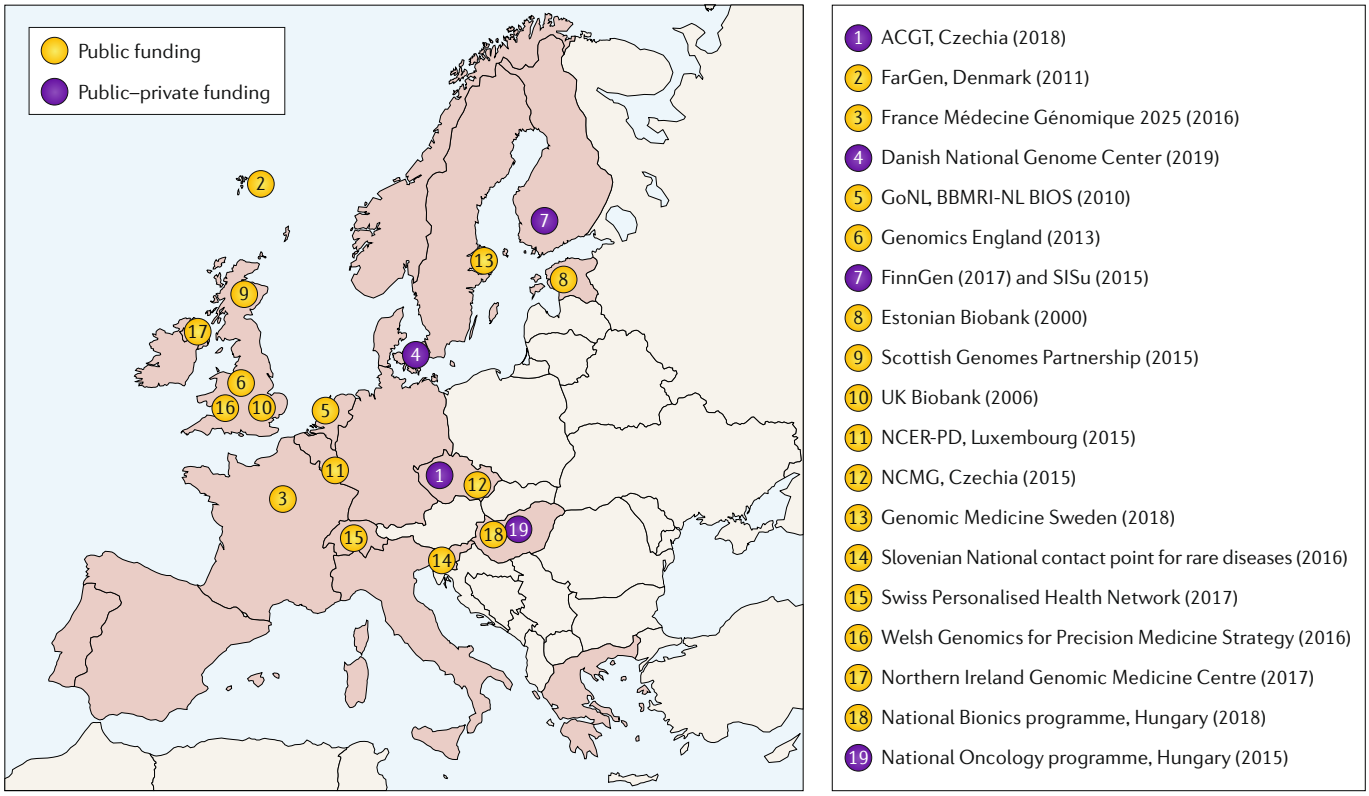

Fig. 1 | **Examples of current health care-focused and genomics-based national initiative projects across ELIXIR members.** In many European countries (for example, Spain and Italy) health care is administered regionally and, until now, genomics-based projects have been linked to the regional health-care 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 use their existing institutional identities 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	EU declaration signatory	BBMRI-ERIC status	ELIXIR status	EMBL status
IARC[a]	–	Full Member	–	–
Austria	Yes	Full Member	–	Full Member
Belgium	No	Full Member	Member	Full Member
Bulgaria	Yes	Full Member	–	–
Croatia	Yes	–	–	–
Cyprus	Yes	Observer	Observer	–
Czech Republic	Yes	Full Member	Member	Full Member
Denmark	No	–	Member	Full Member
Estonia	Yes	Full Member	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	Full Member	Member	Full Member
Ireland	No	Full Member	Member	Full Member
Israel	No	Full Member	Member	Full Member
Italy	Yes	Full Member	Member	Full Member
Latvia	Yes	Full Member	–	–
Lithuania	Yes	–	–	Prospect member
Luxembourg	Yes	–	Member	Full Member
Malta	Yes	Full Member	–	Full Member
Montenegro	–	–	–	Full Member
Netherlands	Yes	Full Member	Member	Full Member
Norway	Yes	Full Member	Member	Full Member
Poland	–	Full Member	–	Prospect member
Portugal	Yes	–	Member	–
Slovakia	–	–	–	Full Member
Slovenia	Yes	–	Member	–
Spain	Yes	–	Member	Full Member
Sweden	Yes	Full Member	Member	Full Member
Switzerland	No	Observer	Member	Full Member
Turkey	–	Observer	–	–
United Kingdom	Yes	Full Member	Member	Full Member

BBMRI-ERIC, Biobanking and Biomolecular Resources Research Infrastructure; EMBL, European Molecular Biology Laboratory; EU, European Union; IARC, International Agency for Research on Cancer. [a]The IARC is a unique organization. For almost 50 years, since its creation by the World Health Assembly of the World Health Organization, the IARC has been making important contributions to the global fight against cancer, notably through its capacity to bring together people and organizations from across the world that share common values and objectives.

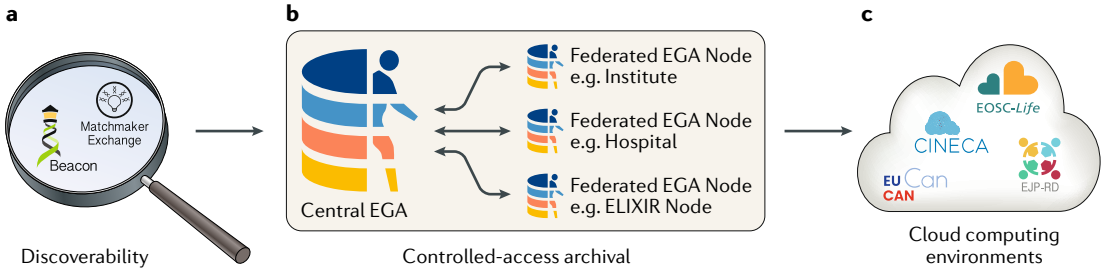

Fig. 2 | **The concept of EGA federation — from data discoverability to raw sensitive human data access.** **a** | Discoverability. Metadata is shared from each of the sensitive data archives to a centralized database upon which query interfaces can be built; these can be project-specific portals or 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)[13]. 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[14] 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 Tryggve 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 supercomputing 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 AAI.

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[15]. 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[16,17]. 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[18]. 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[19]. By contrast, in Sweden, consent can be flexible under the condition that an ethics approval is obtained[20]. 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[a]	Implemented at scale[a]
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

[a]'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)[21]. 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[22] 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[4], 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[5]) 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[23] or the NIH Big Data to Knowledge initiative[24], 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

1. Lochmüller, H. et al. RD-Connect, NeurOmics and EURenOmics: collaborative European initiative for rare diseases. *Eur. J. Hum. Genet.* **26**, 778–785 (2018).
2. Horgan, D. From here to 2025: personalised medicine and healthcare for an immediate future. *J. Cancer Policy* **16**, 6–21 (2018).
3. Auffray, C. et al. Making sense of big data in health research: towards an EU action plan. *Genome Med.* **8**, 71 (2016).
4. 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).
5. 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 (2018).
This declaration from the European Commission posits the provision of transnational access to at least 1 million human genomes by 2022.
6. Philippakis, A. A. et al. The Matchmaker Exchange: a platform for rare disease gene discovery. *Hum. Mut.* **36**, 915–921 (2015).
7. Lappalainen, I. et al. The European Genome-phenome Archive of human data consented for biomedical research. *Nat. Genet.* **47**, 692–695 (2015).
8. Durinx, C. et al. Identifying ELIXIR core data resources. *Version 2. F1000Res.* **5**, 2422 (2016).
9. 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.
10. Holub, P. et al. BBMRI-ERIC directory: 515 biobanks with over 60 million biological samples. *Biopreserv. Biobank.* **14**, 559–562 (2016).
11. Litton, J. E. Launch of an infrastructure for health research: BBMRI-ERIC. *Biopreserv. Biobank.* **16**, 233–241 (2018).
12. 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).
13. 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).
14. Linden, M. et al. Common ELIXIR service for researcher authentication and authorisation. *F1000Res.* **7**, 1199 (2018).
15. 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).
16. 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).
17. 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).
18. 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).
19. Government of the Netherlands. Regels ter uitvoering van Verordening (EU) 2016/679 van het Europees Parlement en de Raad van 27 april 2016 [Dutch]. *Rijksoverheid* 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).
20. 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).
21. Mayrhofer, M. & Schlünder, I. Mind the gap: from tool to knowledge base. *Biopreserv. Biobank.* **16**, 458–462 (2018).
22. 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).
23. Rozenblatt-Rosen, O. et al. The Human Cell Atlas: from vision to reality. *Nature* **550**, 451–453 (2017).
24. Paten, B. et al. The NIH BD2K center for big data in translational genomics. *J. Am. Med. Inform. Assoc.* **22**, 1143–1147 (2015).

Acknowledgements
The authors thank D. Lloyd (ELIXIR-Hub), U. Gerst-Talas (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. Ardeshirdavani (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, GlaxoSmithKline 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.

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.ee/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-rivr.si/home_en/
Innovative Medicines Initiative: https://www.imi.europa.eu/
Matchmaker Exchange: https://www.matchmakerexchange.org/
Minimum Information About Blobank 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
RItrain: http://ritrain.eu/
Training eSupport System: https://tess.elixir-europe.org/
Tryggve project: https://neic.no/tryggve/