European infrastructures to access one 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], **Steven Newhouse**[8], **Tommi Nyrönen**[20], **Angela Page**[15,24], **Bengt Persson**[25], **Aarno Palotie**[26], **Helen Parkinson**[8], **Jordi Rambla**[27], **David Salgado**[6], **Erik Steinfelder**[13], **Morris A. Swertz**[28], **Susheel Varma**[8], **Niklas Blomberg**[1], **Serena Scollen**[1,*]

[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 [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, 13005 Marseille, France [7]Département de Génétique Médicale et de Biologie Cellulaire, APHM, Hôpital d'Enfants de la Timone, 13385 Marseille, France [8]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom [9]Department of Health Technology, Technical University of Denmark, Kemitorvet, DK-2800 Lyngby, Denmark [10]Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Blegdamsvej 3A, DK-2200 Copenhagen, Denmark [11]Cancer Centre, Epidemiology and Public Health, Sciensano, 1050 Ixelles, Belgium [12]Barcelona Supercomputing Centre (BSC), Barcelona, 08034, Spain [13]BBMRI-ERIC, 8010 Graz, Austria [14]Ontario Institute for Cancer Research, 661 University Ave, Suite 510, Toronto, ON M5G 0A3, Canada [15]Global Alliance for Genomics and Health, Toronto, Ontario, Canada [16]Department of Computer Science, Vrije Universiteit, De Boelelaan 1105, 1081 HV Amsterdam, Netherlands [17]Flemish Institute for Technological Research, VITO, Boeretang 200, Mol B-2400, Belgium [18]School of Computer Science, The University of Manchester, Manchester, UK [19]Genome Biology Unit, EMBL, Heidelberg, Germany [20]CSC - IT Center for Science, Espoo, Finland [21]IBMI, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia [22]Katholieke Universiteit Leuven, 3000 Leuven, Belgium [23]Estonian Genome Center, University of Tartu, Tartu, Estonia [24]Broad Institute of MIT and Harvard, MA 02142, USA [25]Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala, Sweden [26]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland [27]Centre for Genomic Regulation, Carrer del Dr. Aiguader, 88, 08003 Barcelona, Spain [28]BBMRI-NL/University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

---

[*] serena.scollen@elixir-europe.org.

Abstract

Many countries in Europe have nascent personalized medicine programmes, and human genomics is undergoing a step change from being a predominantly research-driven activity to one driven through healthcare. To maximise the value of the generated genomic data, these data will need to be shared between institutions and across countries. In recognition of this challenge, a declaration has recently been signed by 20 European countries to share at least one million human genomes by 2022, transnationally. In this Roadmap, we identify challenges of data sharing and demonstrate that European research infrastructures are well-positioned to support the rapid implementation of widespread genomic data access.

Introduction

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]. However, realizing the potential of personalized medicine beyond rare diseases will require systematic access and integration of research and healthcare data at a greater scale, for example, across countries[2–4].

Across Europe, several national initiatives are being established to generate genomic data (Fig. 1), most of which are disease-agnostic, although some initiatives focus on cancer, infectious disease and/or rare disease. Recently, representatives of 20 member states of the European Union signed a joint declaration to deliver cross-border access to human genomes by the end of 2022[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 personalized 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 healthcare systems. Currently, each European country sets its own regulatory framework for processing health and genetic data, and enabling access to these data for research. Moreover, genetic and associated data generated through healthcare are not shared as widely as research data; given that healthcare 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-science 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; and by providing 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 genomics data at this scale is possible by building on established European research infrastructures. 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 data access via federated data sharing models, data discovery, data standards, computing, regulatory frameworks, and training. By leveraging existing services to achieve this ambitious aim, Europe can be positioned as a global leader in this field.

European Unionwide infrastructure

Access and management of genomics data is now more of a challenge than generation of the data itself. 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 (IMI) [https://www.imi.europa.eu/], have made and continue to make considerable investments in data and knowledge management infrastructure. However, efforts are mostly independent and uncoordinated, resulting in fragmented and overlapping investments in data management. By implementing a Europe-wide framework of experts and long-term services, the European Strategy Forum on Research Infrastructures (ESFRIs) [https://www.esfri.eu/roadmap-2018], which includes, for example, ELIXIR and the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-ERIC) (Box 1), can drive the coordination of efforts at both the national and international levels, as illustrated by the impact that the provision of infrastructure has already had on the rare diseases (Box 2) and cancer research communities (Box 3).

One possible solution to 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 they can respond to queries as if they were in a single database. Matchmaker Exchange [https://www.matchmakerexchange.org/][6] is an example of a federated data sharing platform, successfully facilitating the matching of rare disease cases with similar phenotypic and genotypic profiles. Rare disease patient willingness to share data has driven earlier implementation compared to models that are being established for data sharing/access beyond rare disease. We present two platforms in mature stages of development that are moving towards use case driven implementation - the European genome–phenome archive (EGA; [https://ega-archive.org/])[7] and the Personal Health Train (PHT)[8]

The EGA is a resource for 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 (EMBL-EBI), UK, and the Centre for Genomic Regulation (CRG), Spain, hosts over 1,700 studies that comprise more than 4,000 data sets from >900 data providers, and has served data to over 10,000 requestors since 2008. Key data collections for human genetics research are hosted by the central EGA, such as those from RD-Connect, BLUEPRINT, UK10K, UKBiobank, the Human Induced Pluripotent Stem Cells Initiative (HipSci), Wellcome Trust Case Control Consortium (WTCC), and the International Cancer Genome Consortium (ICGC) (full list of EGA studies: [https://ega-archive.org/studies]). The EGA is an ELIXIR Core Data Resource and

the recommended database for deposition of controlled-access human data[9]. The EGA is now being extended to a federated model, which will enable local implementations at research institutes in the 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 interface. After data discovery, access to the data itself can be requested from the source e.g. via application to a Data Access Committee, to establish agreements for data use. The EGA participates in the large funded projects euCanSHare [http://www.eucanshare.eu/], EUCANCan [https://eucancan.com/], and CINECA [https://edukad.etag.ee/project/4011?lang=en]. The CINECA project, will work with 18 organizations representing European, Canadian and African 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) regulations.

Another possible solution being developed by consortia in The Netherlands and Germany is the Personal Health Train (PHT), which is a concept for the (re)use of personal data in health care, 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 - achievable using a suite of standardised 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

An essential element to unlock access for authorized researchers to one million human genomes across the EU is the awareness of the existence and location of these data. This requires the provision of metadata characterizing the samples and genomes, such as their association with a certain disease, as well as their registration in a searchable database that allows data to be found by both humans and computers.

Metadata, such as data set descriptors, can be shared and made searchable through a common interface even when data is hosted locally, as demonstrated by the EGA. The findability 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[10]. To maintain participant anonymization, only the presence or absence of the specified variant in data collections is reported. This information allows the researcher to contact the person(s) 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 Global Alliance for Genomics and Health (GA4GH). Currently, nine ELIXIR member countries have launched national Beacons.

A large part of the data and samples needed to sequence one million genomes is already stored in biobanks, and is searchable, for example, via the European research infrastructure for biobanking, BBMRI-ERIC directory[11]. BBMRI-ERIC facilitates access to high-quality samples and data by networking 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[12]. The biobank information standard group (MIABIS) 2.0 (Minimum Information About BIobank data Sharing)[13] and BBMRI-ERIC interoperability forum groups are working on developing a common application programming interface (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) [https://elixir-europe.org/about-us/what-we-do/elixir-programme] will see the generation of the necessary interfaces and data models to allow the biobanks to become interoperable with the Beacon discovery protocol for the genetic data component. 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 is often heterogeneous and recorded using varied standards and ontologies. The communities working with these data need coordinated expert advice on which standards to adopt for federated data access. To facilitate reuse, data producers must have compatible (interoperable) interfaces and provide computational services that allow data integration.

GA4GH is a policy framing and standards setting organisation for genomics and has a multi-year plan to provide standards upon which federated data sites (including research, healthcare and commercial organizations as well as individuals) use, analyse and store the data needed to drive precision medicine. Going forward, the vast majority of these data are expected to come from healthcare rather than research, and they span individuals of many national and ethnic origins. Harmonized data governance architectures allow for broad spheres of responsible data access, allowing researchers to perform analysis on virtual cohorts of populations, or the use of virtual analytical tools, without data movement. To meet the aims of the EU declaration it will be necessary to establish coordinated European collaboration with GA4GH. ELIXIR and GA4GH collaborate, as the long-term goals of ELIXIR and GA4GH are aligned. ELIXIR contributes resources to the development and implementation of GA4GH standards via implementation studies and infrastructure projects which fund GA4GH Driver Projects [https://www.ga4gh.org/how-we-work/driver-projects/] — real world genomic data initiatives that have signed on to help scope, develop, and pilot GA4GH standards.

For example, the ELIXIR Beacon project is a GA4GH Driver Project that actively contributes to four of the eight GA4GH Work Streams [https://www.ga4gh.org/how-we-work/workstreams/]: Discovery, Data Use and Researcher Identities (DURI), Clinical and Phenotypic Data Capture, and Genomic Knowledge Standards (GKS). Additionally, ELIXIR delegates co-lead four of the Work Streams: Discovery, DURI, GKS, and Large Scale Genomics Work Streams.

In another example, the EGA actively contributed to the development of - and has now deployed - the Data Use Ontology (DUO), an approved GA4GH standard that provides a computable representation of data use requirements (https://ga4gh-duri.github.io). 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 human data that has been consented for research, and access to these data in a timely manner for approved researchers.

Finally, 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 ELIXIR Authentication and Authorization Infrastructure (AAI). The vision 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 Project(s) could provide alone. The intention is to coordinate and position ELIXIR to provide a gateway for GA4GH into Europe.

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 quality of the biological material and associated data, BBMRI-ERIC leads work on interoperable provenance information model within International Organization for Standardization (ISO) Technical Committee 276. 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 these samples and associated data are used effectively, appropriate efforts should be taken to define needed specifications for sample quality and select the data of designated samples. 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 realizing the potential of cloud services across Europe so they can be used in seamless transnational workflows. The restrictions on export of human genomic data derived from healthcare means that we need to develop 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 with harmonization of task and workflow execution systems is required. Furthermore, the General Data Protection Regulation (GDPR) allows individual EU member states to define their

safeguards required to process health and genetic data (Article 9.4 GDPR)[14]. 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 (e.g. European Open Science Cloud (EOSC)[15]), as well as commercial clouds (e.g. Google Cloud, Microsoft Azure, Amazon Web Service), whilst simultaneously meeting full individual and national level identity and access requirements. Therefore, data could be organized as a federation, where data processors can get access to data sets, computational tools to process them and scalable compute resources, with a linked electronic identity provided from technologies, such as ELIXIR AAI[16] or BBMRI-ERIC AAI[17]. 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 the data, in turn increasing trust in the overall ecosystem.

Within EOSC, the BioMedical Science Research Infrastructures (BMS RI) aim to connect existing national cloud infrastructures associated with BMS RI Nodes, adopt interoperable AAI services, such as the ELIXIR AAI service, provide secure data transfers between BMS RIs to facilitate sensitive data processing, such as the Reference Data Set Distribution Service[18], and implement agreed standards for workflow and task execution, such as the GA4GH Workflow Execution Service (WES)[19] and Task Execution Service (TES)[20] 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 (e.g. by using the GA4GH cloud standards) and with other science domains in EOSC.

National and regional capacities are actively developing necessary software layers that enable genomics data management to leverage investments made in e-Infrastructures. For example, the partners in the Tryggve project will invest 6 million Euros from 2017 to 2020 to develop and facilitate access to secure e-Infrastructure for human data, suitable for hosting large-scale cross-border biomedical research studies[21]. Services will be based on key ELIXIR technologies such as the EGA, cloud capacities of the Nodes, and federated AAI. Another example is HPC RIVR (High Performance Computing, Research Infrastructure Eastern Region)[22] project that will invest 20 million Euros into a secure national supercomputing centre in Slovenia to support national and regional research infrastructures, including life science ESFRIs with HPC services. Services will be aligned with ELIXIR key technologies such as cloud/container capacities of the Nodes and federated AAI.

Regulatory issues

The declaration states that shared access to one million human genomes transnationally across borders will be achieved by 2022, and this will require regulatory issues to be resolved within the community. Rules will be needed to implement procedures that can be efficient and still privacy preserving (e.g. inclusion criteria for participants, 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 harmonise data protection law in the European Union. The principle setup allows the possibility for flexibility given to scientific research purposes and poses practical challenges[234]. Such flexibility includes broad consent as one possible legal basis for data processing. A condition is that organisational and technical safeguards are put in place to protect the rights and the freedom of the data subjects in research. This is combined with a higher responsibility and accountability of the data controller, which leads to extensive documentation requirements. While the GDPR is directly applicable in all member states of the European Economic Area, it leaves a high degree of freedom to the countries in the implementation of many of these research relevant provisions[24,25]. Therefore, even once the national implementations of the GDPR are fully established, and the possible national derogations are clarified, cross-border data access will possibly still suffer. Following GDPR Art. 9(4), [https://gdpr-info.eu/] each country is free to set its own rules for processing health and genetic data as well as for the derogations for research. In each case this will have an effect on 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 Arts. 6 and 9[24]. Correspondingly, the legislation to process genetic data for research requires, for example, in Ireland that explicit consent is obtained[26], in the Netherlands an 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[27] or in Sweden, it can be flexible under the condition that an ethics approval is obtained[28]. Such different requirements for processing the same data may 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 genome 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 e.g. 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 BBMRI-ERIC coordinates the GDPR Code of Conduct for Health Research initiative, bringing together more than 130 individuals (such as legal and ethics experts, researchers, patient advocates, industry representatives and BMS RIs) that represent more than 80 organizations in the field of health research[29]. The aim of the Code of Conduct is to provide an instrument following GDPR Art. 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 and to find the right balance in enabling research whilst 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 ELSI including a Helpdesk and Knowledge Base[30,31]. Within the CORBEL project[32], an initiative of 13 BMS RIs to create a platform for harmonised user access to biological and medical technologies, biological samples and data services required by cutting-edge biomedical research, these services have been broadened to support the broader BMS research infrastructure community and are set-up to address the ethical, legal and societal challenges of genomic research.

Bioinformatics training

Bioinformatics is a rapidly evolving field. 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 these 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.

Training and corresponding materials are already available and could be utilised. For example, the ELIXIR training platform is an interactive training community that spans all member states and offers a seamlessly integrated technical infrastructure, including the flagship Training eSupport System (TeSS [https://tess.elixir-europe.org/]). TeSS is a training toolkit that can be adopted and implemented by all ELIXIR Nodes and contains guidelines, metrics, training descriptors, as well as a course portfolio to support the training needs of the ELIXIR community. For example, within the ELIXIR framework, EMBL-EBI's training programme 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 unite to deliver innovative 'business process' training programmes for managers and operators of research infrastructure, such as the Executive Master's in Management of Research Infrastructure[33] developed by the RItrain project[34]; enabling managers of research infrastructures across all domains to gain expertise on compliance, data coding (for example, using DUO - an

approved GA4GH standard), governance, organization, financial and staff management, funding, intellectual property, service provision and outreach in an international context. Additionally, the 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, engage with patients and citizens as well as experts, and are beginning to set Europe apart from the rest of the world.

Conclusions

Our understanding of the human genome is recognized as a primary factor for improvement in healthcare. Initiatives on a national scale are being established to generate genomic data to realize the benefits of personalized medicine. The most advanced, Genomics England in the UK, has now completed full genome sequencing for more than 100,000 participants, and has already demonstrated benefits by providing a diagnosis for one in four participants of the rare disease component of the initiative[35]. 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 has been generated. As the majority of genomics data generation shifts to the healthcare 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 one million human genomes cross-border by 2022 (as proposed by the EU declaration[5]) will be even more so. We must also bear in mind that beyond the technical capabilities, patients need to be satisfied that their data is shared securely, 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 4).

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 organisations that exist within the transnational regulatory and institutional framework of the European Union. 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 (e.g. ICGC-ARGO, Human Cell Atlas[36], or the NIH BD2K initiative[37]) 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 is a first step towards this global vision.

Acknowledgements

References

1. Lochmuller H, et al. RD-Connect, NeurOmics and EURenOmics: collaborative European initiative for rare diseases. Eur J Hum Genet. 2018; 26:778–785. [PubMed: 29487416]

2. Horgan D. From here to 2025: Personalised medicine and healthcare for an immediate future. Journal of Cancer Policy. 2018; 16:6–21.

3. Auffray C, et al. Making sense of big data in health research: Towards an EU action plan. Genome Med. 2016; 8:71. [PubMed: 27338147]

4. Birney E, Vamathevan J, Goodhand P. Genomics in healthcare: GA4GH looks to 2022. bioRxiv. 2017

5. Declaration of Cooperation. [Accessed: 15th May 2019] Towards access to at least 1 million sequenced genomes in the European Union by 2022. 2018. http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50964

6. Philippakis AA, et al. The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery. Human Mutation. 2015; 36:915–921. [PubMed: 26295439]

7. [Accessed: 15th May 2019] Personal Health Train - Dutch Techcentre for Life Sciences. Dutch Techcentre for Life Sciences. Available at: http://www.dtls.nl/fair-data/personal-health-train/

8. Lappalainen I, et al. The European Genome-phenome Archive of human data consented for biomedical research. Nat Genet. 2015; 47:692–695. [PubMed: 26111507]

9. Durinx C, et al. Identifying ELIXIR Core Data Resources. Version 2. F1000Res. 2016; 5

10. Fiume M, et al. Federated discovery and sharing of genomic data using Beacons. Nat Biotechnol. 2019; 37(3):220–224. [PubMed: 30833764]

11. Holub P, et al. BBMRI-ERIC Directory: 515 Biobanks with Over 60 Million Biological Samples. Biopreserv Biobank. 2016; 14:559–562. [PubMed: 27936342]

12. Litton JE. Launch of an Infrastructure for Health Research: BBMRI-ERIC. Biopreserv Biobank. 2018

13. 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. 2016; 14:298–306. [PubMed: 26977825]

14. [Accessed: 15th May 2019] 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 - 32016R0679 - EN. http://data.europa.eu/eli/reg/2016/679/oj

15. European Open Science Cloud (EOSC). [Accessed: 15th May 2019] Open Science - Research and Innovation - European Commission. 2018. Available at: https://www.eosc-portal.eu/

16. Linden M, et al. Common ELIXIR Service for Researcher Authentication and Authorisation. F1000Res. 2018; 7

17. [Accessed: 15th May 2019] Index of Policies. Available at: https://web.bbmri-eric.eu/Policies/

18. [Accessed: 15th May 2019] Data-repository-service-schemas. data-repository-service-schemas. Available at: https://github.com/ga4gh/data-repository-service-schemas

19. [Accessed: 15th May 2019] Workflow-execution-service-schemas workflow-execution-service-schemas. Available at: https://ga4gh.github.io/workflow-execution-service-schemas/

20. [Accessed: 15th May 2019] Task-execution-schemas task-execution-schemas. Available at: https://github.com/ga4gh/task-execution-schemas

21. [Accessed: 15th May 2019] Tryggve - NeIC web. Available at: https://neic.no/tryggve/

22. [Accessed: 15th May 2019] HPC RIVR project. Available at https://www.hpc-rivr.si/home_en/

23. 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. 2019; 170(5):332–334. [PubMed: 30776795]

24. Dove ES. The EU General Data Protection Regulation: Implications for International Scientific Research in the Digital Era. The Journal of, Law, Medicine & Ethics. 2018; 46(4):1013–1030.

25. Shabani, et al. Rules for processing genetic data for research purposes in view of the new EU General Data Protection Regulation. European Journal of Human Genetics. 2017; 26doi: 10.1038/s41431-017-0045-7

26. [Accessed: 15th May 2019] DATA PROTECTION ACT 2018 (SECTION 36(2)) (HEALTH RESEARCH) REGULATIONS. 2018. http://www.irishstatutebook.ie/eli/2018/si/314/made/en/pdf

27. [Accessed: 15th May 2019] Regels ter uitvoering van Verordening (EU) 2016/679 van het Europees Parlement en de Raad van 27 april 2016 betreffende de bescherming van natuurlijke personen in verband met de verwerking van persoonsgegevens en betreffende het vrije verkeer van die gegevens en tot intrekking van Richtlijn 95/46/EG (algemene verordening gegevensbescherming) (PbEU: 2016, L 119) (Uitvoeringswet Algemene verordening gegevensbescherming). 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

28. [Accessed: 15th May 2019] Lag (2003: 460) om etikprovning av forskning som avser manniskor. http://rkrattsbaser.gov.se/sfst?bet=2003:460

29. A Code of Conduct for Health Research. [Accessed: 15th May 2019] A Code of Conduct for Health Research regarding the EU GDPR. Available at: http://www.code-of-conduct-for-health-research.eu

30. [Accessed: 15th May 2019] ELSI: Ethical, Legal, and Social Issues in Biobanking. BBMRI-ERIC: Making New Treatments Possible. Available at: http://www.bbmri-eric.eu/services/common-service-elsi/

31. Mayrhofer M, Schlunder I. Mind the, Gap: From Tool to Knowledge Base, Mind the, Gap: From Tool to Knowledge Base. Biopreservation and Biobanking. 2018; 16(6)

32. ZweiEinzB. About Corbel | CORBEL Project. [Accessed: 15th May 2019] Coordinated Research Infrastructures Building Enduring Life-science Services. 2019. (2019). Available at: http://www.corbel-project.eu/about-corbel.html

33. [Accessed: 15th May 2019] Executive Masters in management of research infrastructures. Available at: http://www.emmri.unimib.it/en/

34. [Accessed: 15th May 2019] RItrain. Available at:http://ritrain.eu/

35. [Accessed: 15th May 2019] Mark Caulfield (GA4GH news). 2018. https://www.genomicsengland.co.uk/the-uk-has-sequenced-100000-whole-genomes-in-the-nhs/

36. Rozenblatt-Rosen O, et al. The Human Cell Atlas: from vision to reality. Nature. 2017 Oct; 550(7677):451–453. [PubMed: 29072289]

37. Paten B, et al. The NIH BD2K center for big data in translational genomics. J Am Med Inform Assoc. 2015; 22(6):1143–7. [PubMed: 26174866]

Box 1

Description of the European Strategy Forum Research Infrastructures ELIXIR and BBMRI-ERIC (full list of Research Infrastructures can be found here: [http://roadmap2018.esfri.eu/media/1044/part1-project-landmarks-list.pdf])

Biobanking and biomolecular resources research infrastructure (BBMRI-ERIC)

BBMRI-ERIC is a research infrastructure for biobanking that brings together all the main players from the biobanking field — researchers, biobankers, industry and patients — to boost biomedical research. BBMRI-ERIC offers management services, support with ethical, legal and societal issues, and a number of online tools and software solutions. Ultimately, the goal is to make new treatments possible.

ELIXIR

ELIXIR is an intergovernmental organization that brings together life science resources from across Europe. These resources include databases, software tools, training materials, cloud storage and supercomputers. The goal of ELIXIR is to coordinate these resources so that they form a single infrastructure. This infrastructure makes it easier for scientists to find and share data, exchange expertise, and agree on best practices. Ultimately, it will help them gain new insights into how living organisms work.

Box 2

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 roughly 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 (https://www.elixir-europe.org/communities/rare-diseases) promotes and funds activities between ELIXIR platforms and relevant rare disease infrastructures and initiatives. This community provides a strong example of how a coordinated infrastructure can provide direct, tangible benefits to healthcare systems and patients. For example, the RD-Connect platform[1] includes a Biobank and Registry Finder, a Sample Catalogue (integrated with BBMRI-ERIC) and the Genome–Phenome Analysis Platform (GPAP). The genomic data available in GPAP is processed through a validated standard pipeline, and the raw data is deposited at the EGA for long-term storage. GPAP is part of the International Rare Diseases Research Consortium (IRDiRC), 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 one million human genomes. In this sense, other RD-Connect based local systems have already been deployed using dockers, enabling full control on data discovery and access and allowing data to be kept within national boundaries (for example, NaGen Navarra 1000 Genomes [https://www.nagen1000navarra.es/en/home], URD-Cat). GPAP is working towards providing tiered discoverability and data access between local instances based on user permissions.

Box 3

Transnational data access will enable pan-cancer analysis and potentially impact personalized cancer treatment

Cancer is the leading cause of death in several European countries, results in €126 billion in annual health-related costs in Europe, and its relative impact on society is expected to grow due to demographic changes[3]. Progress in DNA sequencing technology has revolutionized the field of personalized cancer treatment, and hundreds of thousands of cancer genomes will be sequenced across Europe by the end of this decade, resulting in exabytes of data. Complex international translational and clinical research programmes will require compatible clinical molecular profiling, robust computational analysis pipelines, and standardized descriptions of molecular and imaging diagnostics. Progress will also need a culture of data collection, international access and storage that allows construction of large-scale longitudinal and transversal data sets. This level of coordination requires utilization of the European Health Research Infrastructures (http://roadmap2018.esfri.eu/media/1044/part1-project-landmarks-list.pdf).

Pan-cancer analysis is a powerful driver for large scale genome access, and provides synergy with the Pan-Cancer Analysis of Whole Genomes (PCAWG) initiative[5], a recognized international pioneer in meeting the challenges related to the accessing, secure storage and cloud-based processing of sensitive patient genomic data. PCAWG has developed transnationally and uses interoperable workflows for cancer genome analysis that are now being incorporated into clinical sequencing efforts through the International Cancer Genome Consortium "Accelerating Research in Genomic Oncology" initiative (ICGC-ARGO, a Global Alliance for Genomics and Health (GA4GH) Driver Project),. A European Open Science Cloud Pilot (EOSCPilot) science demonstrator project - involving ELIXIR and BBMRI-ERIC - on pan-cancer analysis offers provisioning of these workflows on European clouds.

The Colorectal cancer (CRC) cohort, developed within the EU-funded project ADOPT BBMRI-ERIC (H2020), is a use case for piloting access to European biobanks. The building of the CRC-cohort will enable high-quality research and innovation to improve treatment of colorectal cancer. The procedures and IT tools developed within the CRC-cohort will be reusable for similar future efforts on different disease entities, implemented using BBMRI-ERIC as an infrastructure, and directly applicable to improving healthcare for European citizens, and beyond.

Box 4

Summary of recommendations

A coordinated, secure, federated environment that enables population-scale genomic, phenotypic and biomolecular data to be accessible across international borders will be required to enable the committed EU Member States to achieve their goal to access 1M genomes and other health related data.

Research infrastructures such as ELIXIR and BBMRI-ERIC 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 Global Alliance for Genomics and Health (GA4GH) to enable data access to authorized researchers.

Glossary terms

Term	Definition
Application Programming Interface (API)	APIs allow applications to communicate with one another. An API is not a database. It is an access point that an application can utilise in order to access a database.
Biobanks	A biobank is a type of biorepository that stores biological samples (usually human) for use in research.
BioContainers	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: https://biocontainers.pro/#/
BioMedical Science Research Infrastructures (BMS RI)	ESFRI RIs that provide support to life scientists (listed under Food and Health): [http://roadmap2018.esfri.eu/media/1044/part1-project-landmarks-list.pdf] - including ELIXIR, BBMRI-ERIC.
ELIXIR Nodes	An ELIXIR Node is a collection of research institutes within a member country that run the resources and services that are part of ELIXIR. There are currently 23 ELIXIR Nodes.
ELIXIR Communities	ELIXIR Communities bring together experts across Europe to develop standards, services and training within specific life science domains.
ESFRI	European Strategy Forum on Research Infrastructures
Federated	Federated is used in this context to describe an enterprise architecture that allows interoperability and information sharing between semi-autonomous de-centrally organized lines of business information technology systems and applications.
GDPR	The General Data Protection Regulation 2016/679 is a regulation in EU law on data protection and privacy for all individuals within the European Union and the European Economic Area. It also addresses the export of personal data outside the EU and EEA areas
Global Alliance for Genomics and Health (GA4GH)	The Global Alliance for Genomics and Health (https://www.ga4gh.org) is a policy-framing and technical standards-setting organization, seeking to enable responsible genomic data sharing within a human rights framework. 5-year strategic plan: https://www.ga4gh.org/wp-content/uploads/GA4GH-Connect-A-5-year-Strategic-Plan.pdf
HPC	High Performance Computing most generally refers to the practice of aggregating computing power in a way that delivers much higher performance than one could get out of a typical desktop computer or workstation in order to solve large problems in for example science, engineering, or business.
Metadata	Metadata is a set of data that describes and gives information about other data, or dataset(s).
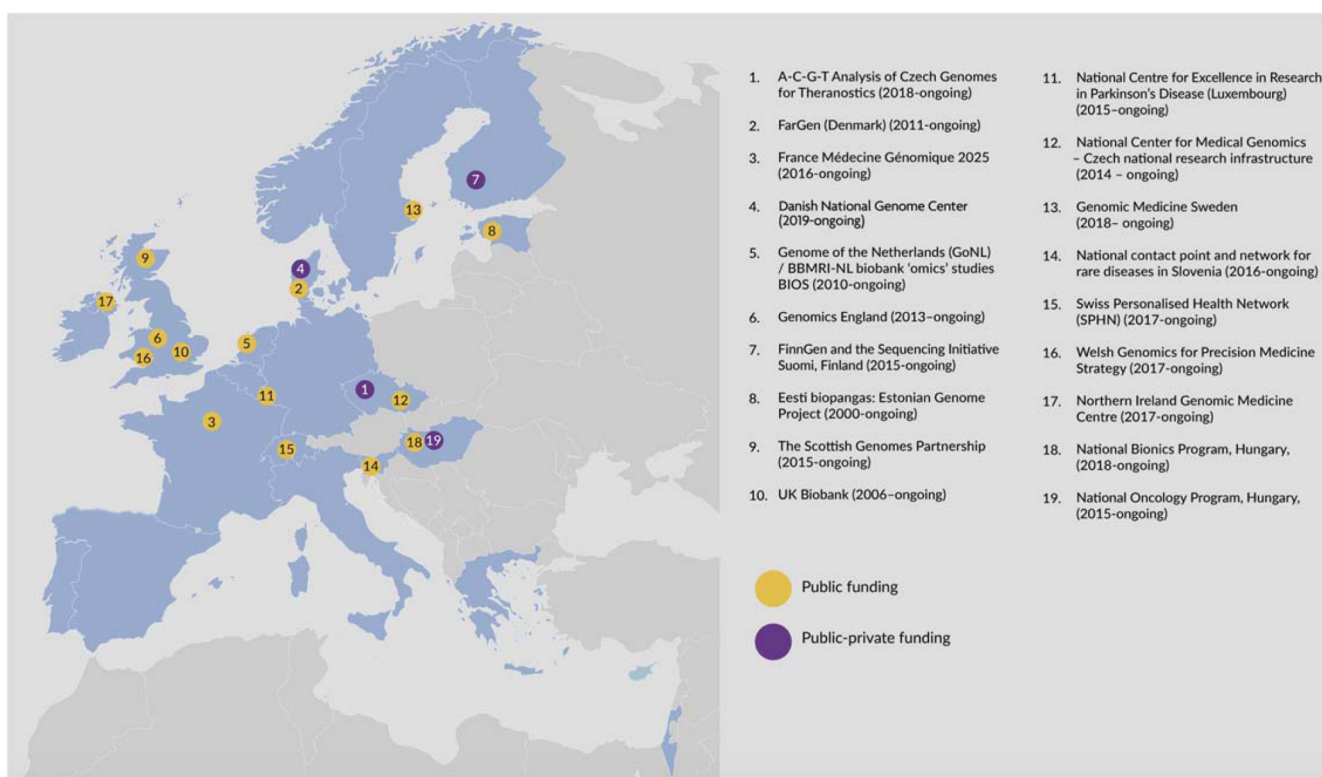
Figure 1. Current healthcare-focussed and genomics-based national initiative projects across ELIXIR members.
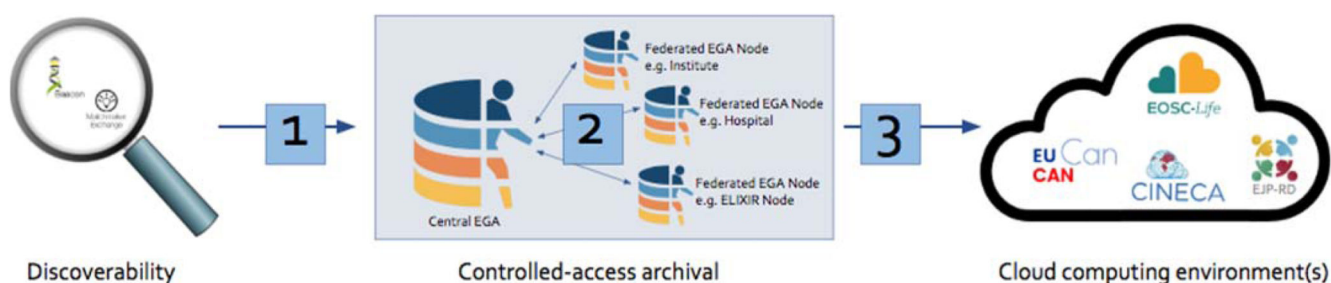
Figure 2. The concept of EGA federation, including a potential user workflow from data discoverability to raw sensitive human data access.

1 = Discoverability: metadata is shared from each of the sensitive data archives to a centralised database upon which queryable interfaces can be built; these can be project specific portals, or interfaces to query the metadata associated with all datasets across a federated network. The GA4GH Driver Projects ELIXIR Beacon and MatchMaker Exchange, for example, provide standards and interfaces to query such metadata in order to aid discoverability. 2 = Controlled-access archival: as a GA4GH Driver Project EGA/ENA/EVA provides interoperable programmatic interfaces that are required to enable metadata transfer and user authentication and authorisation (provided by ELIXIR AAI, for example) across the federated network of controlled-access archives. 3 = Cloud computing environment(s) = community-curated workflows (e.g. those found in BioContainers) are able to be executed remotely and run locally at one or more sensitive data archives utilising the standards from the GA4GH Cloud and Large Scale Genomics Work Streams.

Table 1
EU declaration signatory and membership status

Country	EU declaration signatory*	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	Observer	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	Prospect 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

EU: European Union; BBMRI-ERIC: Biobanking and Biomolecular Resources Research Infrastructure; EMBL: European Molecular Biology Laboratory. Hyphens stand for "not applicable".

*
The initiative is also open to countries of the European Economic Area and the European Free Trade Association.

Table 2

Status of necessary infrastructure for accessing and analysing genomics data at European scale

Necessary minimal infrastructure component:	In development*	Implemented at scale*
Genomics data and clinical information standards, geared towards specific disease communities	Yes	No
Common application programming interfaces (APIs) 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 for enabling 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 healthcare as well as shift the culture towards openness and integration of research data across national boundaries	Yes	Yes

*"In development" and "Implemented at scale" refers to locally defined status within the ELIXIR and/or BBMRI-ERIC Research Infrastructures.