*D980–D987* *Nucleic Acids Research, 2022, Vol. 50, Database issue* *Published online 17 November 2021*
*https://doi.org/10.1093/nar/gkab1059*

# The European Genome-phenome Archive in 2021

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

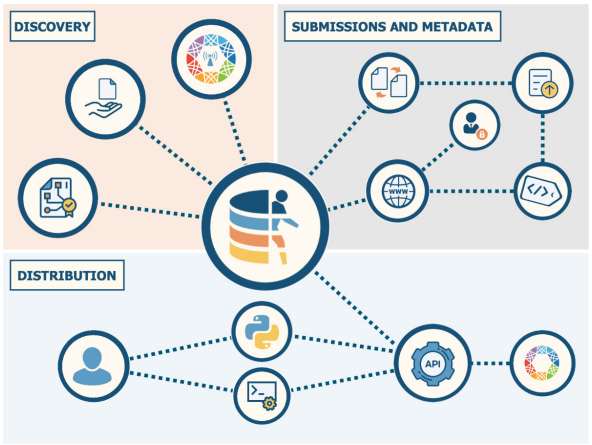
[1]European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Hinxton CB10 1SD, UK and [2]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr Aiguader 88, Barcelona 08003, Spain

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

## ABSTRACT

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

## GRAPHICAL ABSTRACT




## INTRODUCTION

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

*To whom correspondence should be addressed. Email: jordi.rambla@crg.eu
Correspondence may also be addressed to Thomas Keane. Email: tk2@ebi.ac.uk
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

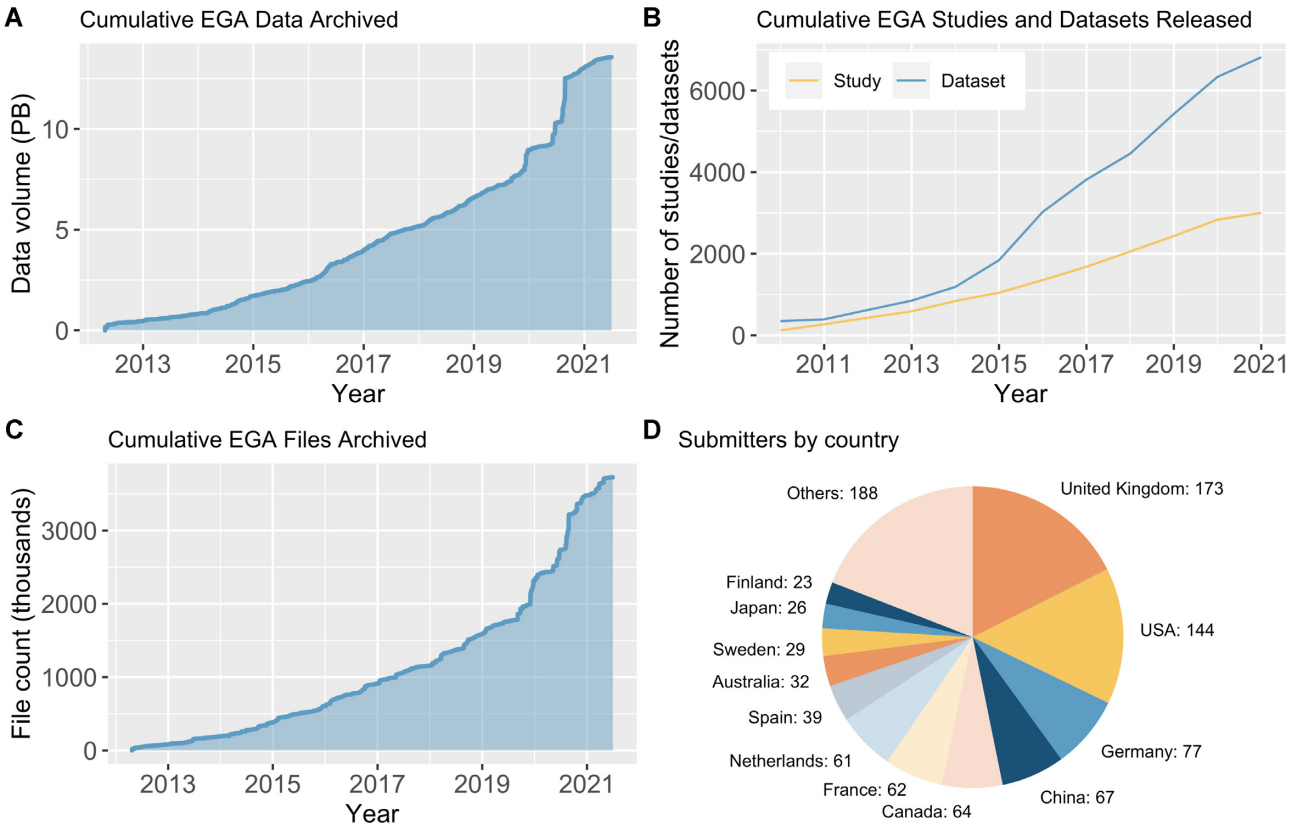



**Figure 1.** Data archived at EGA between 2013–2021. Cumulative size of data (**A**), number of studies and datasets (**B**), and number of files (**C**) archived and available for download from EGA per year. (**D**) Number of institutes per country that have archived data at the EGA.

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

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

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

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

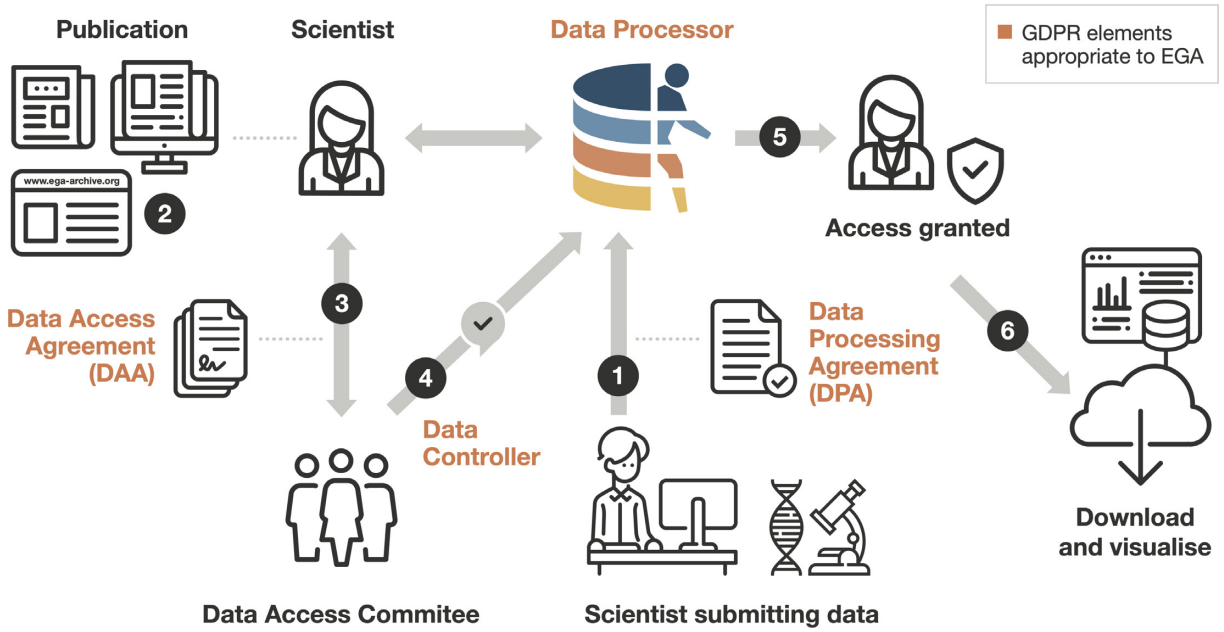



**Figure 2.** EGA facilitates the submission, discovery, access, and distribution of sensitive human data. A researcher submits controlled access human genetic, phenotypic and clinical data to EGA after signing a Data Processing Agreement (1). EGA processes, archives, and releases the dataset to be findable.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

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

## DEPOSITING DATA AT THE EGA

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

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

## DATA DISCOVERY AT THE EGA

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

### Discovery by publication

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

### Discovery by variants

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

### Discovery by public metadata

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

### Discovery by data use ontology

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

### Discovery by data quality

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

### Discovery through linked resources

To broaden data discoverability, the EGA has established links with other public resources. For example, EGA samples are accessioned by BioSamples (12) which stores information about biological samples used in research. Within BioSamples, researchers can link samples from the same study even if the data generated from those samples are in different archives. By linking samples, researchers can discover, for example, viral sequences archived at the European Nucleotide Archive (13) that have corresponding host genomic data archived at the EGA. In response to the COVID-19 pandemic, the European COVID-19 Data Portal (https://www.covid19dataportal.org) was established to accelerate COVID-19 research through data sharing. COVID-19 and SARS-CoV-2 studies archived at the EGA are indexed and displayed in the COVID-19 Data Portal, providing an additional route by which EGA data can be discovered by researchers. Finally, through daily synchronization with a metadata exchange server, the EGA provides summary information for and links to studies archived at the database of Genotypes and Phenotypes (dbGaP, https://www.ncbi.nlm.nih.gov/gap/). In this way, the EGA serves as a global hub for discovery of human data under access control.

## ACCESSING DATA AT THE EGA

### Data access model

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

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

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

### Authentication and authorisation

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

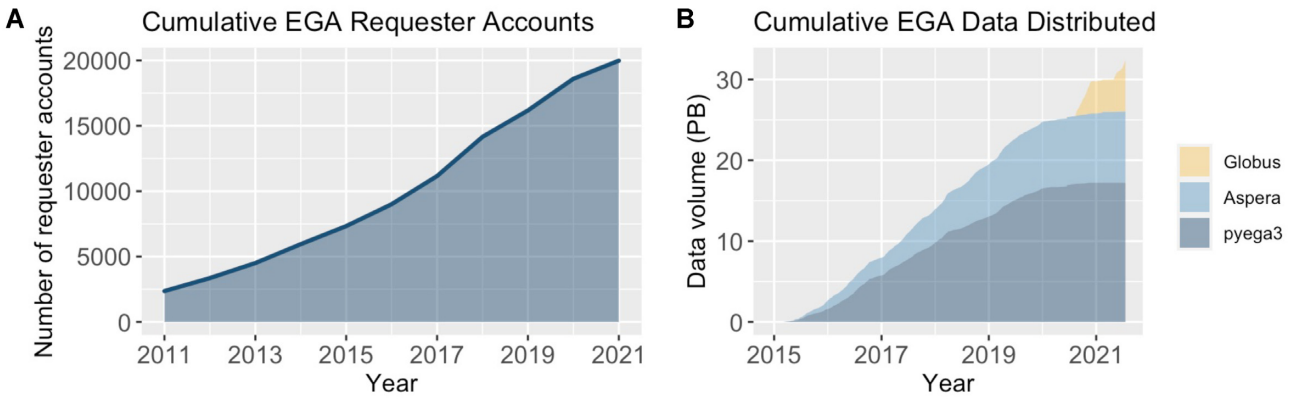



**Figure 3.** EGA data distribution to approved researchers between 2011 and 2021. (**A**) Number of EGA data requester accounts created over time. (**B**) Amount of data distributed to approved researchers over time.

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

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

## Data distribution

Genomic and phenotypic datasets archived at the EGA can be composed of a few files to hundreds of thousands of files ranging in size from very small to quite large. Importantly, these files must be encrypted at rest and during download over secure channels to prevent unauthorised access. EGA brings value to this process by offering users a diversity of options based on their needs (Figure 4).

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

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

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

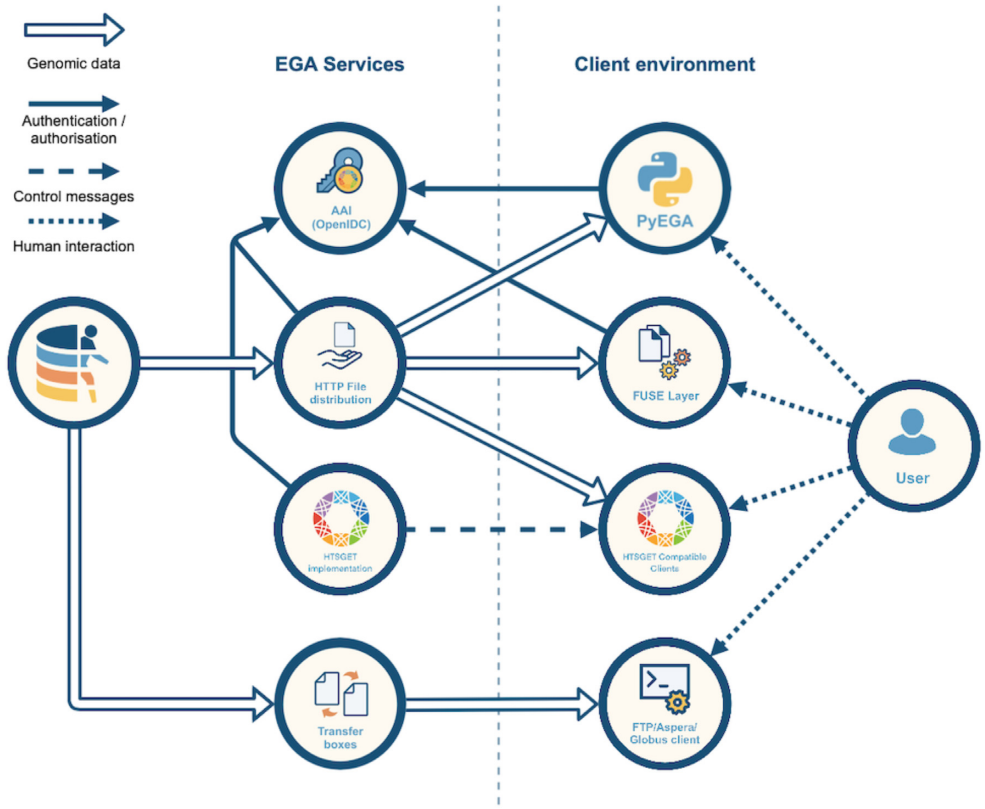



**Figure 4.** The EGA offers a variety of secure data access and download services to meet user needs, many of which implement GA4GH standards. *FUSE: Filesystem in Userspace. AAI: Authentication and Authorization Infrastructure. OpenIDC: OpenID Connect, an open standard and decentralized authentication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

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

## DISCUSSION

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

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

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

## DATA AVAILABILITY

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

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

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

## FUNDING

Horizon 2020 Programme of the European Union [CORBEL [654248], ELIXIR-EXCELERATE [676559], Solve-RD [779257], EASI-Genomics [824110], EJP-RD [825575], CINECA [825775], EuCanCan [825835], EUCanshare [825903], ELIXIR-CONVERGE [871075]]; Wellcome Trust Global Alliance for Genomics and Health [201535/Z/16/Z]; UK Biobank; Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation [2017-171304 (5022)]; European Molecular Biology Laboratory (EMBL); LaCaixa Foundation [004745/008034]; [LCF/PR/CE20/50740008]. Funding for open access charge: LaCaixa Foundation [LCF/PR/CE20/50740008].
*Conflict of interest statement.* Paul Flicek is a member of the Scientific Advisory Boards of Fabric Genomics, Inc. and Eagle Genomics, Ltd.

## REFERENCES

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

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